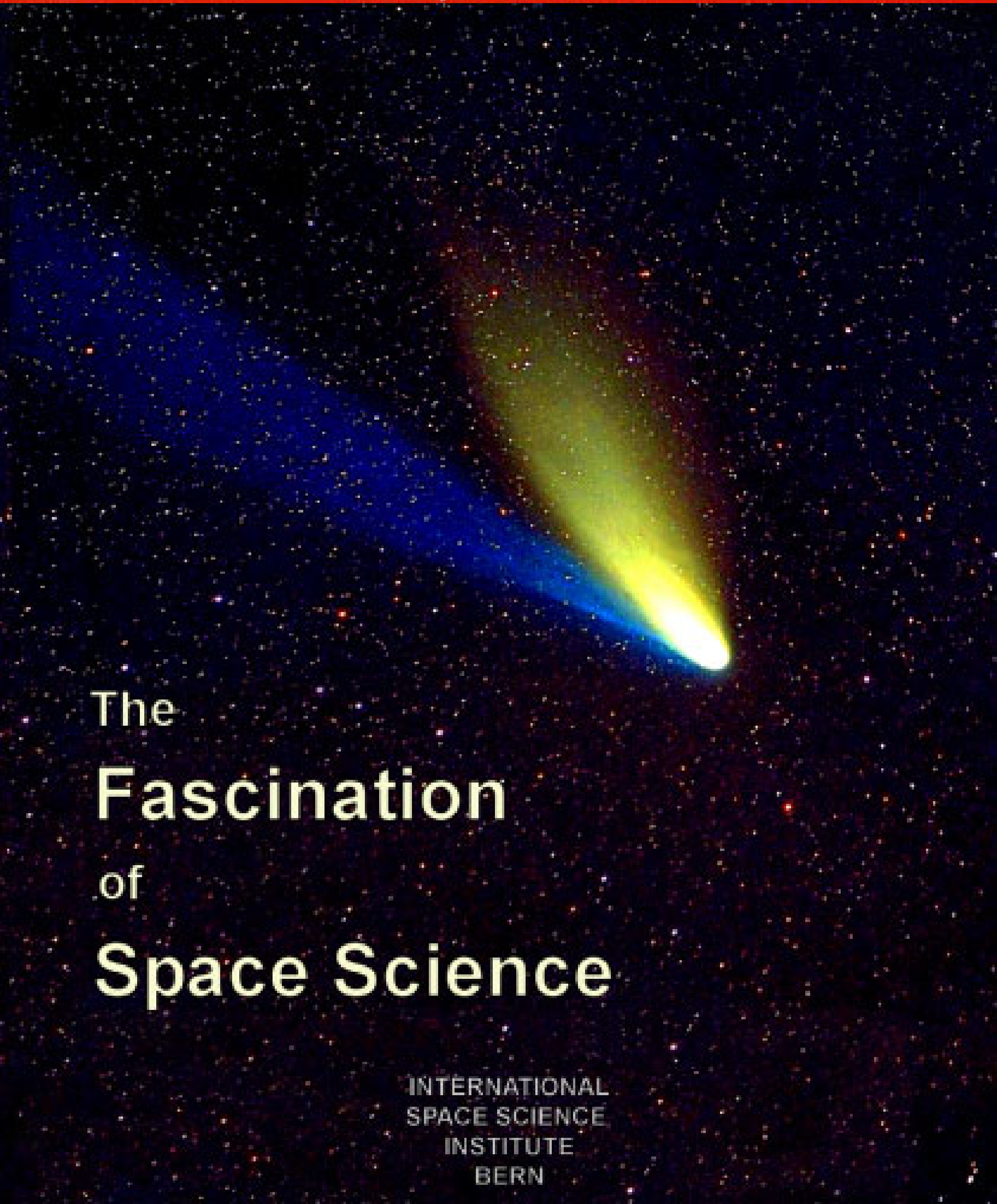
INTERNATIONAL SPACE SCIENCE INSTITUTE

SPATIUM

Published by the Association Pro ISSI



The Fascination of Space Science

INTERNATIONAL
SPACE SCIENCE
INSTITUTE
BERN

INTERNATIONAL SPACE SCIENCE INSTITUTE

SPATIUM

Published by the Association Pro ISSI

The Fascination of Space Science

INTERNATIONAL
SPACE SCIENCE
INSTITUTE
BERN

Cuvillier Verlag Göttingen

Bibliografische Information der Deutschen Nationalbibliothek
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.ddb.de abrufbar.
1. Aufl. - Göttingen : Cuvillier, 2008

978-3-86727-754-9

Nonnenstieg 8, 37075 Göttingen
Telefon: 0551-54724-0
Telefax: 0551-54724-21
www.cuvillier.de

1. Auflage, 2008
Gedruckt auf säurefreiem Papier

978-3-86727-754-9

Editorial

What a wonderful time are we experiencing! Armadas of scientists all over the world are discovering this marvellous world: a generous Mother Nature endows her inexhaustible richness to those who spare no effort to come by.

Scientific progress takes place on many frontiers but certainly one of the most fascinating is space research that aims at uncovering the secrets of the Universe and answering some of mankind's oldest questions: Where do we come from? Where are we going?

This is why the space agencies' scientific programmes are amongst their most breathtaking endeavours. In the frame of its own Science Programme, the European Space Agency (ESA) together with the Swiss Confederation sponsors the International Space Science Institute (ISSI) in Bern to offer the science community a forum of international co-operation and exchange of ideas in the field of space science.

But the outcome of scientific endeavours needs not to be retained within the scientific community but rather be communicated to let the public participate at one of the most brilliant parts of our civilisation. This is the mission stipulated for the Association Pro ISSI to divulge the results of space science to interested parties outside the science community. The present collection of the first twenty-two issues of Spatium is amongst its tangible results achieved so far.

We are indebted to the sponsors of ISSI, to its Directorate and its staff, to all the authors, and specifically to the Pro ISSI association: without their continued support, this book would never have seen the light of the day. May it help to convey the fascination of space science to a wide audience!

Cinuos-chel, September 2008

Hansjörg Schlaepfer

Impressum

SPATIUM
Published by the
Association Pro ISSI

Association Pro ISSI
Hallerstrasse 6, CH-3012 Bern
Phone +41 (0)31 631 48 96
see
www.issibern.ch/pro-issi.html
for the whole Spatium series

President
Prof. Klaus Pretzl
University of Bern
Layout and Publisher
Dr. Hansjörg Schlaepfer
CH-8173 Neerach
Front Cover
Beautiful space:
comet Hale Bopp approaching
Earth in 1997.
(Credit: J. C. Casado, Astronomy
Picture of the Day,
http://antwrp.gsfc.nasa.gov/apod,
Hansjörg Schlaepfer)

INTERNATIONAL SPACE SCIENCE INSTITUTE
No. 1, April 1998

SPATIUM

Publikationsorgan des Vereins Pro ISSI

Entstehung des Universums

Faszination des Ursprungs: Die drei ersten Minuten vor rund 14 Milliarden Jahren waren bestimmend für unser Universum. Experimente der Weltraumwissenschaft ermöglichen uns Einblicke in die Tiefen von Raum und Zeit.

Editorial

Die vorliegende erste Nummer des SPATIUM ist den ersten drei Minuten unseres Universums gewidmet. Sie stellt aber auch den Auftakt für eine lockere Folge von Mitteilungen des Vereins Pro ISSI dar. Dieser Verein hat sich die Förderung der Raumfahrt in der Schweiz zur Aufgabe gemacht. Seine Mitglieder sind Einzelpersonen und Firmen, die sich für die Erforschung des Sonnensystems, der Milchstrasse und des gesamten Universums interessieren. Sein Mitteilungsorgan ist das vorliegende SPATIUM, dessen Name sich vom lateinischen Raum ableitet.

Herrn Professor Johannes Geiss, dem Gründer unseres Vereins und geschäftsführenden Direktor des International Space Science Instituts, sind wir dankbar, dass er mit dieser Arbeit nicht nur den Anfang macht, sondern gleich auch zeigt, was wir wollen: Forschungsergebnisse in diesen Wissensbereichen so vermitteln, dass auch Nichtfachleute berührt werden von der Faszination, die der Himmel ausübt, seitdem es Menschen gibt, die sich Fragen stellen zu ihrem Woher und Wohin.

Hansjörg Schlaepfer
Bern, im März 1998

Inhalt

Impressum

SPATIUM
Publikationsorgan des Vereins Pro ISSI
Erscheint ein- bis zweimal jährlich

INTERNATIONAL SPACE SCIENCE INSTITUTE

Verein Pro ISSI
Hallerstrasse 6, 3012 Bern
Tel. 031 631 48 96
Fax 031 631 48 97

Präsident:
Prof. Hermann Debrunner, Universität Bern
Herausgeber / Quästor:
Dr. Hansjörg Schlaepfer, Contraves Space, Zürich
Redaktion dieses Beitrags:
Peter Abgottspon, Thun
Graphik:
Gian-Reto Roth, Contraves Space, Zürich
Druck:
Drucklade AG, 8008 Zürich

Die ersten Minuten und das weitere Schicksal des Universums

Johannes Geiss, International Space Science Institute, Bern
Vortrag für den Verein Pro ISSI am 30. 10. 1997

Der Weg zur naturwissenschaftlichen Kosmologie

Die Kosmologie im Sinne einer naturwissenschaftlichen Gesamtsicht des Universums ist ein Produkt des 20. Jahrhunderts. Dies gilt sowohl für die relevanten astronomischen Beobachtungen und physikalischen Experimente, wie auch für die mathematische Theorie des Kosmos, denn erst seit 1915 haben wir mit Einsteins Allgemeiner Relativitätstheorie die Möglichkeit einer mathematischen Formulierung für den Raum, die Zeit, das Licht und die Materie des Kosmos.

Die Kosmologie ist die Wissenschaft vom Universum als Ganzem. Man muss daher aus der Fülle der Beobachtungen und Messungen diejenigen herausfinden, welche für das Verständnis des gesamten Kosmos relevant sind. Für die moderne Kosmologie sind drei Beobachtungsbereiche wesentlich:

– Die Expansion des Universums
– Charakter und Zusammensetzung der Materie, die beim Urknall (Big Bang) entstanden ist.
– Die Restwärme im heutigen Universum

Bis 1960 standen für die Kosmologie ausschliesslich Beobachtungen von der Erde aus zur Verfügung. Im Laufe der letzten Jahrzehnte haben jedoch Beobachtungen vom Weltraum aus an Bedeutung gewonnen. Aus der kosmologischen Forschung sind heute das Hubble Space Telescope, das einen tiefen Blick in das Universum erlaubt, der NASA-Satellit COBE, mit dem die Restwärme im Universum genau bestimmt wird, und der europäische Satellit HIPPARCOS, der das System der kosmischen Distanzen auf eine sichere Basis gestellt hat, nicht mehr wegzudenken. Auch bei den Beobachtungen und Messungen, die uns Aufschluss über Art und die Zusammensetzung der Materie geben, wie sie im Urknall entstanden ist, stehen Ergebnisse aus dem Weltraum gleichrangig neben Beobachtungen vom Boden aus. Es seien hier etwa die Beobachtungen mit dem Hubble Telescope, Ergebnisse der Apollo-Landungen auf dem Mond, oder Resultate der europäisch-amerikanischen Weltraumsonde Ulysses erwähnt.

Abbildung 1: Das alte Weltbild
Das Weltbild des Aristoteles war geozentrisch. Demzufolge kreisen der Mond, die Sonne und die Planeten um die Erde. Bis ins 18. Jahrhundert hinein galt der Saturn als der entfernteste Planet. Im Jahre 1781 wurde Uranus und später noch zwei Grossplaneten, Neptun und Pluto, entdeckt. Im Weltbild des klassischen Altertums lagen die Sphären der Fixsterne jenseits des Saturn. Heute wissen wir, dass alle Fixsterne, auch die nächsten unter ihnen, mindestens 20 000 Mal weiter von der Erde entfernt sind als der Saturn.

Obwohl ein Produkt des 20. Jahrhunderts, steht die Kosmologie auf dem Fundament der allgemeinen Naturgesetze, welche die Forscher über Jahrhunderte bis in die neueste Zeit hinein mit Hilfe von Naturbeobachtungen, Experimenten im Labor und theoretischen Ansätzen herausgefunden haben.

Vom Geozentrischen zum Heliozentrischen System

Bis ins 17. Jahrhundert hinein prägte das geozentrische System, bei dem die Erde im Mittelpunkt des Universums steht, das menschliche Denken **(Abbildung 1).** Es entsprach der philosophisch-religiösen Einstellung, dass der Mensch und seine Lebenswelt im Zentrum des Universums angesiedelt sei. Dieses Weltbild wurde von Aristoteles und anderen Philosophen des klassischen Altertums auf der Grundlage von Beobachtungen des Himmels mit blossem Auge geschaffen.

Schon Aristarch von Samos und dann später, im 15. Jahrhundert, Kopernikus hatten ein heliozentrisches System eingeführt. Schliesslich waren es aber die Beobachtungen Galileis mit dem von Holländern neu erfundenen Fernrohr und Galileis physikalischen Argumente, welche eine breite Öffentlichkeit und schliesslich auch die römische Kirche davon überzeugten, dass die Planeten – einschliesslich der Erde – um die Sonne kreisen. Diese revolutionäre Änderung des Weltbildes steht am Anfang der erfolgreichen naturwissenschaftlichen Methodik, die für das Abendland charakteristisch ist.

Das expandierende Universum

Entfernungsmessung im Kosmos

Der Übergang vom geozentrischen zum heliozentrischen System hatte aber noch eine andere, ebenso wichtige Auswirkung: zum ersten Mal wurde es möglich, Distanzen zu Fixsternen zu bestimmen. Im heliozentrischen System bewegt sich die Erde, und dies ermöglicht die Betrachtung des Sternenhimmels von verschiedenen Positionen aus und damit eine Bestimmung der Parallaxen (griechisch: Abweichungen) von Fixsternen **(Abbildung 2)**. Die ersten Messungen gelangen im 18. Jahrhundert, und es zeigte sich, dass selbst die nächsten Fixsterne in einer Entfernung von einigen Lichtjahren liegen: Sie sind ungefähr fünftausendmal weiter entfernt von der Erde als Pluto, der fernste Planet des Sonnensystems. So konnte man erstmalig ahnen, wie riesig gross das Universum ist, sogar im Vergleich zum gesamten Sonnensystem. Diese Erkenntnis bewirkte eine ebenso fundamentale Änderung des Weltbildes wie der Übergang vom geozentrischen zum heliozentrischen System.

Im Vergleich zu kosmischen Dimensionen ist die Parallaxen-Methode auf ein kleines, begrenztes Umfeld der Sonne beschränkt. Wie kommt man aber weiter? Wie be-

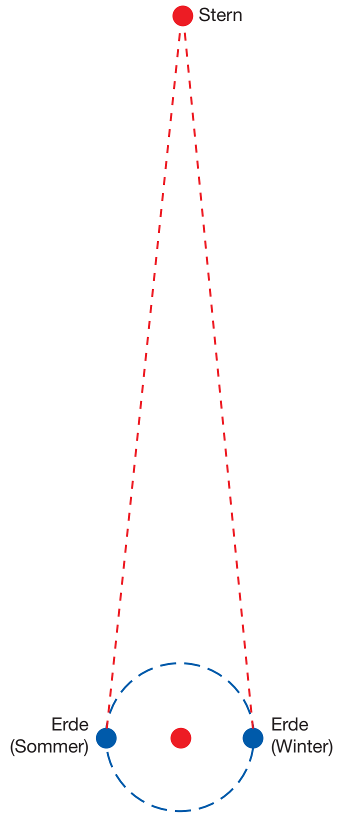


Abbildung 2: Die Entfernungsmessung mit der Parallaxen-Methode
Von 2 Punkten der Erdbahn aus gesehen verschiebt sich die scheinbare Position eines näher liegenden Sterns vor dem Hintergrund sehr weit entfernter Sterne. Aus der Änderung der Richtung (Parallaxe), in welcher der Stern gesehen wird, lässt sich dessen Entfernung genau bestimmen. Obwohl der Durchmesser der Erdbahn 300 Mio Kilometer beträgt, misst man als Winkeländerung (Parallaxe) selbst zu den nächsten Sternen nur einige Bogensekunden. Daraus folgt unter Anwendung einfacher geometrischer Gesetze, dass die entsprechenden Entfernungen ungefähr 100 000 Mal grösser sind als der Erdbahndurchmesser, das heisst die nächsten Fixsterne finden wir in einer Entfernung von einigen Lichtjahren. Der Satellit Hipparcos der europäischen Weltraumbehörde ESA hat mit dieser Methode die Entfernung von 20 000 Fixsternen mit einer Genauigkeit von besser als 10% bestimmt. Dies ist erstaunlich, wenn man bedenkt, dass man mit blossem Auge nur etwa 5000 Sterne erkennen kann.

Die gleiche Parallaxen-Methode wird auch bei der Erdvermessung verwendet. Es sei hier ein einfaches Beispiel genannt: Wenn zwei Personen auf der grossen Schanze in Bern im Abstand von 200 Metern stehen, sehen sie die Sphinx auf dem Jungfraujoch in verschiedenen Richtungen. Der Unterschied beträgt 11 Bogenminuten. Daraus ergibt die einfache Rechnung eine Distanz zur Sphinx von etwa 60 Kilometern.

stimmt man die Distanzen zu den entferntesten Objekten im Kosmos? Die Antwort ist im Prinzip einfach: Die Helligkeit einer jeden Lichtquelle nimmt mit dem Quadrat ihrer Entfernung ab. Wenn die absolute Helligkeit einer Quelle bekannt ist, kann man daher deren Entfernung bestimmen. Diese «Methode der Standardkerzen» funktioniert sowohl auf der Erde als auch im Kosmos. Die Astronomen haben verschiedene solcher Standardkerzen gefunden. Besonders wichtig sind die Cepheiden, das sind Sterne variabler Leuchtkraft. Aus der Frequenz ihrer Lichtschwankungen kann ihre absolute Helligkeit abgelesen werden. Für grössere Distanzen sind Supernova-Explosionen eines bestimmten Typs verwendbar, und für die grössten Entfernungen nimmt man das Licht ganzer Galaxien oder deren Durchmesser. Auf diese Weise wird die kosmische Distanzleiter Stufe um Stufe aufgebaut. Das Ganze klingt heute sehr einfach. Es handelt sich aber um eine hochentwickelte Methodik, die nach vielen Jahren der Grundlagenforschung zu einer Ausmessung des Kosmos bis zu Distanzen von Milliarden von Lichtjahren geführt hat(1).

Sterne, Galaxien und Galaxienhaufen

Ein Blick zum Himmel zeigt uns: Die Materie im Universum ist sehr ungleichmässig verteilt. Wir finden sie konzentriert in Sternen, und diese sind in grösseren Gebilden, den Sternsystemen oder Galaxien organisiert. Galaxien kommen in verschiedenen Formen vor, am bekanntesten sind die «Spiralnebel» **(Abbildung 3).** Doch auch die Galaxien sind nicht gleichmässig, d.h. rein statistisch im Raume verteilt. Sie kommen vielmehr in Haufen (Clusters of Galaxies) vor, von denen die grössten mehr als tausend Galaxien enthalten. Der uns nächstgelegene Haufen ist der Virgo-Haufen **(Abbildung 4),** an dessen Rand sich die Galaxis, die unser Sonnensystem beherbergt, befindet.

Aber auch jenseits der Dimensionen von Galaxienhaufen beobachten wir noch Inhomogenitäten in der Materieverteilung. Diese werden mit exotischen Namen wie «Great Attractor» belegt. Trotzdem herrscht bis heute die Ansicht vor, dass mit zunehmender Grösse des betrachteten Volumens die Verteilung der Materie im Verhältnis immer homogener wird.

Abbildung 3: Die Galaxie M66
Das Licht der Galaxien ist das Licht ihrer Sterne. Wie andere Galaxien besteht M66 aus schätzungsweise 100 Milliarden Sternen.

Abbildung 4: Der Virgo-Haufen
Unsere Galaxis liegt am Rande des Virgo-Haufens und wird daher von diesem merklich angezogen. Daraus resultiert eine lokale Bewegung, die der grossräumig beobachteten Flucht der Spiralnebel überlagert ist. So kommt es, dass unsere Galaxis und der Andromedanebel sich nicht voneinander entfernen, sondern sogar nähern.

(1) Das Heft Nr. 78 (Februar 1997) von Uninova, dem Wissenschaftsmagazin der Universität Basel, ist der Astronomie gewidmet. Der Artikel von Gustav Andreas Tammann in diesem Heft beschreibt die Methoden und Ergebnisse der kosmischen Distanzmessungen. Prof. Tammann hat zur Bestimmung der Dimensionen und des Alters des Universums bahnbrechende Beiträge geleistet.

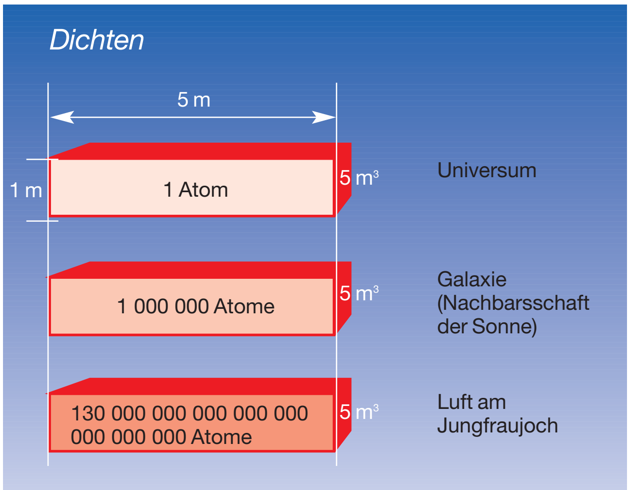


Abbildung 5: Vergleich von Dichten
Man sagt, die Luft auf dem 3475 m hohen Jungfraujoch sei schon recht dünn, ihre Dichte ist aber im Vergleich zur Dichte in einer Galaxie und erst recht zur Dichte im Universum unglaublich gross. Dieser Vergleich veranschaulicht die sprichwörtliche Leere des Universums.

Im Raum zwischen den Galaxien und noch mehr zwischen den Haufen ist die Materiedichte äusserst gering **(Abbildung 5).** «Leere herrscht im Universum!»

Die «Flucht der Spiralnebel»

Im Jahre 1929 entdeckte der amerikanische Astronom Edwin Hubble die «Flucht der Spiralnebel»: Er beobachtete, dass die Galaxien sich systematisch von uns entfernen: Die von ihm gemessene Fluchtgeschwindigkeit der Galaxien wird mit zunehmendem Abstand immer grösser **(vgl. Abbildung 6)**. Diese Entdeckung ist heute nachdrücklich bestätigt und erhärtet[(1)]. Heute sind wir sicher, dass das Universum wirklich expandiert.

Die Darstellung in **Abbildung 6** veranschaulicht, dass sich die fernen Spiralnebel systematisch nicht nur von uns entfernen, sondern auch von jedem anderen Punkt des Universums. Dies bedeutet: Das Universum hat kein Zentrum: Es ist demnach gleichgültig, ob ich ein Bewohner des Andromedanebels oder in unserer Galaxis zu Hause bin. In jedem Falle fliegt sozusagen alles von mir weg.

Durch die Entdeckung Hubbles wurde auch das heliozentrische System des Universums endgültig widerlegt. Wir betonen aber, dass die Sonne selbstverständlich nach wie vor das Zentrum des Planetensystems bildet.

Dimensionen im Universum

Die **Abbildung 7** gibt einen Eindruck von der Tiefe des Universums. Die grösser erscheinenden Galaxien weisen geringere Rotverschiebungen auf als die kleineren, das heisst, die ersteren liegen uns näher. Der entfernteste Galaxienhaufen ist in **Abbildung 8** gezeigt. Seine Distanz wurde zu 12 Milliarden Jahre bestimmt. Das Universum kann aber noch grösser sein, und wir können zur Zeit nicht ausschliessen, dass es unendlich gross ist. Wir wollen diese Frage nicht weiter verfolgen, weil wir dann näher auf die Allgemeine Relativitätstheorie Einsteins eingehen müssten. Diese Theorie lässt verschiedene Lösungen zu; man muss die Frage nach der Homogenität und der Unendlichkeit des Universums durch Vergleich zwischen diesen Lösungen und den Beobachtungen beantworten.

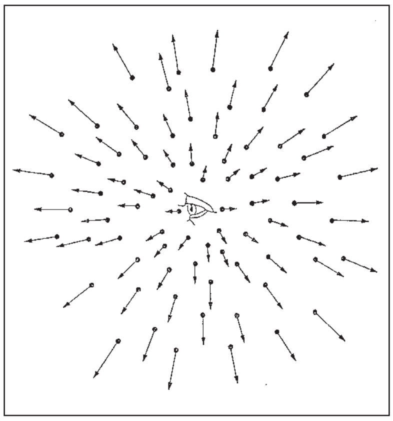


Abbildung 6: «Flucht der Spiralnebel»
Von uns aus gesehen fliegen die fernen Galaxien davon. Die Fluchtgeschwindigkeit wird mit zunehmendem Abstand grösser, woraus folgt, dass das Universum kein Zentrum hat. Die Fluchtbewegung wird nämlich von jedem Punkt des Universums aus in gleicher Weise wahrgenommen. Dieses Bild (nach Hubert Reeves) lässt dies leicht erkennen: Nehmen wir irgendeine der als Punkt dargestellten Galaxien und beziehen – mit Hilfe der Vektor-Additionsregeln – alle Geschwindigkeitspfeile auf diesen Punkt, so sehen wir, dass sich alle anderen Galaxien auch von diesem Bezugspunkt entfernen. Jeder Punkt, jede Galaxie ist also gleichberechtigt bezüglich der Relativbewegungen im Raume.

Im **Kasten «Vergleich der Dimensionen»** werden die Dimensionen des Kosmos, der Galaxis und des Sonnensystems verglichen. Die Hierarchie der Grössen ist beeindruckend. Die nächsten Sterne sind 3 Lichtjahre von uns entfernt, im Vergleich zu kosmischen Dimensionen ist das sehr nah. Doch wie winzig klein nimmt sich erst unser Sonnensystem mit seinem Durchmesser von 0.002 Lichtjahren aus! Und dann die Distanz Erde – Mond, die doch die grösste Entfernung ist, die je zwischen Menschen und ihrem Mutterplaneten lag. Die erste Reise zum Mond war eine bahnbrechende Leistung. Aber, wie der Vergleich zeigt, ist es nicht einmal ein erster Schritt zur Eroberung des Weltalls.

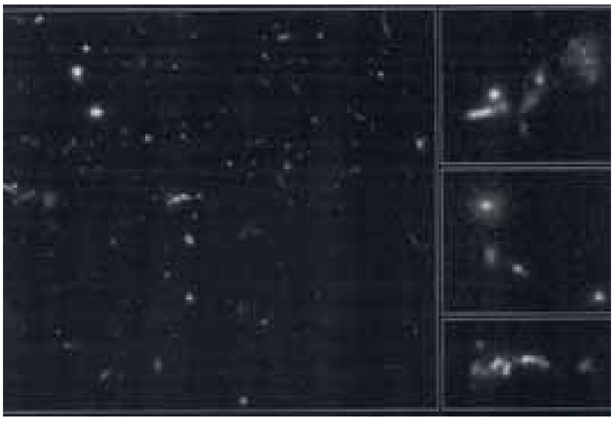


Abbildung 8: Der entfernteste Galaxienhaufen
Derartig entfernte Objekte können nur mit dem Hubble Space Telescope beobachtet werden. Dieses wurde von der NASA konstruiert und in den Weltraum gebracht, wobei die europäische Weltraumbehörde ESA einen Anteil von 15% hat.

Anfangs war die Bildschärfe des Teleskops durch einen Konstruktionsfehler beeinträchtigt. Im Jahre 1993 wurde dieser Fehler anlässlich des Shuttle-Fluges STS 61 korrigiert. Hierbei hatte der Schweizer Astronaut Claude Nicollier einen hervorragenden Anteil.

Abbildung 7: Hubble Deep Field
Es handelt sich hier um eine der schönsten und wissenschaftlich ergiebigsten Aufnahmen mit dem grossen Hubble Space Telescope. Fast alle Objekte, die man sieht, sind Galaxien. Die gross erscheinenden liegen näher bei uns als die kleineren. Sterne sind an der kreuzförmigen Erscheinung ihres Lichts erkennbar. Es gibt nur ein oder zwei Sterne auf diesem Bild. Sie liegen vergleichsweise sehr nahe bei uns und gehören unserer Galaxis an.

Alter des Universums

Rechnet man die Flucht der Spiralnebel zurück bis zu dem Zeitpunkt, da alles noch beieinander lag, und berücksichtigt dabei noch Korrekturen für die Anziehung, so kommt man auf ein Alter von zirka 14 Milliarden Jahren. Diese Zahl ist mit gewissen Unsicherheiten behaftet, es können 15 Milliarden oder auch nur 12 Milliarden Jahre sein. Den Grund für die Ungenauigkeit bilden zwei Faktoren: die zu unpräzisen Distanzangaben der Millionen von Lichtjahren entfernten Galaxienhaufen und die Verlangsamung der

Vergleich der Dimensionen

Beobachtetes Universum	>	12 000 000 000	Lichtjahre
Distanz zum Andromeda Nebel		2 000 000	Lichtjahre
Nächste Sterne		3	Lichtjahre
Sonnensystem		0,002	Lichtjahre
Distanz Erde-Mond		0,00000005	Lichtjahre

Zeiten und Alter

Universum	ca. 14	Milliarden Jahre
«Dark Age of the Universe»	um 13	Milliarden Jahre
Erste Galaxien	ca. 12	Milliarden Jahre
Sonne & Erde	4,6	Milliarden Jahre
Kambrium	0,5	Milliarden Jahre
Letzte Eiszeit	0,00002	Milliarden Jahre

Expansion des Universums im Laufe der Zeit, die durch die gegenseitigen Anziehungskräfte bewirkt wird. Bevor man derartige Rückwärtsberechnungen mit genügender Genauigkeit anstellen konnte, schätzte man das Alter des Universums auf etwa 2 1/2 Milliarden Jahre, während man das Alter der Erde mit 4 1/2 Milliarden Jahren schon recht genau bestimmt hatte. Dieser offensichtliche Widerspruch ist seit etwa 35 Jahren durch immer bessere Messungen vom Tisch.

Der **Kasten «Zeiten und Alter»** gibt einen Überblick über kosmische und erdgeschichtliche Zeiten. Zwischen dem Big Bang und der Entstehung der Galaxien lag wahrscheinlich ein Intervall von 1 bis 2 Milliarden Jahren. Man nennt diese Epoche auch «Dark Age of the Universe», etwa mit «Finsterem Mittelalter» zu übersetzen. Dies hat zwei Gründe: Erstens weiss man wenig über diese Epoche, und zweitens war dazumal das Universum nach dem unglaublich hellen und heissen Big Bang unter 0° Celsius abgekühlt. Aber die ersten Sterne hatten noch nicht begonnen zu leuchten. Diese bildeten sich erst etwa 1–2 Milliarden Jahre später. **(Abbildung 9).**

Zur Entschlüsselung der Chronologie des Sonnensystems und der Erde hat auch das Physikalische Institut der Universität Bern seit Jahrzehnten mit Altersbestimmungen an Meteoriten-, Mond- und Erdproben beigetragen. Wir wissen heute, dass die Sonne und alle Planeten vor 4,6 Milliarden Jahren während einer relativ kurzen Zeitspanne entstanden sind. Die Erde hat also ein respektables Alter, immerhin sind es 30% des Alters des Universums. Dagegen ist die Epoche der klassischen Geologie, die mit dem Kambrium begann, doch schon relativ kurz.

Abbildung 9: Kugelsternhaufen
Die Kugelsternhaufen sind in ihrer Gesamtheit innerhalb relativ kurzer Zeit entstanden. Dies gibt uns die Möglichkeit, aus einem Vergleich des Entwicklungsstadiums von Sternen verschiedener Grösse das Alter des Haufens, welches dem Alter seiner Sterne entspricht, zu bestimmen. Für die Sterne des gezeigten Haufens ergibt sich so ein Alter von 12 Milliarden Jahren.

Materie im Universum (2)

Art und Verteilung der Materie ermöglichen uns, die Vorgänge im Frühstadium des Universums zu rekonstruieren (2). Sie geben uns Aufschluss über den Bildungsprozess und die Weiterentwicklung der Galaxis, aber auch über den Ursprung des Sonnensystems sowie über die Geschichte von Sonne, Planeten und Monden.

Die uns im heutigen Universum bekannten Teilchen sind in **(Abbildung 10)** zusammengestellt. Wir stellen damit drei grundsätzlich verschiedene Materietypen vor:

1 Die Atome bilden die «gewöhnliche» Materie.
2 Die Photonen oder Lichtteilchen

(2) Über dieses Thema fand im Mai 1997 am International Space Science Institute in Bern ein Workshop statt. Die Ergebnisse erscheinen im April 1998 unter dem Titel «Primordial Nuclei and their Galactic Evolution» (eds. Nikos Prantzos, Monica Tosi und Rudolf von Steiger) als Band 4 der Space Science Series of ISSI, Kluwer Academic Publishers.

Die Bausteine der Atome

Baustein	Atomgewicht	Ladung	Ort im Atom
Proton	1,0073	positiv	Atomkern
Neutron	1,0087	neutral	Atomkern
Elektron	0,0005	negativ	Elektronenhülle

sind zwar sehr zahlreich, aber ihre Gravitationskraft spielt im heutigen Universum nur eine geringe Rolle, da sie masselos sind.

3 Die Neutrinos gehören zu einem dritten Typus, den wir «exotische» Materie nennen wollen.

Die «exotische» Materie zeichnet sich dadurch aus, dass sie mit der gewöhnlichen Materie und mit dem Licht nur sehr schwach in Wechselwirkung tritt. Deshalb wird die «exotische» Materie von uns nicht direkt wahrgenommen. Sie kann nur mit den subtilsten Methoden der modernen Physik nachgewiesen werden. Die Massen der einzelnen Neutrinos sind sehr klein, wir wissen aber nicht, ob die Neutrinos vollständig masselos sind. Deshalb sind wir noch nicht sicher, ob ihre Gravitationskräfte im heutigen Universum eine wesentliche Rolle spielen. Wir werden später auf die «exotische» Materie zurückkommen und zunächst näher auf die «gewöhnliche» Materie eingehen.

Die «gewöhnliche» Materie umfasst alle chemischen Elemente mit ihren stabilen und radioaktiven Isotopen. Sonne, Erde, Wasser und Luft sowie alle Lebewesen bestehen aus dieser «gewöhnlichen» Materie.

Die Bausteine der «gewöhnlichen» Materie sind die Protonen, Neutronen und Elektronen, deren Eigenschaften in **Kasten «Die Bausteine der Atome»** zusammengestellt sind. Sitz der Protonen und Neutronen sind die Atomkerne, die Elektronen bilden die Atomhülle. In **Abbildung 11** ist angegeben, wie die einzelnen Kerne entstanden sind. Im Big Bang werden nur die 6 leichtesten synthetisiert, nämlich die zwei Isotope von Wasserstoff (^{1}H und ^{2}H), von Helium (^{3}He und ^{4}He) und von Lithium (^{6}Li und ^{7}Li). Vollständig aus dem Big Bang stammt allerdings nur der schwere Wasserstoff (^{2}H, auch Deuterium genannt).
Wie wir später zeigen werden, ist der Prozess der Elementsynthese im Big Bang nach etwa drei Minuten abgeschlossen. Lithium ist das schwerste Element, das dabei entstanden ist. Zu den extrem seltenen Elementen Lithium, Beryllium und Bor trägt die Zertrümmerung schwerer Kerne durch die kosmi-

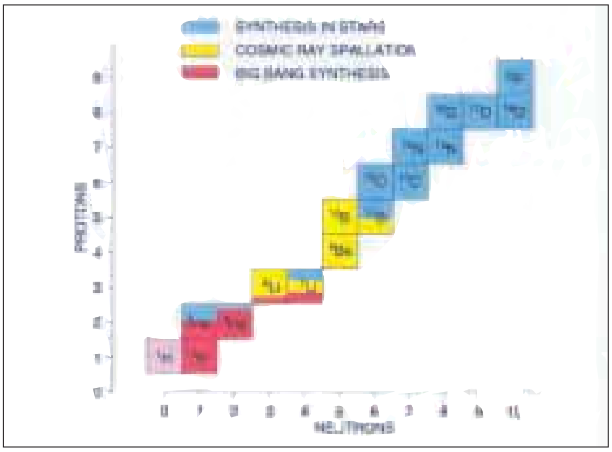


Abbildung 11: Die Herkunft der chemischen Elemente und ihrer Isotope
Im Big Bang werde nur die Isotope der drei leichtesten Elemente synthetisiert. Isotope eines Elements (z. B. ^{3}He und ^{4}He) haben ungefähr gleiche chemische Eigenschaften. Sie unterscheiden sich nur durch ihr Atomgewicht, dessen abgerundeter Wert links vom Element-Symbol angegeben wird.

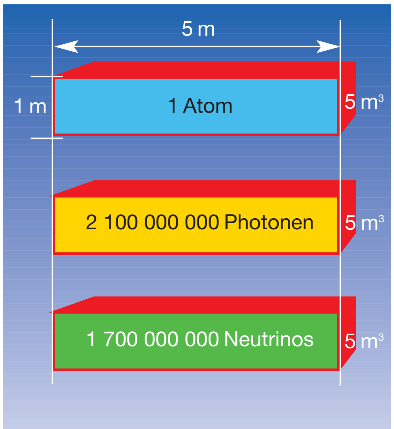


Abbildung 10: Welche Bestandteile kennen wir im Universum?
Die mittlere Anzahl Teilchen pro 5 m^3 im heutigen Universum. Die Atome und die Photonen lassen sich aus Beobachtungen direkt «abzählen». Die Anzahl Neutrinos erhält man, wenn man die Theorie der Elementarteilchen, die Relativitätstheorie und die Thermodynamik auf die Prozesse im frühen Universum anwendet. Da die Photonen masselos sind, tragen sie im heutigen Universum zur Dichte (in kg / m^3) nur wenig bei. Die Massen der Neutrinos sind noch nicht gut bekannt. Falls sie nicht völlig masselos sind, könnte ihr Beitrag zur Gesamtdichte im Universum wichtig sein.

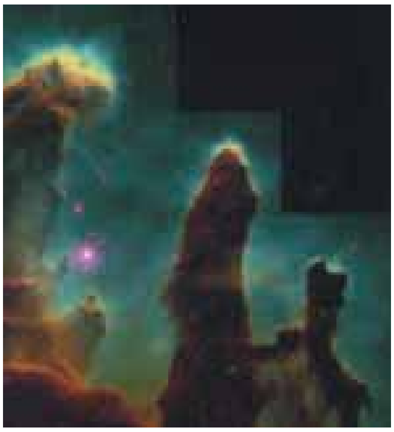


Abbildung 12: Die Bildung der Sterne
Ein Stern entsteht durch Gravitationskollaps eines Fragments einer interstellaren Dunkelwolke. In diesen Wolken ist die Dichte der Materie hoch und die Temperatur gering, so dass die Gravitationskräfte den Gasdruck überwinden können. Es kommt örtlich zum Kollaps, und es entstehen Sterne. Die oberen Bereiche der hier abgebildeten Dunkelwolke «Adler Nebel» sind durch neugebildete Sterne beleuchtet. Beobachtungen, chemische Daten und Isotopenanalysen in unserem Sonnensystem lassen darauf schliessen, dass Sonne und Planeten vor 4,6 Milliarden Jahren auf ähnliche Weise entstanden sind.

Abbildung 13: Sterne als Quelle schwerer Elemente
Bei den hohen Temperaturen und Dichten im Innern der Sterne werden aus den leichteren Elementen schwerere aufgebaut. Grössere Sterne verlieren im fortgeschrittenen Alter einen Teil ihrer Materie, manchmal relativ sanft, wie im Falle des hier abgebildeten Ringnebels, oder, im Falle sehr grosser Sterne, durch Super Novae Explosion. Gas und Staub in der Galaxis werden so allmählich immer reicher an schweren Elementen.

sche Strahlung bei. Der Aufbau aller schweren Elemente wie Kohlenstoff, Sauerstoff oder Eisen wird im Big Bang durch die Instabilität der Zwischenprodukte verhindert. Diese Elemente werden ausschliesslich bei den extrem hohen Dichten, die im Inneren der Sterne herrschen, synthetisiert **(Abbildungen 12 und 13).**

Kernsynthese im Big Bang

Die Expansion des Universums bewirkt eine kontinuierliche Abkühlung der Materie. Diese erreichte eine Zehntelsekunde nach dem Big Bang (Universalzeit 0,1 Sekunde) eine Temperatur von 40 Milliarden Grad. Bei dieser Hitze sind alle zusammengesetzten Atomkerne instabil, d.h. unsere gewöhnliche Materie bestand dazumal aus freien Neutronen, Protonen und Elektronen. Neutronen und Protonen wurden laufend ineinander umgewandelt, durch Austausch mit den in Überzahl vorhandenen Elektronen und Neutrinos. Hierbei resultierte ein Überschuss an Protonen, da diese etwas leichter und stabiler sind als die Neutronen. Expansion und Abkühlung des Universums bewirkten eine rasche Abnahme der Dichte von Elektronen und Neutrinos, so dass der genannte Austauschprozess bei der Universalzeit von 1 Sekunde praktisch zum Stillstand kam. Das Neutronen / Protonen-Verhältnis «fror ein». Nach etwa 100 Sekunden, als die Temperatur auf eine Milliarde Grad abgesunken war, begannen sich Neutronen und Protonen miteinander zu verbinden, es entstanden zusammengesetzte Atomkerne[(3)]. Dabei verbanden sich je zwei Neutronen und Protonen zu ^{4}He, dem sehr stabilen, schweren Heliumisotop. Fast alle Neutronen wurden so im Helium fixiert, die Ausbeute an leichteren und schwereren Kernen ist sehr gering. Die Synthese im Big Bang ergab also im wesentlichen schweres Helium (^{4}He) und, wegen des Protonenüberschusses, auch leichten Wasserstoff (^{1}H).

Gültigkeitsbereich unserer physikalischen Gesetze: Test beim Universalalter von einer Sekunde

Die Messung der Menge an Helium im Kosmos stellt einen kritischen Test dar für unser Verständnis der Vorgänge im frühen Universum, zu einer Zeit, da die Temperatur bei 1–10 Milliarden Grad lag, und der Druck bei 10^{14}–10^{18} Atmosphären. Die Übereinstimmung von theoretischer Vorhersage und Beobachtung ist verblüffend: Mit den aus Versuchen im Labor und Beobachtungen von Planetenbahnen hergeleiteten physikalischen Gesetzen berechnen wir, ohne irgendwelche Zusatzannahmen, dass der Big Bang 24–25 Gewichtsprozent an Helium produzieren sollte. Die astronomische Beobachtung ergibt 24.3–24.6 Prozent. Diese perfekte Übereinstimmung zwischen theoretischer Vorhersage und Himmelsbeobach-

(3) Diese Synthese von Atomkernen im Big Bang erfolgte bei wesentlich höheren Temperaturen als die Synthese von chemischen Stoffen. Ansonsten gibt es aber Analogien. Nehmen wir zum Beispiel die Verbrennung von Kohle: hier wird aus den Bestandteilen C (Kohlenstoff) und O_2 (Sauerstoff) das sehr stabile CO_2-Molekül (Kohlendioxyd) synthetisiert.

tung lässt sich in ihrer Tragweite durchaus mit Newtons mathematischer Herleitung der Gesetze Keplers vergleichen. In beiden Fällen haben wir den Beweis des Wirkens der gleichen Gesetze am Himmel und auf der Erde, nur dass jetzt unter Himmel nicht nur unser Planetensystem, sondern das gesamte Universum zu verstehen ist. Umgekehrt deutet für sich genommen diese keineswegs selbstverständliche, sondern eher erstaunliche Gültigkeit der Naturgesetze an allen Orten und zu allen Zeiten im gesamten Universum darauf hin, dass dieses einen wohldefinierten Anfang hatte.

Wieviel Materie gibt es im Universum?

Die Synthese von ^4He im Big Bang ging über Zwischenprodukte. Zu diesen gehört der schwere Wasserstoff (^2H), auch Deuterium genannt. Je geringer die Dichte zur Zeit der Elementsynthese war, desto unvollständiger verlief die Synthese zum ^4He, d.h. desto mehr an Deuterium und anderen Zwischenprodukten blieb übrig. Man kann also aus der kosmischen Häufigkeit des Deuteriums die Menge an «gewöhnlicher» Materie im Universum berechnen.

Abbildung 14: Die Erde aus grosser Distanz gesehen

Wasser und andere wasserstoffhaltige Substanzen enthalten neben dem normalen, leichten Wasserstoffisotop (^1H) eine kleine Beimischung des schweren Isotops Deuterium (^2H). Auf der Erde, also in den Gesteinen, im Meerwasser und in den Lebewesen, haben wir ungefähr ein Deuteriumatom unter 6000 Wasserstoffatomen.

Wie alles Deuterium im Universum wurde auch das irdische im Big Bang erzeugt. Oder anders ausgedrückt, fänden wir überhaupt kein Deuterium im Meer, so wäre bewiesen: es gab den Big Bang nicht. Also auch die gesamte Deuteriummenge, die wir in uns tragen – 2 bis 3 Gramm, je nach Körpergewicht – wurde im Big Bang synthetisiert, hat 14 Milliarden Jahre überdauert und ist unverfälscht in unseren Körper gelangt, eine faszinierende Vorstellung!

Der irdische Deuteriumgehalt im Wasserstoff von $^1/_{6000}$ ist aber nicht repräsentativ für das Universum, weil in dem Wasserstoff, der auf die Erde gelangte, das Deuterium durch chemische Reaktionen angereichert wurde. Am genauesten lässt sich die universelle Deuteriumhäufigkeit bisher aus Sonnenwindmessungen und aus Absorptionsspektren des interstellaren Gases bestimmen.

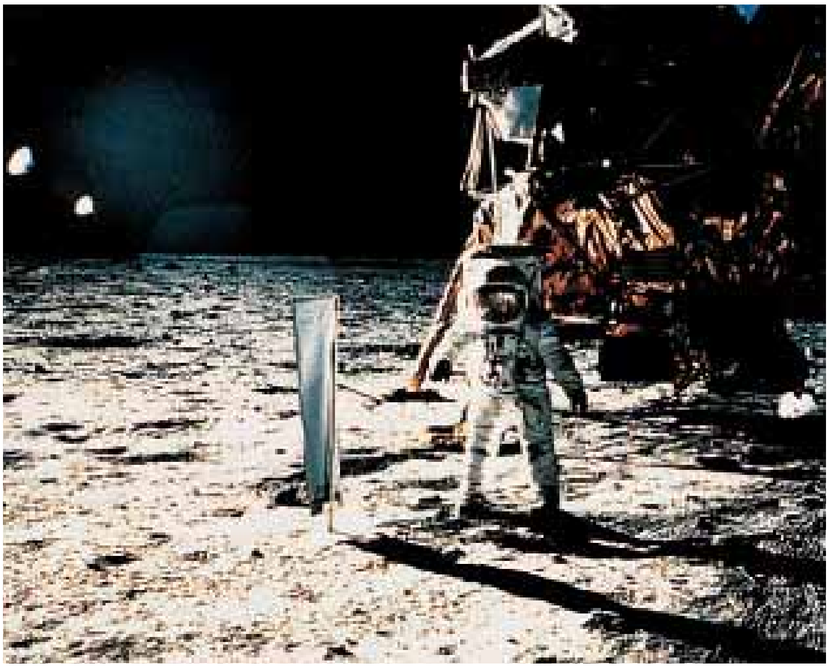


Abbildung 15: Das schweizerische Sonnenwindexperiment

Die Abbildung zeigt das schweizerische Apollo-Experiment SWC, welches kurz nach der ersten Mondlandung am 21. Juli 1969 von Edwin Aldrin aufgestellt worden ist. Bei vier weiteren Mondlandungen kam dieses Experiment zum Einsatz. Da Deuterium unmittelbar nach der Entstehung der Sonne in das leichte Heliumisotop ^3He verwandelt wurde, konnte damit schon 1972 am Physikalischen Institut der Universität Bern aus dem gemessenen Isotopenverhältnis des Heliums im Sonnenwind der Deuteriumgehalt der Ur-Sonne und damit die Dichte der «gewöhnlichen» Materie im gesamten Universum berechnet werden.

Wie können wir nun die Menge an schwerem Wasserstoff im Universum bestimmen? Das Wasser auf der Erde gibt keine schlüssige Antwort **(Abbildung 14).** Aus Isotopenmessungen im Sonnenwind wurde schon Anfang der 70er Jahre das Verhältnis von schwerem zu leichtem Wasserstoff in der ursprünglichen Sonnenmaterie ermittelt **(Abbildung 15).** Inzwischen ist das Deuterium auch im interstellaren Gas, das uns heute umgibt, gemessen worden. Berücksichtigt man die Zerstörung von Deuterium in den Sternen unserer Galaxis, so ergibt sich, dass es im Universum direkt nach dem Big Bang 1 Deuteriumatom unter 30 000 Wasserstoffatomen gab. Hieraus kann man nun, wie oben erläutert, den Gehalt an «gewöhnlicher» Materie im Universum berechnen. Es ergibt sich 1 Atom in 5 m^3 **(vgl. Abbildung 10).**

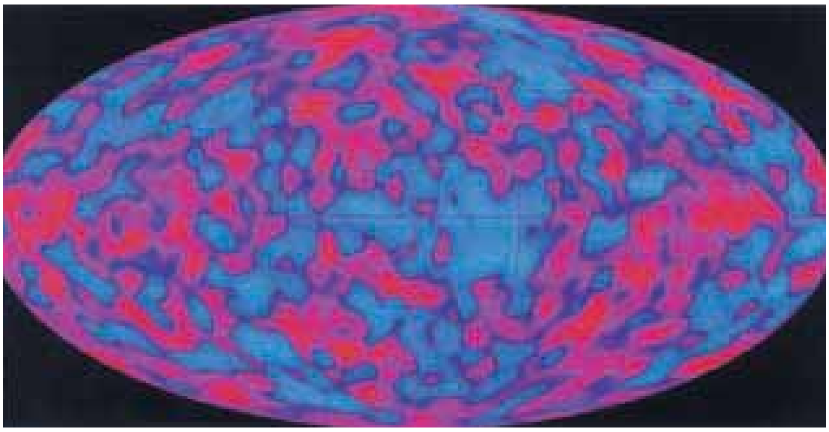


Abbildung. 16: Die Temperaturstrahlung («Hintergrundstrahlung») im Universum
Das Bild zeigt die Temperaturstrahlung aus der gesamten Himmelskugel. Die kosmische Temperatur, die im gesamten Universum herrscht, beträgt heute 2.73 K (2.73 Grad über dem absoluten Nullpunkt). Aufgezeichnet sind hier nicht die Temperaturen selbst, sondern deren winzige Anisotropien. Zwischen blau und rot besteht ein Temperaturunterschied von einem hunderttausendstel Grad. (COBE data, provided by the COBE Science Working Group, NSSDC, NASA Goddard Space Flight Center).

«Exotische» Materie im Universum

Die Materiedichte im Universum lässt sich aber auch auf ganz andere Art bestimmen, nämlich aus den Anziehungskräften zwischen den Galaxien untereinander oder zwischen den Galaxien und dem Licht. Dabei ergeben sich aber 3–5 mal grössere Materiedichten! Wie kann man sich diese Diskrepanz erklären? Hierüber ist viel diskutiert worden, aber es scheint nur die eine Möglichkeit zu geben: neben der uns vertrauten «gewöhnlichen» Materie gibt es im Universum noch eine andere, «exotische» Art von Materie, die in ihrer Dichte und damit ihrer Gravitationswirkung die «gewöhnliche» sogar übertrifft. Zwischen der «gewöhnlichen» Materie und den Teilchen der «exotischen» Materie besteht nur eine sehr schwache Wechselwirkung. Die «exotischen» Materieteilchen laufen durch unsere Körper, ja durch die gesamte Erde ungehindert hindurch, man nennt sie daher weakly interacting particles. Es ist schwer vorstellbar: im Universum dominiert eine Form der Materie, von der wir direkt nichts merken.

Die Neutrinos, die wir aus Laboratoriumsexperimenten kennen, und deren Häufigkeit im heutigen Universum wir berechnen können **(vgl. Abbildung 10)** gehören in diese Klasse der «exotische» Materie. Falls Neutrinos eine wenn auch nur geringe Masse besitzen, könnten sie die beobachtete, zusätzliche Anziehungswirkung erzeugen. Es könnten aber auch viel schwerere «exotische» Teilchen existieren(4).

Wie können Galaxien entstehen?

Für die Zusammenballung der Materie in Galaxien und Galaxienhaufen sind zweifellos Gravitationskräfte verantwortlich. Berechnungen zeigen jedoch, dass diese Zusammenballung gar nicht so einfach zu bewerkstelligen ist. Während der ersten hunderttausend Lebensjahre unseres Universums war die gewöhnliche Materie nämlich an das Photonengas gekoppelt, und dieses bestimmte dannzumal den Verlauf der Expansion. Weil sie masselos sind, zeigten die Photonen jedoch keinerlei Neigung zur Zusammenballung und verhinderten dadurch auch das Zusammenballen der «gewöhnlichen» Materie. Als sich dann, nach etwa hunderttausend Jahren der Expansion, die Kopplung an das Photonengas abschwächte, war die «gewöhnliche» Materie

(4) Vgl. den Artikel «Bringing Dark Matter in from the Dark» von Prof. Klaus P. Pretzl, Leiter des Laboratoriums für Hochenergiephysik an der Universität Bern, in der Zeitschrift «Europhysics News», Band 24, S. 167–171 (1993).

schon zu stark verdünnt, um noch eine für die Bildung von Galaxienhaufen ausreichende Gravitationswirkung zu entfalten.

Die Existenz von «exotischer» Materie im Universum könnte die Entstehung der Galaxien und der Galaxienhaufen entscheidend begünstigt haben. Erstens scheint, wie oben gezeigt, ihre Dichte und damit ihre Gravitationswirkung stärker zu sein als diejenige der «gewöhnlichen» Materie. Zum anderen, und das ist noch wichtiger, ist die Kopplung der «exotischen» Materie an das Photonengas gering. Deshalb kann sie schon relativ früh «ausflocken» und Gravitationszentren bilden, die dann auch die Zusammenballung der «gewöhnlichen» Materie begünstigen.

Die Restwärme im Universum

Wir haben oben erwähnt, dass das Photonengas die Dynamik des frühen Universums entscheidend beeinflusst hat. Durch die mit der Expansion während 14 Milliarden Jahren einhergehende Abkühlung sind die Photonen heute für die Dynamik recht bedeutungslos. Ihre Zahl und mittlere Energie entspricht einer Temperatur von 2.73 Kelvin (d.h. 2.73 Grad über dem absoluten Nullpunkt). Obwohl die Photonen – weil masselos – nicht zur Zusammenballung neigen, beobachten wir doch ganz geringe Anisotropien in der Temperatur des Universums **(Abbildung 16).** Diese spiegeln den Beginn der Zusammenballungen der «gewöhnlichen» Materie und der «exotischen» Materie wider. Mit den sich in der Entwicklung befindlichen Satelliten, MAP der NASA und Planck der ESA, sowie mit Beobachtungen vom Boden aus, sollen die Temperaturanisotropien wesentlich genauer ausgemessen werden. Wir hoffen,

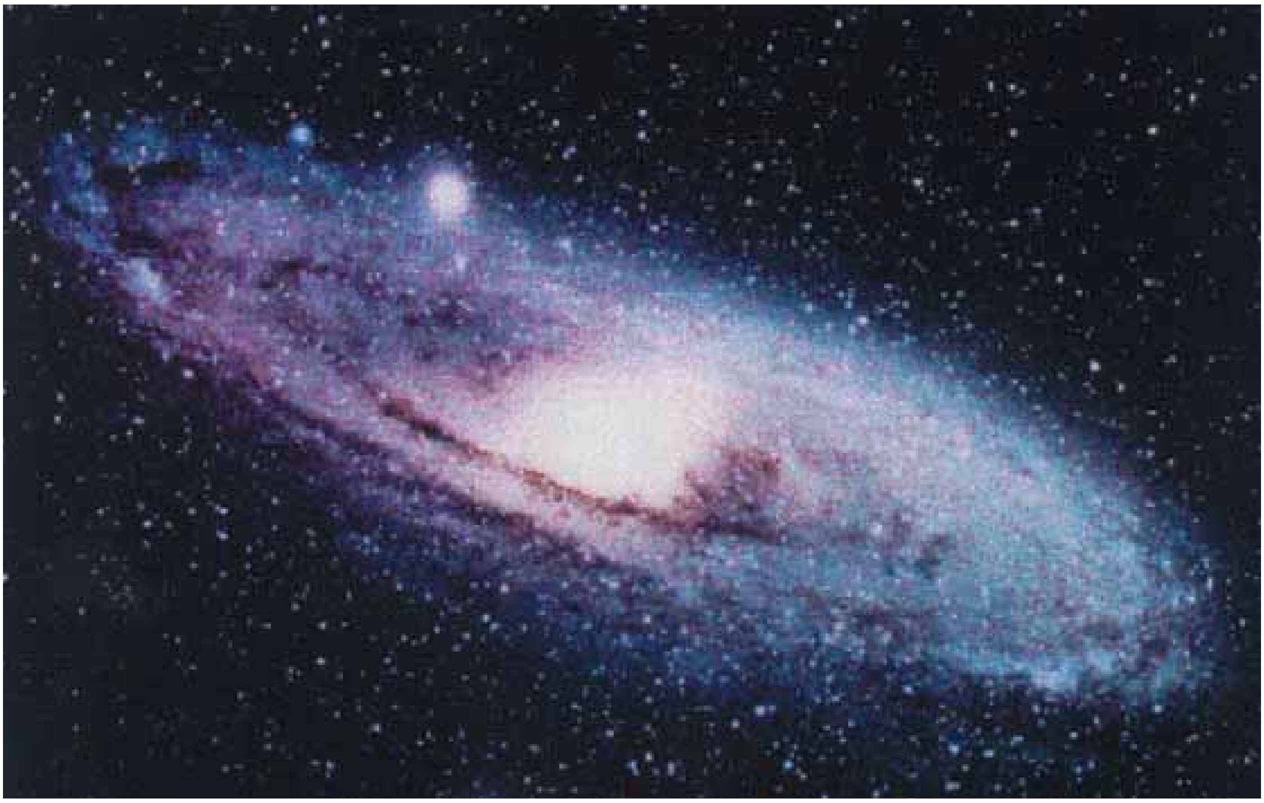


Abbildung 17: Der Andromeda-Nebel
Dies ist die uns am nächsten gelegene voll ausgebildete Galaxie. Sie ist unserer Galaxis bezüglich Grösse und Struktur ähnlich und ist von blossem Auge im Sternbild Andromeda sichtbar. Die Bezeichnung «Nebel» ist irreführend, denn das Licht stammt von 100 Milliarden Sternen.

Zukunftsperspektiven

Verdoppelung des CO_2 in der Luft	in	60 Jahren
Nächste Eiszeit	in	20 000 Jahren
Die Sonne wird zum Roten Riesen	in	5 000 000 000 Jahren
Kollision oder «near miss» mit dem Andromeda-Nebel	in	6 000 000 000 Jahren
Unser Universum fällt in sich zusammen	wahrscheinlich nie	

dass die Resultate eindeutig Auskunft geben werden über die Entstehung der Galaxienhaufen. Insbesondere darüber, ob «exotische» Materie dabei eine wesentliche Rolle gespielt hat, und welcher Art diese «exotische» Materie ist, die wahrscheinlich heute noch das Universum füllt.

Wo stehen wir, was bringt die Zukunft?

Ein paar zukünftige Ereignisse sind im **Kasten «Zukunftsperspektiven»** zusammengestellt.

Von unmittelbarem Interesse für die Menschheit sind die beiden ersten Aussagen. Trotzdem wollen nicht nur Wissenschafter liebend gerne wissen, was eigentlich mit der Sonne und der Erde langfristig passieren wird. Die Wahrscheinlichkeit, dass unser Zentralgestirn einstmals mit einem Stern kollidiert und zerstört wird, ist klein. Hingegen wird sich die Sonne in etwa 5 Milliarden Jahren, wenn ihr Wasserstoffvorrat aufgebraucht ist, in einen Roten Riesen verwandeln. Dabei dehnt sie sich gewaltig aus und verbreitet eine unglaublich starke Strahlung. Das bedeutet das Ende allen Lebens auf der Erde. Als weitere direkte Auswirkung werden die Planeten und deren Atmosphären ihre Eigenschaften ändern, was das Aus für unser Sonnensystem, so wie wir es heute kennen, bedeutet.

Bald darauf, in kosmischer Zeitrechnung gedacht, bahnt sich bereits die nächste Katastrophe an. Der

Abbildung 18: Infrarot-Aufnahme der Galaxis durch den COBE-Satelliten.
Das Bild könnte den Endruck erwecken, es sei von einem Ort ausserhalb unserer Galaxis aufgenommen, was natürlich nicht zutrifft. Aufnahmen dieser Art sind nur deshalb möglich, weil wir uns im äusseren Teil der Galaxis befinden und die Intensität der Infrarotstrahlung gegen das Zentrum hin stark zunimmt.

Abbildung 19: Kollision zweier Galaxien
Sollten unsere Galaxis und der Andromeda-Nebel in sechs Milliarden Jahren kollidieren, könnte sich einem zukünftigen entfernten Betrachter ein Bild dieser Art zeigen. Die Cartwheel Galaxy ist nämlich das Produkt einer Kollision zweier Galaxien.

Andromedanebel **(Abbildung 17)** und unsere Galaxis **(Abbildung 18)** bewegen sich nämlich aufeinander zu. Nach heutigen Beobachtungen werden die beiden in sechs Milliarden Jahren entweder nahe aneinander vorbeifliegen oder gar miteinander kollidieren. Was dann genau passieren wird, ist offen und diese Frage ist ja auch nicht sehr akut. **Abbildung 19** zeigt aber die Auswirkungen einer Kollision zweier Galaxien recht eindrücklich.

Die grosse philosophische Frage, ob sich das Universum immer noch weiter ausdehnen wird oder in sich zurückfällt (der sogenannte Big Crunch) können wir noch nicht schlüssig beantworten. Nach heutiger Einschätzung reichen aber die Anziehungskräfte der Materie nicht aus, um die Expansion des Universums umzukehren.

Schluss-bemerkung

Spezialisierung, aber auch Integration bei den Naturwissenschaften

Es wird heute oft geklagt, die Naturwissenschaft begäbe sich immer mehr auf den Weg der Spezialisierung. Dies ist in mancher Beziehung richtig, aber es sollte auch erkannt werden, dass gegenläufig zur Spezialisierung fortwährend auch ein den Naturwissenschaften inhärenter Integrationsprozess abläuft. Der Grund liegt in der Allgemeingültigkeit der Naturgesetze, die immer wieder zu Vereinheitlichungen von Theorien und zu neuen Querverbindungen zwischen scheinbar streng getrennten Wissensgebieten führt. Eines der wichtigsten Beispiele in neuerer Zeit ist die Molekularbiologie, in der sich Quantenphysik, Chemie und Biologie treffen.

Never say never!

Auch in die Kosmologie werden immer mehr Teildisziplinen und Methoden der Astronomie und Physik einbezogen. Mein Vortrag vermittelt davon vielleicht einen Eindruck. Dies wird sich fortsetzen. Und – wer weiss – vielleicht wird eines Tages ein ganz grosser Schritt getan: Die Suche nach Anzeichen oder Spuren ausserirdischen Lebens wächst allmählich aus dem Stadium der zum Teil wilden Spekulationen und unbegründeten Behauptungen hinaus. Es gibt heute Ansätze und Möglichkeiten der Beobachtung, die es erlauben, mit naturwissenschaftlichen Methoden nach direkten oder indirekten Spuren ausserirdischen Lebens im Sonnensystem und sogar in der Galaxis zu suchen. Sollten diese Versuche konkrete Hinweise oder gar Ergebnisse zeitigen, so würden nicht nur die Biologie, sondern auch viele Geisteswissenschaften näher an die astronomisch-physikalische Kosmologie rücken. Ich persönlich glaube, wir sind noch recht weit davon entfernt, konkrete Evidenz für ausserirdisches Leben zu finden, aber niemand weiss, wie weit.

Ich möchte diesen Artikel Sir Hermann Bondi widmen. Dieser hat während mehr als vierzig Jahren wichtige wissenschaftliche Arbeiten und tiefe Gedanken zur Kosmologie und zur Kosmogonie beigetragen. In seiner Zeit als Director General der Europäischen Weltraumforschungsorganisation ESRO hat er Astronomie und Astrophysik in das Wissenschaftsprogramm aufgenommen. Schliesslich hat Sir Hermann bei der Gründung des International Space Science Institute Pate gestanden, er hielt am Eröffnungsabend die Festrede.

Ich danke Frau Dr. Kathrin Altwegg für ihren Rat und ihre Hilfe bei der Vorbereitung meines Vortrags und Herrn Prof. Rudolf Treumann für die kritische Durchsicht des Manuskripts. Herrn Peter Abgottspon danke ich für die Redaktion und Herrn Dr. Hansjörg Schlaepfer für die Herausgabe dieses Artikels.

Zum Autor

Prof. Johannes Geiss wurde 1926 in Pommern geboren, promovierte 1953 an der Universität Göttingen und habilitierte sich 1957 nach mehreren Forschungsaufenthalten in den USA an der Universität Bern. Nach einer Professur für Ozeanwissenschaften an der Universität von Miami wurde er 1960 Professor, ab 1966 Direktor des Physikalischen Instituts und 1982 / 83 Rektor der Universität Bern.

In diese Zeit fallen auch weitere Forschungsaufenthalte in Frankreich und den USA. Die Öffentlichkeit verbindet den Namen von Prof. Geiss mit dem Sonnenwind-Segel, das die amerikanischen Astronauten 1969 als erstes Experiment auf dem Mond in Betrieb genommen haben.

Seither war und ist Professor Geiss massgeblich an wissenschaftlichen Projekten der NASA und der Europäischen Weltraumorganisation ESA beteiligt.

Johannes Geiss ist Gründer des Vereins Pro ISSI und heute geschäftsführender Direktor des International Space Science Institutes ISSI in Bern. ISSI wird von der gleichnamigen durch Contraves Space in Zürich gegründeten Stiftung getragen. Das Institut wird hauptsächlich von der Europäischen Weltraumorganisation ESA, von der Schweizerischen Eidgenossenschaft, vom Kanton Bern sowie vom Schweizer Nationalfonds finanziert. ISSI widmet sich der interdisziplinären Erforschung des Sonnensystems und des Universums, indem es Wissenschaftler aus aller Welt zu Arbeitstagungen und Workshops einlädt und ihnen Gelegenheit gibt, ihre Erkenntnisse und die neuesten Ergebnisse der Forschung auszutauschen und zu bewerten. Grundlagen dieser Forschungstätigkeiten sind die Weltraummissionen der Raumfahrtagenturen Europas, der USA, Russlands und Japans, aber auch Experimente im Labor und Simulationen auf dem Computer.

INTERNATIONAL SPACE SCIENCE INSTITUTE
Nr. 2, November 1998

SPATIUM

Publikationsorgan des Vereins Pro ISSI

Das neue Bild der Sonne

Seit 4,6 Milliarden Jahren scheint die Sonne, aber erst seit kurzem lernen wir ihre Wirkungsmechanismen zu verstehen und damit auch ihre Entstehungsgeschichte nachzuvollziehen.

Editorial

Die jeweiligen Vorstellungen über das Zentrum der Welt symbolisieren wie kaum eine andere Thematik das Erwachen des menschlichen Geistes. In der Antike war es die Erde, die es zu entdecken galt und die dem Menschen seine volle Aufmerksamkeit abverlangte, um zu überleben und naturgemäss war sie das Zentrum seines Denkens. Kopernikus im 15. Jahrhundert und andere Forscher danach zeigten, dass die Erde nur einer von neun Planeten ist, die um die Sonne kreisen und diese damit als Zentrum zu gelten hat. Später erkannte man, dass auch die Sonne nur einer von Milliarden von Sternen unserer Galaxis und diese wiederum nur eine von unzählbar vielen Galaxien ist. Mit einem überraschenden Ergebnis hat schliesslich Edwin Hubble in der ersten Hälfte dieses Jahrhunderts die Suche beendet, der die Expansion des Universums erkannte und damit auch, dass die Welt gar kein Zentrum hat.

Die Suche nach einem materiellen Zentrum ist ergebnislos verlaufen. Dafür steht fest, dass wir der Sonne die Energie verdanken, die im Anfang aus anorganischen Molekülen immer komplexere Bausteine des Lebens auf der Erde geformt und damit eine Evolution in Gang gebracht hat, deren vorläufiges Ergebnis der Mensch und seine belebte Umwelt ist. In diesem Sinne bleibt die Sonne eben doch Ursache und Zentrum der uns zugänglichen Welt.

Was liegt dem Wissenschaftler daher näher, als die Sonne zu erforschen und ihre Wirkungsweise zu verstehen zu suchen? Die Weltraumwissenschaften haben in den vergangenen Jahren manche Raumsonde geschaffen und damit unsere Kenntnisse auch über die Sonne ganz wesentlich erweitert. Die Sonne ist daher auch ein Schwerpunkt der Forschung am International Space Science Institute in Bern. PD Dr. Rudolf von Steiger, Senior Scientist am ISSI, hat es unternommen, das neue Bild, das wir uns heute von der Sonne machen, für unsere Mitglieder zu skizzieren und uns auf eine Reise mitzunehmen, die vom heissen Innersten der Sonne über ihre Korona bis zu dem auf der Erde beobachtbaren Licht führt.

Hansjörg Schlaepfer
Bern, im November 1998

Inhalt

Impressum

SPATIUM
Publikationsorgan des Vereins Pro ISSI
Erscheint ein- bis zweimal jährlich

INTERNATIONAL SPACE SCIENCE INSTITUTE

Verein Pro ISSI
Hallerstrasse 6, 3012 Bern
Telefon 031 631 48 96
Fax 031 631 48 97

Präsident
Prof. Hermann Debrunner, Universität Bern
Herausgeber / Quästor
Dr. Hansjörg Schlaepfer, Contraves Space, Zürich
Redaktion dieses Beitrags
Peter Abgottspon, Thun
Graphik
Gian-Reto Roth, Contraves Space, Zürich
Druck
Drucklade AG, 8008 Zürich

Titelbild:
Heisse Plasmaströme entweichen aus der Sonne in den Weltraum. Aufnahme: Extreme Ultraviolet Imaging Teleskop des Solar Heliospheric Observatories (ESA/NASA).

Das neue Bild der Sonne

Rudolf von Steiger, International Space Science Institute, Bern
(Vortrag für den Verein Pro ISSI am 1. April 1998)

Die Sonne im Zentrum

Seit urdenklichen Zeiten schon ist die Sonne Gegenstand mystischer Verehrung und religiösen Kultes. Auch in der Gegenwart hat die Sonne nichts von ihrer Faszination verloren. Im Gegenteil, die moderne Weltraumforschung hat deutlich gemacht, welch intensive Beziehungen zwischen unserem Zentralgestirn und den Vorgängen auf der Erde herrschen. Der Sonne war deshalb ein Workshop gewidmet, der Anfang 1998 am ISSI stattfand und über dessen Ergebnisse im folgenden berichtet werden soll.

Wesentliche Fortschritte in der Erforschung der Sonne konnten erzielt werden, seit Sonden im Weltraum zur Verfügung stehen, welche ungestört von der Atmosphäre mit verschiedensten Instrumenten unseren nächsten Stern beobachten und ihre Daten zur Erde melden. Damit können neben dem auf die Erde auftreffenden Sonnenlicht, das seit Jahrhunderten beobachtet wird, nun auch der von der Sonne in den Weltraum abgestrahlte Teilchenstrom und ihr Magnetfeld gemessen werden. Daraus lassen sich Rückschlüsse ziehen auf die Bestandteile der Sonne, die deshalb von grundlegender Bedeutung sind, weil sie – im Gegensatz etwa zu derjenigen der Planeten – die Zusammensetzung der «Ursuppe» darstellen, aus welcher sich vor 4,6 Milliarden Jahren unser Sonnensystem gebildet hat. Abweichungen von dieser Urzusammensetzung, wie sie offensichtlich in den Planeten, aber auch (in wesentlich kleinerem Ausmasse) in Korona und Sonnenwind vorkommen, sind der Schlüssel zu seiner Entstehungs- und Entwicklungsgeschichte.

SOHO, Ulysses und Co.

Die beiden wichtigsten Weltraumsonden für die Sonnenforschung sind SOHO (Solar Heliospheric Observatory) und Ulysses, beides Gemeinschaftsprojekte der europäischen und der amerikanischen Weltraum-Agenturen ESA und NASA, siehe Fussnote. SOHO wurde im November 1995 gestartet und hat seither die Sonne ununterbrochen beobachtet. Es liefert uns Daten vom Sonneninnersten, der Sonnenoberfläche und -atmosphäre im ultravioletten und im sichtbaren Licht, sowie auch vom interplanetaren Raum. Ulysses ist der erste künstliche Himmelskörper, der die Ekliptik (die Ebene, in welcher die Planeten um die Sonne kreisen) verlassen hat und nun bereits seit über sieben Jahren die Sonne aus einer polaren Umlaufbahn beobachtet.

Vom Kern der Sonne bis zu den äussersten Zonen der Korona

Unsere gedankliche Reise beginnt im innersten Kern der Sonne. Doch wie ist es möglich, in den Kern der Sonne hineinzusehen? Natürlich können wir nicht einfach einen Achtel der Sonne wegschneiden, wie uns dies **Abbildung 1** (s. folgende Seite) vormacht.

Es gibt zwei Phänomene, welche uns Auskünfte über das Innere der Sonne geben. Zum ersten Modell verhelfen uns folgende Informationen: im Inneren der Sonne wird durch die Fusion von Wasserstoff Helium produziert. Der Sonnenkern ist also ein riesiger Fusionsreaktor, in welchem bei einer Temperatur von etwa 14 Millionen Grad ungeheure Energiemengen freigesetzt werden. Diese Energie bahnt sich in Form von Strahlung langsam ihren Weg an die Sonnenoberfläche. In einer bestimmten Distanz, etwa bei zwei Dritteln des Sonnenradius, manifestiert sich ein derart grosser Wärmeunterschied zwischen dem inneren und äusseren Bereich, dass der weitere Energietransport an die Oberfläche nur noch mittels Konvektion (zu vergleichen mit brodelndem Wasser, wenn es siedet) erfolgen kann. Dieser kontinuierlich ablaufende Prozess regt die ganze Sonne zu Schwingungen an.

Im Internet sind auf den Websites von SOHO (http://sohowww.nascom.nasa.gov/) und Ulysses (http://helio.estec.esa.nl/ulysses/) jederzeit Informationen und Bilder verfügbar.

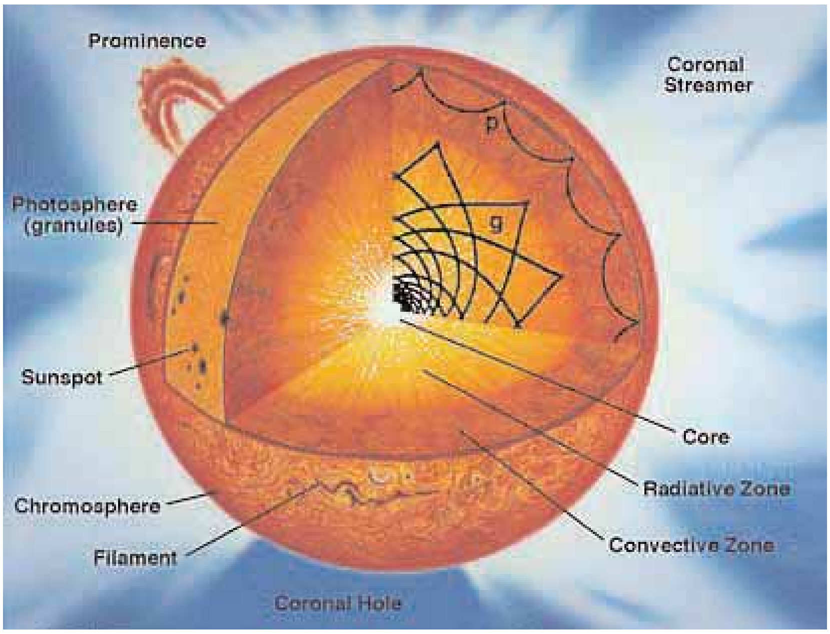


Abbildung 1: Schematischer Blick ins Innere der Sonne, wo einzelne Wege von Schallwellen angedeutet sind. Nicht jede beliebige Welle, sondern nur ganz bestimmte Kombinationen von Wellenlänge und Frequenz bringen den Sonnenkörper in Resonanz, wobei die exakten Werte von den Bedingungen im Sonneninneren abhängen. Sonnenoszillationen liefern damit ein Fenster zum unsichtbaren Innern, zu Struktur und Zusammensetzung des Sonnenkörpers. SOHO/MDI/SOI

Die Sonne schwingt

Unser Muttergestirn verhält sich daher wie ein riesiger, dreidimensionaler Gong. Wie aber kommen diese Oszillationen zustande? Betrachten wir einen Punkt an der Sonnenoberfläche: Er wird, wie erwähnt, von der Konvektion nach oben gedrückt, fällt darauf nach unten und regt so eine Welle an. Diese pflanzt sich ins Innere der Sonne fort, trifft auf immer grössere Dichten und wird deshalb gebeugt. An einer anderen Stelle tritt sie erneut an die Oberfläche. Dort wird die Welle reflektiert und so fort: es entsteht eine in der Sonne umlaufende Welle. Weist die Welle eine bestimmte Länge auf, trifft sie nach einem Umgang genau an der gleichen Stelle wieder auf und überlagert sich sozusagen mit sich selbst. Die Frequenzen dieser Wellenart kann man durch Beobachten der Schwingungen an der Sonnenoberfläche feststellen. Weil hier der Druck gegenüber der Auslenkung die rückstellende Kraft ist, heisst sie Druckwelle oder p-Welle (p ist die übliche Abkürzung für Druck).

Die Sonne als dreidimensionaler Gong könnte in einem speziellen Falle etwa so aussehen wie in **Abbildung 2:** Wir sehen eine p-Welle, welche mit den drei Quantenzahlen n = 14, l = 20 und m = 16 beschrieben ist. Quantenzahlen heissen sie, weil sie die dreidimensionale Schwingung einer Kugel mit denselben Mitteln beschreiben, wie die Atomhülle in der Quantenmechanik. n beschreibt die Richtung von innen nach aussen, l die azimutale Richtung und m die Richtung von Pol zu Pol. Rot sind Gebiete, die auf uns zukommen, blau solche die sich von uns weg bewegen. Auf diese Weise kann man die Schwingung der Sonnenoberfläche beobachten und deren Frequenz messen.

Würde die Sonne nur in einem einzigen der möglichen Moden schwingen, ergäbe sich ein relativ einfaches Bild: in **Abbildung 2** ist ein solcher reiner Schwingungsmode dargestellt. In diesem Fall hat MDI eine Frequenz von 2935,88 ± 0,2 Mikroherz gemessen. Diese unglaubliche Genauigkeit ist nur möglich, weil sich SOHO auf dem Lagrange-Punkt L1 zwischen der Erde und der Sonne befindet und diese ohne zeitliche Unterbrechungen beobachten kann. Nun schwingt die Sonne aber nicht nur

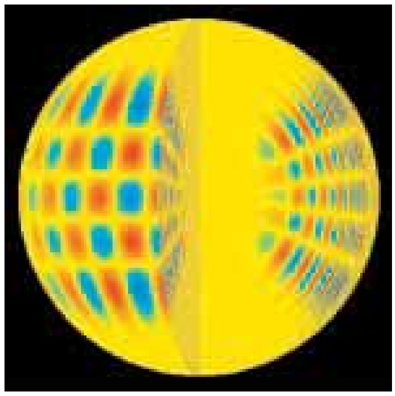


Abbildung 2: Computersimulation einer reinen Sonnenoszillation (p-Welle) der radialen Ordnung n = 14, azimutalem Grad l = 20, und polarer Ordnung m = 16. Rote und blaue Bereiche bedeuten Verschiebungen mit unterschiedlichem Vorzeichen. Die Frequenz dieses Schwingungsmodes, bestimmt durch MDI, beträgt 2935,88 ± 0,2 Mikroherz. SOHO/MDI/SOI

in einem einzigen, sondern in sehr vielen der möglichen Schwingungsmoden gleichzeitig. Die Situation entspricht etwa einer angezupften Gitarrensaite, bei welcher Grund- und Obertöne gleichzeitig mitschwingen. Dies allerdings mit dem Unterschied, dass die Sonne als dreidimensionaler Gong über sehr viel mehr Schwingungsmoden verfügt als die eindimensionale Saite. Wenn MDI die Sonnenoberfläche beobachtet, sieht es also sämtliche Moden gleichzeitig. Daraus entstehen Bilder wie das Full-Disk Dopplergramm in **Abbildung 3.**

Abbildung 3: Dopplergramm
der schwingenden Sonnenoberfläche. Die Schwingungen der Sonne werden durch ihre Konvektion angeregt. Anders als in Abb. 2 sind alle Schwingungsmoden gleichzeitig aktiv, was zu einer komplizierten, rasch veränderlichen Oberflächenstruktur führt.
SOHO/MDI/SOI

Ausser der orangenhautartigen Oberfläche der Sonne ist darin eigentlich nicht viel zu sehen. Doch bei der Live-Aufnahme ist deren rasche Veränderung als «Brodeln» gut zu beobachten. Zudem kommt, bedingt durch die Sonnenrotation, der östliche Rand mit 2 km/sec auf uns zu, und der westliche entfernt sich mit der gleichen Geschwindigkeit. Subtrahiert man die Sonnenrotation, bleiben die Schwingungen übrig. Man muss nun die Frequenzen der einzelnen Schwingungen aus der brodelnden Sonne herausfiltern. Es ist unschwer, sich vorzustellen, welchen Arbeitsaufwand dies mit sich bringt: Jeder Bildpunkt muss einzeln beobachtet und seine Schwingung gemessen werden. Kommt er auf uns zu, sind seine Spektrallinien blau –, entfernt er sich von uns, sind sie rotverschoben. Diese Frequenzverschiebungen zeigen uns, wie schnell die Oberfläche an der fraglichen Stelle oszilliert. Erstellt man eine Analyse dieser Geschwindigkeiten als Funktion der Zeit, ergeben sich damit die Frequenzen aller schwingenden Moden aufs Mal. Es ist zwar unmöglich, einen Modus einzeln zu beobachten, aber durch die Frequenzanalyse können wir sie auseinanderhalten. Eine Zeitreihe solcher Bilder liefert uns also die einzelnen Schwingungsmoden und ihre Frequenzen **(Abbildung 4, 5).**

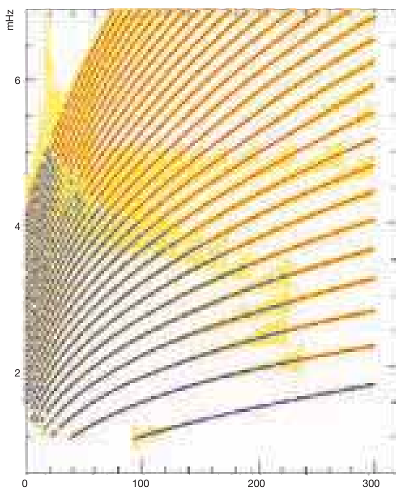


Abbildung 4: Das Frequenzspektrum
der Sonne. Jedem Schwingungsmode (horizontal) mit seiner Frequenz (vertikal) entspricht ein Punkt in diesem Diagramm. Die Fehlerbalken (gelb) sind stark überzeichnet und mit einem Faktor 100, teilweise gar mit einem Faktor 1000, multipliziert. Die mit SOHO erzielte Genauigkeit ist also phantastisch.
SOHO/MDI/SOI

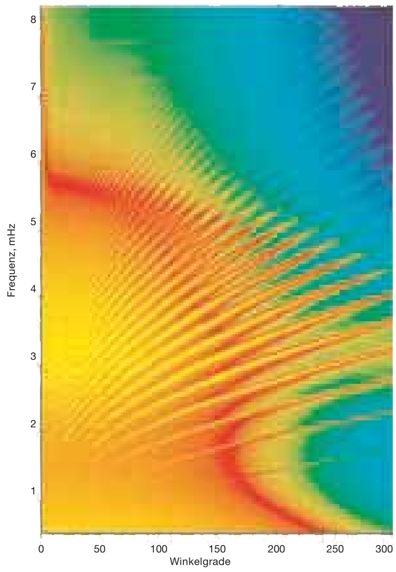


Abbildung 5:
Dieses Bild entspricht dem Diagramm in Abb. 4, wobei aber an jeder Stelle die Energie farbcodiert aufgezeichnet wurde, welche im dortigen Mode steckt. In den hellen Bereichen steckt viel Energie, in den dunklen weniger. Man sieht, dass die Schwingungen hauptsächlich im Bereich zwischen 2 und 4 Millihertz auftreten, was ungefähr einer Schwingungsdauer von etwa 5 Minuten entspricht.
SOHO/MDI/SOI

Der Weg ins Innere

Bisher haben wir nur über Messungen an der Oberfläche der Sonne gesprochen. Wie aber kann man daraus auf ihr Inneres schliessen? Jeder Schwingungsmode pflanzt sich durch den Sonnenkörper fort und wird von den Bedingungen beeinflusst, die er unterwegs antrifft. Er enthält insbesondere Information über den Zustand der Sonne in der

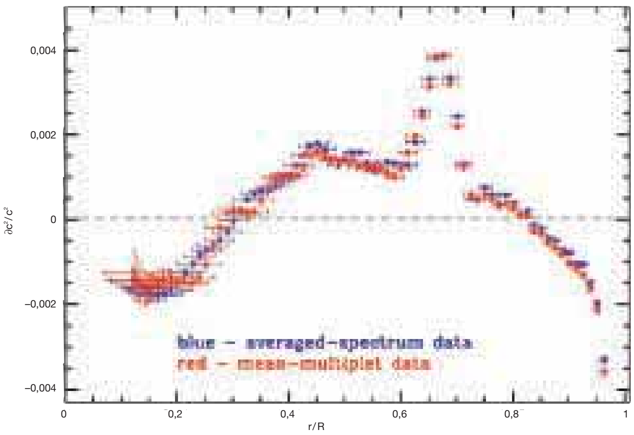


Abbildung 6:
Auf der vertikalen Achse ist das Quadrat der Schallgeschwindigkeit in Funktion vom Zentrum (links) bis zur Oberfläche der Sonne (rechts) dargestellt. Die Schallgeschwindigkeit ist deshalb gewählt, weil sie die Parameter Dichte, Temperatur und Heliumgehalt zusammenfasst. Die roten und blauen Punkte unterscheiden sich nur durch die Art wie die Daten gemittelt wurden, was hier nicht bedeutungsvoll ist.
SOHO/MDI/SOI

tiefsten Schicht, bis zu welcher er vordringt, bevor er wieder zur Oberfläche hin gebeugt wird. Weil verschiedene Moden bis zu grösseren oder kleineren Tiefen vordringen, ergibt sich so ein Tiefenprofil der Bedingungen im Sonneninneren.

Damit präsentiert sich uns folgende Situation: die Sonnenbeobachter sagen uns, wie die Oberfläche schwingt. Jetzt treten die Theoretiker oder «Sonnenmodellierer» in Aktion und konstruieren ein mathematisches Modell, welches den Sonnenkörper bestmöglich beschreibt. Es muss das Temperatur- und das Dichteprofil als Funktion des Radius vom Sonneninnersten nach aussen bis zur Oberfläche enthalten. Besonders wichtig ist auch die Zusammensetzung der Sonne, vor allem das Verhältnis von Helium zu Wasserstoff. Aus diesen beiden Elementen besteht die Sonne ja zur Hauptsache, dazu aus einer kleinen Beimischung von schwererem Material, welches die Astronomen unter dem Begriff «Metalle» zusammenfassen.

Nach dem Bau eines Sonnenmodells lässt der Theoretiker sein Werk auf dem Computer schwingen und vergleicht danach das Ergebnis mit den Messungen von SOHO. Dabei muss er den Heliumanteil, die Dichte und die Temperatur in seinem Modell so festlegen, dass die Übereinstimmung zwischen Modell und den gewonnenen Daten möglichst hoch sind. In **Abbildung 6** ist die Abweichung zwischen Modell und Beobachtung als Funktion des Sonnenradius dargestellt. Auf den ersten Blick hat man den Eindruck, es bestehe eine grosse Abweichung zwischen modellierten und gemessenen Werten. Man beachte aber die vertikale Achse: Es handelt sich hier um Bruchteile eines Prozents! Es darf füglich als sensationell bezeichnet werden, dass die besten Sonnenmodelle mit den Resultaten von SOHO bis auf ein paar Promille übereinstimmen! Die grösste Abweichung beträgt 0,4%, und zwar gerade dort, wo die Sonne von Strahlungsenergietransport auf Konvektionsenergietransport umschaltet, also am unteren Rand der Konvektionszone. Es ist gut möglich, dass die Modelle an dieser Stelle nicht vollständig ausgereift und deshalb noch verbesserungsfähig sind.

Es gibt noch eine weitere Möglichkeit, ins Sonneninnere hinein zu sehen. Ich muss hier eingestehen, dass ich vorhin nicht die vollständige Wahrheit gesagt habe. Ein Schwingungsmode, wie er in **Abbildung 2** mit den roten und blauen Bäuchen und Wellentälern gezeigt wurde, schwingt nicht mit einer einzigen Frequenz, sondern man erkennt bei genauer Beobachtung, dass diese sich aus mehreren, ganz fein aufgespaltenen Frequenzen zusammensetzt. Die Ursache dieser Aufspaltung liegt in der Rotation der Sonne. Stände sie still, würde sie mit einer einzigen exakt definierten Frequenz pro Schwingungsmode schwingen. (Genauer: alle Moden mit den selben Quantenzahlen n und l schwingen unabhängig von m mit der selben Frequenz). Sie rotiert aber (einmal in fünf- bis sechsundzwanzig Tagen) um ihre eigene Achse, was die Aufspaltung der Schwingungen mit verschiedenen Quantenzahlen m bei gleichem n und l bewirkt. Man kann diesen Effekt mit einem einfachen Experiment zeigen. Schlägt man eine ruhende Kuhglocke an, laufen zwei Wellen gleicher Frequenz gegenläufig rund um die Glocke. Sie überlagern sich gegenseitig und bilden eine sogenannte stehende Welle, welche man als reinen Glockenton wahrnimmt. Schlägt man aber die rotierende Glocke an, läuft die eine Welle mit der Glockenrotation (und damit etwas schneller), und die andere ihr entgegen (somit langsamer). Deshalb haben beide ganz leicht unterschiedliche Frequenzen, was man als Schwebung (d. h. ein an- und abschwellen der Lautstärke) wahrnimmt. Indem man die Frequenzaufspaltung beobachtet, kann man also auf die Rotation schliessen **(Abbildung 7).**

Die Sonnenoberfläche rotiert am Äquator etwas schneller als an den Polen und diese differentielle Rotation setzt sich im Sonneninneren fort. Etwas unterhalb der Konvektionszone verschwindet sie aber und wird zu einer stabilen Rotation mit einer einzigen Rotationsgeschwindigkeit. Man kann also sagen, dass der Kern der Sonne wie ein fester Körper rotiert und zwar mit einer Frequenz, die jener an der Sonnenoberfläche bei 30 Grad Breite entspricht.

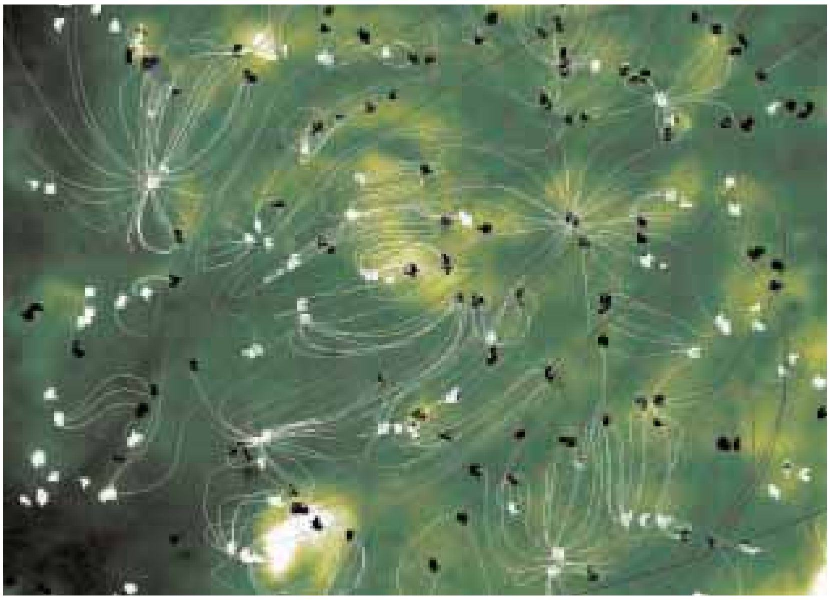


Abbildung 8: Die Sonnenoberfläche
wird auf kleinsten räumlichen Skalen von magnetischem Fluss durchdrungen (schwarze und weisse Stellen im kleinen Bild). Das Magnetfeld bildet so einen eng geknüpften Teppich. Die beständige Bewegung ihrer Fusspunkte bringt gegenläufige Feldlinien in Kontakt, und dies bewirkt Rekonnektion, eine Neuordnung der Feldlinien. Die dadurch freigesetzte Energie ist wahrscheinlich für die Heizung der Korona auf 1 bis 2 Millionen Grad verantwortlich. SOHO/MDI/SOI

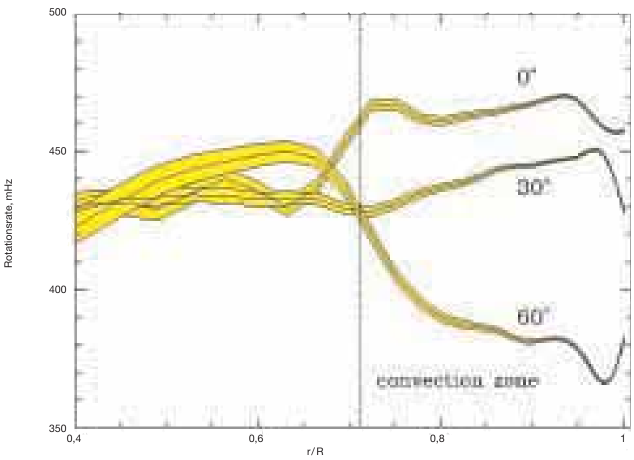


Abbildung 7: Rotationsgeschwindigkeit im Sonneninneren,
ermittelt aus den Frequenzaufspaltungen gemessen mit MDI. An ihrer Oberfläche (rechts im Bild bei r/R = 1 rotiert die Sonne, abhängig von der heliographischen Breite, mit unterschiedlichen Geschwindigkeiten, doch etwa am Grund der Konvektionszone (Mitte, bei r/R = 0,7) verschwinden diese Unterschiede, und der Kern (links) rotiert wie ein fester Körper überall gleich schnell. SOHO/MDI/SOI

Die innere Korona

Verlassen wir nun das Sonneninnere und kommen an die Oberfläche. Ein schönes Resultat von SOHO soll hier speziell erwähnt werden: mit MDI kann man von der Sonnenoberfläche Magnetogramme, d. h. Karten der Stärke und Richtung des Magnetfeldes, erstellen und zwar mit einer Auflösung, die von der Erde aus wegen der Luftunruhe in der Atmosphäre nicht möglich wäre.

Wie ein Teppich …

An den weissen Stellen in **Abbildung 8** treten Magnetfeldlinien aus der

Sonne heraus, an den schwarzen Stellen dringen sie wieder in die Oberfläche hinein. Offenbar ist dies jeweils an einzelnen konzentrierten Stellen der Fall. Versucht man sich das Bild dreidimensional vorzustellen, bedeutet dies, dass dazwischen Magnetfeldschleifen die Sonnenoberfläche bedecken. Das Magnetfeld scheint demnach einem eng geknüpften Teppich zu gleichen, der die Sonnenoberfläche bedeckt. Neu an dieser Beobachtung ist, dass die Knüpfdichte dieses Teppichs offenbar sehr kleinräumig ist. Von der Erde aus wäre das Magnetogramm lediglich als graue Fläche erkennbar, weil wegen der Luftunruhe die weissen und schwarzen Flecken ineinander verschwimmen würden. Dank SOHO ist es nun aber möglich, diese einzelnen Magnetfeldkonzentrationen aufzulösen. Das ist deshalb von Interesse, weil diese Fusspunkte der Magnetfeldschleifen natürlich nicht ruhig bleiben. Die Konvektion spielt auch hier eine wichtige Rolle: sie verwirbelt diese «Knoten» des Teppichs andauernd untereinander. Die Teppichschlaufen verwirren sich mit der Zeit immer stärker miteinander, was unweigerlich dazu führen wird, dass gegenläufige Magnetfeldlinien miteinander in engen Kontakt kommen. Nun können sich aber die Feldlinien – genau wie Teppichschlaufen – nicht gegenseitig durchdringen, sondern sie werden sozusagen entzweigeschnitten und in einfacherer Konfiguration wieder zusammengesetzt (Rekonnektion). Dadurch wird magnetische Energie in Wärme umgewandelt, und genau diese thermische Energie ist wahrscheinlich für die Aufheizung der Korona verantwortlich. Gemäss den Beobachtungen von MDI existieren zu jeder Zeit etwa 10 000 Schleifen, von denen jede im Durchschnitt etwa 40 Stunden bestehen bleibt (der Teppich wird also etwa alle 2 Tage vollständig erneuert). Jede einzelne Schleife enthält etwa soviel Energie, wie sämtliche AKW der Schweiz zusammen in 1 Million Jahren liefern würden. Es war bisher ein grosses Rätsel, wie es die «kalte» Sonnenoberfläche (sie ist «nur» ein paar Tausend Grad heiss!) schafft, die darüberliegende Korona auf 1 bis 2 Millionen Grad aufzuheizen. Wahrscheinlich wird magnetische Feldenergie durch Rekonnektion oder Neukonfiguration des magnetischen Teppichs in thermische Energie umgewandelt.

Löcher in der Korona

Das «Extreme Ultraviolet Imaging Telescope» EIT beobachtet die Sonnenatmosphäre in vier ganz schmalen Frequenzbändern im extremen Ultraviolettbereich **(Abbildung 9).** Die im Bild sichtbaren Farben sind künstlich erzeugt, um die einzelnen Frequenzbänder sichtbar zu machen. Sie zeigen uns die etwa 1 Million Grad heisse untere Korona. Die aktiven Regionen sind darauf unschwer als helle Gebiete zu sehen.

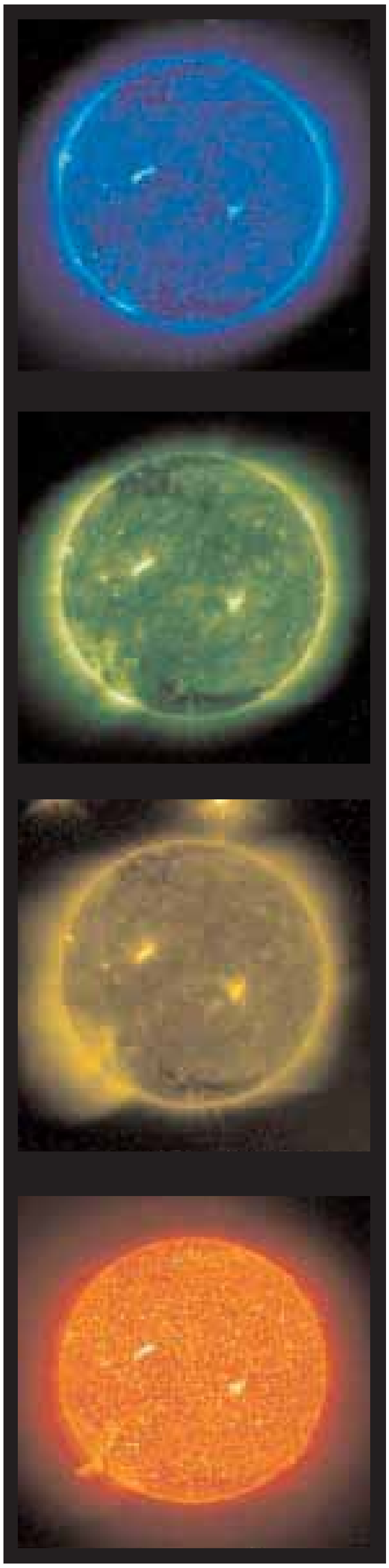


Abbildung 9: Beobachtung von 4 schmalen Frequenzbändern im extremen Ultraviolett durch EIT, nämlich bei einer Frequenz von 171 Angström im Licht von Eisen IX und X (d. h. acht- und neunfach geladenen Eisenionen), bei 196 A (Eisen XII), bei 284 A (Eisen XV), und bei 304 A (Helium II, also einfach geladenem Helium). Deutlich sind die Koronalöcher (Coronal Holes) als dunkle Gebiete an den Polen zu erkennen. SOHO/EIT

Besonders schön sind in diesem Licht die koronalen Löcher zu erkennen: Wie bei der Erde sind an den Polen der Sonne Polkappen erkennbar, nämlich die dunklen, scharf begrenzten Koronalöcher. (Der Vergleich mit der Erde ist entfernt zutreffend. Zwar ist die Korona an den Polen immer noch 1 Million Grad heiss, aber die Polkappen sind, ähnlich wie bei uns, deutlich kühler als die äquatorialen Gebiete. Dies allerdings aus ganz anderen Gründen als bei uns, wo die flache Sonneneinstrahlung an den Polen für die Kälte verantwortlich ist.) EIT liefert aber nicht nur ein stehendes Bild, sondern Bildfolgen, welche die ganze Dynamik dieser Korona ausgezeichnet wiedergeben. Man kann verfolgen, wie sich grosse Strukturen in der Korona mitunter innert Stunden völlig verändern und neu formieren.

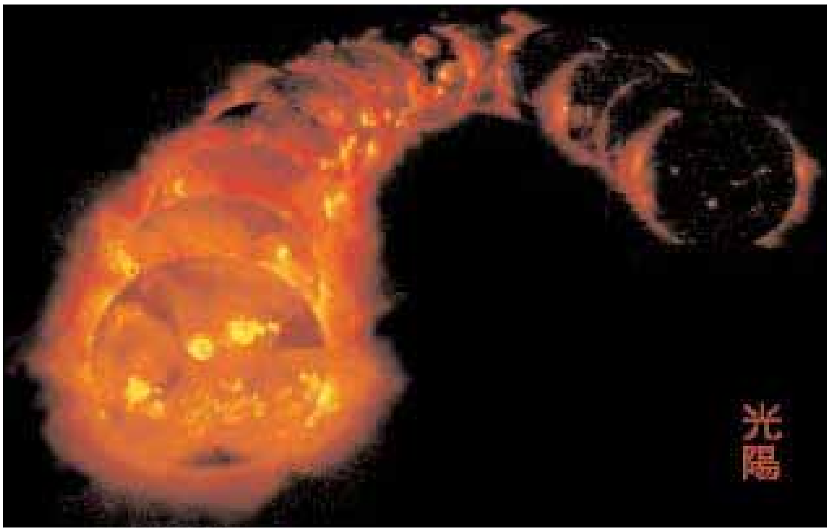


Abbildung 11: Aufnahmen der japanische Weltraumsonde Yohkoh
im weichen Röntgenlicht. Die Bilder entstanden in Abständen von 120 Tagen zwischen 1991 (links) und 1995 (rechts). Deutlich ist die abnehmende Sonnenaktivität zu erkennen. Yohkoh

Die Bilder von EIT können sogar für eine direkte Temperaturmessung in der Korona herangezogen werden: Das Verhältnis zweier Bilder, jener der Eisen-Spektrallinien bei 171 und 195 Angström liefert ein Bild der Korona, in welchem die heissen Gebiete heller, die kühleren dunkler erscheinen **(Abbildung 10).** Deutlich sind die aktiven Gebiete zu erkennen, und ihre Struktur tritt im Differenzbild besonders schön hervor. Zur Zeit existieren jedoch nur wenige solcher Aktitvitätszonen, da sich die Sonne praktisch in der ruhigsten Phase ihre Zyklus befindet. Eine andere Weltraumsonde, der japanische Satellit Yohkoh (was soviel wie «Sonnenlicht» heisst), liefert bereits seit vielen Jahren Bilder der Korona im weichen Röntgenlicht (also etwas kurzwelliger als die vorhergehenden Bilder von EIT), die ähnliche Informationen enthalten.

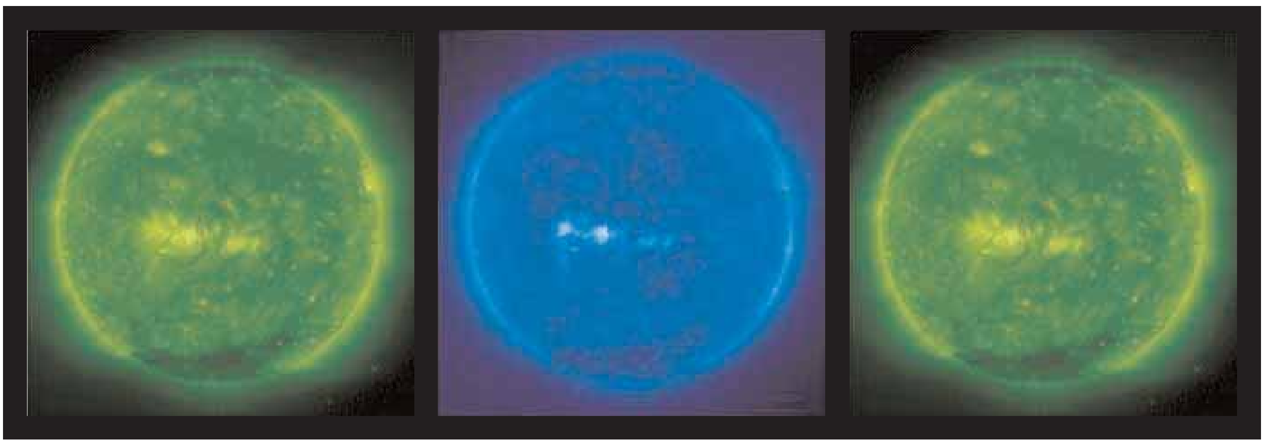


Abbildung 10: Das Verhältnis
der beiden Bilder von EIT bei 195 A (links) und bei 171 A (Mitte) liefert ein Bild der Temperaturverteilung in der Korona (rechts), wobei kühlere Regionen dunkler, heissere dagegen heller erscheinen. In der Mitte des Bildes ist deutlich eine grosse, aktive Region zu erkennen, mitsamt ihrer zusammengesetzten Struktur aus heissen Bogenstrukturen und kühleren Zwischengebieten. SOHO/EIT

Abbildung 12: Aktive Gebiete
auf der Sonne können auch im sichtbaren Licht als Sonnenflecken ausgemacht werden. Das Bild links zeigt die Sonne im sichtbaren Licht, rechts eine Aufnahme von Yohkoh zur gleichen Zeit im weichen Röntgenlicht. Die hellen aktiven Gebiete entsprechen eindeutig den Sonnenfleckengruppen. Yohkoh

Man benötigt jedoch nicht ein Röntgenteleskop, um die Phase des Sonnenzyklus festzustellen. Die Sonnenflecken sind seit Jahrhunderten bekannt und seit den Beobachtungen von H. Schwabe um 1830 auch ihr elfjähriger Zyklus. Seit damals und bis heute ist die Zürcher Sonnenfleckenrelativzahl ein wichtiger Indikator der Sonnenaktivität.

Die äussere Korona

Der Koronagraph LASCO (Large Aperture Solar Coronagraph) lässt in **Abbildung 13** die äussersten Schichten der Sonnenkorona hervortreten. Die Sonne selbst ist mit einer Scheibe abgedeckt, welche hier etwa den dreifachen Sonnendurchmesser aufweist. Bei einer natürlichen Sonnenfinsternis kann man die Korona auf ähnliche Weise beobachten, wobei der Mond die Aufgabe der Scheibe übernimmt. Dieses Bild der Korona ist für die gegenwärtig minimale Sonnenaktivität charakteristisch: Man erkennt grosse koronale Strukturen, «Streamers» genannt, über den äquatorialen Gebieten der Sonne einerseits und eine Leere oberhalb der koronalen Löcher andererseits. Man sieht dort schwache Strahlen, die darauf schliessen lassen, dass das Magnetfeld der Sonne dort offen ist und die Korona frei in den interplanetaren Raum ausströmen kann. Die von links unten gegen die Koronagraphenscheibe hin gerichtete gekrümmte Spur ist übrigens ein Komet, der von SOHO ent-

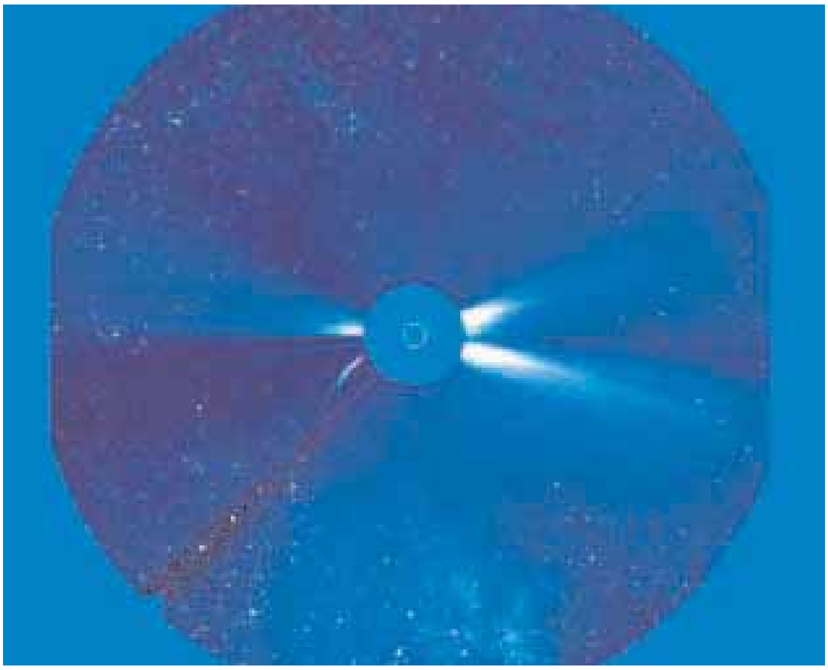


Abbildung 13: Sonnenkorona
(mit Komet!), aufgenommen von LASCO. Die flammenförmigen Objekte («Streamers») sind Quellen des (langsamen) Sonnenwindes. Gelegentlich werden sie als «Coronal Mass Ejections» abgestossen. SOHO/LASCO


Abbildung 14: UVCS-Bilder,
aufgenommen im Licht von neutralem Wasserstoff (links) und von fünffach geladenem Sauerstoff (rechts). SOHO/UVCS

deckt wurde. Dieser stürzte um Weihnachten 1996 in die Sonne und wurde von ihr verschlungen.
Ein letztes Bild der Sonnenkorona **(Abbildung 14)**, durch den Ultraviolett-Koronaspektrometer (UVCS) registriert, zeigt einen Streamer im Licht von Wasserstoff (links) und von fünffach geladenem Sauerstoff (rechts). Die quantitative Analyse solcher Bilder zeigt, dass der Sonnenwind nicht aus dem Inneren des Streamers stammt und von seiner Spitze wegströmt, sondern dass er von der Basis her um den Streamer herum fliesst. Der schwerere Sauerstoff scheint nämlich wegen der Schwerkraft der Sonne im Inneren des Streamers nur spärlich vorhanden zu sein, der leichtere Wasserstoff hingegen häufiger. Wir wissen aber, dass der Sonnenwind Sauerstoffionen enthält. Er muss folglich aus den peripheren Gebieten des Streamers stammen, denn in der Mitte ist ja kaum Sauerstoff vorhanden. Der langsame Sonnenwind muss also um den Streamer herumströmen bevor er in den interplanetaren Raum gelangt. Diese Beobachtung löst ein weiteres Rätsel, die Frage nämlich, woher die Sonne ihren langsamen Sonnenwind, der von einem magnetisch scheinbar geschlossenem Gebiet in der Korona kommt, zu uns hinaus sendet. Sie ist auch ein gutes Beispiel dafür, dass Beobachtungen von seltenen Elementen (das Verhältnis von Sauerstoff zu Wasserstoff ist nur etwa ein Promille) entscheidende Hinweise liefern können.

Der Sonnenwind

Als Sonnenwind bezeichnet man den von der Sonne ausgehenden Materiestrom. Das Vorhandensein eines Sonnenwindes wurde bereits vor dem Raumfahrtzeitalter aus der Existenz von Kometenschweifen vermutet, die immer von der Sonne weg gerichtet sind **(Abbildung 15).**

Abbildung 15: Kometenschweife
waren der erste Hinweis für die Existenz eines Sonnenwindes, da sie immer direkt von der Sonne weg weisen, zumindest der feine, bläuliche Plasmaschweif. Das Sonnenlicht kann dafür nicht verantwortlich sein, da der Schweif sonst auch das Sternlicht absorbieren müsste, was eindeutig nicht der Fall ist. Wally Pacholka

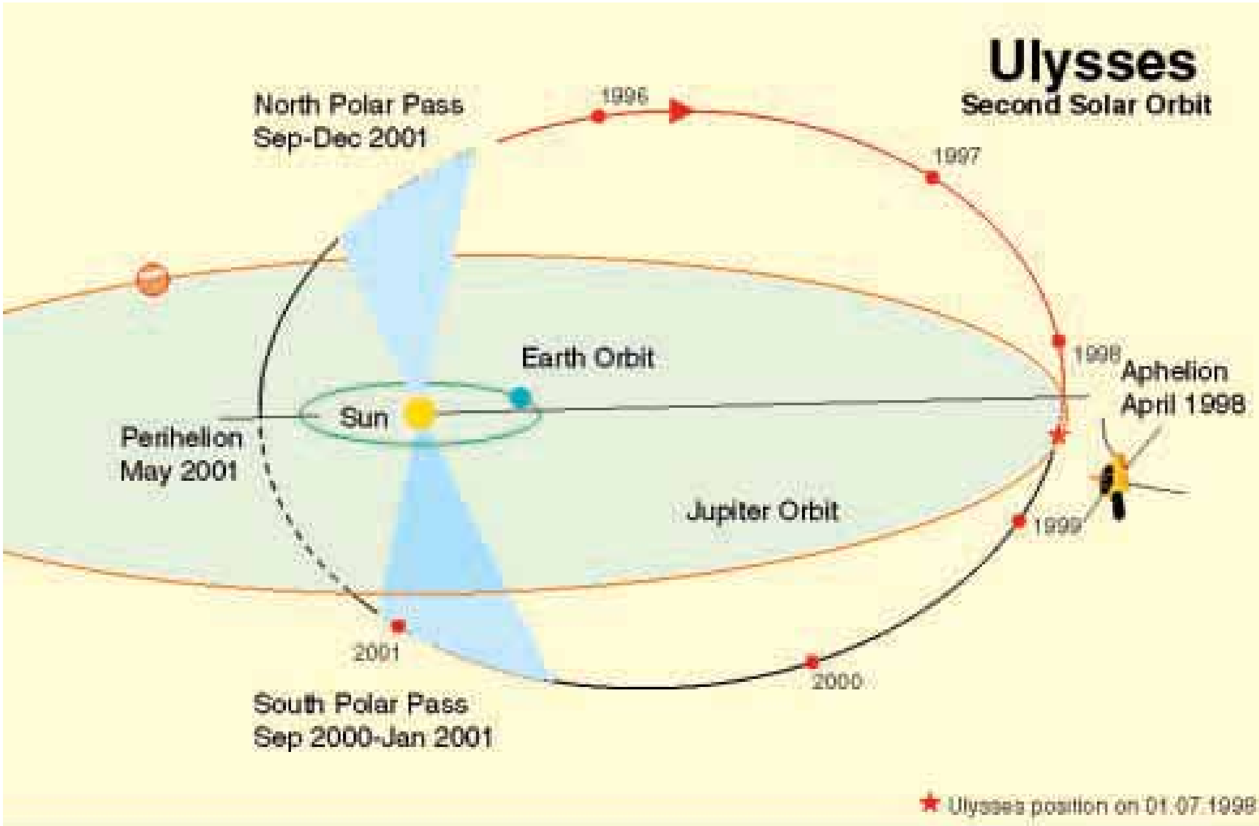


Abbildung 16: Bahn der Weltraumsonde Ulysses.
Nach ihrem Start im Oktober 1990 verlief sie zunächst in der Ebene der Planetenbahnen, bis Ulysses im Februar 1992 durch einen nahen Vorbeiflug bei Jupiter auf eine stark geneigte, nahezu polare Bahn gebracht wurde. In den Jahren 1994 und 1995 überflog die Sonde ein erstes Mal die polaren Gebiete der Sonne und ist heute wieder fast in ihrem sonnenfernsten Punkt angelangt.
Ulysses/ESTEC

Ulysses ist die erste (und bis auf Weiteres wohl einzige) Weltraumsonde, die sich auf einer stark geneigten Bahn gegenüber den Planetenbahnen und dem Sonnenäquator befindet **(Abbildung 16).** Sie kann deshalb auch die bisher unbekannten Gebiete über den Polen der Sonne erforschen. Ulysses ist also die erste Polarexpedition über der Sonne – und dies weniger als 100 Jahre nach den grossen Polarexpeditionen von Nansen, Amundsen und Scott auf der Erde.

Wie aber kann Ulysses den Sonnenwind messen? Beim Sonnenwind haben wir es mit einem sehr feinen Plasmastrom zu tun, dessen Zusammensetzung uns im Besonderen interessiert. Diese kann durch Massenspektrometer ermittelt werden, welche das Plasma in seine Bestandteile zerlegen können. Wir unterscheiden drei Generationen solcher Messgeräte, welche mit der Zeit zunehmend komplexer und leistungsfähiger geworden sind. Die ersten waren elektrostatische Kondensatoren. Sie lenkten die einfallenden Ionen (geladene Teilchen) ab und sortierten sie lediglich nach dem Verhältnis von Energie zur Ladung. Die nächste Generation kombinierte die elektrostatische mit einer elektromagnetischen Ablenkung, wodurch neben dem Verhältnis Energie / Ladung zusätzlich auch das Verhältnis Masse / Ladung bestimmt werden konnte. Die dritte Generation, welche auch auf Ulysses zur Anwendung gelangt, ist (wie auch die zweite Generation) am Physikalischen Institut der Universität Bern, in Zusammenarbeit mit anderen Instituten, entwickelt und gebaut worden. Ein solches Gerät misst zusätzlich zu Energie pro Ladung und Masse pro Ladung auch die Gesamtenergie der Ionen, also die drei wesentlichen Grössen, je separat. Damit sind die Teilchen vollständig charakterisiert.

Es gibt neben den Weltraummassenspektrometern eine zweite wichtige Art des Sonnenwindnachweises: Meteoriten und Mondproben enthalten seit Millionen und Milliarden von Jahren aufgesammelten Sonnenwind. Dieser kann durch Erhitzen oder Ätzen extrahiert und mit hochpräzisen Labor-Massenspektrometern analysiert und in seiner Zusammensetzung untersucht werden. Dasselbe kann sogar mit Folien realisiert werden, wie sie anlässlich der Mondlandungen verwendet wurden, um den Sonnenwind einzufangen **(Abbildung 17).**

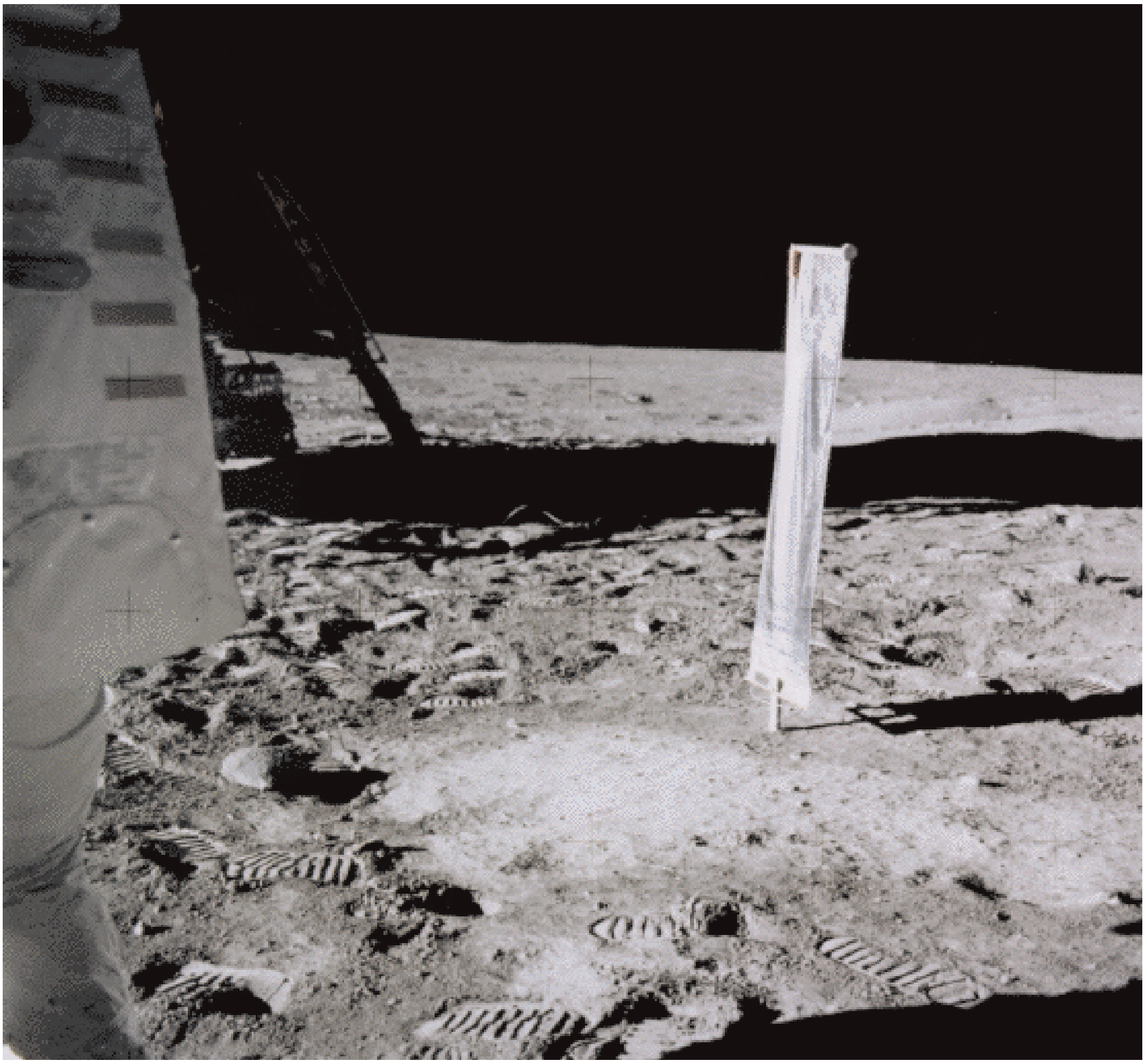


Abbildung 17: Sonnenwindsegel von J. Geiss,
aufgestellt auf der Mondoberfläche von den Astronauten der Apollo-Missionen. Mit dieser einfachen Aluminiumfolie wurden für wenige Stunden bis Tage Sonnenwindteilchen aufgesammelt, um sie nach dem Rückflug im Labor zu analysieren. Für einzelne seltene Elemente (z. B. Argon) sind dies noch heute die besten Messungen.
NASA

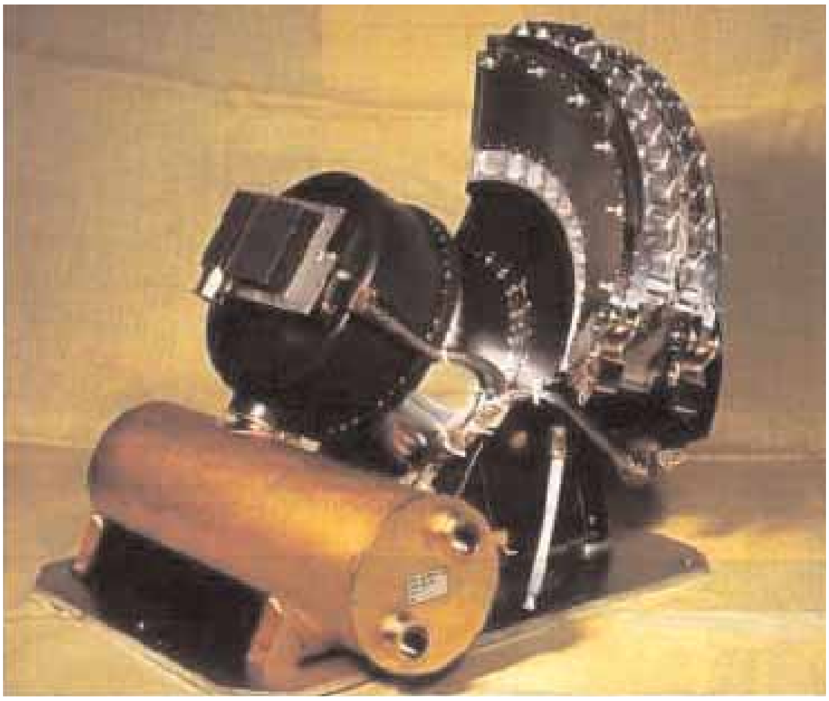
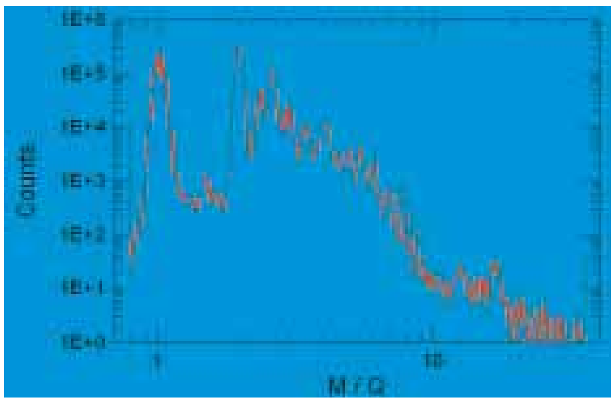


Abbildung 18: Das Massenspektrometer SWICS,
das heute auf Ulysses im Einsatz steht. Die Ionen treten durch die fächerförmige Struktur links ins Instrument und werden im Zylinder in der Mitte analysiert. Der goldene Zylinder im Vordergrund ist eine Hochspannungsquelle von 30000 Volt; sie wird benötigt, um die eintretenden Ionen zusätzlich zu beschleunigen und so leichter nachweisbar zu machen. Ulysses/SWICS

Doch zurück zu Ulysses und dem Weltraummassenspektrometer SWICS **(Abbildung 18).** Eine typische Darstellung seiner Messungen ist in **Abbildung 19** gezeigt: Wie bereits erwähnt, wird jedes einfallende Ion nach seiner Energie, Masse und Ladung klassifiziert. Die Energie ist in dieser Darstellung nicht aufgezeichnet, weil es hier vor allem um die Unterscheidung jedes Ions nach seiner Masse und seiner Ladung geht. (Traditionellerweise zeigen wir auf der horizontalen Achse nicht die Masse, sondern das Verhältnis der Masse pro Ladung, aber die Information ist die gleiche). Die Ionen im Sonnenwind sind offenbar nicht zufällig verteilt, sondern es kommen ganz bestimmte Elemente in bestimmten Ladungszuständen vor. Zu sehen sind hier unter anderem Kohlenstoff (4-, 5-, und 6-fach geladen), Sauerstoff (6-, 7-, und 8-fach geladen) usw. bis hinauf zum Eisen.

Die Informationen im obersten Bild zeichnete Ulysses während eines Jahres bei seinem Überflug über den Südpol der Sonne auf. Das mittlere Bild in **Abbildung 19** dagegen ist das Ergebnis von Aufzeichnungen, die während über einem Jahr gemacht wurden, als sich Ulysses in der Nähe der Äquatorebene der Sonne, also auch in der Nähe der Ekliptik, befand. Das unterste stammt wiederum aus dem Polargebiet, diesmal aber dem nördlichen.

In **Abbildung 19** sieht man deutlich, dass sich die Bilder von Nord- und Südpol offenbar sehr gleichen, das Bild vom Äquator aber davon klar verschieden ist. Durch weitere Analyse solcher Bilder kann nun die Geschwindigkeit und die Häufigkeit jedes Ions gemessen werden. Indirekt lässt sich damit sogar auf die Temperatur der Korona schliessen, indem wir die Häufigkeiten benachbarter Ladungszustände eines Elements miteinander vergleichen. Stellvertretend für andere Parameter sehen wir in **Abbildung 20** die Geschwindigkeit von Helium, die Koronatemperatur aus den Ladungszuständen von Sauerstoff, sowie das Verhältnis von Magnesium zu Sauerstoff. Aufgezeichnet, und zwar



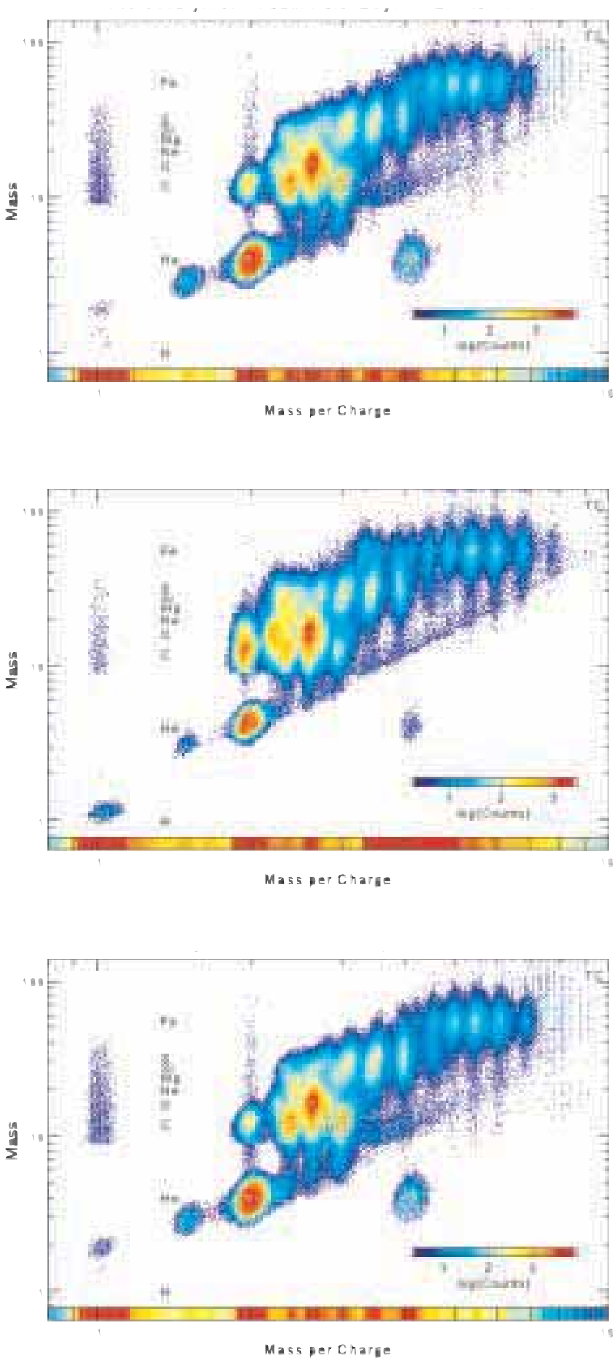


Abbildung 19: Zusammensetzung des Sonnenwindes, gemessen von SWICS auf Ulysses. Die drei Teilbilder wurden zu verschiedenen Zeiten aufgenommen: Südpolgebiet (ganz oben), Ekliptikebene (Mitte) und Nordpolgebiet (unten). In diesen Diagrammen wird jede Ionensorte gemäss ihrer Masse und ihrer Ladung auf eine bestimmte Stelle abgebildet, ähnlich wie Berggipfel auf einer Landkarte, und kann so einzeln identifiziert werden. Frühere Instrumente dagegen lieferten lediglich ein Panorama (Bild links). Ulysses/SWICS

als Funktion der heliographischen Breite, ist die Zeitspanne vom Start bis zum ersten Quartal 1998 des Ulysses-Projektes. Die Messungen vom Südpol befinden sich links im Bild, jene vom Äquator in der Mitte und jene vom Nordpol rechts.

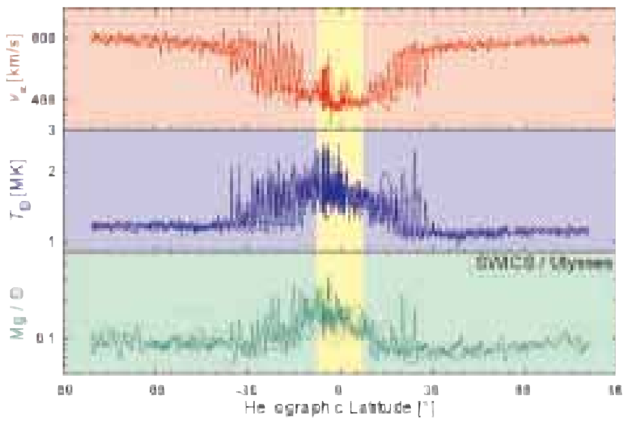


Abbildung 20:
Zusammenfassung dreier Sonnenwindparameter als Funktion der heliographischen Breite: Geschwindigkeit von Helium (oben), die Koronatemperatur aus den Ladungszuständen von Sauerstoff (Mitte), sowie das Verhältnis von Magnesium zu Sauerstoff (unten). In allen drei Parametern bestehen offensichtliche Unterschiede zwischen den polaren Gebieten (links und rechts) und den äquatorialen Zonen (Mitte). Ulysses/SWICS

Der gelbe Streifen in der Mitte zeigt den Bereich der Ekliptik und illustriert eindrücklich, welch limitierte Perspektive wir von der Erde aus im Vergleich zu Ulysses haben. Es ist in dieser Darstellung offensichtlich, dass der Sonnenwind bei hohen Breiten schnell und fast konstant, bei tiefen Breiten dagegen langsam und variabel ist. In einer Übergangsregion zwischen ca. 15–30 Grad Breite wechselt seine Geschwindigkeit im Rhythmus mit der Sonnenrotation. Deutlich wird auch, dass in hohen Breiten die Korona kühl ist, was nicht erstaunt, wenn man sich an Koronalöcher in den EIT-Bildern zurückerinnert Auch die Zusammensetzung ist unterschiedlich, wie hier am Beispiel Mg/O gezeigt.

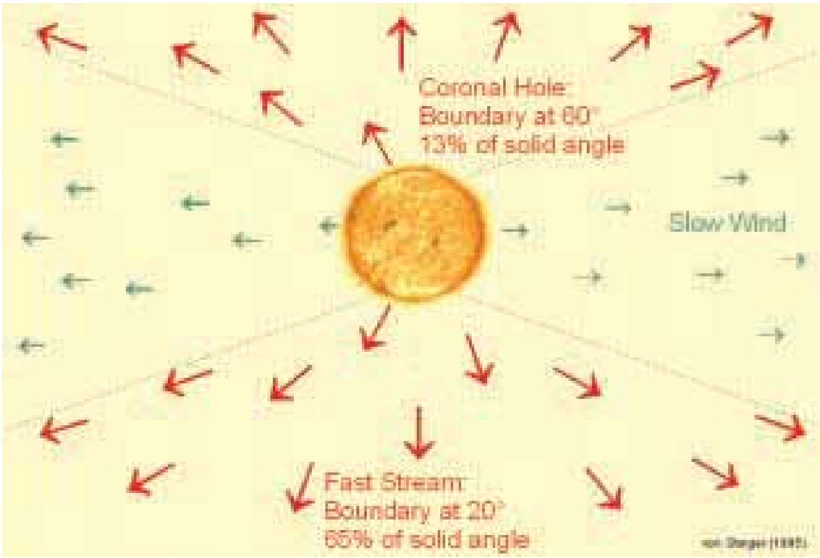


Abbildung 21: Überradiale Expansion des Sonnenwindes im Aktivitätsminimum: Die polaren koronalen Löcher bedecken lediglich ca. 13% der Sonnenoberfläche (kleines Bild von EIT, Mitte), während der schnelle Sonnenwind bis hinunter zu 20 Grad beobachtet werden kann, also etwa 65% der Heliosphäre einnimmt, und den langsamen Wind auf ein Band um den Äquator komprimiert. Die Dynamik der Heliosphäre wird also (zumindest während des Aktivitätsminimums) von den schnellen Strömen dominiert. Ulysses/SWICS

Zusammengefasst könnte dies zu folgendem vereinfachten Bild der Sonne und ihres Windes führen **(Abbildung 21):** Wir kennen einerseits den langsamen Sonnenwind, der von den tiefen Breiten mit den aktiven Regionen der Sonne abströmt. Das Gebiet, in dem er dominiert, reicht grob gesagt von 20 Grad Nord, bis 20 Grad Süd heliographischer Breite. Auf den EIT-Bildern haben wir gesehen, dass die Koronalöcher an den Polkappen sitzen. Ihre Grenzen liegen bei etwa 60 Grad Breite. Zusammengesetzt bedeutet dies, dass der schnelle Sonnenwind aus den Koronalöchern sehr stark expandiert bis hinunter auf 20 Grad. Die Koronalöcher selber bedecken nur etwa 13% der Sonnenoberfläche, die schnellen Ströme nehmen aber etwa 65% der Heliosphäre ein. Es findet also eine überradiale Expansion des schnellen Sonnenwindes statt, welche die Heliosphäre zumindest zu Zeiten des Aktivitätsminimums dominiert.

Abschliessend noch ein Blick auf die neueste Generation von Sonnenwindinstrumenten. Auf SOHO fliegt ein Massenspektrometer, CELIAS / MTOF, welches ebenfalls in Bern mitentwickelt und -gebaut wurde. Gegenüber SWICS auf Ulysses besitzt es eine phantastische Massenauflösung (siehe **Abbildung 22).** Die quantitative Auswertung dieser Daten wird in naher Zukunft unser Bild der Sonne verfeinern und womöglich erweitern.

Unser neues Bild der Sonne

Fassen wir zum Schluss kurz zusammen:

- Die Resultate der Helioseismologie stimmen mit den besten Sonnenmodellen auf Bruchteile eines Prozentes überein.
- Dies liefert eine sehr genaue Messung des solaren Heliumgehalts.
- Die Lösung des Neutrinoproblems (d.h. dass von der Sonne weniger Neutrinos produziert werden als erwartet) muss im Rahmen der Elementarteilchenphysik gesucht werden.
- Die «kalte» Photosphäre (5000 K) heizt die Korona durch Rekonnektion im magnetischen Teppich auf über 1 000 000 K auf.

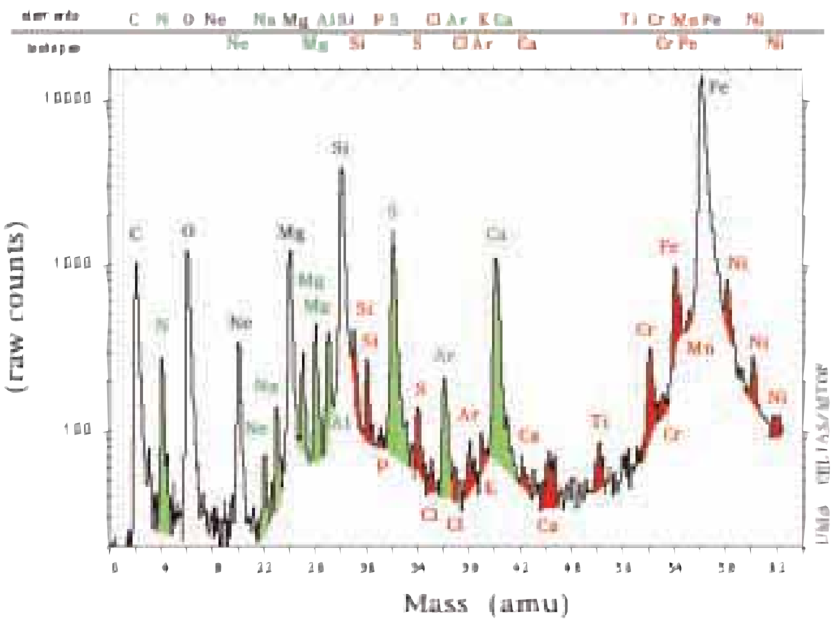


Abbildung 22: Die nächste Generation
von Weltraummassenspektrometern ist auf SOHO bereits im Einsatz. Einer der drei Sensoren von CELIAS, das MTOF, besitzt eine phantastische Massenauflösung (dies auf Kosten der Information über die Ladungszustände der gemessenen Ionen). Es kann bisher nicht beobachtete Elemente (grün) und Isotope (rot) nachweisen und voneinander separieren. Die quantitative Auswertung dieser Daten wird in naher Zukunft unser Bild der Sonne verfeinern und womöglich erweitern.
SOHO / CELIAS / MTOF

– Es gibt zwei fundamental verschiedene Sonnenwindtypen:

– Der schnelle Wind strömt ungehindert aus den Koronalöchern und liefert ein nahezu treues Abbild der solaren Zusammensetzung.

– Der langsame Wind strömt um die Streamers herum, seine veränderte Zusammensetzung liefert Hinweise über die Bedingungen und Prozesse in der Chromosphäre und Korona.

Die beiden Weltraummissionen Ulysses und SOHO der europäischen Raumfahrtagentur ESA haben uns ein neues Bild der Sonne geliefert. Wir dürfen aber nicht vergessen, dass dieses bei abnehmender und minimaler Sonnenaktivität aufgenommen wurde. Es ist daher leicht möglich, dass die aktive Sonne in ihrem nächsten Maximum in einigen Jahren noch grosse Überraschungen bereithält. Eine Fortsetzung beider Missionen bis ins Jahr 2001 oder länger wäre daher von grösster Bedeutung. Ulysses scheint dafür bestens gerüstet zu sein, die Sonde sowie alle ihre Instrumente funktionieren acht Jahre nach dem Start (und drei Jahre nach dem ursprünglich geplanten Ende der Mission) tadellos. Die Verbindung zu SOHO dagegen ging leider infolge eines fehlerhaft durchgeführten Manövers im Juni 1998 verloren. Inzwischen (August 1998) konnte sie für kurze Zeit wieder hergestellt werden, aber die Sonde trudelt weiterhin unkontrolliert vor sich hin. Hoffen wir, dass sie durch geeignete Befehle wieder in ihre richtige Orientierung gebracht werden kann und dass die empfindlichen Instrumente an Bord inzwischen keinen wesentlichen Schaden erlitten haben. Dann wird auch SOHO weiterhin wertvolle Beiträge zu unserem neuen Bild der Sonne liefern können.

Redaktion
P. Abgottspon und H. Schlaepfer

Nachtrag vom Oktober 1998

Die Operation ist geglückt! SOHO befindet sich wieder in seiner Normallage und auch die Instrumente scheinen kaum in Mitleidenschaft gezogen worden zu sein.

SPATIUM

Zum Autor

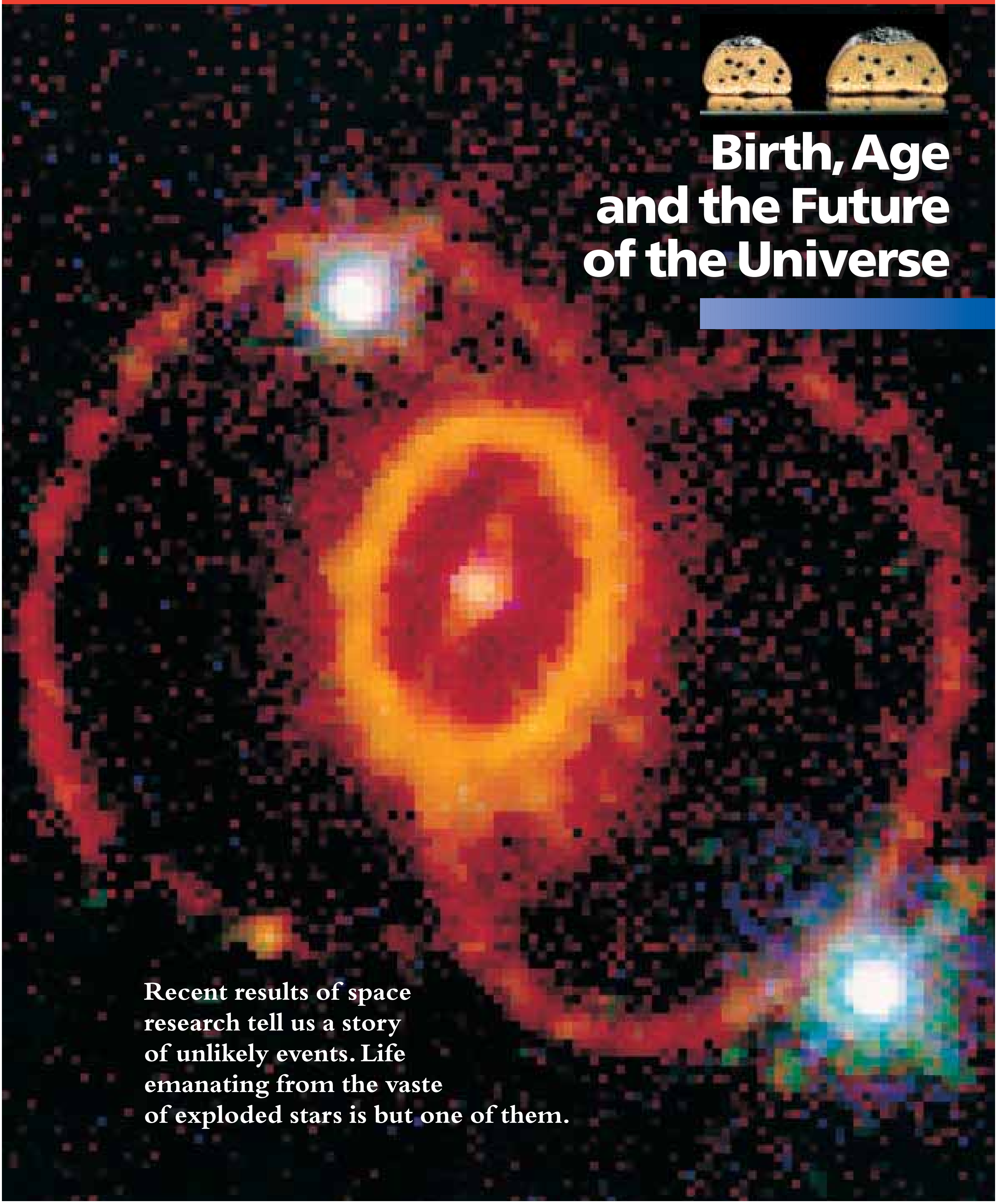
Rudolf E. von Steiger kam am 21. Juni 1957 als Berner Stadtbürger zur Welt. Er besuchte das Gymnasium Kirchenfeld in Bern, wo er 1976 die Maturität Typus B erlangte. Physik, Mathematik und Astronomie waren die Fächer, die er im anschliessenden Studium an der Universität Bern belegte. Er schloss mit einem Diplom in theoretischer Physik 1984 ab und doktorierte 1988 in experimenteller Physik bei Prof. J. Geiss ebenfalls an der Universität Bern. Dort war er anschliessend Forschungsassistent; ein Stipendium des Nationalfonds ermöglichte ihm einen Aufenthalt am Institut für Physik und Astronomie an der Universität von Maryland in College Park (USA). Im Jahre 1991 kehrte er als Forschungsassistent an die Universität Bern zurück, wo er sich der Analyse und Interpretation der Daten des SWICS-Instrumentes auf Ulysses widmete. In diese Zeit fallen auch Lehraufträge, insbesondere in experimenteller Kosmologie und in Physik Praktiken, an der phil.-nat. Fakultät, wo er sich 1995 habilitierte. Seit der Schaffung des International Space Science Institutes im Mai 1995 wirkt R. von Steiger Senior Scientist, wo er neben der eigenen wissenschaftlichen Arbeit für Teile des Wissenschaftsprogramms und die Edition der daraus resultierenden Buchserie zuständig ist. Zur Zeit arbeitet er am Institut für atmosphärische, ozeanische und Weltraumwissenschaften der Universität von Michigan in Ann Arbor an Fragen zur Feinstruktur des Sonnenwindes.

Die humanistische Ausbildung mag einer der Gründe für die sprachlichen Interessen von Rudolf von Steiger sein, die von der Muttersprache Deutsch über Englisch, Französisch, Italienisch bis hin zum Japanischen reichen. Als Berner Bürger ist ihm die Aare natürlich nicht nur ein Fluss, sondern auch ein Zentrum geselliger Aktivität, wenn es in den Sommermonaten gilt, in ihren kühlen Fluten durch die Stadt zu schwimmen. Er ist verheiratet und lebt mit seiner Frau und zwei Katzen in der Nähe von Bern.

INTERNATIONAL SPACE SCIENCE INSTITUTE
No 3, May 1999

SPATIUM

Published by the Association Pro ISSI



Birth, Age and the Future of the Universe

Recent results of space research tell us a story of unlikely events. Life emanating from the vaste of exploded stars is but one of them.

Editorial

We are here. There is no doubt about it. –

Research, however, has shown that life, our solar system and all the galaxies interrupting locally the voidness of space have been caused by tiny unbalances, which seem very unlikely a priori. Recent results of space research tell us a story of unlikely events.

This fascinating history of our universe was the subject of the key note address by Professor Gustav Andreas Tammann, Director of the Institute for Astronomy of the University of Basle on the occasion of the third anniversary of the International Space Science Institute on the 27th of November 1998.

It is with greatest pleasure that we devote this issue to Professor Tammann, who has actively contributed to the understanding of our universe over the past 40 years.

Hansjörg Schlaepfer
Berne, May 1999

Impressum

SPATIUM
Published by the
Association Pro ISSI
once or twice a year

INTERNATIONAL SPACE SCIENCE INSTITUTE

Association Pro ISSI
Hallerstrasse 6, CH-3012 Berne
Phone ++41 31 631 48 96
Fax ++41 31 631 48 97

President
Prof. Hermann Debrunner,
University of Berne
Publisher
Dr. Hansjörg Schlaepfer,
Contraves Space, Zurich
Layout
Gian-Reto Roth,
Contraves Space, Zurich
Printing
Drucklade AG, CH-8008 Zurich

Front cover
Rings of glowing gas encircling the supernova 1987A.
(Hubble Space Telescope, NASA/ESA, February 2, 1994)

Birth, Age and the Future of the Universe

G. A. Tammann, Director, Institute for Astronomy, University of Basle

Key note address at the third anniversary of the International Space Science Institute, November 27, 1998

The sciences have fundamentally changed the views of our world during this Century. They have brought a cultural revolution by affecting technology, medicine, civilisation and, perhaps most importantly, the thinking of men.

Man has explained the origin of all life, i. e. the energy source of the Sun which had remained a total mystery up to 1930. He has established a worldwide net of communication and transportation such that he partakes almost instantly in occurrences all over the world. Humans have left the Earth's gravity field and have seen that the Earth is round. The Moon has been occupied – so far temporarily – by men and the planets have been visited by spacecraft. The triumph of rational thinking has even brought an unprecedented 50-year period of peace in Western Europe.

For thousands of years the human life has been an interval of 40 years of suffering of hunger and freezing. Science has doubled its length and it is now expected to be painless, comfortable and pleasurable. At the end of this century man finds himself in a world which has been fundamentally changed by the sciences from what it was at the beginning of the Century, – and not only the world has changed but also man himself, – and were it just in physical size.

It may seem surprising that this radical change has been brought about by the sciences which are believed to progress step by step by a painstaking method dominated by mathematical logic and bare of fantasy. While it is true that all sciences proceed in minute steps, it happens once in a while that the understanding jumps suddenly on a higher level. A broader insight is gained, a new paradigm is born. These are the great discoveries which influence our understanding, thinking, and culture.

A few examples will be given in the following from the realm of astronomy. They are to illustrate how the scientist works on a gigantic mosaic that suddenly reveals a face or contour.

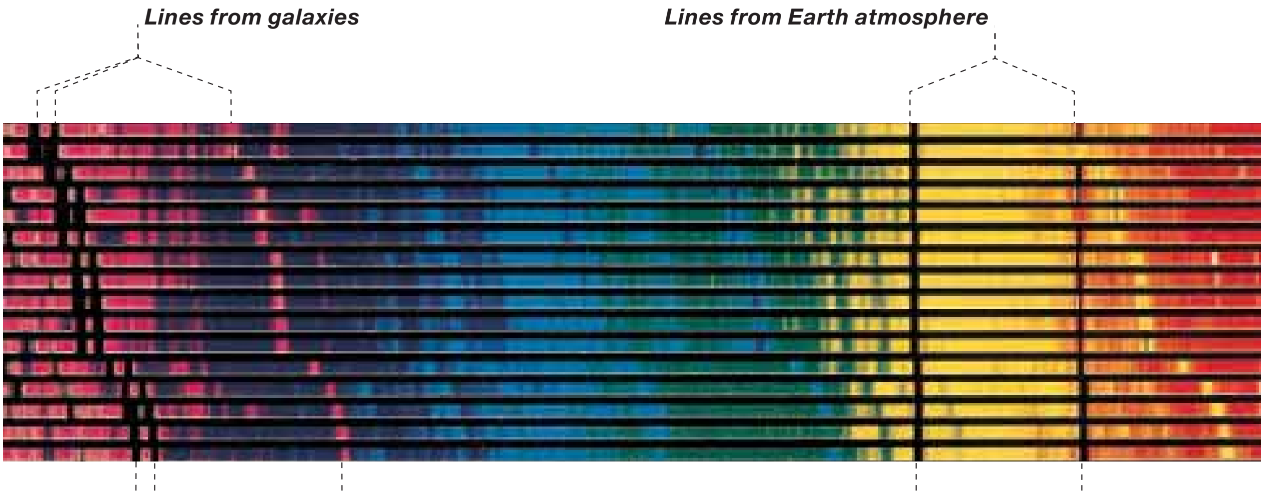


Figure 1. The spectra of 16 galaxies. They are perfectly aligned as seen from the lines originating in the Earth's atmosphere. However, the galaxy lines (originating from different chemical elements) are progressively shifted toward the red. This redshift effect is caused by the expansion of space stretching the light to longer wavelength. The different redshifts correspond here to recession velocities of 3000–16,000 km s^{-1}. (From P. Stein, Basle).

The expanding Universe

Figure 3. A yeast cake as a model for the expanding Universe. As the cake increases in size the distances between the raisins become larger. Close raisins get separated by small (absolute) amounts, distant raisins by large amounts. The aspect is the same for *all* raisins. (Neglect the rim). The model has the disadvantage that one can oversee it instantaneously while the concept of simultaneity fails in the Universe (see page 5).

In 1912 Vesto M. Slipher took the first spectra of what was then called spiral "nebulae" (now known to be galaxies). It was a tedious process with small telescopes and slow emulsions. He found the spectral lines shifted toward the red **(Figure 1)** and he concluded correctly that the objects were hence receding from us, – yet faster than any known star in our Galaxy. The large recession velocities became a puzzle which many astronomers tried to solve. Finally Edwin Hubble, after having proved in 1925 that the "nebulae" are distant galaxies, consisting of hundredthousand million stars like our Milky Way, realised in 1929 that not only (almost) all galaxies move away from us but also that their velocities were proportional to their distances! **(Figure 2)** The picture was like that any raisin has in a growing yeast cake: all other raisins are moving away and the faster the further away they are **(Figure 3)**. Hubble concluded that the whole Universe was expanding like a yeast cake and that there must have been a time where the Universe was arbitrarily small. The Universe had a beginning, now known as the Big Bang!

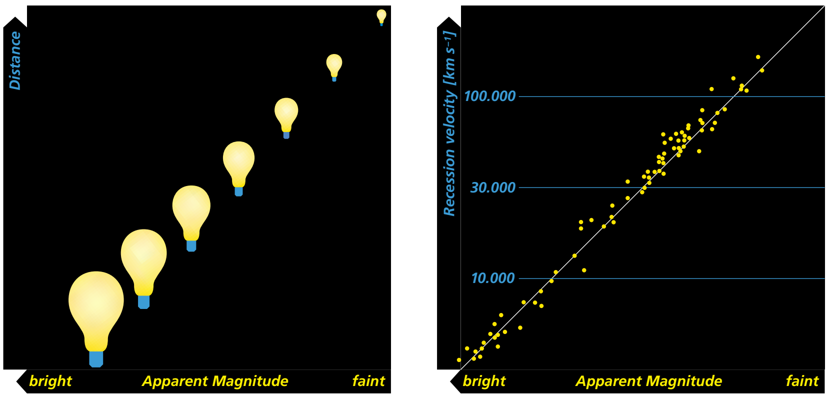


Figure 2. The demonstration of the linear expansion of the Universe. Left: Standard candles (e.g. 60 W light bulbs) get fainter as the distance increases. Right: The brightest galaxies in clusters of galaxies are good standard candles (note the relatively small scatter!). As they get fainter they must be more distant, but at the same time their recession velocities increase. (*Linear* expansion requires in this logarithmic plot a slope of 0.2, which is shown).

In the last seventy years the Big Bang has become a physical fact. The most distant galaxies have redshifts corresponding to *almost* the speed of light. Several independent experiments prove beyond doubt that the young Universe must have been tiny, extremely dense, and excessively hot.

The expansion of the Universe gives us a simple tool to determine its age. At a very early epoch all galaxies (or the matter or energy they were later made of) were compressed in one place. When the expansion began some were slowly carried away; they are today our neighbors. Others were carried away with high speed; they populate now our horizon. All had the same travel time, i. e. the age of the Universe. Therefore one must determine "only" the distance to one galaxy – or much better to many galaxies, which is technically not easy **(Figure 4)** – and divide this distance by the recession velocity. The result is nothing else but the expansion age of the Universe.

Measured redshifts and the corresponding recession velocities plus the best distance determinations give an expansion age of 14 (±2) Gigayears (1 Gigayear = 1000 million years). This time may appear long, but when one realizes that life began 3 Gigayears ago, that the oldest rocks on Earth have an age of almost 4 Gigayears, and that the oldest stars in our Galaxy were formed 12 Gigayears ago, one cannot but be impressed by the youth of the Universe. Imagine you take a very old rock in Greenland into your hand and it has almost one third of the Universe's age!

The expanding Universe revises in several ways our thinking. For instance, the fact that all galaxies are receding from us makes us falsely believe that the Big Bang has taken place *here*. In reality *any* galaxy, like any raisin in the yeast cake, sees *all* other galaxies (raisins) in recession. There is no preferred, absolute point in the Universe. Our thinking is unjustifiably egocentric. After some more thought (concerning the "edge" problem of the Universe) one realizes that the picture of galaxies travelling through space cannot be correct. In reality space expands and carries the galaxies along. This is perfectly reasonable with Einstein and is again exemplified by our yeast cake: not the raisins move apart, but the dough expands and increases the distances between the raisins.

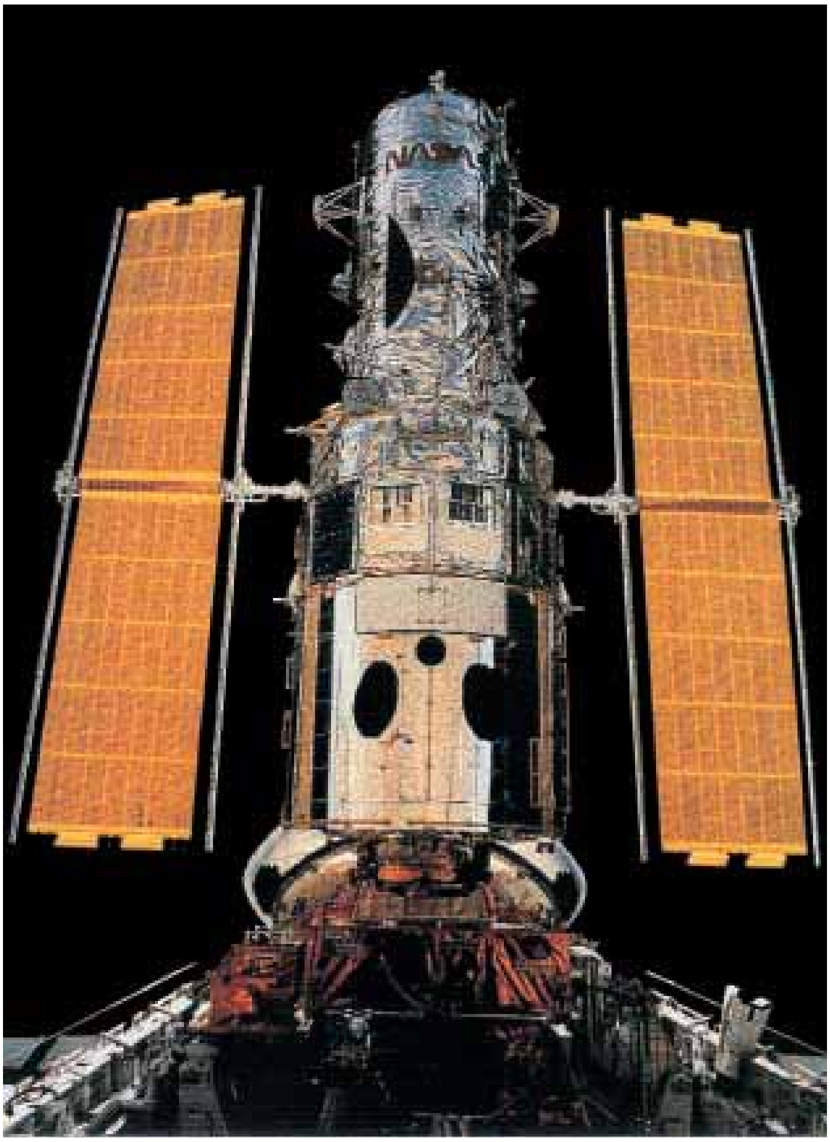


Figure 4. The Hubble Space Telescope has a mirror of "only" 2.5 m diameter, but since it orbits the Earth above the atmosphere it provides exceptionally sharp images. It has greatly helped to determine galaxy distances and hence to determine the expansion age of the Universe. The telescope was built and is operated by NASA with a 15% share of the European Space Agency (ESA). The telescope is here docked to the Space Shuttle during a repair mission involving the Swiss astronaute Claude Nicollier.

Evolution in the Universe

Still hundred years ago there was an embittered debate between creationists and evolutionists. Did God create the world as we see it today? Were the large erratic blocks placed by him into the midst of green meadows, or where they transported there by the ice? Do the fossils give clues to the evolution of life, or did God make the rocks including the fossils? But as correct as the evolutionists were, they could not answer the question where evolution began.

The starting point of evolution became clear only with the discovery of the Big Bang. At the beginning, an unthinkably small fraction of a second after the Big Bang, the Universe was unmeasurably hot and it contained nothing but immensely condensed energy. As it expanded and cooled the energy transformed into matter, first into exotic, short-lived particles, then at an age of $^1/_{10,000}$ seconds into protons and neutrons, i.e. the matter we know today. At that time the temperature had fallen to a few times 10^{12} degrees and the density had decreased to 1000 million tons per thimbleful.

The creation of matter is no trivial thing. When one produces matter from energy in the great particle accelerators, e.g. at CERN, equal amounts of matter and antimatter are produced. But the two constituents annihilate each other to form energy again. It therefore seems that the Universe cannot ever create lasting matter. However, there is a very slight unbalance favoring matter over antimatter. The excess of newly created matter is only one part in 2000 million. The bulk of all matter has actually been annihilated. Only the tiny excess matter has survived and is nowadays what we see as matter. The unbalance between matter and antimatter is called the symmetry breaking; it is still not fully understood. And yet it is decisive for our existence.

Hundred seconds after the Big Bang the temperature had dropped to 1000 million degrees and for the first time protons and neutrons could stick together to form the most simple elements, deuterium and helium and traces of lithium. Theory predicts that 24 percent of all matter was transformed into helium, and it is a triumph for Big Bang theory that one has never found a gas cloud in our Galaxy or other galaxies with less than this amount of helium. Also the observed abundances of deuterium and lithium agree well with theory. The chemical composition

Figure 5. A part of the large "Lagoon Nebula" in our Galaxy (here about 1 lightyear x 1 lightyear). The bright hot star in the lower right ionizes and excetes atoms of sulphur, oxygen, and hydrogen which radiate at different wavelengths (here in false colors red, blue, and green, respectively). The heated gas is in tornado-like turbulence and triggers star formation in the surrounding cold, dark cloud of molecular gas and dust. This is a typical cradle of newly born stars. (Space Telescope Science Institute).

Figure 6. Structure formation has produced galaxies like this beautiful spiral (NGC 1232). The Milky Way seen from outside would offer a similar view. (Picture taken by the European Southern Observatory, ESO).

of these gas clouds is telling because they have yet experienced little chemical mixing and they reflect still today the primordial composition. The relevance of the abundances of the lightest elements for cosmology was first recognized at the University of Berne, and it is still an important research topic at ISSI.

The "primordial nucleosynthesis" of the lightest elements raises a fundamental question. Why was not *all* matter transformed into helium? In that case all "fuel" for later stars (which shine because they convert hydrogen [protons] into helium in their interior) would have been exhausted early on. The sky would have remained dark and life would have become impossible. The reason that most hydrogen (protons) has survived is that the neutron is more massive than the proton by 0.14 percent. Hence more energy is required to create a neutron and there are fewer of them. When all neutrons were used up the helium production stopped. Again we realize that life hangs on a thin thread, i.e. on the tiny mass difference between the neutron and the proton.

In its early phases the Universe was very simple; one says that it was in thermodynamical equilibrium. That means it was everywhere the same. The whole Universe could be fully described by a few numbers. One might now expect that the expanding Universe would just cool and thin out. Nothing interesting would ever happen.

The opposite is the case. That is, the Universe formed *structures.* There were regions where by chance more matter was than at others. In these dense regions gravity had its way to *locally* slow down the expansion and even to reverse it into a contraction. Gigantic, contracting clouds of hydrogen (and helium) were formed, which fractionated into smaller, still contracting clouds. The latter evolved into what we see today as galaxies **(Figure 6)**. The galaxies had spun up during their contraction, and their rotation protected them from further collapse. But individual cloudlets, still with thousands of solar masses, could continue to contract and form stars. The formation of stars is an *ongoing* process in galaxies **(Figure 5)**. New stars are born and old stars die continuously until all gas is used up. Some galaxies have already exhausted their gas supply; our Galaxy can still go on for a long time.

A contracting star heats up until its interior has reached a temperature of a few million degrees. At this moment a hydrogen bomb is ignited and hydrogen is fused into helium, a process which releases enormous amounts of energy. The energy protects the star against further contraction and lets the star shine. In the case of the Sun this energy is also the basis of all life.

Low-mass stars die as so-called White Dwarfs **(Figure 7)** when they have converted all their hydrogen into helium. More massive stars can then take regress to "burning" helium to carbon, oxygen and other more complex elements up to iron. The production of still more heavy elements does not release energy but costs energy. This energy can only be provided by the most massive stars when they die in the gigantic explosion of a supernova **(Figure 8)**. It is an unfamiliar thought that the gold on our finger was once produced in a supernova explosion.

Yes, all chemical elements in the Universe and on Earth have once been produced in stars. When stars die they eject part of their mass, either peacefully as a Planetary Nebula or explosively in a supernova. In this way some of the

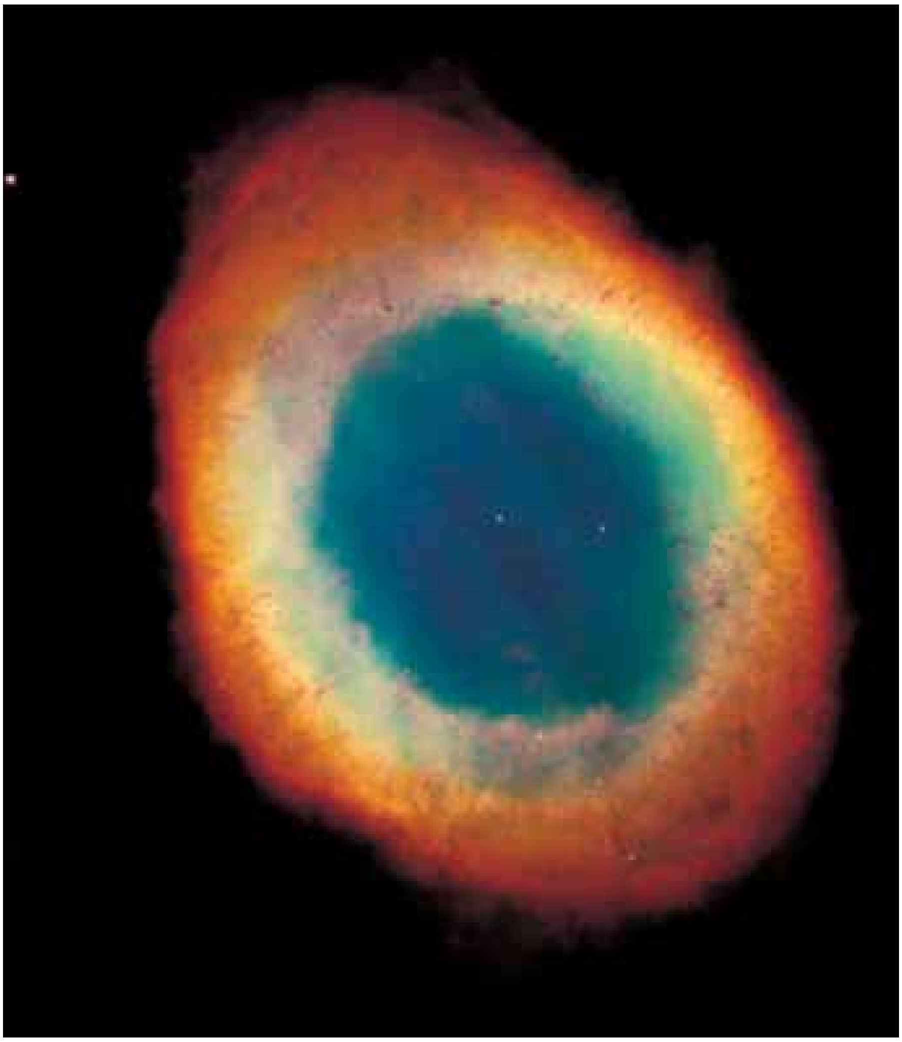
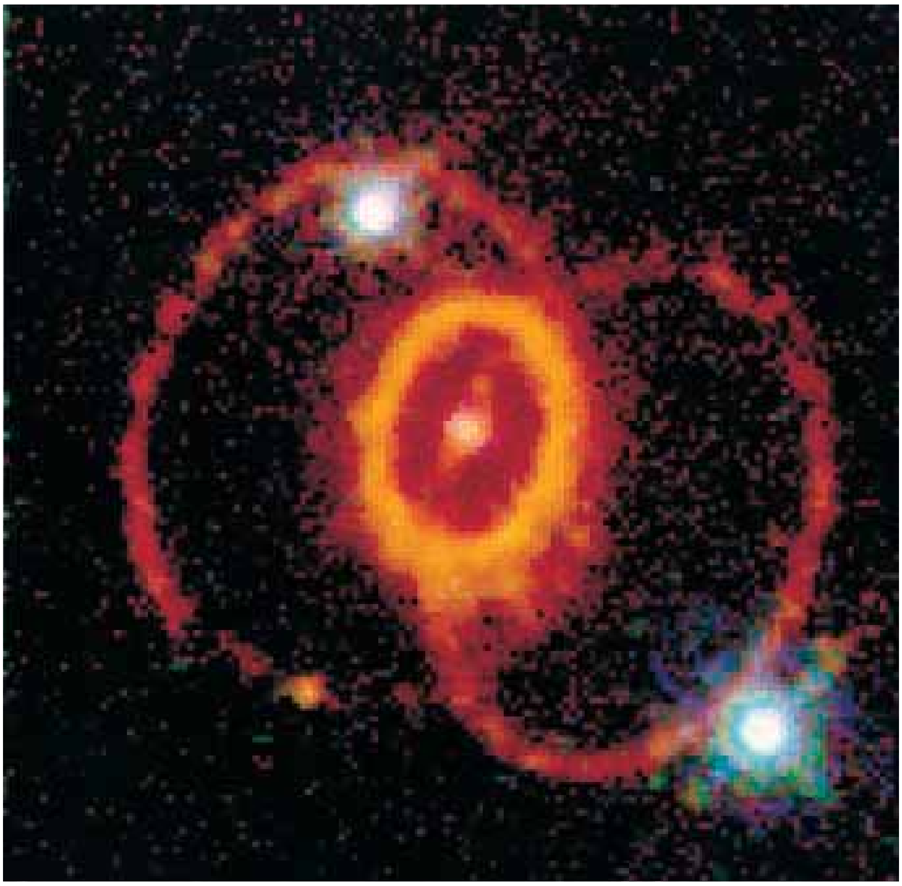


Figure 7. The Planetary Nebula in the constellation Lyra. The faint central star has lost its envelope which is enriched by the chemical elements formed during the lifetime of the star. The central star is only the left-over nucleus with a surface temperature of 120,000 degrees. It is going to become a White Dwarf.

Figure 8. The remnant of the supernova 1987A in the Large Magellanic Cloud, taken seven years after explosion with the Hubble Space Telescope. Large amounts of freshly formed heavy elements are expelled into the interstellar gas of the galaxy. Pre-ejected matter echoes the flash of the explosion.



chemically processed material comes back into the interstellar gas. Newly forming stars are in this way "contaminated" with carbon, oxygen, iron etc. When our Solar System formed 4.6 Gigayears ago all 92 elements were already present **(Figure 9)**. This was possible only because the most massive, chemically most productive stars are paradoxically quite short-lived, and most elements were actually built up long before the Solar System isolated itself from the interstellar gas.

The conclusion is that the chemical variety on Earth is possible only because previous generations of stars have produced all elements, except the primordial hydrogen and helium. We are made of stellar material.

The most puzzling point of this story is that the formation of structures (galaxies and stars) went so fast. The oldest stars in our Galaxy have ages of 12 Gigayears, i. e. they were formed within the first 2 Gigayears after the Big Bang. Computer models fail to form structures in so short an interval, unless one assumes that more than one half of all matter is not in form of protons and neutrons, but is "exotic". This so-called Dark Matter consists of unknown particles with unknown properties, but it seems unavoidable to explain structure formation. Great efforts are presently undertaken to detect this elusive form of matter. There are still things in the Universe we only can dream of.

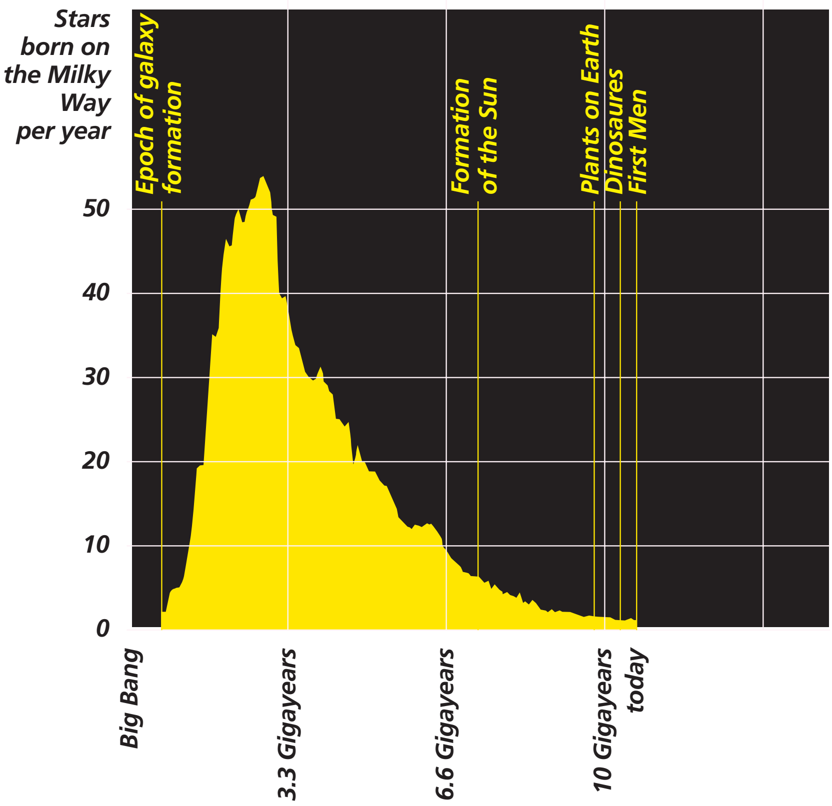


Figure 9. The star formation rate in the Milky Way in function of time. Most stars were formed about 3 Gigayears after the Big Bang. When the Solar System, including the Earth, was formed about 5.4 Gigayears later, all 92 chemical elements had already been formed in stars and expelled into the interstellar gas. This explains why the Earth is relatively rich in processed material. (From M.Samland, Basle).

The Astronomer as Perfect Historian

In the 1660's G.D. Cassini observed in Paris night for night the four Galilean moons of Jupiter as they revolved around the planet and disappeared and reappeared behind or in front of the planet's disk. As Galilei before him he hoped to find in this way a perfect clock which would have been of upmost practical importance for navigation. Yet he noticed irregularities: the predicted occurrences happened sometimes too early or too late by several tens of minutes. The puzzle was solved by Olaus Römer in 1677 who realized that the occultations occurred early when Jupiter was relatively close to the Earth and late when the planet was distant. This led him to one of the most fundamental discoveries of the millennium, i.e. the speed of light is *finite*.

We know today that the speed of light is c = 300,000 km s^{-1} and, according to Einstein, that this is the largest velocity possible because the light particles (photons) have zero rest mass. Any particle with mass can travel only slower. In particular this means also that no signal can be transmitted faster than with the speed of light. (So-called phase velocities, e.g. the movement of a shadow, can exceed c, but they involve neither mass motion nor signal transmission).

Figure 10. The Very Large Telescope of the European Southern Observatory (ESO) in the Atacama Desert in Chile where the observing conditions are optimal. With four 8 m-telescopes it will be the largest telescope in the world. The first 8 m-telescope is already successfully working. A major goal is to study the most distant (i.e. youngest) galaxies during the formation process.

Within human dimensions light travels *almost* infinitely fast. But over cosmic distances light is a slow messenger. It takes 8 minutes from the Sun to the Earth, 4 hours from Pluto, 10,000 years from a distant star in our Galaxy, and one million years from our nearest large neighbor galaxy, the Andromeda nebula. The light of the most distant known galaxies takes more than 10,000 million years!

This implies immediately that we can never have an instantaneous picture of the World. The information we receive here and now is staggered in time according to distance. Any object in the Universe gives us information about the appearance it had when it emitted the light we receive today. The time interval between light emission at the object and reception on Earth is obviously just the (distance-dependent) light travel time. Looking at (very) distant galaxies means therefore to investigate (very) *young* galaxies. The evolution of galaxies can therefore directly be observed.

The enormous progress in the construction of modern telescopes in space and on the ground **(Figures 4 and 10)** is therefore driven not only by the aim to use them as "space ships", bringing distant objects closer to the observer, but also as "time ships", bringing past events into the present.

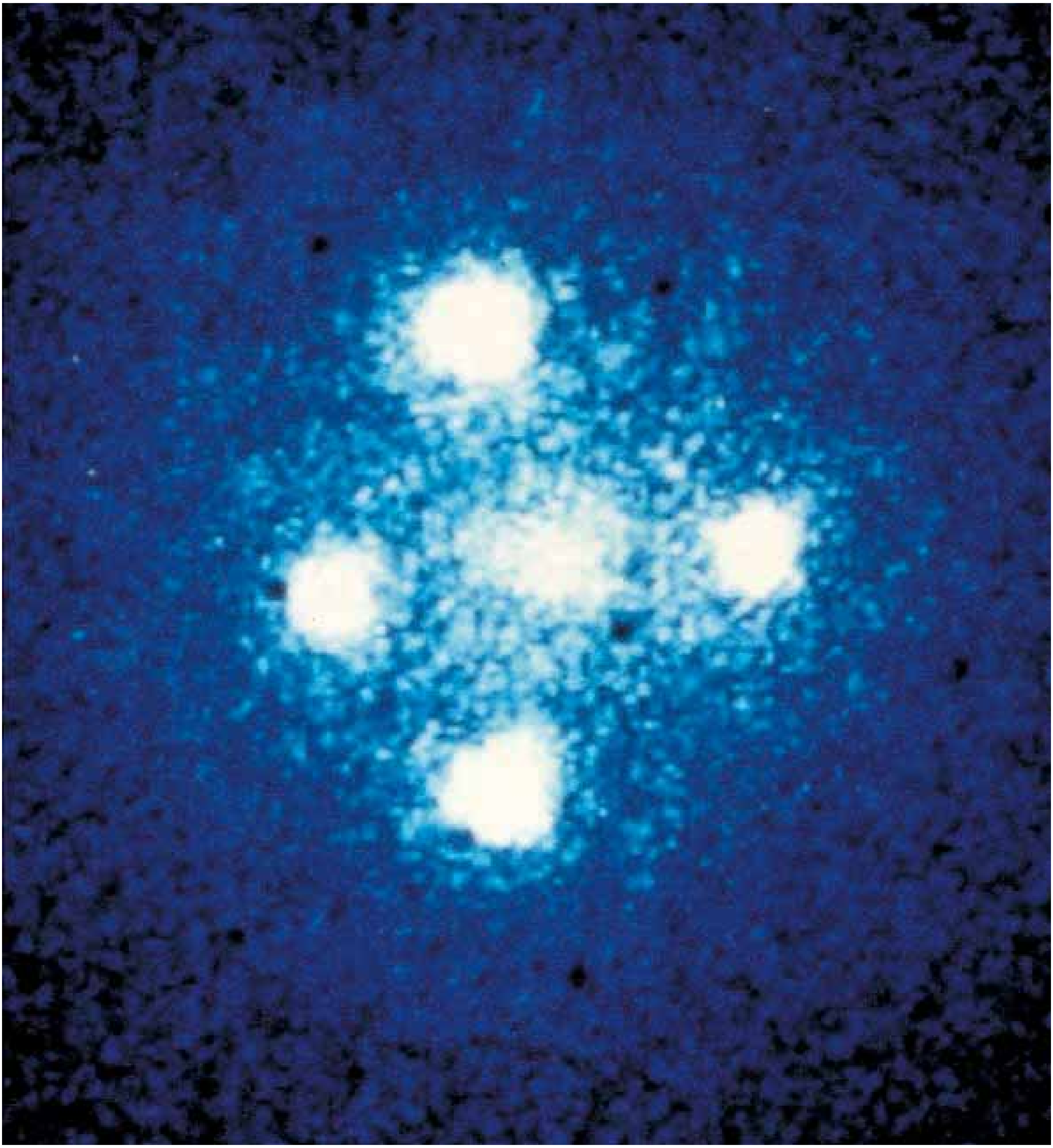
The unique ability of the astronomer to observe the past is not a free ride. The prize is that he cannot observe distant objects as they are today. Here theory jumps in. One *assumes* as a first step that one has observed a time series and then checks with more and more sophisticated computer models, always maintaining the universality of the laws of physics, whether the information from the young Universe does indeed lead to the present (nearby) Universe. With gigantic computers this process can be carried now to such detail that there is no doubt that the observed differences between "out there" and "here" are simply due to *evolution*.

If one looks into extremely large distances one will see no more galaxies because the look-back time brings the observer back into an epoch *before* structure formation when the stars had not yet lit up and the Universe was still dark. The observational proof of this distant cutoff of galaxies seems presently at hand.

There is, however, one still earlier source of radiation which can be observed today. It comes from a time, barely 500,000 years after the Big Bang, when the entire Universe was a single "primordial fireball" with a temperature of 3000 degrees. At this temperature the Universe became transparent for the first time, because the free electrons, which had made it opaque, could now take their places around the protons. Once the Universe was transparent, the glow of the primordial fireball could not vanish and it must still fill the present Universe. Yet the Universe has expanded since by a factor of about 1000 and the radiation's wavelength must have been stretched by the same factor such that – as *predicted* by G. Gamow in 1946 – it must now lie in the mm wavelength region. The actual observation of the so-called "Cosmic Microwave Background" (CMB) with a radio telescope by A. Penzias and R. Wilson in 1965 is considered as the definite proof for a once very hot and tiny Universe, and the discoverers were accordingly awarded the Nobel Prize.

The CMB radiation not only supports the Big Bang Universe, but also shows *minute* temperature variations from one place in the sky to another. These fluctuations are the very early seeds of the subsequent structure formation. The smallness of the fluctuations is the most dramatic demonstration of evolution in the Universe, because they reveal that the Universe at an age of 500,000 years was almost structureless and has since mysteriously evolved to ever higher complexity.

Although electromagnetic radiation cannot bring us information from still earlier epochs, there are other ways to investigate the young Universe. The primordial nucleo-synthesis 100 seconds after the Big Bang has already been mentioned. Theory supported by direct experiments at CERN, can describe the Universe back to an age of 10^{-10} seconds. Then one depends on certain observed boundary conditions and our knowledge becomes more and more speculative with every factor of 10 one goes back in time. At



an age of 10^{-42} seconds our understanding of physics breaks down, because General Relativity and Quantum Physics have to be applied simultaneously, which is still impossible. Even if this barrier will eventually be overcome there is no hope to ever describe the time zero. The Big Bang itself will always remain hypothetical.

Figure 11. The so-called Einstein Cross. It shows in the center a galaxy which acts as a gravitational lens. The four surrounding images come from *one* quasar in the background. As the quasar varies in luminosity the four images react at different times (!) because the light paths differ by some light-days (i. e. several 25,000 million km).

The Concept of Time

In daily life one has a rather unrealistic concept of time. Time is considered as connecting the past eternity with the future eternity in uniform steps. In reality time began at zero in the Big Bang. There was no time *before* the Big Bang, and the question what was before the Big Bang has for the physicist no meaning. Whether, on the other hand, time will lead into a future eternity depends on whether the cosmic expansion will eventually be halted, in which case the Universe will collapse under its own gravity into the "Big Crunch" and this would also be the end of time, or whether it will expand forever. Astronomers believe there is not enough matter in the Universe (including the above-mentioned Dark Matter) to ever halt the expansion. This believe was strengthened in 1998 when rather strong evidence was found for a fifth force in the Universe (Einstein's so-called "Cosmological Constant") which drives the Universe into an accelerated expansion. It would correspond in the yeast cake to an *increasing* power of the yeast. Therefore the view is favored that the Universe will expand forever and time will never come to an end. It is, however, imprudent to extrapolate our limited knowledge into eternity, and one should say more carefully that the Universe is likely to expand into a very distant future.

In addition time is uniform only for an observer who travels with constant speed through space. Any acceleration will change his clock rate. Consequently there are hardly any two observers in the Cosmos who have the same clock rate. In our technical age this relativistic effect has already *practical* consequences. Clocks on satellites orbiting the Earth run slightly differently than on the ground. If this difference was not accounted for communication with these satellites would soon be lost. This leads also to the famous "twin paradoxon" which requires that a twin travelling fast through space will return younger than his brother left on the ground.

The variability of the clock rates implies immediately that the idea of simultaneity is a misconcept. A stunning example is the Einstein Cross **(Figure 11)**, where the gravity field of a foreground galaxy bends the light rays of a distant quasar (i. e. a particularly luminous galaxy) such that the same quasar is seen four times! Quasars have variable luminosities and the four images vary accordingly in brightness. However, the four light paths differ somewhat in length, and as a consequence a given luminosity outbreak of the quasar is observed in the four images at *different* times! If the Einstein Cross was a prominent object in the sky this would have been noticed long ago and the concept of simultaneity would never have been born.

In addition to the Einstein Cross many other gravitational lenses are known today **(Figure 12)**. They all are impressive demonstrations of Einstein's prediction that light rays are deflected by gravity. Therefore light propagates through space along curved lines. But since nothing is more straight than a light ray, our believe in the existence of straight lines is also a misconcept.

In principle the entire space could be positively curved, flat, or negatively curved. Even if space is flat, as presently believed, the existence of structures and the accompanying variations of the gravity field, force every light ray on a curved path.

Figure 12. All the blue images in the picture come from the same galaxy or from parts of it. In about half the galaxy's distance is a large cluster of galaxies whose tremendous gravitational field bends the light and causes the multiple, distorted images of the background galaxy. (Picture taken with the Hubble Space Telescope).

Epilogue

The few examples from the field of astronomy show that our thinking is fundamentally influenced by science. Nature is infinitely more complex than conceived in daily life. The human mind tends to extrapolate daily experiences into natural laws which have nothing to do with reality. Apparent facts like the absoluteness of time or the existence of straight lines are erroneous. On the other hand speculations about a single creation event and subsequent evolution are factual.

It is here the place still to counter another misconcept. We have today not much difficulty to accept the fact of evolution. But we make easily the mistake to assume that the present World and present man are the end product of evolution. In reality evolution must continue. In 100,000 years we will not recognise our descendants, and in say 50 Gigayears the last star will extinguish and thus terminate all possibility for life. Organic life, which we value so highly, is nothing but an intermezzo in the evolution of the Universe.

The author

Gustav Andreas Tammann, born in Göttingen on the 13th of June in 1932, spent his early years in Basle, where he attended the schools and then registered at the University of Basle in astronomy. Indeed, this proved to be a wise decision. On one hand, the old humanist tradition of this University provided G.A.Tammann with essential cultural ingredients. On the other, it was the key to a field of science which was due to bring forth many spectacular discoveries in the years to come. Further study years saw him in Freiburg i. Br., Göttingen and Basle again, where, in 1961, he acquired the Dr. phil degree.

In 1963, the great American astronomer Professor Allan Sundage offered G.A.Tammann the opportunity to join his research team in the frame of a post-doctoral employment in Pasadena (USA) introducing him into the research of the extragalactic Universe. The focus of research was on the determination of extragalactic distances and the expansion of the Universe. This subject, and the still lasting friendship with Allan Sandage, are among the most important cornerstones of his stay in Pasadena. In 1972, he returned to Europe and was appointed Professor of Astronomy in Hamburg and in 1977 in Basle, where he became Director of the Astronomical Institute. In these years, G. A. Tammann began with furthering scientific activities in Switzerland. Among many other honours, he was elected scientific representative in the European Space Agency's Science Programme Committee and Chairman of the Swiss Committee of Space Research. In both functions, he followed Professor Johannes Geiss. When it came to decide on the Swiss membership to the European Southern Observatory, it was up to J. Geiss and G. A. Tammann to develop the necessary lobbying activities.

G.A.Tammann is not only a wonderful teacher, but also an outstanding communicator to the public at large. In countless interviews on radio and television but also in many articles, he reported on the findings of modern astronomy.

A curriculum of G. A. Tammann would not be complete if one would not briefly mention his interests outside the boundaries of natural sciences: music, history and art. Classical humanistic tradition and the paradigms of modern scientific research have been the grounds allowing G. A. Tammann to grow to one of this country's best-known scientists.

INTERNATIONAL SPACE SCIENCE INSTITUTE
Nr. 4, Oktober 1999

SPATIUM

Publikationsorgan des Vereins Pro ISSI

Kometen

Kometen sind beredte Zeugen aus der Zeit vor 4,6 Milliarden Jahren, als aus dem solaren Nebel unser Sonnensystem entstand. Die modernen Mittel der Weltraumforschung ermöglichen, ihre vielfältigen Geheimnisse zu entziffern.

Editorial

Es ist ein alter Traum der Menschheit, in die Zukunft sehen zu können. In der griechischen Antike hat es Delphi zu Berühmtheit gebracht, wo Eingeweihte den aus Felsspalten aufsteigenden Dämpfen Hinweise über Krieg und Frieden, Hochzeiten in Königshäusern und andere wesentliche Ereignisse zu entnehmen wussten. Kometen haben in allen Kulturen als Boten der Zukunft wichtige Dienste geleistet. Ihr während langer Zeit nicht voraussagbares Erscheinen wurde mit Bevorstehendem in Zusammenhang gebracht und als Zeichen überirdischer Mächte gedeutet. Computersimulationen haben heute die alten Dämpfe weitgehend ersetzt; doch geblieben ist das Bedürfnis, in die Zukunft zu sehen.

Auch die Bedeutung der Kometen hat sich mittlerweile geändert. Ihre Bahnen sind berechenbar geworden, und wir wissen heute, dass sie Materie in ursprünglicher Zusammensetzung aus der Entstehungszeit unseres Sonnensystemes vor etwa 4,6 Milliarden Jahren aufbewahrt haben. Im Gegensatz dazu ist die Materie, aus welcher sich damals Planeten und Monde bildeten, durch vielfältige Prozesse so weit umgeformt worden, dass sie keine Spuren ihrer ursprünglicher Eigenschaften mehr besitzt. Kometen sind damit zu Boten der Vergangenheit geworden.

Frau PD Dr. Kathrin Altwegg-von Burg ist Experimentalphysikerin. Sie arbeitet am Physikalischen Institut der Universität Bern und am International Space Science Institute. Wir freuen uns, mit dieser vorliegenden vierten Nummer des Spatium wieder eine prominente Persönlichkeit unseren Lesern vorstellen zu dürfen, die weit über unsere Landesgrenzen hinaus Ansehen erlangt hat.

Bern, im Oktober 1999
Hansjörg Schlaepfer

Impressum

SPATIUM
Publikationsorgan des Vereins Pro ISSI
erscheint ein- bis zweimal jährlich

INTERNATIONAL SPACE SCIENCE INSTITUTE

Verein Pro ISSI
Hallerstrasse 6, CH-3012 Bern
Tel. ++41 31 631 48 96
Fax ++41 31 631 48 97

Präsident
Prof. Hermann Debrunner, Universität Bern
Herausgeber / Quästor
Dr. Hansjörg Schlaepfer, Contraves Space, Zürich
Layout
Marcel Künzi, Contraves Space, Zürich
Druck
Drucklade AG, CH-8008 Zürich

Titelbild
Der Komet Hale-Bopp, 1997
(A. Dimai und D. Ghirardo, Col. Druscie Obs.)

Von Giotto bis Rosetta –

Kometen als Schwerpunkt der europäischen Weltraumforschung

PD Dr. Kathrin Altwegg, Physikalisches Institut der Universität Bern

Vortrag für den Verein Pro ISSI am 24. März 1999

Kometen: unheimliche Boten des Himmels

Kometen haben die Menschheit schon immer fasziniert. Ihr plötzliches Auftreten am Nachthimmel, ihre ungewöhnlichen Bahnen und das Erscheinungsbild mit dem langen Schweif, das sie so sehr von Sternen und Planeten unterscheidet, haben die Phantasie der Leute angeregt. Sie tun es noch heute, wie das Auftreten von Komet Hale-Bopp 1997 (Titelblatt) einmal mehr gezeigt hat. Schon im alten China, lange vor Christi Geburt, wurden Kometen wissenschaftlich genau registriert. Ihre Bahnen, die Länge des Schweifes und ihre Helligkeit wurden gewissenhaft aufgeschrieben. Kometen galten lange Zeit als Boten einer überirdischen Macht. Meist verkündeten sie ein kommendes Unheil, den Untergang eines Königreiches, den Tod eines Herrschers, sie waren Vorboten einer Naturkatastrophe. Auf dem Teppich von Bayeux **(Figur 1)** ist z.B. der Normannenkönig Harold dargestellt, wie er beim Erscheinen des Kometen Halley 1066 die kommende Niederlage bei Hastings gegen den englischen König Edward vorausahnt. Manchmal galten aber Kometen auch als besondere Glücksbringer wie in der **Figur 2**, einem Gemälde des italienischen Malers Giotto di Bondone, 1321, wo der gleiche Komet Halley als Stern von Bethlehem dargestellt ist. Heute wissen wir, dass Kometen keineswegs bedrohliche Boten sind, dass sie vielmehr zu unserem Sonnensystem gehören und wir können auch den langen Schweif durchaus erklären. Wir wissen aber auch, dass Kometen zwar nicht direkt unser Leben beeinflussen, dass auch grosse politische Ereignisse nicht von Kometen vorausgesagt werden, dass sie aber durchaus eine Botschaft in sich tragen, die sich zu erforschen lohnt. Sie können uns etwas über die Vergangenheit unseres Sonnensystems und damit unserer Erde erzählen, über die Entstehung der Sonne und der Planeten und über eine längst verschwundene dunkle Molekülwolke, aus der unser Sonnensystem entstanden ist. Damit können sie uns letztendlich auch Hinweise geben über den Ursprung des Lebens. Dies möchte ich im folgenden erläutern und damit erklären, warum Kometen im Zentrum des Wissenschaftsprogramms der Europäischen Weltraumagentur ESA stehen.

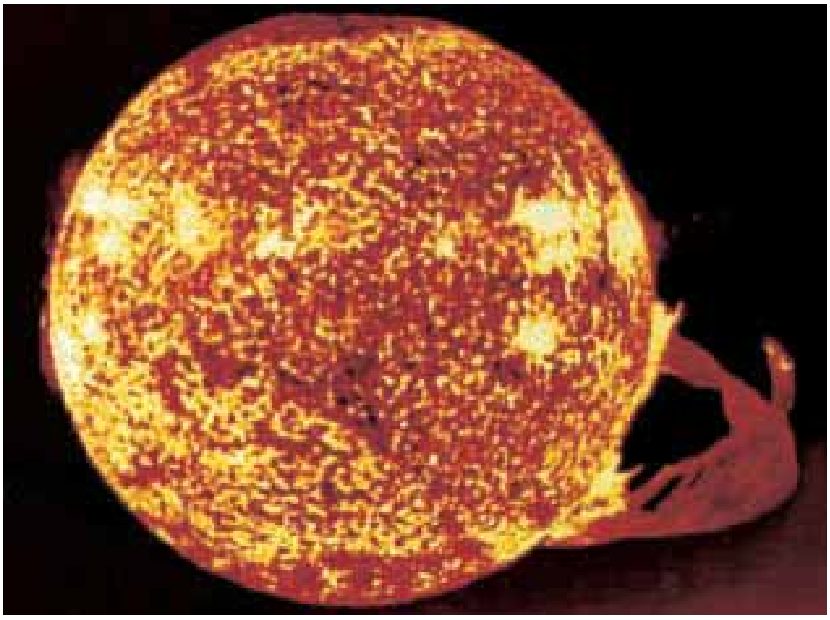
Figur 1
Der Teppich von Bayeux, Schlacht von Hastings (British Museum)

Figur 2
Die Anbetung der Weisen, Giotto di Bondone, 1321

Der Stammbaum unseres Sonnensystems

Die Kinder

Betrachten wir unser Sonnensystem einmal als Familie und beginnen wir mit den Kindern. Die Sonnenfamilie hat viele verschiedene Kinder. Alle Kinder sind gleich alt, nämlich rund 4,6 Milliarden Jahre. Das prominenteste Kind ist sicher die Sonne **(Figur 3)**. Sie hat mehr als 99% aller Materie in sich vereinigt und ihren Geschwistern nur sehr wenig übriggelassen. Die Planetengeschwister variieren in der Grösse und in ihrem Abstand zur heissen Sonne. Je nach Abstand und Grösse hat sich ihr Material in den 4,6 Milliarden Jahren ihres Daseins stark verändert. Dies gilt natürlich vor allem für die inneren Planeten, wie Erde, Merkur, Venus und Mars **(Figur 4)**. Die übrigen Geschwister, die Meteoriten, Asteroiden und Kometen **(Figur 5)** haben eine unbedeutende Grösse. Während die Asteroiden zwischen Mars und Jupiter beheimatet sind, haben Kometen von den 4,6 Milliarden Jahren praktisch die gesamte Zeit tiefgekühlt ausserhalb des äussersten Planeten, des Pluto, verbracht. Darum konnten sie ihre Geheimnisse bis auf den heutigen Tag bewahren.



Figur 3
Die Sonne (NASA)

Die Mutter

Alle diese Körper sind aber aus derselben Urmasse entstanden, nämlich aus dem solaren Nebel (in unserem Stammbaum also die Mutter). Dieser solare Nebel hat sich längst aufgelöst. Kometen entstanden einerseits zwischen Uranus und Neptun. Durch Gravitationskräfte wurden die kleinen Körper gestreut. Ein Teil ging dabei verloren, ein Teil aber findet sich am Rande unseres Sonnensystems, in der sogenannten Oort'schen Wolke wieder. Aus der Häufigkeit von Kometenbeobachtungen schätzt man, dass die Oort'sche Wolke ungefähr 10^{13} Kometen enthält. Die Wolke befindet sich zwischen 300 und 10 000 astronomischen Einheiten

Figur 4a
Der Mars (HST, NASA / ESA)

Figur 4b
Die Erde (NASA)

Figur 5a
Der Asteroid Ida mit dem Mond Dactyl (NASA)

Figur 5b
Der Komet Hyakutake (R. Scott und J. Orman in the Night of the Comet)

Figur 5c
Der Mars-Mond Deimos (HST, NASA / ESA)

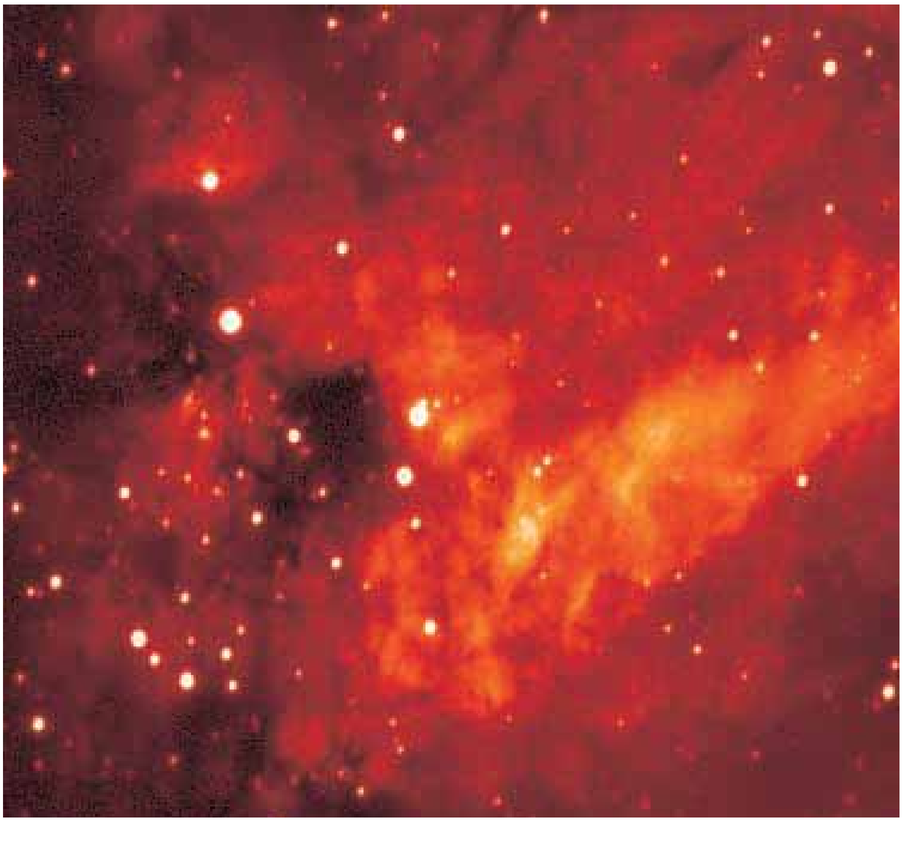
(1 Astronomische Einheit entspricht dem Abstand Sonne–Erde). Die Wolke ist nicht nur in der Ekliptik (Ebene, in der sich die Planeten bewegen) zu finden, sondern umspannt unser Sonnensystem mehr oder weniger isotrop. Oort'sche Kometenbahnen haben deshalb häufig einen relativ grossen Winkel zur Ekliptik und der Orbit kann durchaus auch retrograd sein (z.B. Halley), also umgekehrt verlaufen als die Planetenbahnen. Andere Kometen wurden vermutlich zwischen Jupiter und Saturn gebildet. Ihre Streuung erfolgte vorwiegend in der Ekliptik und die Kometen befinden sich heute im sogenannten Kuiper-Gürtel bei etwa 300 astronomischen Einheiten, immer noch weit ausserhalb von Pluto.

Figur 4c
Der Saturn (HST, NASA / ESA)

Die Grossmütter

Gehen wir jetzt eine Generation zurück. Über den solaren Nebel, die Mutter unseres Sonnensystems, ist recht wenig bekannt. Insbesondere wissen wir wenig über die chemischen und physikalischen Randbedingungen, wie z.B. Temperaturen und Moleküldichten. Geburtsstätten solcher solarer Nebel und damit in unserem Bild Grossmütter sind sogenannte dunkle Molekülwolken **(Figur 6)**.

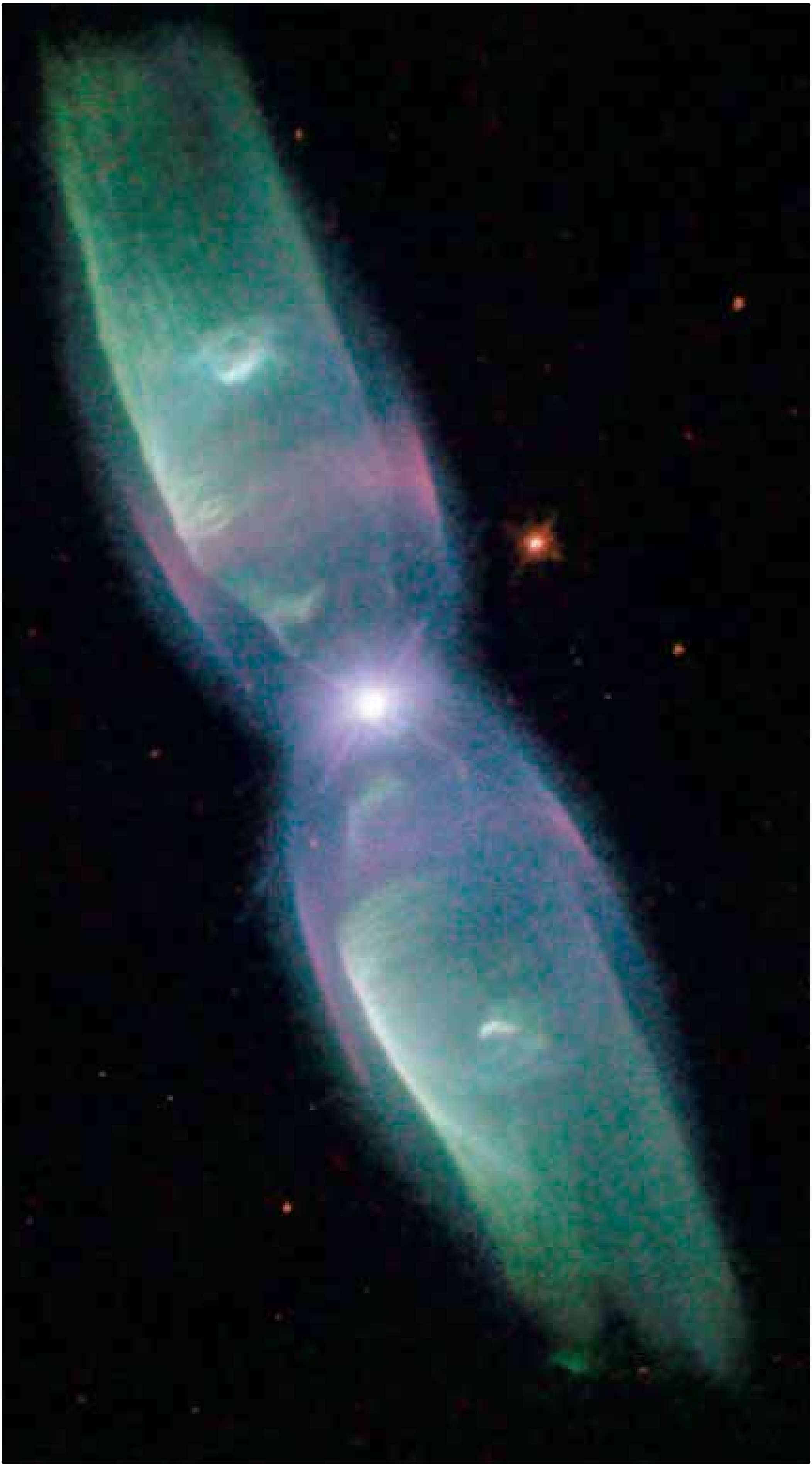
Figur 6
Dunkle Molekülwolke, wie der hier abgebildete Schwan-Nebel, sind die Geburtsstätten neuer Sterne. Von einigen der sichtbaren Sterne vermutet man auf Grund ihrer ungewöhnlichen Farben, dass es sich um ganz junge Sterne handelt, die (noch) von der Gaswolke abgeschirmt werden, aus welcher sie entstanden sind. (B. Wallis und R. Provin, HST NASA / ESA)



Diese Wolken enthalten sehr viel Staub. Das Licht wird in der äusseren Schicht deshalb absorbiert. Im Innern dieser Wolken herrschen ungewöhnliche Bedingungen: grosse Kälte und Dunkelheit. Durch Kollaps gewisser Teile einer solchen Wolke entstehen Gebiete, eben solare Nebel, die dicht genug sind, um ganz zu kollabieren und dabei Sterne und Planeten zu bilden.

Die Vorfahren

Wer liefert den Staub, der die dunklen Molekülwolken bildet? Staub kann in Sternhüllen, Sternatmosphären und bei Supernovae Explosionen entstehen. Diese sind also ihrerseits die Vorfahren der Molekülwolke, aus der unser solarer Nebel und schlussendlich die Planeten und Kometen entstanden sind. In **Figur 7** sieht man als Beispiel die Gashülle eines Sterns, der seine Materie nach einer Explosion weit in den Weltraum schleudert. Dieser Zyklus wurde wahrscheinlich seit der Entstehung des Universums vor etwa 13 Milliarden Jahren mehrmals durchlaufen. Es entstanden und vergingen Sterne, die Staub produzierten, daraus entstanden dunkle Molekülwolken und daraus wiederum solare Nebel und neue Sterne, unter anderem auch unsere Sonne.



Figur 7
Die Flügel des Schmetterlingsnebel sind durch eine Explosion der Gashüllen des Sterns M-29 entstanden. Er befindet sich etwa in 2100 Lichtjahren Entfernung. (B. Balick, Universität Washington, Hubble Space Telescope, NASA / ESA)

Die Geschichte der Materie unseres Sonnensystems

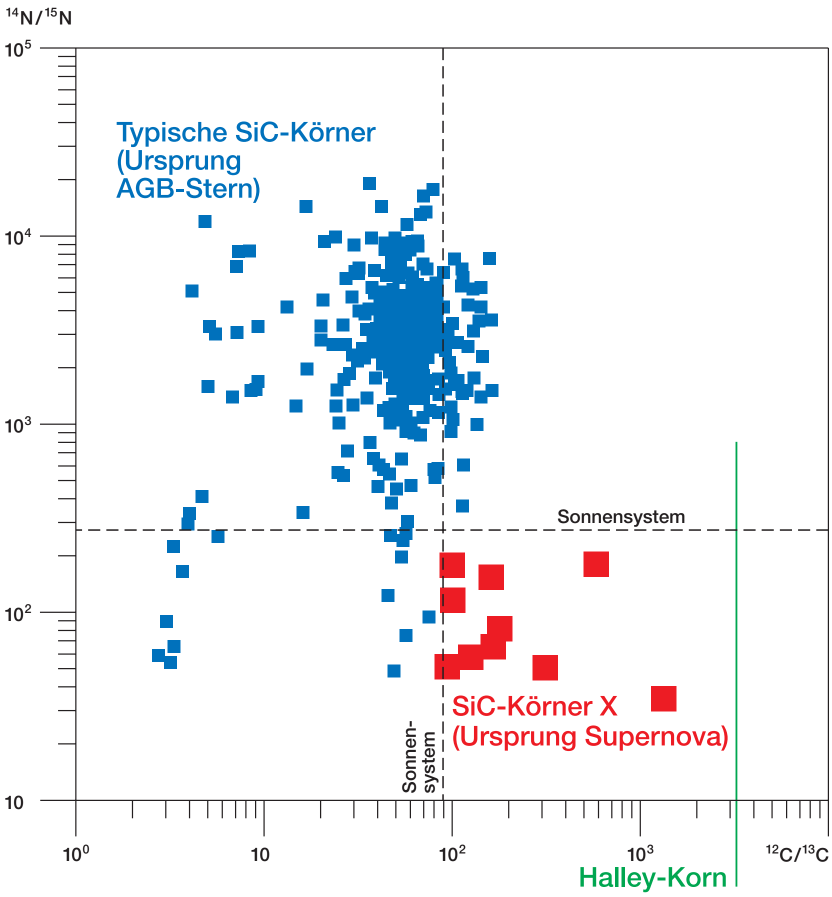
Die Kernsynthese

Woher stammt das Material, aus dem unser Sonnensystem besteht? Verfolgen wir einmal die Materie zurück zu ihrem Ursprung. Dabei müssen wir unterscheiden zwischen Elementen und Molekülen. Beim Urknall wurden die leichten Elemente Wasserstoff, Helium und Lithium geschaffen. Insbesondere für den Wasserstoff ist dies die einzige Quelle, d.h. jedes Wasserstoffatom, das es heute gibt, ist 12–13 Milliarden Jahre alt. Anders für schwerere Elemente, Kohlenstoff bis Uran. Diese stammen aus der Nukleosynthese in Sternen und Supernovae, können also jünger sein und werden unter anderem auch in unserer Sonne hergestellt. Die Nukleosynthese braucht sehr hohe Energien. Molekülbindungen hingegen verlangen nach tieferen Temperaturen, sonst sind sie nicht stabil.

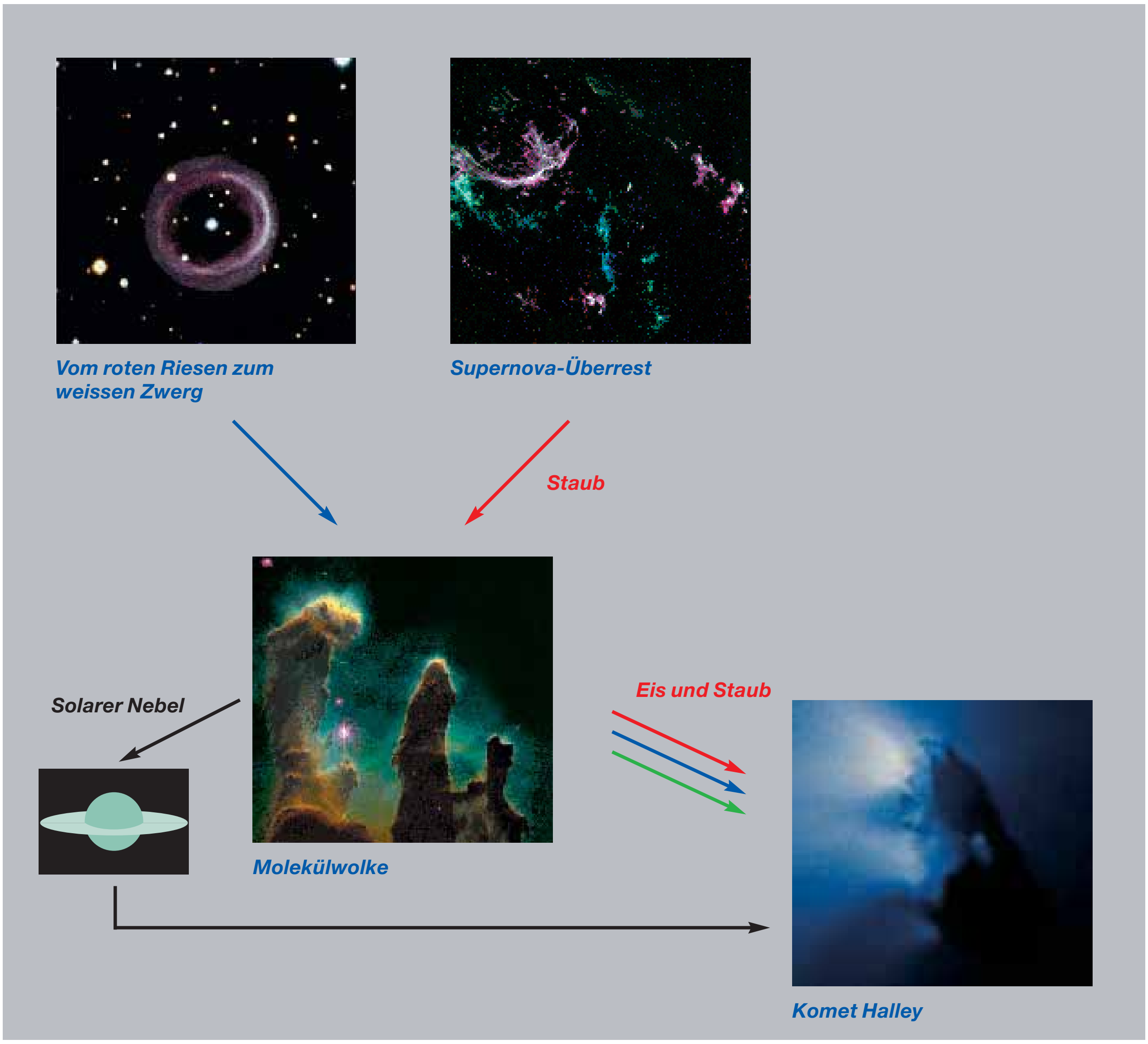
Wie oben erwähnt, entstand und entsteht Staub, d.h. sehr stabile, chemische Verbindungen wie z.B. Eisenoxyd oder Siliziumcarbid, in Sternhüllen, Sternatmosphären oder bei Supernovae-Explosionen. Dieser Staub hat zum Teil seit seiner Entstehung bis in unsere Zeit überlebt. Dies kann man an Meteoritenstaub zeigen. **Figur 8** zeigt dazu Isotopen-Messungen an einem Meteoriten. Der Meteorit wurde im Labor in viele kleine Bruchstücke zerlegt. Misst man nun Isotopenverhältnisse in den einzelnen Bruchstücken, erwartet man eigentlich eine statistische Verteilung um den typischen Wert unseres Sonnensystems, da der Meteorit ja aus dem gleichen Material entstand wie die Sonne. Aus der Figur geht aber klar hervor, dass der Meteorit Körner aus zwei verschiedenen Populationen besitzt. Die eine Population hat ein erhöhtes $^{14}N/^{15}N$-Verhältnis, was typisch ist für einen Ursprung in einem älteren Stern. Die andere Population hat ein erhöhtes $^{12}C/^{13}C$-Verhältnis, was auf einen Ursprung in einer Supernova hindeutet. Im gleichen Meteorit befinden sich also Staubkörner, die unterschiedliche Vorfahren haben. Das heisst, bei der Bildung des Sonnensystems wurden diese Körner nicht aufgebrochen und neu gebildet, sondern sie wurden im Originalzustand eingebaut.

Die Synthese der Moleküle

Etwas anders sieht es bei Molekülen aus, deren Bindung schwä-



Figur 8
Isotopenmessungen an Körnern des Murchinson Meteoriten (J. Hoppe et al. 1996)



Figur 9
Der Ursprung der Moleküle, des Eises und des Staubes in den Kometen

cher ist, z.B. bei Wasser oder organischen Molekülen. In der Umgebung von Sternen oder Supernovae sind die Temperaturen viel zu hoch und solche Moleküle haben keinen Bestand. Wie oben schon erwähnt, sind die Temperaturen in den dunklen Molekülwolken so tief, dass keine «normale» Chemie mehr möglich ist. Trotzdem laufen chemische Prozesse ab, allerdings sehr langsam. Die Molekülwolken existieren mehrere Millionen Jahre. Mit Hilfe der Staubkörner, die katalytisch wirken, können in diesen Wolken flüchtige Moleküle entstehen. Aus den Elementen H, C, N und O kann an der Oberfläche der Staubkörner eine grosse Anzahl an Molekülen gebildet werden wie z.B. Wasser, Formaldehyd (H_2CO), Kohlenmonoxyd und -dioxyd (CO, CO_2), Amoniak (NH_3), Methanol (CH_3OH), aromatische Kohlenwasserstoffe (siehe **Figur 9**). Viele dieser Moleküle wurden denn auch schon in dunklen Molekülwolken durch Radioastronomie identifiziert. Aus Laborexperimenten ist zudem bekannt, dass aus diesen Molekülen zusammen mit etwas Ultraviolettstrahlung sehr viel komplexere Moleküle bis hin zu Aminosäuren gebildet werden können. Die Frage ist nur: Können diese Moleküle dann die Bildung eines solaren Nebels, den Kollaps und die Bildung von Sonne und Planeten überleben? Wenn wir diese Frage beantworten wollen, müssen wir charakteristische Moleküle und Isotope untersuchen, die darauf hinweisen können, ob das Molekül erst bei oder nach der Bildung des Planeten oder Kometen entstanden ist oder aus einer Zeit vor unserem Sonnensystem stammt.

Kometen: Zeugen der Vergangenheit

Wie schon oben erwähnt, waren die Planeten immer in der Nähe der Sonne während den ganzen 4,6 Milliarden Jahren. Die Temperaturen selbst auf Saturn oder Neptun sind recht hoch. Das Material auf diesen Himmelskörpern hat sich somit chemisch verändert. Wenn wir also möglichst ursprüngliches Material erforschen wollen, das uns Einblick in die Zeit vor der Entstehung der Sonne geben soll, müssen wir Himmelskörper untersuchen, die möglichst weit von der Sonne weg sind, wie z.B. den Pluto oder eben Kometen. Der Mars dürfte kaum mehr Originalmaterial haben, auch wenn Marsforschung heute sehr populär ist. Die Kometen hingegen, die praktisch die ganze Zeit ihres Bestehens weit weg von der Sonne in der kalten Oort'schen Wolke oder im Kuiper-Gürtel verbracht haben, und die praktisch keine interne Heizung durch Radioaktivität besitzen, bieten sich als Reservoir für gut konserviertes Material an. Kometen besitzen zudem noch – im Gegensatz zu Pluto – die praktische Eigenschaft, dass sich hie und da durch Gravitationsstösse einer von ihnen in die Nähe der Sonne, und damit der Erde, verirrt, wo wir ihn mit nicht allzu grossem technischen Aufwand erforschen können.

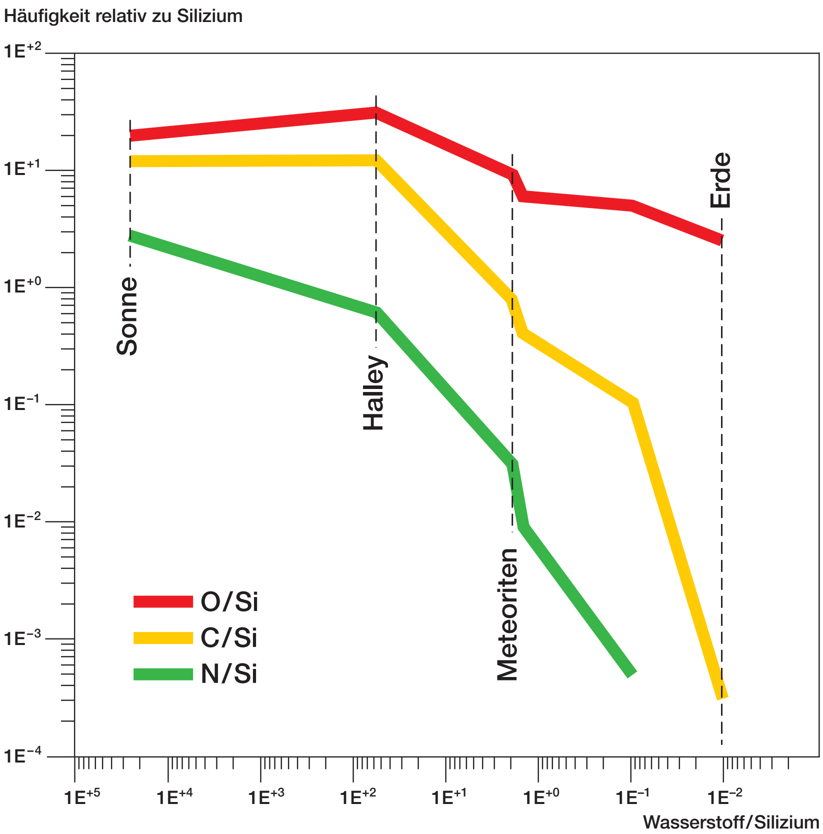
Dies führt dazu, dass Kometen in der Weltraumforschung trotz ihrer geringen Grösse einen sehr wichtigen Platz einnehmen. Die Europäische Weltraumforschung trug diesem Umstand Rechnung mit der erfolgreichen Giotto-Mission zum Kometen Halley und zu Grigg-Skjellerup, aber auch mit der sich im Bau befindlichen Rosetta-Mission.

Im folgenden Abschnitt möchte ich zeigen, wo man heute in der Kometenforschung steht, und was wir von der zukünftigen Kometenforschung noch lernen möchten.

Resultate aus der Kometenforschung

Was wir wissen

In **Figur 10** ist die Elementhäufigkeit von Sauerstoff O, Stickstoff N und Kohlenstoff C im Verhältnis zu Silizium für einige Körper unseres Sonnensystems dargestellt. Da die Sonne 99% der Materie des Sonnensystems besitzt, repräsentiert sie die ursprüngliche Elementhäufigkeit des solaren Nebels. Von den dargestellten Körpern sind ihr die Kometen am ähnlichsten. Der Sauerstoffgehalt und der Kohlenstoffgehalt sind praktisch gleich hoch. Nur Stickstoff scheint Komet Halley etwas verloren zu haben während seiner Existenz, oder er hat den Stickstoff von Anfang an nicht eingebaut. Meteoriten, die ebenfalls als sehr ursprüngliches Material gelten, sind schon deutlich abgereichert in allen Elementen, und die Erde scheint, was Kohlenstoff und Stickstoff anbelangt, den meisten Teil schon verloren zu haben. Kometen sind also, wie postuliert, bei weitem die ursprünglichsten Körper unseres Sonnensystems.



Figur 10
Die Elementhäufigkeiten nach Geiss, 1991

Das Resultat des Vorbeifluges der europäischen Sonde Giotto am Kometen Halley, das am meisten Aufsehen erregte, war natürlich die Photographie des Kerns **(Figur 11)**.

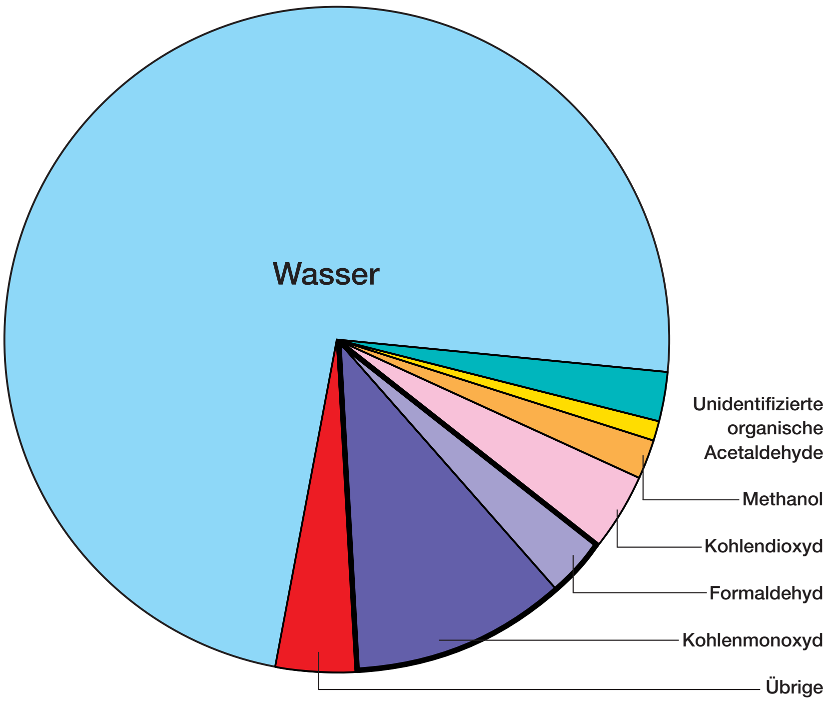
Figur 11
Der Kern des Halley'schen Kometen (Giotto, ESA)

Verblüfft war man vor allem über die schwarze Farbe. Während bisher die Vorstellung vorherrschte, dass ein Komet aus flüchtigem Eis und solidem Staub besteht, weiss man jetzt, dass man es eigentlich mit drei verschiedenen Komponenten zu tun hat: nämlich mit flüchtigem Eis (Wasser, Methanol, Kohlendioxid, etc.), einer zähen Komponente, teerähnlich (evtl. polymerisiertes Formaldehyd; schwere Kohlenwasserstoffe?), die wahrscheinlich die schwarze Farbe ausmacht und einer festen Komponente, dem Staub (Eisenoxyd, Siliziumkarbid). Die teerartige Komponente besteht aus den leichten Elementen C, H, N, O (sie wird deshalb auch CHON genannt), während beim Staub die schwereren Elemente wie Magnesium und Eisen überwiegen. In der Koma des Kometen beobachtet man relativ viel Formaldehyd (CH_2O), das langsam freigesetzt wird, und von dem man annimmt, dass es aus dieser teerähnlichen Substanz stammt. Ob die Oberfläche des Kometen aber wirklich aus polymerisiertem Formaldehyd besteht, oder ob die dunkle Farbe vom Staub stammt, wie dick diese Schicht ist und welche Mechanismen dahinter stecken, dass bei Halley nur ca. 5% der Oberfläche wirklich ausgasen, diese Fragen sind weiterhin offen.

Was wir noch wissen möchten

Es wäre sehr interessant zu wissen, ob der Staub in den Kometen, der ja wahrscheinlich unverändert die 4,6 Milliarden Jahre im Kometen überlebt hat, auch Isotopenunterschiede aufweist wie der im vorherigen Abschnitt gezeigte Meteorit. Leider hat man darüber nur Messungen eines einzigen Korns und damit natürlich keine signifikante Aussage. Immerhin scheint das Halley-Korn ein viel zu hohes $^{12}C/^{13}C$-Verhältnis zu haben und stammt deshalb eher aus einer Supernova.

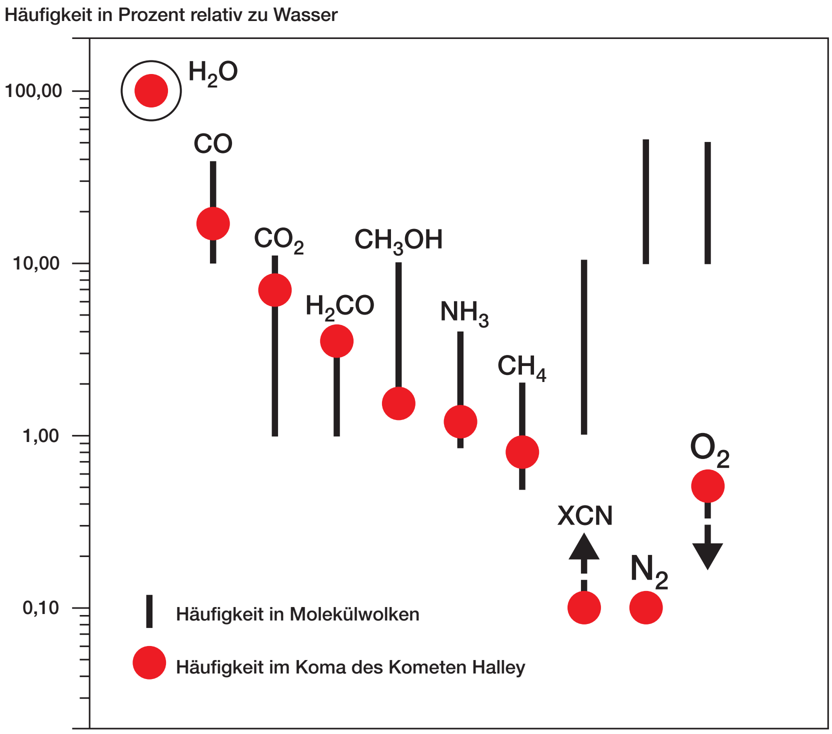
Wenn wir uns jetzt den volatilen Teil der Kometenkoma anschauen **(Figur 12)**, ist klar, dass Wasser das häufigste Molekül ist, wie bereits viel früher durch Whipple (1950) postuliert. Allerdings überrascht die Fülle an anderen Mo-



Figur 12
Flüchtiges Material im Koma des Kometen Halley

lekülen, vor allem natürlich an organischen Substanzen. Die Massenspektrometer bei Halley konnten nur Substanzen bis zu einem Molekülgewicht von 56 atomaren Masseneinheiten (amu) identifizieren. Man weiss aber, dass Molekülgewichte bis mehr als 100 amu vorkommen und eine der grossen Aufgaben für die zukünftige Kometenforschung wird die Identifikation dieser Moleküle sein. Eine äusserst wichtige Messung war die Bestimmung des Deuteriumgehaltes im kometären Wasser. Deuterium hat die gleichen chemischen Eigenschaften wie Wasserstoff, ist aber doppelt so schwer. Das Verhältnis D/H in einem Molekül kann deshalb Aufschluss geben über die physikalischen Bedingungen bei der Herstellung des Moleküls. Im heutigen Sonnensystem ist das Verhältnis von Deuterium zu Wasserstoff D/H etwa 3×10^{-5}. Auf der Erde ist dieser Wert etwa 5 mal höher. Die Messungen im Kometen Halley und vor kurzem ebenfalls in den Kometen Hale-Bopp und Hyakutake ergeben einen Wert von 3×10^{-4}, also rund doppelt so hoch wie auf der Erde. Ein so hohes Deuteriumverhältnis herzustellen, ist nicht einfach und kann nur durch Ionen-Molekülreaktionen bei tiefen Temperaturen geschehen. Dafür braucht man aber eine sehr hohe Ionendichte, was im solaren Nebel sehr unwahrscheinlich ist, oder aber sehr viel Zeit. Die Sonne, Planeten und Kometen haben sich aus dem solaren Nebel in wenigen 100 000 Jahren gebildet. Diese Zeit reicht nicht aus, um das beobachtete Deuterium im Kometenwasser zu erklären. Man muss deshalb postulieren, dass das kometäre Wasser älter ist als unser Sonnensystem und schon in der dunklen Molekülwolke, der Grossmutter, entstanden ist.

Wenn aber das Wasser die Bildung und den Kollaps des solaren Nebels überstanden hat, können dann nicht auch organische Moleküle aus der Molekülwolke stammen? Und wenn wir wissen, dass in solchen Wolken komplexe organische Moleküle entstehen bis hin zu Vorläufern von Aminosäuren, den Grundsteinen der Organismen, könnte dann nicht auch der Ursprung des Lebens in solchen Molekülwolken zu suchen sein? Dies alles ist im Moment noch Spekulation. Noch wissen wir zu wenig über die Vorgänge während der Bildung unseres Sonnensystems, über die physikalischen und chemischen Bedingungen im solaren Nebel. Wenn wir allerdings die bis jetzt bekannte Zusammensetzung von Molekülwolken vergleichen mit der im Kometen Halley gemessenen Zusammensetzung, ist eine gewisse Verwandtschaft nicht abzuleugnen **(Figur 13)**.



Figur 13
Vergleich der Häufigkeiten einiger Moleküle in Molekülwolken und im Koma des Kometen Halley

Dabei muss man noch wissen, dass Sauerstoffmoleküle sich nur in sehr kaltem, apolarem Eis einbetten, das sich nur im innersten Teil einer Molekülwolke befindet, und dass Stickstoffmoleküle erst bei sehr tiefen Temperaturen kondensiert werden, so dass sie in Kometen entweder nicht eingebettet wurden oder seither verloren gegangen sind. Die Häufigkeit der übrigen Moleküle ist in bemerkenswerter Übereinstimmung. Dies ist bei weitem kein Beweis, dass flüchtige organische Moleküle in Kometen älter sind als die Sonne, aber doch immerhin eine Möglichkeit. Mehr Untersuchungen an Kometen sind sicher nötig.

Was geplant ist

Aus den vorherigen Erläuterungen sollte hervorgehen, weshalb Kometenforschung auf grosses Interesse stösst. Dies lässt sich auch aus der Zahl der Forschungsprojekte, die sich mit den physisch zwar sehr kleinen, aber von der Zusammensetzung her sehr wichtigen Himmelskörpern befassen, ablesen. In **Tabelle 1** ist eine Zusammenfassung der wichtigsten laufenden Projekte gegeben. Dabei planen sowohl die NASA wie auch die ESA Kometenmissionen. Stardust ist im Februar 1999 gestartet und soll im Schweif des Kometen Wild 2 Staubkörner sammeln und diese zur Erde zurückbringen. Allerdings werden sich diese Untersuchungen auf den steinigen Anteil des Kometen beschränken müssen. Organisches Material dürfte dabei nicht entdeckt werden, da es zu flüchtig ist und sich auf diese Art nicht einfangen lässt. Mit Contour plant NASA eine Mission, die gleich mehrere Kometen besucht, allerdings jeweils nur im Vorbeiflug. Sicher die grösste und vollständigste Mission wird aber die ESA-Rosetta Mission sein. Darüber im nächsten Abschnitt mehr.

Tabelle 1

Zukünftige Kometenforschung

Beobachtung von vorbeiziehenden Kometen von der Erde aus und mit Satelliten-basierten optischen Instrumenten (Very Large Telescope [VLT], Hubble Space Telescope, etc.); z. B.	**Hale-Bopp**
Stardust, NASA: Start 1999, Einsammeln von Staubkörnern im Schweif des Kometen	**Wild 2** **2006**
Contour (Comet Tour), NASA, Vorbeiflug an mehreren Kometen,	**Encke** **2003** **Schwassmann-Wachmann-3** **2006** **d'Arrest** **2008**
Rosetta, ESA: Rendez-vous mit Komet Wirtanen, Begleitung des Kometen auf seiner Bahn zum Perihel, Absetzen eines Landers	**2011–2013**

Rosetta

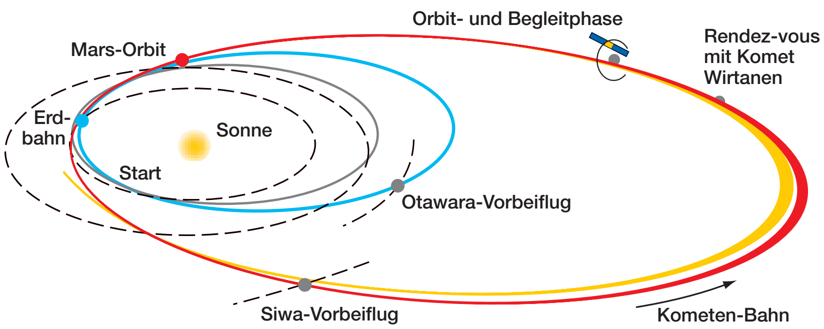
Von Hieroglyphen und Kometen

Die Mission ist nach dem Rosetta-Stein benannt **(Figur 14)**. Dieser Stein wurde 1799 durch die Truppen Napoleons in Ägypten gefunden. Eingemeisselt ist dreimal derselbe Text, nämlich in Hieroglyphen, in demotischer Schrift und in Griechisch. Dadurch wurde es möglich, Hieroglyphen zu entziffern und damit viele Geheimnisse der ägyptischen Kultur zu lüften. Dasselbe hofft man nun von der Rosetta-Mission. Sie soll ermöglichen, die Geheimnisse der Kometen zu entziffern und damit die Geschichte unseres Sonnensystems zu verstehen. Sie wird im Januar 2003 mit einer Ariane 5 von Kourou gestartet. Ihr Ziel ist der Komet Wirtanen, ein kurzperiodischer Kuiper-Gürtel-Komet, der alle 5 Jahre einmal um die Sonne kreist. Der Radius dieses Kometen ist nur ca. 600 m. Die Sonde soll den Kometen nahe beim sonnenfernsten Punkt, dem Aphel treffen und dann bis zum sonnennächsten Punkt, dem Perihel, begleiten. Wieso geht die Reise zu einem solch unbedeutenden Kometen? Die Bahn von Wirtanen ist gut bekannt, der Winkel zur Ekliptik relativ klein, und es ist im Gegensatz zu Halley, Hale-Bopp und Hyakutake ein Kuiper-Gürtel-Objekt. Die ersten beiden Bedingungen sind insbesondere wichtig, da die Raumsonde nicht sehr viel Brennstoff mitnehmen kann. Auch so dauert die Reise zum Kometen acht Jahre und bedingt einen Ausflug zum Mars und dann zweimal zur Erde zurück, um genügend Schwung zu holen. In **Figur 15** ist die komplizierte Bahn der Raumsonde dargestellt. Auf dem Weg zum Kometen wird die Sonde zusätzlich zwei Asteroiden, den kleinen Otawara und den doch etwa 100 km grossen Siwa, besuchen. Die Sonde wird beim Start etwa 3000 kg wiegen. Davon ist ca. 250 kg für die Nutzlast reserviert. Der Schwerpunkt der Forschung liegt auf der Zusammensetzung und dem Ursprung der Materie. Sowohl der flüchtige Anteil wie auch der Staub sollen gründlich analysiert werden. Selbstverständlich will man aber auch wissen, wie der Komet von der Nähe aussieht, wie sich seine Aktivität als Funktion der Sonnennähe entwickelt, und nicht zuletzt will man auch ein Landemodul absetzen, das seinerseits 9 Experimente besitzt, die den Nukleus des Kometen bis in eine Tiefe von ca. 1 m erforschen werden. Ein Instrument auf der Sonde, das Gas Massenspektrometer, ist für die Universität Bern von besonderer Bedeutung. Nachdem Bern schon bei der Giotto-Mission massgeblich an den Massenspektrometern beteiligt war, liegt die Hauptverantwortung für das Massenspektrometer ROSINA wiederum bei Prof. Balsiger und seinem Team in Bern.

Figur 14
Der Rosetta-Stein (British Museum)

Rosina

Das Instrument soll das Gas des Kometen vom Beginn der Akti-

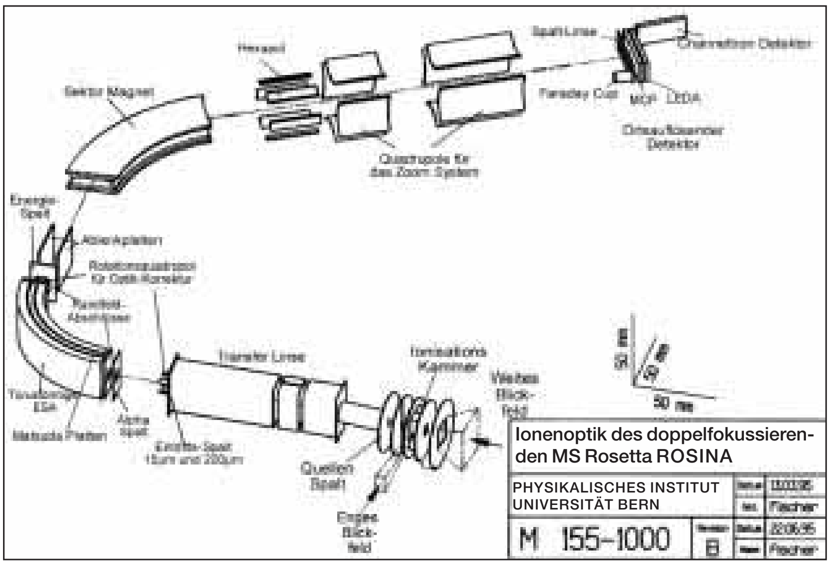


Figur 15
Die Bahn der Rosetta-Mission

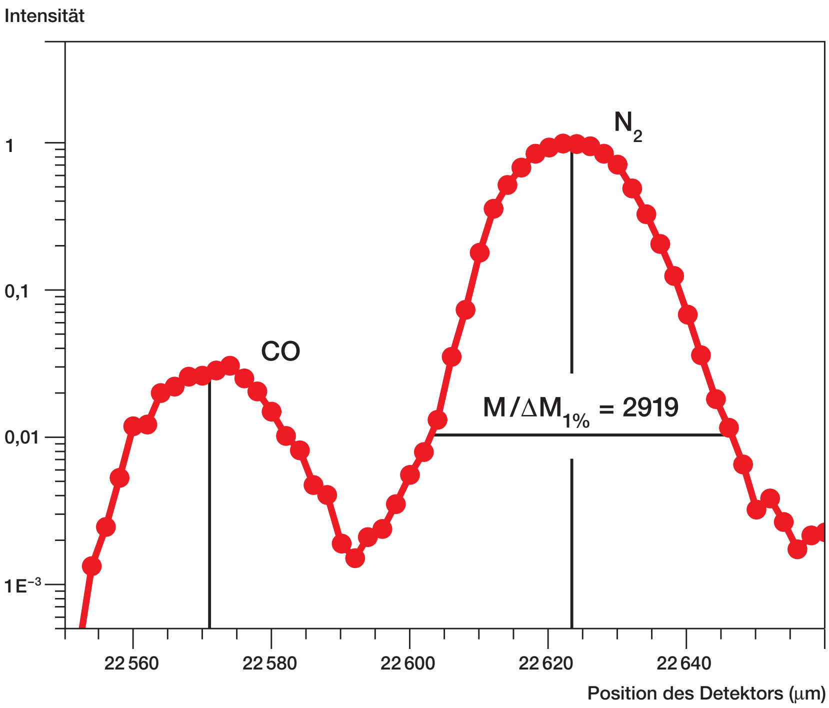
vität bei 3,5 Astronomischen Einheiten bis zum sonnennächsten Punkt untersuchen, und zwar im Hinblick auf die Elementzusammensetzung, die Moleküle hin bis zu schweren organischen Molekülen und die Isotopenverhältnisse. Damit sollte man der Frage nach dem Ursprung des Materials und der Geschichte des Sonnensystems einen grossen Schritt näher kommen. Dies alles kann nicht mit einem einzigen Sensor erreicht werden. Insbesondere schliessen sich der grosse Massenbereich und die verlangte Massenauflösung aus. ROSINA wird deshalb drei Sensoren enthalten, die sich ergänzen, die aber auch wegen der Länge der Mission zum Teil redundant sind.

In Bern steht ein Massenspektrometer, das genügend Massenauflösung besitzt, sehr empfindlich ist und auch zulässt, dass schwere Massen identifiziert werden können. In **Figur 16** ist die Ionenoptik dieses Instruments dargestellt. Das zu untersuchende Neutralgas wird in einer Elektronenstossquelle ionisiert und beschleunigt. Durch einen schmalen Spalt gelangt es in einen toroidförmigen, elektrostatischen Analysator. Dabei wird die Energie der Ionen separiert. Anschliessend gelangen Ionen gleicher Energie in einen Magneten, der den Impuls und damit die Masse der Ionen separiert. Eine Zoom-Optik hilft, die Massenlinien auf einen ortsauflösenden Detektor abzubilden.

Das abgebildete Massenspektrometer füllt allerdings ein ganzes Zimmer, der Magnet allein ist etwa 150 kg schwer. Für eine Weltraummission kann man jetzt zwar alles verkleinern, dabei werden dann aber auch die mechanischen Toleranzen entsprechend kleiner. Es braucht die ganze Erfindungskraft erfahrener Ingenieure, um ein Instrument zu fertigen, das wissenschaftlich ebensoviel leistet, dessen Gesamtgewicht aber nicht mehr als 15 kg ist, das zudem Hitze und Kälte überstehen kann und auch beim Start der Ariane 5 nicht auseinanderfällt. Dass das Prinzip funktioniert, sieht man in **Figur 17**. Dieses Massenspektrum wurde mit einem Prototypen aufgenommen. Man sieht ein Spektrum der Masse 28. Kohlenmonoxyd und Stickstoffmoleküle sind dabei ganz klar getrennt, obwohl sich die Massen nur um ca. $^1/_{3000}$ unterscheiden. Damit sollte es möglich sein, beim Kometen den vorhandenen Stickstoff zu messen und zu entscheiden, ob er wirklich fehlt, oder ob er in einer anderen Form doch vorhanden ist.



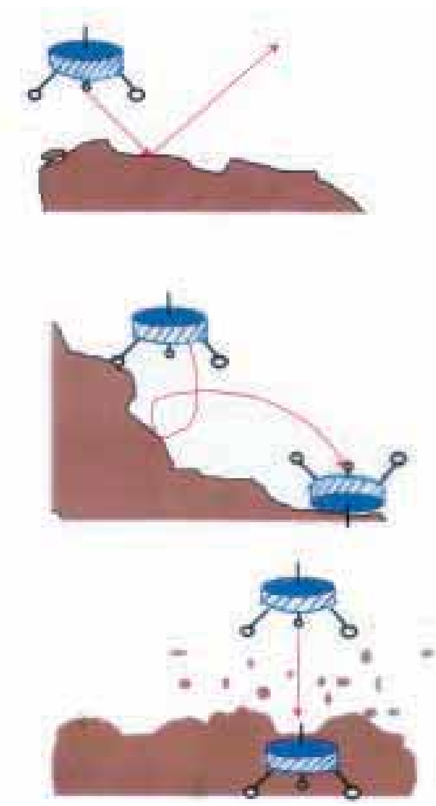
Figur 16
Das Prinzipschema des Rosina-DFMS-Instrumentes



Figur 17
Messungen mit dem Prototypen des Rosina-Instrumentes: Nachweis des massenspektralen Auflösungsvermögens an Hand von CO und N_2

Ein weiteres faszinierendes Experiment ist sicher das Landemodul, das in deutsch-französischer Kooperation gebaut wird **(Figur 18)**. Dabei sind ganz besondere Schwierigkeiten zu überwinden. Zum Beispiel schwanken die Einschätzungen der Festigkeit der Kometenoberfläche um mehrere Grössenordnung je nach Modell. Die Landung auf dem Kometen dürfte dabei besonders schwierig und unsicher werden. In **Figur 18** sind verschiedene Szenarien gezeichnet. Da der Komet so klein ist, hat er praktisch keine Anziehungskraft. Ist die Oberfläche also hart, wird der Lander aufprallen und wie ein Gummiball gleich wieder wegfliegen. Diesem Umstand wird mit Harpunen Rechnung getragen, die bei der Landung ausgeworfen werden, um das Landemodul zu verankern, Die Oberfläche des Kometen ist wahrscheinlich alles andere als flach. Damit bleibt zu hoffen, dass die Landung nicht an einem zu steilen Abhang geschieht. Ist aber die Oberfläche, wie zum Teil vermutet wird, viel flaumiger als der leichteste Pulverschnee, könnte das Modul auch gleich vollständig im Kometen versinken. Es gibt Abschätzungen, die postulieren, dass das Modul eigentlich 17 m lange Füsse haben müsste.

Zahlreiche Wissenschafter und Ingenieure arbeiten zur Zeit intensiv an der Rosetta-Mission. Mit der Länge der Mission und den extremen Umweltbedingungen, der sie ausgesetzt wird, ist die Rosetta-Mission sicher eines der ambitiösesten Projekte der Europäischen Weltraumforschung. Sollte sie gelingen, und davon sind wir alle überzeugt, wird sie einen Meilenstein in der Erforschung des Sonnensystems darstellen. Sicher ist auch, dass uns einige spannende Augenblicke bevorstehen, bevor das Geheimnis der Kometen und damit ein weiteres Kapitel unserer Vergangenheit offen gelegt wird.



Figur 18
Szenarien für die Landung auf dem Kometen Wirtanen

Zur Autorin

Kathrin von Burg ist am 11. Dezember 1951 in Balsthal als Tochter eines Ärztepaares zur Welt gekommen. Sie hat dort die Primarschule besucht, später die Kantonsschule in Solothurn, wo sie 1970 die Maturität vom Typus A (mit Latein, Griechisch und Hebräisch) erworben hat. Nach einem Exkurs in praktische archäologische Arbeiten führte sie ihr weiterer Weg in ihre Vaterstadt Basel, wo sie sich an der Universität in Physik mit den Nebenfächern Mathematik und Chemie immatrikulierte. 1976 erhielt sie das Diplom in Experimentalphysik mit einer Arbeit über das Thema «Bau eines Wattmeters zur Leistungsmessung von Lichtquellen». Eine Arbeit auf dem Gebiet der Optik anorganischer Molekülkristalle trug ihr 1980 die Würde eines Doktors der Physik der Universität Basel ein. In den folgenden beiden Jahren arbeitete Kathrin von Burg als Research Assistant an der Universität New York auf dem Gebiet der Photoelektronenspektroskopie an Flüssigkeiten. Wieder in die Schweiz zurückgekehrt galt es, eine Stelle zu finden und zufälligerweise war beim Physikalischen Institut der Universität Bern eine Stelle frei, und so kam es, dass sie unter der Leitung von Prof. Johannes Geiss Arbeiten im Bereich der Erforschung von Kometen aufnahm. Damals wurde gerade das Massenspektrometer für die Sonde Giotto gebaut, welches später den Halley'schen Kometen und den Kometen Grigg-Skjellerup beobachtete. Ab 1996 übernahm sie die Projektleitung für eine neue Generation von Massenspektrometer für die Sonde Rosetta der europäischen Raumfahrtagentur, welche im Zeitraum 2011 bis 2013 den Kometen Wirtanen aus nächster Nähe beobachten wird. Die Universität Bern erteilte der inzwischen mit dem Physiker Laurenz Altwegg verheirateten Autorin 1996 die Venia Legendi auf dem Gebiet der Experimentalphysik, insbesondere in der Physik des Sonnensystemes. Parallel dazu arbeitet Kathrin Altwegg-von Burg als Beraterin des International Space Science Institutes, wo sie Workshops zum Thema Kometen leitet.

Nicht nur Studenten an den verschiedenen Hochschulen ihres bisherigen Wirkens kamen in den Genuss des umfassenden Wissens von Kathrin Altwegg. Sie ist begeisterte Reiterin und manches junge Pferd hat sich von ihr in die hohe Kunst der Dressur einführen lassen. Im Weiteren erwandert sie gerne mit ihrem Mann und ihren beiden Töchtern die irdische Natur, sei es zu Fuss oder auf den Skiern.

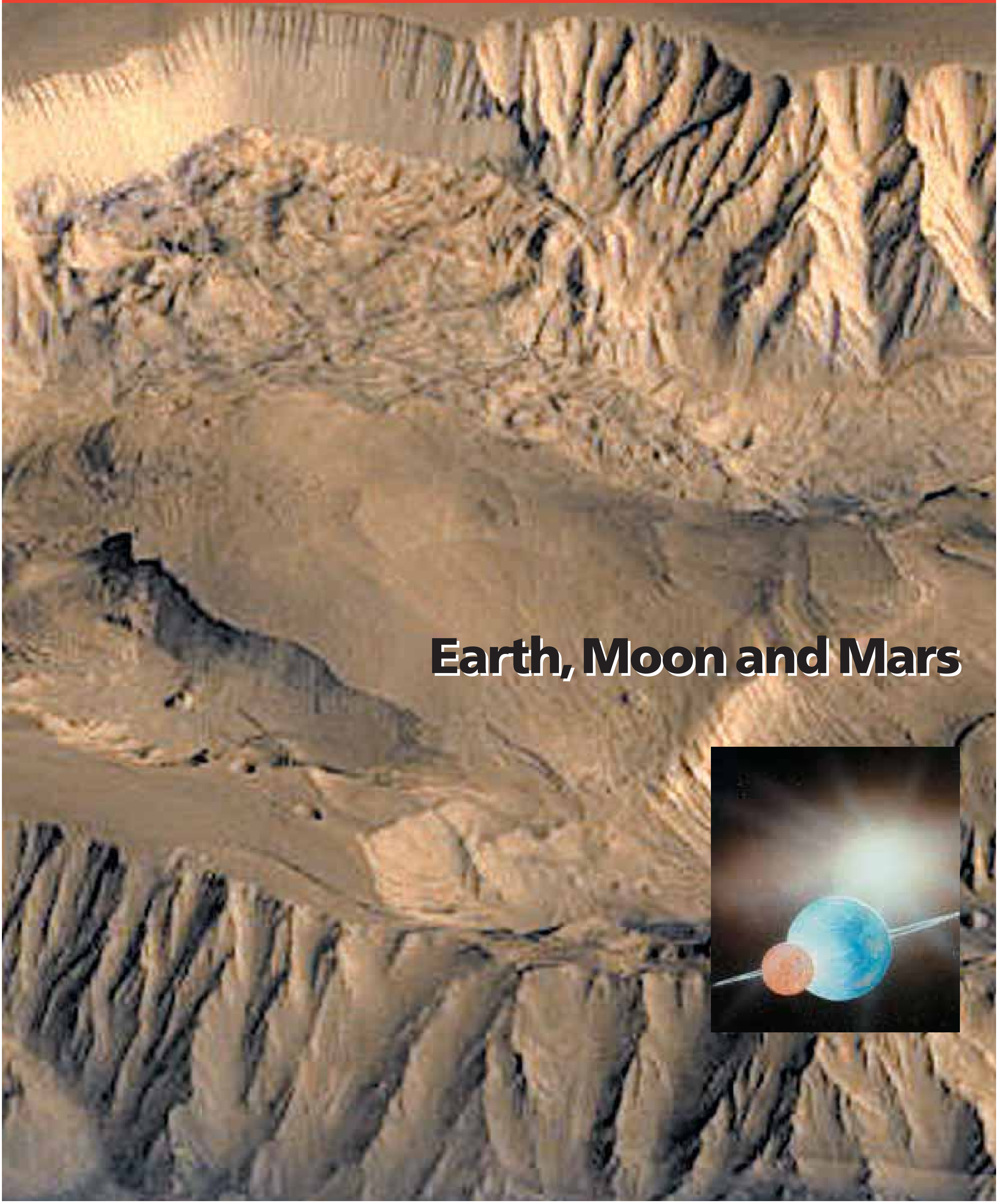
Es ist Kathrin Altwegg-von Burg ein grosses Anliegen, dass Knaben und Mädchen den gleichen Zugang zur Berufsbildung haben. Meitschi und Giele sind gleich intelligent. Vor allem die Bilder, welche die Gesellschaft in unsere junge Generation projiziert, sind die Ursachen für die Absenz von Frauen in Technik und Wissenschaft. An der Astrophysik faszinieren sie die Freiräume des Denkens, die sich in den Tiefen des Alls noch finden lassen. Da ist der Weg zum Aufbau einer neuen Theorie auf Grund der erst in den Anfängen steckenden Grundlagen wesentlich weiter als in anderen Bereichen der Naturwissenschaften. Da ist Kreativität gefragt.

INTERNATIONAL SPACE SCIENCE INSTITUTE

No 5, June 2000

SPATIUM

Published by the Association Pro ISSI



Earth, Moon and Mars

Editorial

Mankind finds itself thrown into a hostile universe. While our bodies are subject to the soulless laws of chemistry and physics governing physical processes on even the most distant star, our minds, our hopes and fears find no spiritual equivalent. The intriguing sketches on the walls of the Altamira cave, for example, may have been the attempt of our early ancestors to create humanlike beings simply in order to alleviate their spiritual loneliness. The Greeks later filled their heavens with a great variety of deities who, by virtue of their human features, clearly betray the intent of their creators.

Little has changed since. The past two thousand years have seen mankind struggling to find a rationale for its existence, hopefully given by an entity beyond its transitory existence. These days, evidence seems to take precedence over belief. Hence, modern science seeks to provide answers to these age-old questions. It is only logical to begin with the nearby celestial bodies, i.e. with the Earth, Moon and Mars, to which the present issue of Spatium is devoted, and to search there for traces of life. But, while this endeavour has provided us with deep insight into the history of our own planet, the feasibility of life on the Moon has meanwhile been excluded, and corroboratory proof of life on Mars has not yet been found.
We are still alone.

Prof. Johannes Geiss is the *spiritus rector* of the International Space Science Institute in Bern and its Executive Director. We are very grateful for his kind permission to publish herewith an adapted version of his address at the Fourth Anniversary of the ISSI foundation.

Bern, June 2000
Hansjörg Schlaepfer

Impressum

SPATIUM
Published by the
Association Pro ISSI
twice a year

INTERNATIONAL SPACE SCIENCE INSTITUTE

Association Pro ISSI
Hallerstrasse 6, CH-3012 Bern
Phone ++41 31 631 48 96
Fax ++41 31 631 48 97

President
Prof. Hermann Debrunner,
University of Bern
Publisher
Dr. Hansjörg Schlaepfer,
Contraves Space, Zurich
Layout
Marcel Künzi, marketing · kommunikation, CH-8483 Kollbrunn
Printing
Drucklade AG, CH-8008 Zurich

Frontcover
Mars: South of Candor Chasma
(NASA/JPL)

Earth, Moon and Mars *)

Johannes Geiss, International Space Science Institute, Bern

Introduction

The ultimate goal of planetary research is to find answers to the big questions: What happened when 4'600 million years ago the Sun and the planets formed out of an interstellar cloud? What were the external circumstances and specific processes that governed the origin of the Earth and determined its evolution to the present state? Is there – or was there – life in the past on another planet or on one of the moons in the solar system? From the presence or absence of signs of life outside the Earth, can we draw conclusions about the origin of life on our own planet?

To an interplanetary traveller, Earth, Moon and Mars would present themselves as entirely different worlds **(Figure 1)**. For earth and planetary scientists, however, these three members of the solar system have important similarities. Their sizes, their chemical compositions and their distances to the Sun are relatively similar, so that comparative studies yield results on their origin and their evolution. Moreover, Earth, Moon and Mars are the only objects in the universe where we have detailed geological observations and at the same time possess rock samples for laboratory analysis.

Nineteenth century scientists realised that evolution is central to biology and geology. On the basis of characteristic fossils, the sequence of geologic epochs from the Cambrian to the Quarternary was identified. Then, at the turn to the 20th century, radioactivity was discovered, opening up a new era for many branches of science, including the earth sciences. The decay of uranium, thorium and potassium isotopes was found to be the main source of energy in the interior of the Earth, maintaining there a high temperature and driving geological processes. The application of radioactive decay for measuring ages of rocks was an equally important breakthrough. By using various "radioactive clocks", the age of the Earth was derived and an absolute time scale was established for the larger part of its history **(Figure 2)**. Thus, it became possible to gain a quantitative understanding of the causes of geological evolution and identify the underlying physical processes.

In the second half of the 20th century, Apollo and other programmes enabled scientists to use the experience gained from the earth sciences and apply their concepts and methods to the investigation of the Moon. By combining observations at the lunar surface and from orbit with the chemical data obtained from the analyses of lunar rocks brought back to Earth, scientists reconstructed a large part of lunar history.

Thirty years after astronauts set foot on the Moon and explored its geology, scientists and space agencies are turning their attention increasingly to Mars, the Red Planet **(Figure 1)**. Let me emphasise that the exploration of Mars for quite a while will be by robots. Human exploration will be a great challenge, because Mars is at least 200 times more distant than the Moon.

Figure 1
Earth, Moon and Mars. The "Blue Planet" Earth, the "Red Planet" Mars and the grey Moon. The relative sizes are to scale. We have evidence that on Mars there were large quantities of water at some time in the past. Today we see only small ice caps of CO_2 and H_2O, which are seasonally distilled from pole to pole, giving rise to strong winds.

*) Keynote address at the Fourth Anniversary of the International Space Science Institute, 21th October 1999

We have on Earth a dozen Martian meteorites. There is little doubt that their origin is the Red Planet, but we do not know the exact location they came from. The Mars meteorites were expelled from the planet by a large impact, as a result of which they were badly shocked. This makes radioactive dating difficult. Still, ages of Mars meteorites have given us a general idea of the times of formation of geological features on Mars. However, samples returned by spacecraft from a few well-defined geological units will probably be needed to obtain a definite time scale for Martian geology.

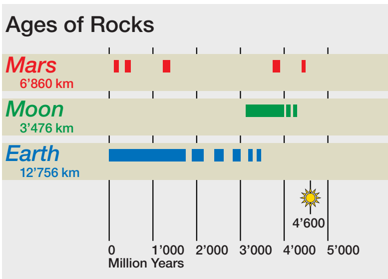


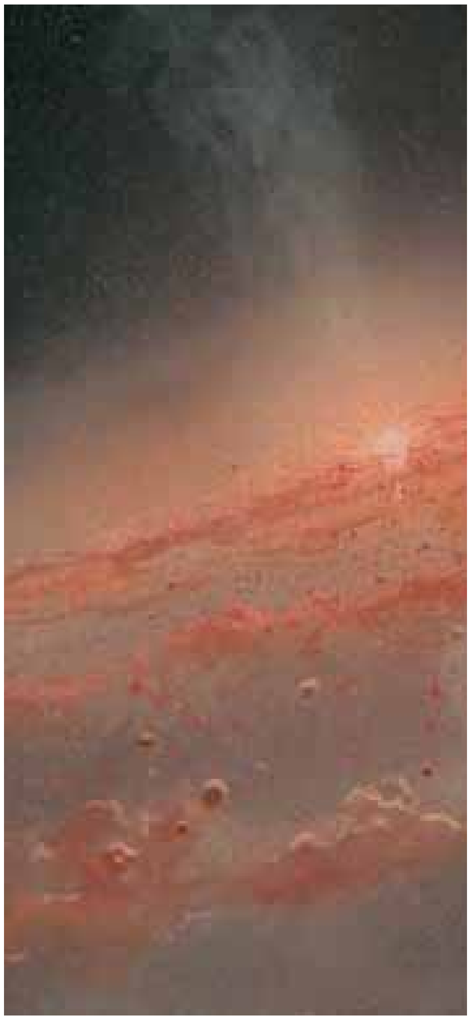
Figure 2
The ages of Earth, Moon and Mars.
This diagram shows the distribution of ages of rocks from Earth, Moon and Mars. Earth has been geologically very active throughout history. As a result, the geological record of the early times has largely been overwritten. The Moon is much smaller than the Earth, and consequently, the geological activity of the Moon ended about 3'000 million years ago. We have, however, a good geological record with absolute times by radioactive dating for the epoch 4'100–3'200 million years ago for our companion. For 3'000 million years, the Moon has been geologically dead. For Mars, we have only a rudimentary time scale of geological events, since we only have 13 Mars meteorites for radioactive dating.

The Early Days of the Solar System

Sun and planets

Stars are created when an interstellar molecular cloud collapses. We can observe this process in the Orion Nebula, where star formation is presently going on. The Sun was born 4'600 million years ago in a collapsing interstellar cloud that was dispersed long ago. Once initiated, it took only a few million years to form the Sun and to allow the nuclear fusion process in its interior to start working and hence for the Sun to shine.

As predicted by Pierre Laplace early in the 19th century, a portion of the gas and dust of the protosolar cloud remained in a rotating disc from which the planets were formed by accretion **(Figure 3)**. This process followed the principle of "big fish eats small fish", well known in the behavioural and social sciences. It began with dust particles attaching to each other. After they reached a certain size, gravitation took over, and larger bodies attracted and swallowed smaller ones. In the end, a number of planets were left, most of them well separated from each other, so that they could safely orbit the Sun for more than four billion years.



The Earth

In the case of the Earth, the accretion process did not go so smoothly. A second object, about the size of Mars, formed on a nearby orbit, so that the two were bound to collide. All the evidence indicates that this collision did occur, and it created the pair of Earth and Moon as we know it, circling around the common center of mass, thus avoiding any further collision.

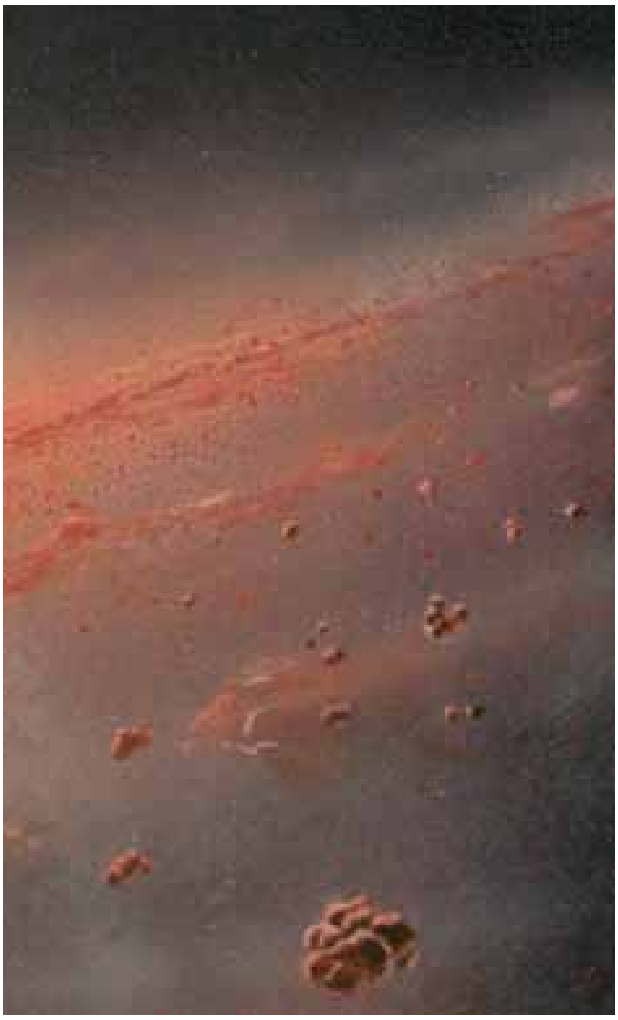


Figure 3
Our Sun is born. The Sun was born by a collapse of an interstellar cloud 4'600 million years ago. Some of the material could not be collected into the central star, the Sun, due to the law of physics, which requires conservation of angular momentum. This material eventually accreted into planets (painting by William K. Hartmann).

The Origin of the Earth–Moon System

For centuries, scientists have wondered about the large size of the Moon. Many planets in the solar system are circled by moons, but, with the exception of the satellites of Earth and Pluto, all these moons are tiny in comparison to the planet they accompany **(Figure 4)**. The giant planets, Jupiter and Saturn, have satellite systems consisting of more than ten moons each. They were probably formed from protoplanetary discs that surrounded Jupiter and Saturn at the time of their formation. The exceptional size of our Moon, however, calls for an entirely different origin.

The Giant Collision

Among the various hypotheses about the origin of our Moon, there is only one, the "Giant Collision Theory" that explains all relevant observations. This theory was originally proposed by William K. Hartmann and colleagues in the United States. The sequence **(Figures 5 to 7)** depicting three stages of the creation of the Earth–Moon system were painted by him. They are based on quantitative modelling by A.G.W. Cameron, W. Benz and others.

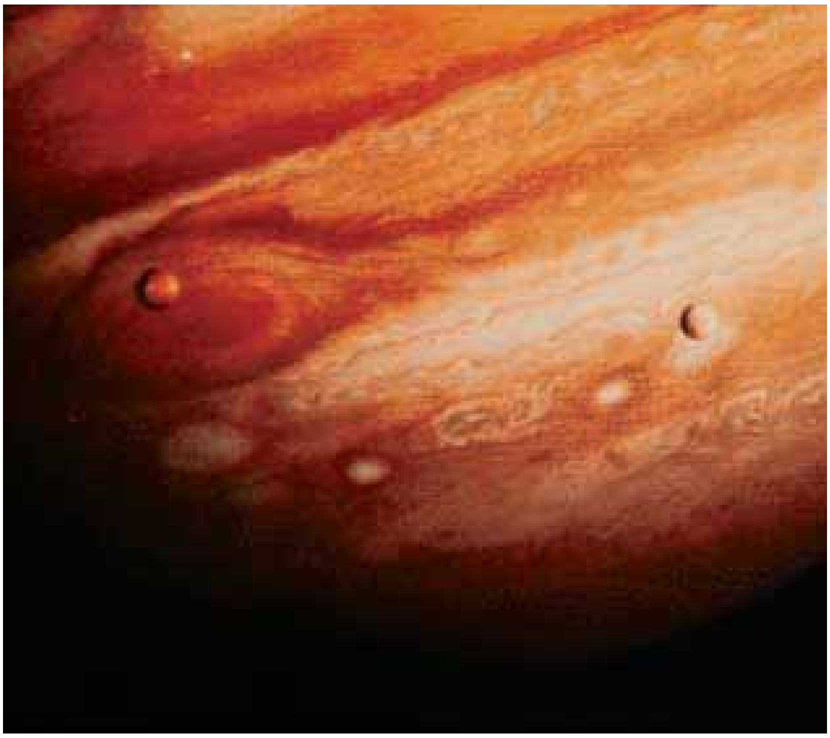


Figure 4
Jupiter with its moons Io and Europa. The giant planet Jupiter with two of its moons, Io (reddish) and Europa (white). While Io and Europa are nearly the same size as our Moon, they are tiny in comparison to the planet they orbit.

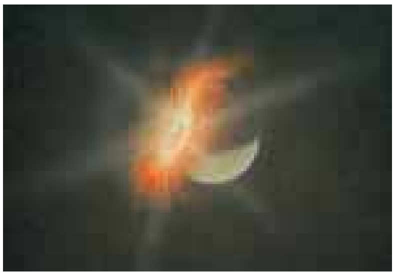


Figure 5
A Giant Collision created the Moon: half an hour after impact. All evidence indicates that our Moon was created by a collision between Earth and another large object, ten percent of the Earth's mass, formed in a nearby orbit (painting by William K. Hartmann).

Figure 6
Five hours after the Giant Collision: Because of the enormous heat created by the impact, all volatiles escaped, so that only non-volatile material remained in Earth orbit (painting by William K. Hartmann).

Figure 7
1'000 Years after the Collision: The formation of the Moon is essentially complete. In its early history, the Moon was molten ("Magma Ocean"). The "tidal forces" at that time were more than a thousand times stronger than they are today and contributed to heating the Moon. Data from the Apollo missions showed us that the Moon is covered by a crust of light material. Thus, the Moon is chemically highly differentiated, and this could only have happened if the Moon was molten early on (painting by William K. Hartmann).

The Giant Collision Theory is consistent with all the relevant pieces of evidence we have. In particular, it explains in a natural way some unexpected results of the analyses of lunar rocks and some observations made by the Apollo astronauts and by photogeologists. These are:

– The lunar rocks are completely free of crystalline water and hydrogen-containing minerals, such as mica or clay minerals **(Figure 8)**. Not only hydrogen, but also the other volatile elements such as carbon and nitrogen are virtually absent. Astronaut observations and all the pictures taken from orbit and on the lunar surface did not reveal any traces of past water flows or products of sedimentation. The complete lack of volatile material and the relatively low abundances of sodium, potassium and other alkalis are readily explained by the extremely high temperatures of the material ejected by the giant collision.

– Relative to the Earth and meteorites, the Moon has much less iron. This would be expected, if the giant impact occurred after the Earth had formed its iron core and if the impact ejected only material from the Earth's crust and mantle.

– The relative abundances of the three isotopes of oxygen show small characteristic differences in the material of different objects in the solar system. Thus meteorites coming from different asteroids, meteorites coming from Mars and rocks on Earth all have their characteristic, but different isotopic signature. The exception is that oxygen in terrestrial and in lunar rocks has exactly the same isotopic composition. The Giant Collision Theory explains this finding, because most of the lunar material was originally part of the Earth. Moreover, the two collision partners accreted from the same region in the protosolar disc.

Figure 8
Landscape on the Moon: This beautiful landscape in Mare Imbrium was visited by the Apollo 15 crew. Like everywhere else on the Moon, the rocks in this mare plain and in the distant highlands are extremely dry. There is no trace of past water flows nor have we found crystalline water in rocks. The rocks and the soil are also devoid of carbon and nitrogen. Thus, there was never any chance for life on the Moon. Small amounts of water ice or other condensates exist in polar areas. They are probably mostly secondary, stemming perhaps from impacting comets (NASA Photo AS-15 9212430).

The accretion process likely left behind some sizeable objects that were bound to collide at the end of the epoch of planet formation. There are indeed indications that large collisions happened elsewhere in the solar system, though the collision producing our Moon may have been the most violent. It has been postulated that the high density of Mercury and the high inclination of the rotation axis of Uranus were caused by large collisions at the end of the accretion process.

According to the Austrian-British philosopher Karl Popper science progresses by *falsification* of theories. The story of the origin of the Moon is a good example of this postulate. As data, observations and theory increased in quality and quantity, one hypothesis after the other about the Moon's origin encountered severe difficulties, the giant collision theory being the only one that remained consistent with observation. Of course, we do not know what kind of new evidence may turn up, but most of us think that with the giant collision theory we have the right scenario for the origin of the Moon.

Evolution of the Moon

The time scale of evolution of the Moon was reconstructed from the radioactive ages of lunar rocks determined by several laboratories in the United States and in Europe.

Accretion of the Moon from the material circling the Earth probably took only a few years. The original heat content, the gravitational energy liberated in the accretion process and heating from the extremely strong tidal forces caused by the proximity of the Earth combined to melt the Moon in its early history. Chemical differentiation set in and produced a crust of light material that has been preserved up to today, especially in the oldest region on the Moon, the highlands on its backside.

The early Moon

Initially, the tidal forces between Earth and Moon were more than a thousand times stronger than they are today, causing significant heating of the Moon and enormous tides on Earth. These tidal forces led to a transfer of angular momentum from the rotation of Earth and Moon to the orbital motion of the Moon. When Moon and Earth were close, the transfer was fast; the Moon receded quickly from Earth, the days lengthened, and soon the Moon's rotation slowed down, so that we always see the same side of it. With the receding of the Moon to its present distance, the tidal forces lost most of their strength. Dangerous tides occur only in a few coastal regions, but there may be another important effect: it has been calculated that the orbiting Moon has stabilized the rotation axis of the Earth, and this in turn has a stabilising effect on the Earth's climate.

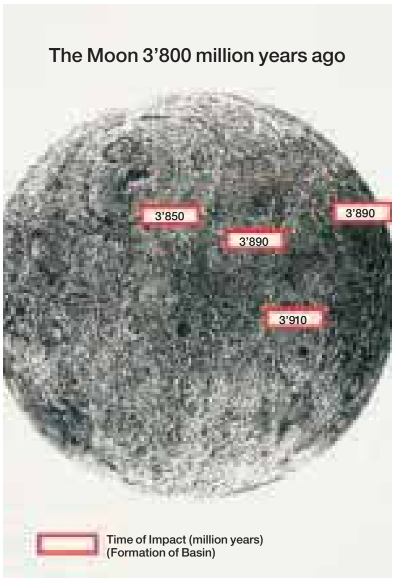
Using various 'radioactive clocks', laboratories in the United States and Europe measured the absolute ages of numerous lunar rocks brought back by the Apollo astronauts and reconstructed the time scale of lunar evolution. The **figures 9, 10 and 11** show the Moon at the end of important epochs in its evolution.

The accretion process in the solar system did not end abruptly: large chunks of material were still roaming around in the spaces between the planets, even several hundred million years after the Sun was born. On the Moon, the last major impacts occurred in the time-span 4'050 – 3'850 million years ago **(Figure 9)**. These late impacts excavated the large mare basins that were later filled with basaltic lava rising up from below. More than ten impacts, forming craters with diameters of more than 500 km, occurred during this time-span. On Earth, we have no record of this catastrophic epoch. However, taking into account the larger size and gravity field of the Earth, we estimate that more than a hundred craters of these dimensions were excavated on our plan-

Figure 9
The Moon 3’800 million years ago. 500 to 700 million years after its formation, the Moon was still bombarded by large chunks of material, forming the huge basins that are still distinguishable. Within about 200 million years, more than ten craters with diameters of more than 500 km were formed. On Earth the record of this bombardment has been lost, but we can estimate that more than 100 such craters were formed on Earth during the same epoch. This must have been a fantastically catastrophic period in the Earth’s history, at a time when oceans and atmosphere were in an early stage. We would not know about this catastrophe if we had not investigated lunar rocks in Earth-bound laboratories.

Figure 10
The Moon 3’000 million years ago. At about 3’800 million years ago, the basins began to fill with lava. Shown are the ages of the local mare basalts in million years, determined by radioactive clocks. From West to East: Oceanus Procellarum (3’200 million years), Mare Imbrium (3’300), Mare Serenitatis (3’800), Mare Tranquillitatis (3’600–3’800), Mare Crisium (3’600). These ages show that this process was largely completed 3’000 million years ago. The large plains, which Galileo called Maria, have existed on the Moon since that time. We know now that their dark colour is due to a titanium-iron mineral that is relatively abundant in the mare basalts.

Figure 11
The Moon today. About 3’000 million years ago, the Moon cooled down sufficiently, so that geological activity afterwards did not change the face of the Moon very much. The main changes are from occasional impacts. On the front side of the Moon, the largest ‘fresh’ crater is Copernicus, which was created by an impact about 900 million years ago.



et during that period. For the emerging atmosphere and oceans, this must have been a catastrophic epoch.

Even before the last mare basins were excavated on the Moon, lava had begun to rise and fill these basins. These were basaltic lavas, richer than the crust in iron, magnesium and titanium. The dark colour of the Maria on the Moon is due to minerals containing these elements, in particular the titanium-iron oxide called ilmenite.

The ages of the mare basalts brought back by the Apollo astronauts and Russian Lunar missions were found to range from 3’800 to 3’200 million years (**Figure 10**). A few small rock pieces in the Oceanus Procellarum were as young as 3’000 million years. Further east there may be some areas covered with even younger basalts, but there are no samples and no absolute ages from this region. In any case, the formation of mare basalts that marks the last major geologic epoch on the Moon ended not much later than 3’000 million years ago. Since then, the Moon has been geologically inactive, its energy reserves exhausted.

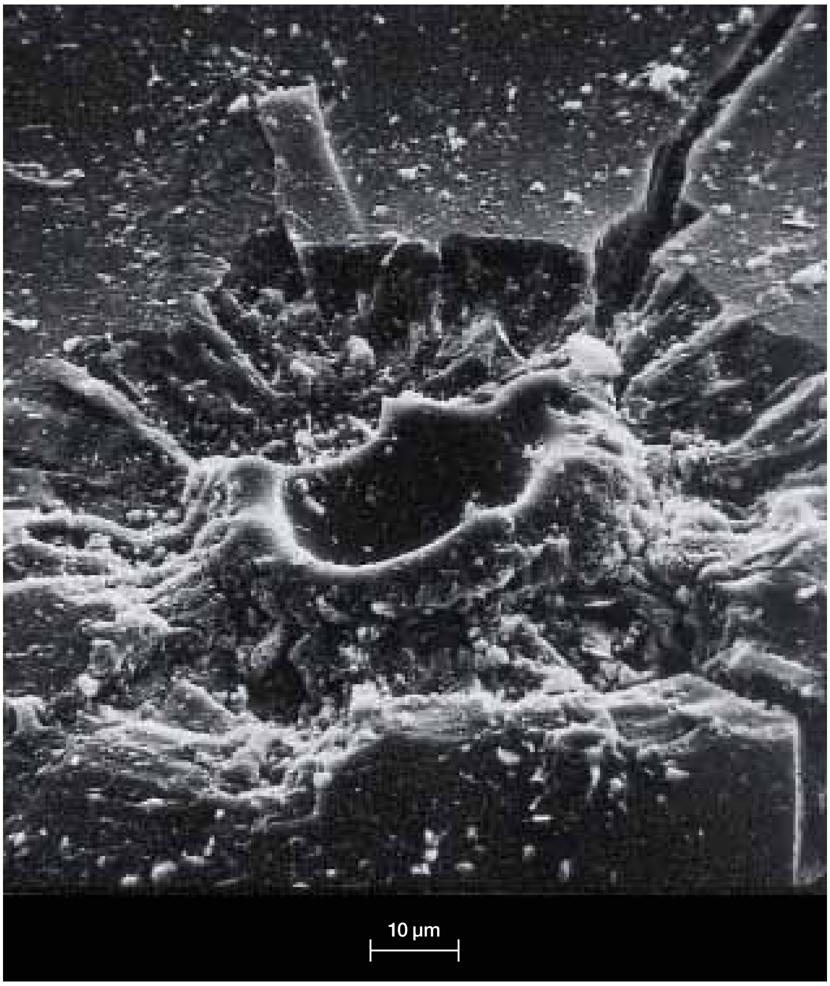


Figure 12
A microcrater in a piece of lunar glass. In the present epoch, small impacts are the main cause of erosion, "gardening" and lateral transport (photo by Norbert Grögler, University of Bern).

Figure 13
The footprint of Neil Armstrong. There are neither winds nor rain on the Moon and erosion is very slow. Thus, the chances are good that this footprint of Neil Armstrong, the first man on the Moon, will still be visible a million years from now. There is, of course, a minute statistical chance that this footprint will be impacted by a larger object, but then other footprints of the Apollo crews would survive (NASA photo).

The serene Moon

In the absence of geological processes and without water or winds, the face of the Moon has not changed very much during the last 3'000 million years, as a comparison of figure 11 with figure 10 shows. Erosion at the lunar surface is now mainly caused by small impacts **(Figure 12)**. This process is extremely slow, so that even small surface features last for a long time **(Figure 13)**. Major impacts were very rare during the last billion years. On the frontside of the Moon, the impact creating the Copernicus crater is the most prominent among them. It occurred about 900 million years ago **(Figure 11)**. Sometimes Copernicus is called a young crater. This expression illustrates the fundamental difference in the evolution of Earth and Moon. Copernicus was formed before the Cambrian, the earliest of the classical geologic epochs. Since on Earth plate tectonics and volcanism continue to the present time, the effect of impacts on the geologic evolution is minor. However, the catastrophes caused by large impacts have had a major influence on biological evolution **(Figure 14)**.

Figure 14
The Cretaceous/Tertiary boundary.
On Earth, impacts did not have a large effect on the geological or geographical evolution of the surface, because volcanism and plate tectonics were dominant. Impacts of meteorites, however, had a major effect on the biological evolution. The best-documented case is the event that occurred 65 million years ago, marking the boundary between the Cretaceous and the Tertiary age. At that time, a large impact – probably located at the edge of the Gulf of Mexico – caused the extinction of many species, giving way to the evolution of new species. Thus, in sedimentary deposits, the CT boundary is marked not only by impact-specific materials (iridium, soot), but also by an abruptly changing micro-fauna over a very short interval (Figure by O. Eugster).

Mars

Mars is intermediate in size between Earth and Moon, and it seems to be intermediate as well in the degree of geological evolution **(Figure 2)**. This is to be expected: Small bodies cool faster than large bodies, since energy production and energy content are basically proportional to the volume, whereas energy losses are proportional to the surface. It is for the same reason that warm-blooded animals need a minimum size to be able to maintain a body temperature that is above the ambient temperature.

Asteroids have diameters below 1'000 km. Thus, they cooled down and became geologically inactive only one or two hundred million years after they were formed. This has been confirmed by radioactive dating of meteorites that are rock pieces from asteroids. The Moon, with a diameter of 3'476 km, became geologically inactive about 3'000 million years ago. Because of its size, the Earth has remained geologically active throughout its history. As a consequence, the geological record of the first 1'000 million years of its lifetime is essentially lost.

The ages of the few Mars meteorites range from 4'500 to 150 million years. Apparently, some rock-forming processes have been going on quite recently. On the other hand, the old ages indicate that the record of the early Martian history was not overwritten as much as on Earth. This is in line with the size of Mars, which is in between the sizes of Earth and Moon **(Figure 1)**.

The Exploration of Mars

The radioactive dating of Martian meteorites gives a general idea of the absolute ages of geologic features on the Red Planet. A relative time scale of the sequence of geologic epochs identifiable on Mars is being set up by scientific groups at the Deutsches Zentrum für Luft- und Raumfahrt, Berlin, and the Planetary Science Institute in Tucson, Arizona. The method is by counting craters, which has been successfully applied on the Moon.

Presently, earth and planetary scientists are engaged in reconstructing the geological evolution of Mars and in estimating absolute time scales for the most prominent events in the Martian history. These issues were discussed in a recent workshop on "the Evolution of Mars" at the International Space Science Institute. The results from Martian meteorites were compared with photogeological evidence obtained from orbit and with the data obtained by the Sojourner, the vehicle landed by NASA on the Martian surface on July 4, 1997. The whole world watched as the Jet Propulsion Laboratory in Pasadena steered the Sojourner to some nearby rocks, so that the chemical sensor of the Max-Planck-Institut für Chemie, Mainz could analyse

The evolution of Mars

The two most outstanding geological features on the Martian surface are the huge volcanoes and the evidence of the existence of liquid water in a past epoch. The largest of the Martian volcanoes, Olympus Mons, is shown in **Figure 15**. It came as a great surprise that the volcanoes on Mars are higher and more massive than the highest mountains on Earth **(Figure 16)**. The principal reason is depicted in **Figure 17**. On Earth, plate tectonics causes continental drifts as predicted by Alfred Wegener early in the 20th century. Thus, plate tectonics determines the global geography on Earth, and it produces the high continental mountain ranges. On Mars, the internal energy is not sufficient to drive plate tectonics. Thus, volcanoes are the dominant mountains, and they can grow without being cut off from their hot spots below.

Figure 15
The Olympus Mons. The most outstanding geographical feature on Mars is the large volcanoes (photo by Viking Orbiter, NASA).

them. At first, the outcome seemed surprising: the chemical abundances were completely different from the abundances in the Martian meteorites. However, the Sojourner investigated rocks in one of the dry riverbeds, the Ares Valley, whereas most of the Martian meteorites probably come from volcanic highlands. Thanks to the Sojourner we now know of the very large variety in the composition of Martian rocks. This indicates that chemical differentiation was much more intense on Mars than on the Moon.

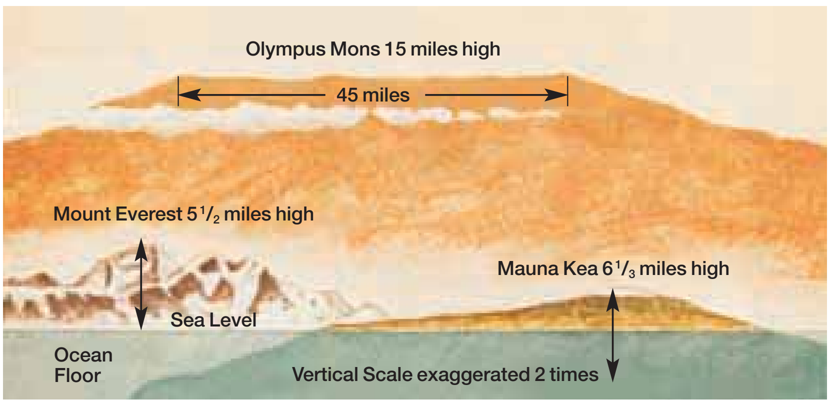


Figure 16
Olympus Mons versus Mauna Kea: Olympus Mons and other volcanoes on Mars are much higher than the largest volcanoes on Earth, even if we measure these from the bottom of the ocean. This may seem surprising, since one might expect less intense volcanic activity on Mars than on Earth, because of the difference in size.

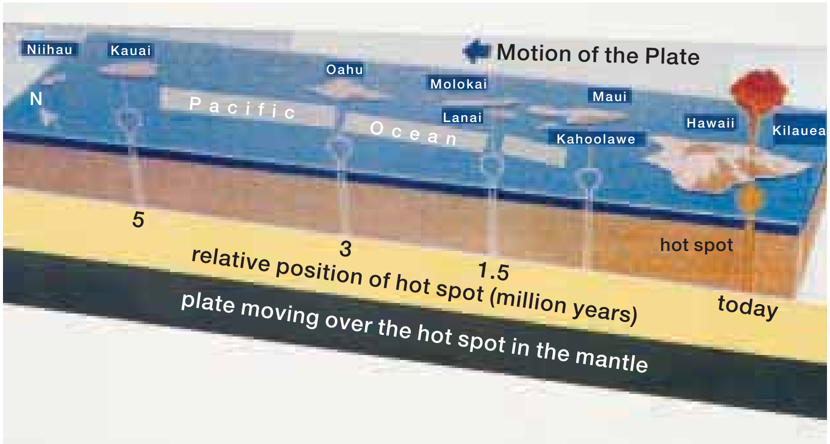


Figure 17
Origin of the Hawaiian volcanic islands: Plate tectonics ("continental drift") on Earth is the answer to the apparent contradiction shown in Figure 16. The oceanic plate moves with a speed of a few centimetres per year over a hot spot in the mantle, creating a chain of volcanic islands. On Mars such plate motions are absent and, therefore, the lava coming from a hot region in the mantle is piled up in one place over a long period of time.

Figure 18
River system on Mars: Sometime in the history of Mars there was a fluvial period. The extensive river systems are evidence of this.

Figure 19
The Hadley Rille: On the Moon there is no indication of liquid water, present or past. But how do we know this? Hadley Rille and the other lunar rills have no tributaries: they were probably made by lava flows. On Mars, the tributary systems of valleys are the most obvious proof that they are dry riverbeds.

Contrary to the Moon, where there are no traces of past water flows, we see large dry riverbeds on Mars. The difference between the lunar rills and the Martian riverbeds is demonstrated in pictures taken from orbit: Martian riverbeds have extensive tributary systems **(Figure 18).** The long lunar rills on the Moon have no tributaries, the rills are thought to stem from lava flows **(Figure 19).**

The larger riverbeds on Mars compare with the Nile valley. At some – as yet – unknown time in the past, large amounts of water were transported, and they must have filled significant portions of the low lands. **Figure 20** is a topographic map, with the area that might have been temporarily filled by the rivers shown in blue. The exact extension of this temporary ocean is still debated, but all indications are that it covered a significant portion of the Martian surface. Jim Head of Brown University, Phillip Masson of the Université de Paris and other photogeologists are working to identify the ancient shorelines. If successful, we can quantitatively determine the amount of water that was present.

Where did these water masses go? A considerable fraction has probably been lost, but some must have leaked through the ocean bottom and is preserved as a permafrost layer.

Life on Mars?

Liquid water is a prerequisite for the emergence of life, at least of the form of life we know. Since water is liquid only in the temperature range of about 0 °C to 100 °C, we could expect life on a planet only in a narrow range of distances from the central star. Thus in the solar system, we have presently only one body, the Earth, with liquid water on its surface. Because of radioactive heating, planetary temperatures increase with depth, and therefore it is believed that water could be liquid at certain layers below the surfaces of Mars and some moons of distant planets with icy surfaces, such as Europa.

We have strong evidence that during an earlier epoch there were large volumes of liquid water on the surface of Mars. It is not yet known during which epoch in Martian history the rivers and the ocean existed **(Figures 18 and 20)**, nor do we know whether there were several fluvial periods or only one. In any case, since Mars is relatively close, the Red Planet offers the best chance for discovering extraterrestrial life or for identifying relics of extinct life.

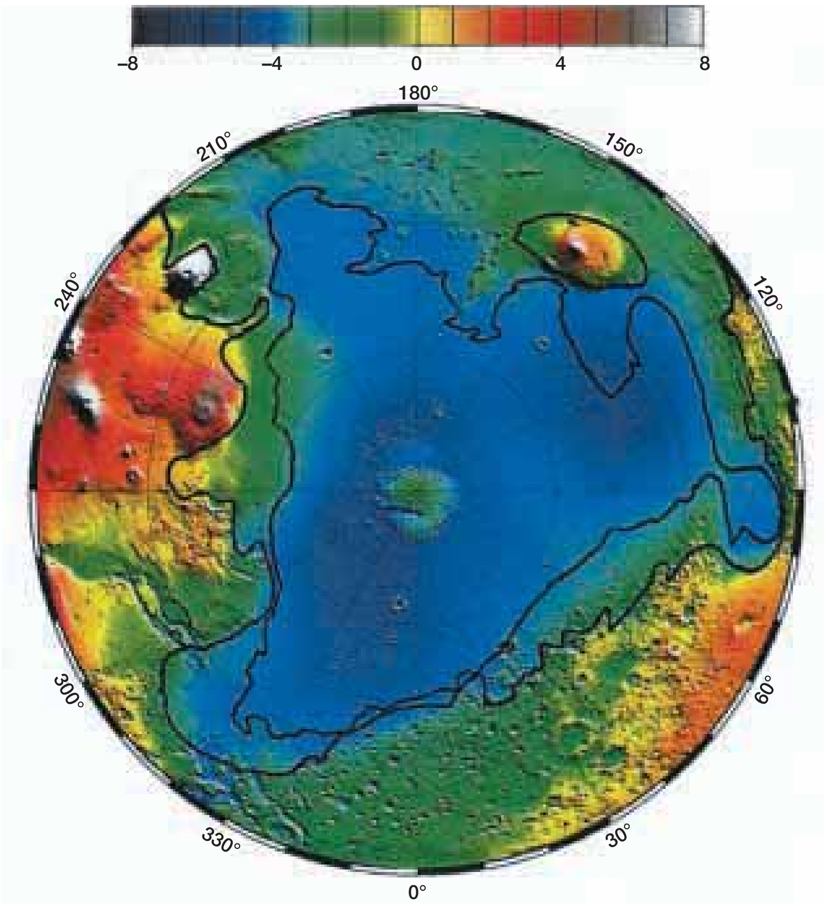


Figure 20
Ancient ocean on Mars. The dry riverbeds on Mars with their tributaries go towards the large low-altitude plains in the north. Considering the amounts of water that have cut the large valleys, the water masses must have fed large oceans. The estimated extent of the ocean is shown here in blue. Elevation is given in kilometers (from J.W. Head, Science, 286, pp. 2134–2137, 1999).

Traces of life in Martian meteorites?

Reports on the observation of long straight "channels" on Mars were not confirmed by orbital photography. The Viking spacecraft, landed by NASA in the 1970s, did not turn up any evidence of life, although they were equipped with some suitable instruments. Then, a few years ago, a group of scientists in the United States found spherules in the oldest Martian meteorite, and after elaborate microscopic investigations and chemical analyses, they concluded that these spherules could very well be relics of life. The evidence is intriguing, but the discussion of its significance is continuing. Indeed, we cannot expect to have a consensus concerning the biological nature of these spherules unless corroborative evidence is found on Mars.

Figure 21
Buhwa iron ore deposit.
This sedimentary iron ore deposit in Zimbabwe is about three billion years old. The iron in this ore is highly oxidised. It is thought, that these ores are the oldest indication for the presence of oxygen produced by photosynthesis of algae (photo by Jan D. Kramers, University of Bern).

If life on a planet were confined to a limited area, a puddle of water or a cave, it would be extremely difficult to discover. After all, the surface of Mars is nearly as large as all continents on Earth taken together. Fortunately, however, once created or imported, life seems to be extremely adaptable and prolific. This opens up another method of searching for life, which we may call the geochemical approach. On Earth, this approach has proven to be valuable for searching for early traces of life. Large ferric iron deposits that formed only in the early Precambrium **(Figure 21)** are thought to be the oldest evidence of life. Most limestone deposits, such as those lifted to the top of Mount Everest or the Bernese Alps, are produced by life. If life existed on Mars during a fluvial epoch, it would have spread over a large area. Thus, searching for geochemical traces of life on Mars should be considered. G. Turner and J.D. Gilmour of the University of Manchester proposed to look for a life-specific iodine distribution on Mars.

Life on Earth!

Let me finish this evening's lecture with a tale that may illuminate the geochemical search for life. The society on a planet orbiting a nearby star decided to explore their stellar neighbourhood. In a planet-wide effort, they built a fleet of extraordinary spaceships that would enable them to reach a dozen or so stars in their neighbourhood during a crew's lifetime. However, by giving so much priority to rocketry, they neglected to develop other equipment, such as high-performance telescopes and sensitive detectors. Ulysses, when cruising in the Mediterranean, found himself in a comparable situation. He had a marvellous ship, but no telescope, poor maps, and no idea of the dangers awaiting him. The spaceship with the extraterrestrials on board going towards the solar system was successful. After short, unrewarding visits to the Jovian moons and to Mars, they turned towards the Earth. In a direct approach and without much "reconnaissance", they chose to land on the western flank of the Jungfrau, because the Bernese Alps were among the few places that protruded through the cloud layer covering all Europe at the time of their landing. All they could see was snow and rocks **(Figure 22)**. One of them tested the rock with a knife: "It's limestone, there is life on this planet or at least there was in the past!" Cautiously, he opened his suit and found he could breathe. So there was oxygen. On their homeplanet the oxygen was produced and continuously replenished by plants. This convinced them that there was a rich flora below the clouds. Slowly, they climbed down. Suddenly, the cloud cover broke up, and there it was: grass, cows and even some people at a distance. When the visitors cautiously approached, one of the earthlings began to talk, and it sounded strange, but friendly. "So that is the language of the people on this planet, for us impossible to understand", they concluded. It was only much later that they found out that almost nobody else on Earth understands these nice people living in the foothills of the Bernese Alps either.

Acknowledgements

I am indebted to Hansjorg Schlaepfer, the editor of Spatium, for his help with the manuscript and his perseverance in getting this issue published. I thank Silvia Wenger, Ursula Pfander and Gabriela Indermuhle for their help with the manuscript and the figures, and Diane Taylor for critically reading the manuscript. With this issue of Spatium I would like to remember two distinguished colleagues and friends, Paul Gast of Columbia University and NASA Johnson Space Center and Norbert Grogler of the University of Bern, with whom I had close cooperation during the Apollo period.

Figure 22
The Bernese Alps shown above the clouds From right to left: Jungfrau, Mönch and Eiger.

The author

Prof. Johannes Geiss received his doctorate in 1953 from the University of Göttingen and, after several research positions in the USA, became a lecturer ("Habilitation") at the University of Bern in 1957. After a professorship in oceanic science at the University of Miami, he returned to the University of Bern as a professor in 1960, where he was named director of the Physikalisches Institut in 1966. In 1982/83 he was rector of the University of Bern, as well.

During his tenure at the University of Bern, he had extended stays as a visiting scientist in France and the USA. The public connects the name of Professor Geiss with the solar wind experiment, the first experiment the American astronauts set up on the Moon in 1969.

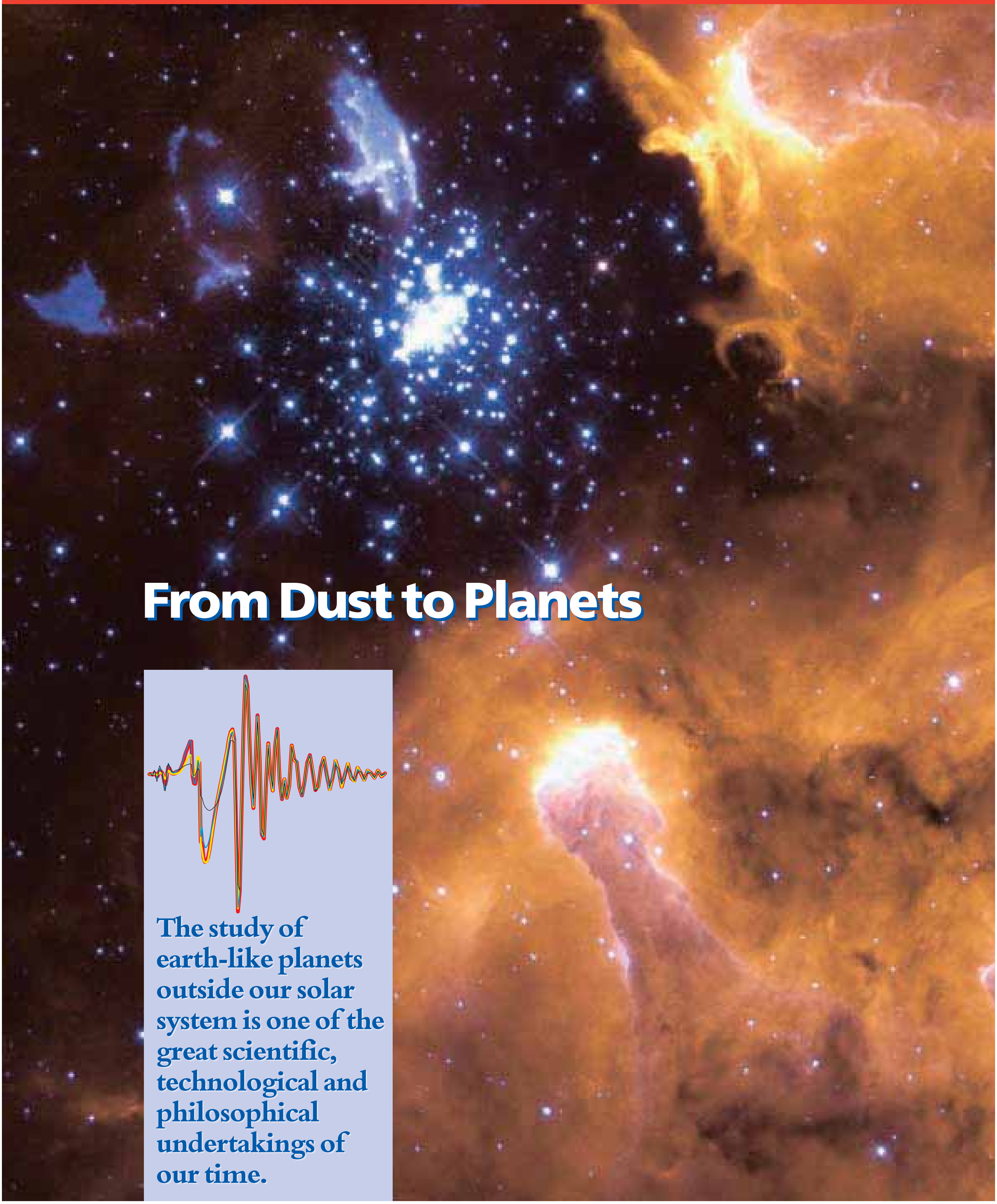
Since then, Professor Geiss has been heavily involved in scientific projects of NASA and the European Space Agency, ESA.

Johannes Geiss is co-founder of the Association Pro ISSI and presently the Executive Director of the International Space Science Institute, ISSI, in Bern. ISSI is supported by a foundation of the same name, endowed initially by Contraves Space in Zurich. The Institute is primarily funded by ESA, the Swiss Confederation, the Canton of Bern and the Swiss National Science Foundation. ISSI is dedicated to interdisciplinary research of the solar system and the universe, inviting scientists from around the world to working group meetings and workshops and offering them the opportunity to exchange ideas and evaluate their newest research results. The basis for ISSI's research activity is the data provided by the scientific missions of the space agencies of Europe, the US, Russia and Japan as well as laboratory experiments and computer simulations.

INTERNATIONAL SPACE SCIENCE INSTITUTE
No 6, October 2000

SPATIUM

Published by the Association Pro ISSI



From Dust to Planets

The study of earth-like planets outside our solar system is one of the great scientific, technological and philosophical undertakings of our time.

Editorial

Chance or necessity? Careful observation of our world provides us a fascinating picture where chance and necessity play the key role. Weather evolution may be one of the most prominent examples in our daily life: While short-term forecasts have become fairly precise, thanks to a wealth of data gathered by spacecraft and ground-based stations and powerful computers processing these data, long-term forecasts remain a dream, because non-linear, chaotic processes become dominant. Interestingly, teachers at every level tend to carefully avoid the topic of chaos, perhaps because they prefer the rigour of order to the freedom of chaos… But order is only half of the truth, as we should have learnt in the last 150 or so years since the publication of the "On the Origin of Species" by Charles Darwin.

The orbits of planets around a central star are another example of a short term deterministic and long term chaotic process. While it is possible to predict for example the orbit of Saturn and its moon Titan over a decade – which is of crucial importance for the common ESA/NASA mission Huygens/Cassini, which is currently on its seven years journey to Saturn and Titan – it is not possible to investigate the formation process of our solar system simply by extrapolating back over billions of years to the era when our sun began to shine.

In recent years, it has become possible to observe – at least indirectly – planets circling around other stars. M. Mayor and D. Queloz of the Observatory of Geneva were the first to announce such a spectacular finding. Observing and comparing solar systems yields a new picture of the star and planet formation process, which itself is the product of interaction between chaos and necessity. It is this topic to which the present issue of SPATIUM is devoted. Prof. Willy Benz of the Physikalisches Institut of the University of Bern gave the members of our association a fascinating lecture on the subject of planet formation. We are pleased to submit herewith the revised version of his lecture to our readers.

Bern, October 2000
Hansjörg Schlaepfer

Impressum

SPATIUM
Published by the
Association Pro ISSI
twice a year

INTERNATIONAL SPACE SCIENCE INSTITUTE

Association Pro ISSI
Hallerstrasse 6, CH-3012 Bern
Phone ++41 31 631 48 96
Fax ++41 31 631 48 97

President
Prof. Hermann Debrunner, University of Bern
Publisher
Dr. Hansjörg Schlaepfer, Contraves Space AG, Zurich
Layout
Marcel Künzi, marketing · kommunikation, CH-8483 Kollbrunn
Printing
Druckerei Peter + Co dpc
CH-8037 Zurich

Frontcover
Birth and death of stars in the NGC 3603 nebula (W. Brander/JPL/IPAC, E.K. Grebel/U. Wash., You-Hua Chu/UIUC, HST/NASA)

From Dust to Planets *)

Willy Benz, Institute for Physics, University of Bern

Introduction

While the search for planets and life outside the solar system has been going on for decades, the discovery in 1995 of the first extra-solar giant planet by Geneva Observatory astronomers Michel Mayor and Didier Queloz followed within months by the discovery of 2 new giant planets by Geoffrey Marcy and Paul Butler of Lick Observatory (USA) has sparked a real revolution. Five years later, over 50 such giant planets have been found **(Figure 1)**. implying that at least 3–5% of all sun-like stars have giant planets.

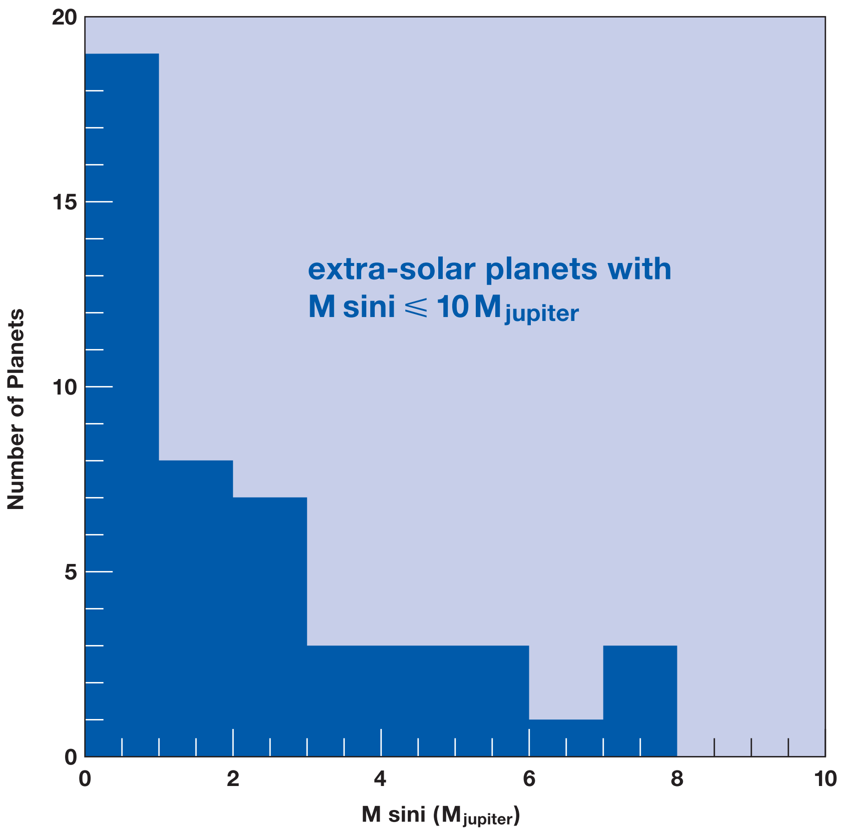


Figure 1
Mass of the extra-solar giant planets discovered thus far. Since more massive planets are easier to detect, the paucity of massive objects indicates that jupiter-sized (or smaller) objects are by far the most abundant ones.

Since this represents only the fraction of stars having planets that could be detected with current instruments, we must conclude that planet formation is not an extraordinary event but rather quite common occurrence.

Our solar system forms the basis for most of our information about how planetary systems must develop. However, the degree to which it is actually representative of all planetary systems is unclear. It now appears to be very different from all those discovered thus far. Indeed, contrary to the giant planets in our own system (Jupiter, Saturn, Uranus, Neptune), the newly discovered planets have much smaller orbits many with sizeable eccentricities. Although there is clearly a strong observational bias against detecting distant and/or small planets, it is significant that none of these newly discovered objects should have existed according to conventional formation theory!

Does this mean that our solar system is unusual or maybe even unique or have we simply not yet been able to detect the right kind of systems elsewhere? To answer this central question requires getting a full measure of the possible diversity between existing systems as well as a much better understanding of the physical processes underlying planet formation and evolution.

*) Pro ISSI lecture, Bern, 3rd November 1999

Planet formation: The conventional picture

Planets are likely nothing else than a by-product of star formation stemming from the necessity of conserving angular momentum. Indeed, stars form through the gravitational collapse of interstellar matter over more than 8 orders of magnitude in size. In the presence of rotation and/or magnetic field, the collapse must results in a star/disk structure with the star having most of the mass and the disk most of the angular momentum. Thanks to the incredible resolution of the Hubble Space Telescope (HST), a few of these disks orbiting nearby young stars could even be imaged **(Figure 2)**.

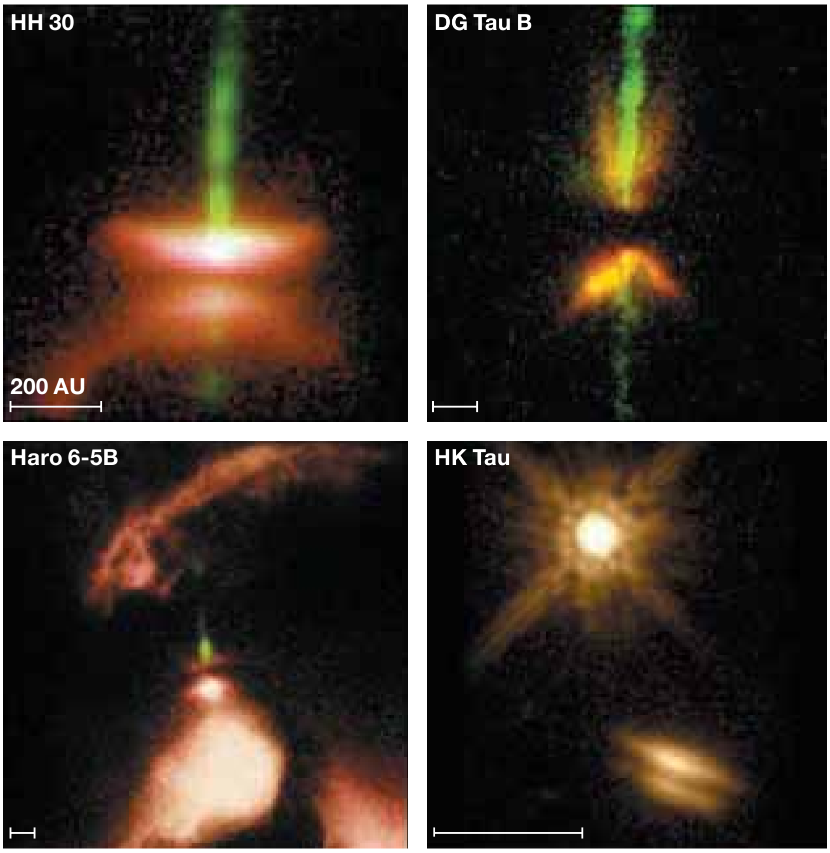


Figure 2
Images of circumstellar disks around a few nearby young stars imaged by HST. Note the the bipolar jets emerging on either side of the disk.

In the case of the solar system, the disk is generally taken to have a mass of a few percent of a solar mass and to be less than 100 AU (1 AU is the distance between the Earth and the sun) in size.

Planets subsequently form in this disk probably mostly through collisions at first between dust grains and as time goes by between larger and larger bodies. Earlier theories in which planets form through the gravitational collapse of patches in the disk have grown out of favor. The flow diagram in **Figure 3** illustrates these concepts.

This picture provides a simple explanation why all planets in our solar system not only orbit the sun nearly in the same plane but also in the same direction. The nearly coeval formation of the planets and other small bodies and the sun is actually supported by comparing the ages of the oldest Moon rocks and the sun.

The basic challenge of planet formation consists therefore of assembling in a disk orbiting a central star micron-sized or smaller dust grains in bodies with over 10^4 km in diameter **(Figure 4)**, a growth by nearly a factor 10^{13} in size or 10^{40} in mass! Since giant planets are mainly gaseous planets, their formation must take place while gas supply lasts. From studies of disks around other young stars, it is believed that typical lifetime of disks are of order a few million years. Hence, as paradoxical as it sounds, giant planets must be formed in less than ten million years while forming terrestrial planets may take much longer.

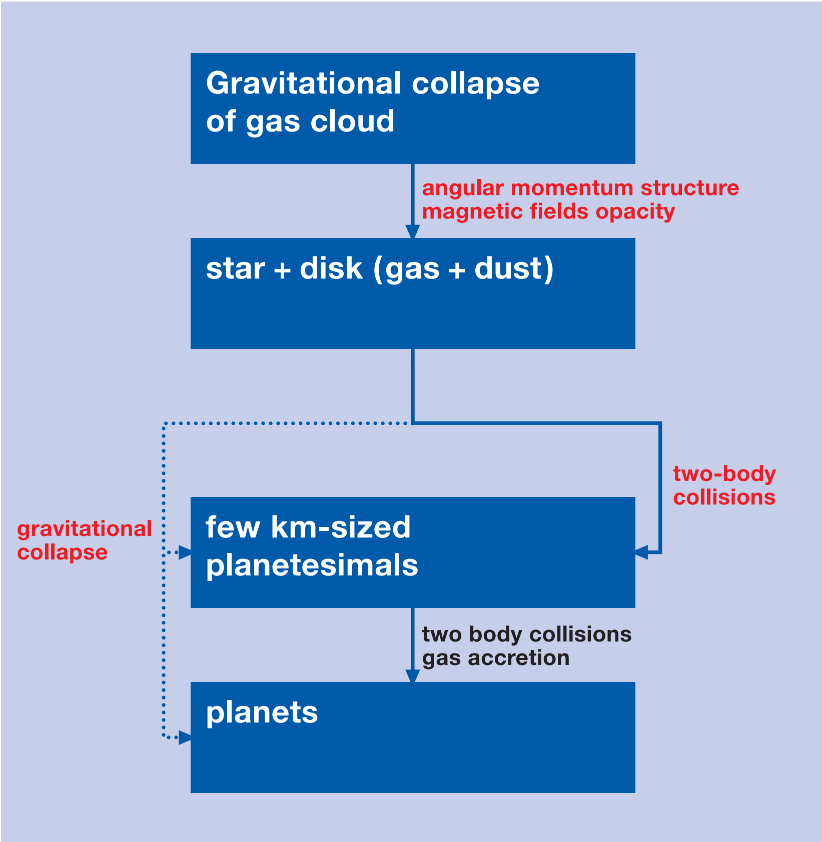
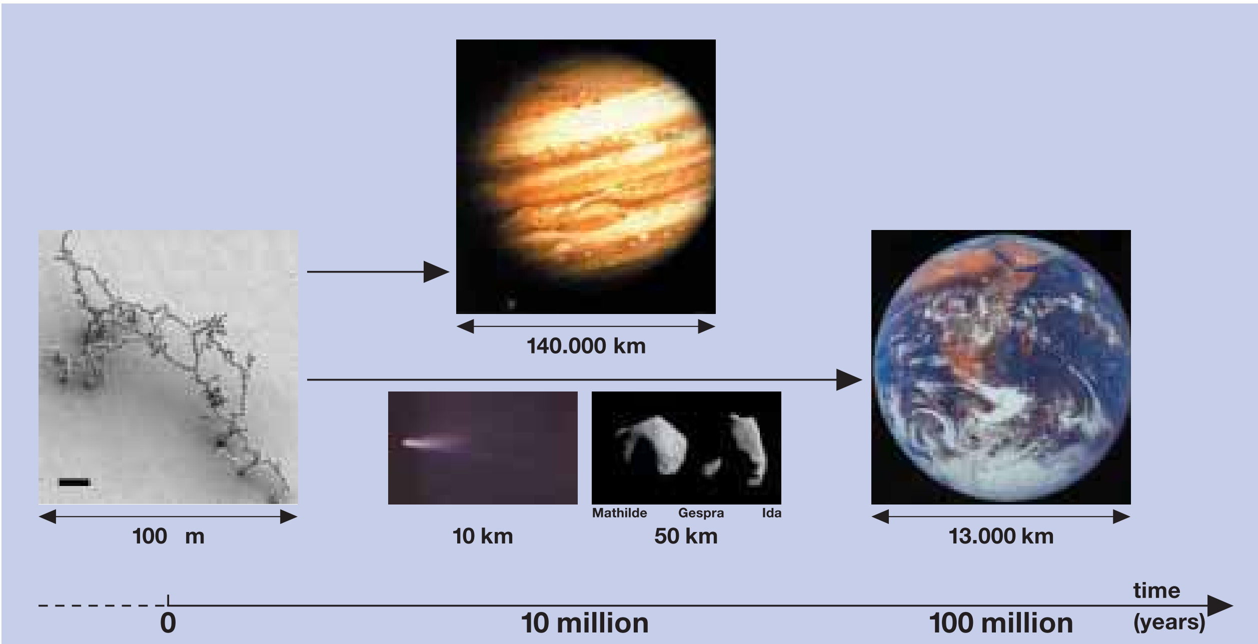


Figure 3
Planetary formation as a by-product of star formation. In the standard model, the planets form entirely through collisions (solid line) while in other models gravitational collapse is invoked at various stages of the formation process. The main physical processes at play are shown in red.

Figure 4
The challenge of planetary formation: Assembling micro-meter dust grains in planets through collisions in an amazingly short timescale.



On their path to becoming a planet, dust grains reach the size of comets and asteroids which, if they can avoid being incorporated in a larger object, are left behind like crumbs on a table after a good meal.

The early phases: The first million year

The growth of planet-sized bodies in this disk is thought to occur essentially through collisions. Earlier models relied on gravitational instabilities in the dust layer to rapidly grow objects several kilometers in size (dotted line in **Figure 3**). However, it has been pointed out that the velocity shear between the dust and the gas will stir up the dust sufficiently to make instabilities impossible. While the extend of this turbulence of the dust layer is still debated, collisional growth from the smallest sizes on has become the favorite scenario.

At first, dust grains collide at relatively gentle velocities which are determined by size and shape dependent gas drag. As bodies grow larger, the importance of gas drag diminishes to vanish completely by the time bodies reach several tens of meters in size. With subsequent increase in mass, the collisional cross section of these planetesimals increases due to gravitational focusing yielding to the so-called runaway growth phase during which the larger bodies sweep-up all the smaller ones within their gravitational reach.

This phase is not without problems. Laboratory experiments have shown that at the very small scale dust aggregates readily (**Figure 5**).

On the very large scales, various impact simulations have shown that self-gravity will ensure growth. However, the situation is much less clear for objects ranging in size from a centimeters to kilometers since in this size range no real "sticking" mechanism has yet been found. Indeed, at this size, the forces operating at the micron size level are no longer effective and gravity is still much too weak.

The escape velocity of a 1 meter-sized rock is of order 1 mm/s while the typical collisional velcity between these objects is of order 100 m/s. Hence, for sticking the bodies involved have to be able to dissipate all but 10^{-10} of the incoming kinetic energy. Whether this can be achieved by purely mechanical structures or requires the presence of a "glue" whith special visco-elastic properties remains to be seen.

Figure 6 summarizes the main three stages of growth, the relevant physical mechanism operating and the main study tools.

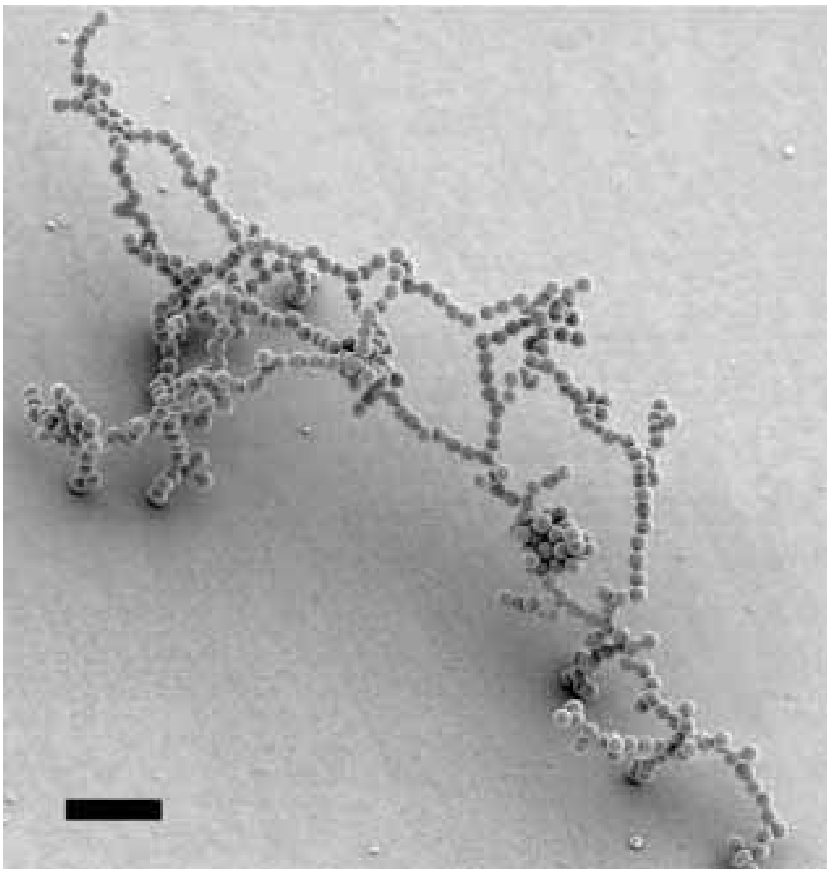


Figure 5
Aggregate obtained in laboratory dust coagulation experiments by J. Blum and collaborators. The scale is given by the 10 µm black line.

The late phases: 100 million years

This early planetary accretion phase was, over a few million years, replaced by an even more violent period when growing bodies encountered one another at increasingly high velocities boosted by mutual gravitational interactions. This phase last for another 100 to 200 million years until all remaining bodies have been swept up by the planets. Collisions occur in a random fashion involving objects of different masses, structures, composition and moving at different speeds. Thus, this phase of planet formation must not be viewed as a monotonic process by which material is incrementally added to a growing planet. Instead, accretion must be viewed as a long chain of stochastic events in which non-disruptive infall exceeds, over time, violent dispersal.

The so-called giant impacts in which proto-planets of comparable size collide represent the ultimate in violence during planetary accretion. While they can lead to the total destruction of the planets involved they can also leave scares arguably the best evidences remaining today of such a violent past. The Earth's Moon, for example, is believed to originate from the debris ejected after such a giant impact and subsequently reaccreted in Earth's orbit **(Figure 7)**.

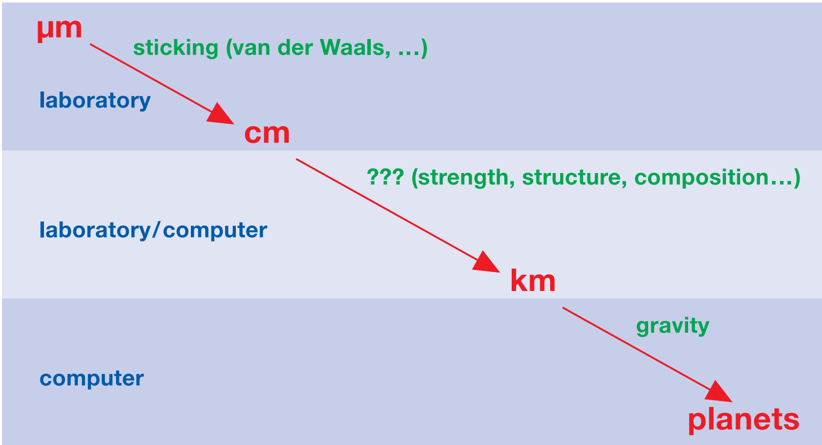


Figure 6
Planetary growth stages. In green the main physical mechanism ensuring growth and in blue the main tool used to study these phases. Note that the growth mechanism in the meter size range is still unknown.

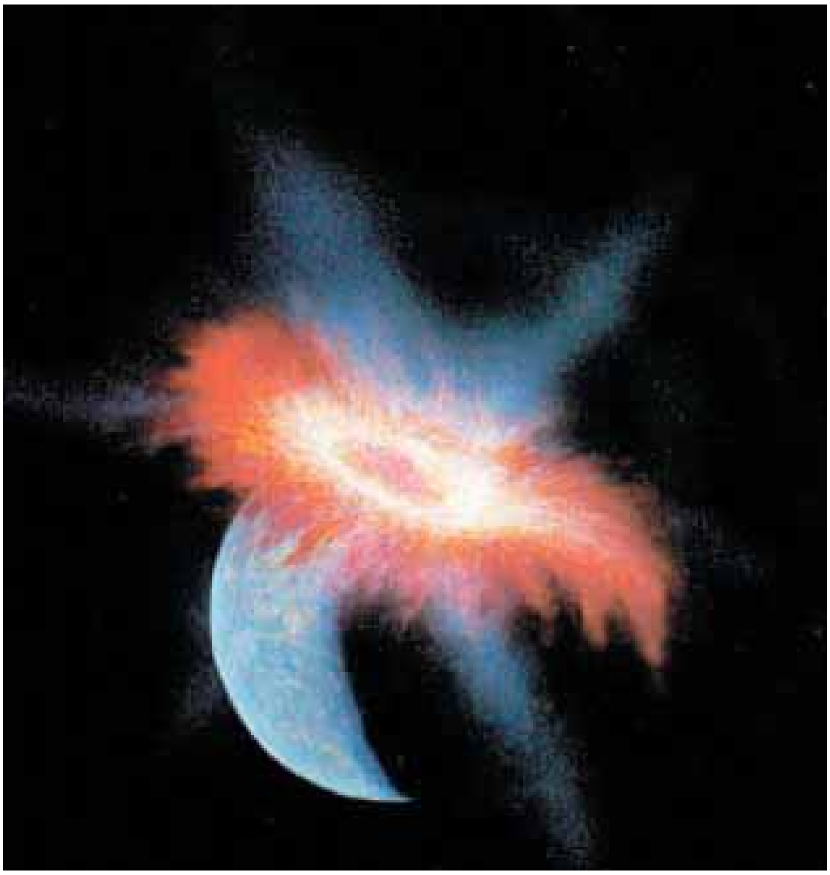


Figure 7
The Moon may have originated from the debris ejected in orbit following a giant impact of the sort depicted in this painting by W. Hartmann.

Simulations of both impact and re-accumulation have not only shown that such a scenario is possible but have made it today's favorite theory of lunar origin. Studies of lead and tungsten isotopic composition of the silicate Earth have even allowed to date the giant impact to about 50 million years after the start of the solar system!

Mercury's anomalous composition can also be explained in terms of a giant impact which ejected most of the mantle of the planet leaving behind essentially the iron core. A similar event could have caused the large obliquity of Uranus. Giant impacts, by explaining many individual planetary characteristics as outcome of a general process rather than the result of unique and ad hoc local conditions, have undoubtedly become a central characteristic of the modern paradigm of planetary formation.

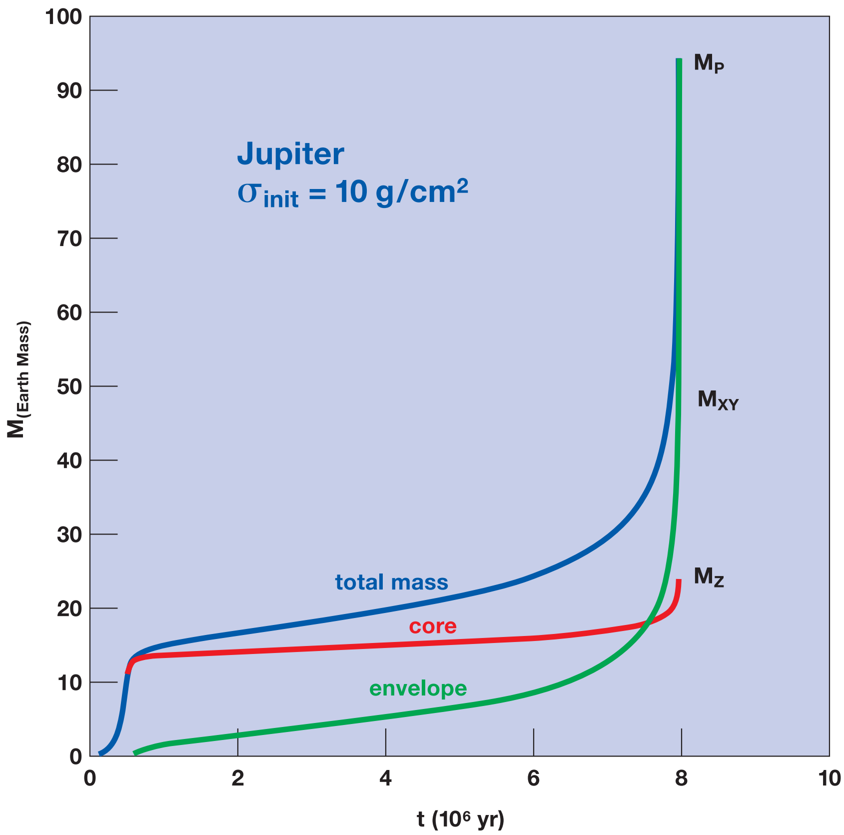


Figure 8
Accretion history of Jupiter. Note the rapid accretion of the gaseous envelope (green) onto the core (red) once the system has reached the critical state. (adapted from Pollack et. al 1996). With the surface density adopted here (10 g/cm²), Jupiter reaches its final mass in about 8 million years.

Giant planets

If a body grows beyond a critical mass of about 10 times the Earth's mass while still embedded in a gaseous disk, it will be able to accrete dynamically a considerable amount of surrounding gas eventually becoming a giant gaseous planet such as Jupiter or Saturn **(Figure 8)**.

In comparison to terrestrial planets, giant planets formation must proceed very rapidly since observations of many young stars as well as some theoretical considerations imply that circumstellar disks have lifetimes ranging from one to ten million years. The time available is therefore relatively short especially since the envelope accretion begins rather slowly at first **(Figure 8)**. It is therefore important that the seed body reaches critical mass rapidly hence relatively high surface densities of solids are required. This explains why giant planets were believed to form only sufficiently far away from the star where ices and not just silicates are present.

Planet formation: The problem

With the discovery of the first extra-solar giant planet, we have learned that some stars have giant planets orbiting at distances up to 10 times closer to their star than Mercury to the sun. While not all are that close, a significant number of them orbit within 0.1 AU of their star! In addition, except for these very close planets for which tides circularize the orbit, the eccentricity of all extra-solar planets is rather large. To illustrate to what extend these systems differ for our own solar system, we have plotted in **Figure 9** the eccentricity of all extra-solar giant planets as well as solar system planets as a function of their semi-major axis.

The presence of these giant planets at close orbital distances requires significant modifications and/or extensions to the standard formation model outlined above for two major reasons. First, the mass of typical proto-planetary disk within the orbit of the closest objects observed would not amount to a jupiter mass by a large factor even assuming 100% efficiency in collecting the matter. Second, even if there was sufficient mass available, the young 51 Peg B for example would be torn apart by the star's gravitational forces at its current location.

To reconcile theory and observations different mechanisms have been considered which essentially allow planets to migrate from their birth place to where they are observed today. This planetary migration is not a new idea, but it was never considered before as an essential ingredient in planet formation.

Most migration scenarii consider the gravitational interactions between the growing planet and the gaseous disk. When a massive objects orbits inside a gaseous disk, gravitational interactions between the two give rise to significant perturbations in the disk. In particular, if the planet is massive enough a gap in the disk opens while density perturbations extend further inwards and outwards **(Figure 10)**.

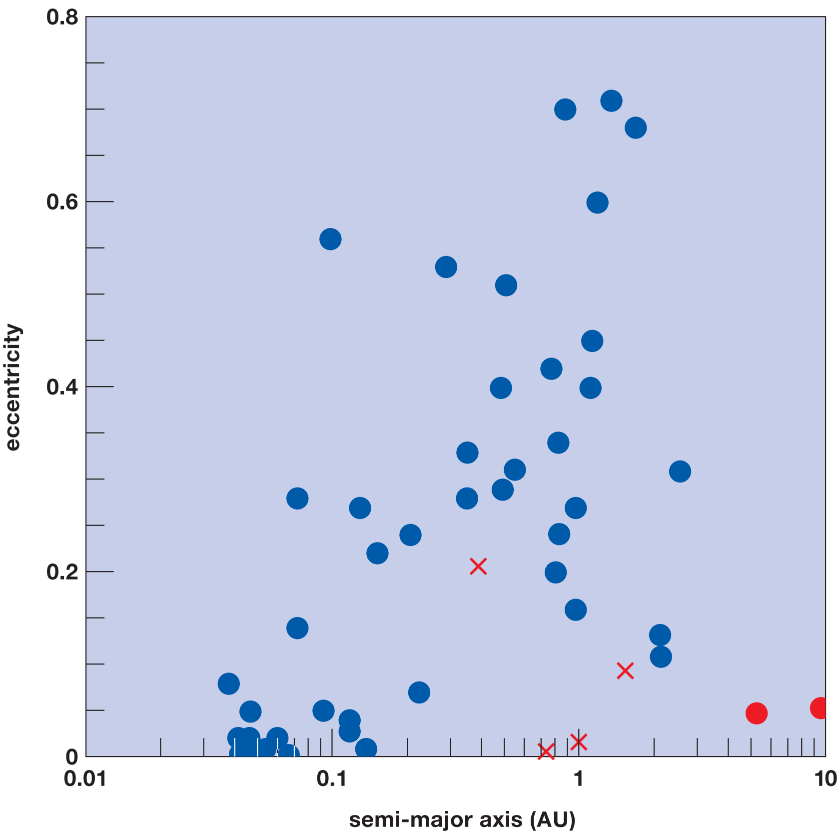


Figure 9
Eccentricity as a function of semi-major axis for giant extra-solar planets (blue dots), giant (red dots) and terrestrial planets (red crosses) of our own solar system. Note the difference in orbital parameters of giant planets in and outside our own planetary system. The absence of extra-solar giant planets beyond 3 AU is due to observational biases.

The tides raised in the disk by the planet result in a non-axisymmetric density distribution which in turn exerts a torque on the planet. The magnitude as well as the sign of the net torque is determined in a relatively complicated manner by the overall structure of the disk itself. Sophisticated multi-dimensional numerical simulations are required to actually compute this torque. **Figure 11** displays an example the torque exerted on a planet from the different regions of the disk.

The result is a transfer of angular momentum from the disk inside the planet's orbit to the planet as well as a transfer from the planet to the disk outside its orbit. As a result of this transfer, the planet opens a gap in the disk and spirals slowly inwards.

Figure 10
Simulation of a giant planet embedded in a gaseous disk. The local surface density is represented as the third dimension with red implying a high density. Notice the non-axisymmetric density perturbations in the disk induced by the presence of the planet (from G. Bryden).

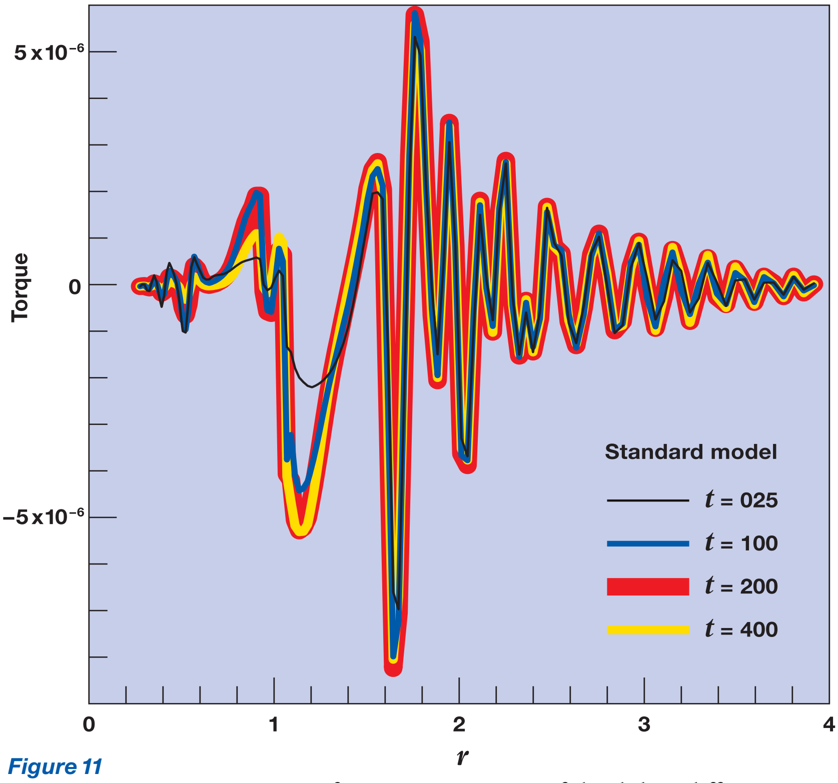


Figure 11
Torques exerted on a planet from various regions of the disk at different times during the simulation during which the planet was not allowed to move (units are arbitary). The net torque determining the radial migration of the planet is the sum of the local torques (from W. Kley).

While migration appears to solve some of the problems raised by the new systems, other issues remain puzzling and may hint to more fundamental problems in our understanding. For example, the migration timescale appears to be quite short (a few 100'000 years) therefore, why didn't the planets "fall" into their star but stop after having traveled 99% of the distance? A hint for the existence of a "stopping" mechanism is given by the apparent pile-up of planets visible in **Figure 9** at the shortest radii. Even more puzzling maybe is the fact that there are no signs of extensive inward migration in our own solar system. In particular, Jupiter does not appear to have migrated significantly.

In summary, while a few years ago it was believed that a relatively consistent working paradigm for the formation of planetary systems existed, today we are left with pieces of theories that do no longer provide a physically coherent picture!

For our understanding to make progress, further detections of extra-solar planets are required in order to have a statistically meaningful sample of objects to analyze. Ideally, this sample must include planets of all sizes not just giant planets so that the full extend of the existing diversity among planetary systems can be grasped.

Finding and studying earth-like planets

Indirect detections

So far the discovery of extra-solar giant planets has been made only through indirect methods in which the perturbations in the motion of the star induced by the presence of a planet are detected. The fact that these perturbations correspond indeed to orbiting planets has been confirmed recently by the detection of a transit, that is the decrease in stellar light as the planet passes between the star and us. These observations also yielded the radius and mean density of the object confirming its giant gaseous planet nature.

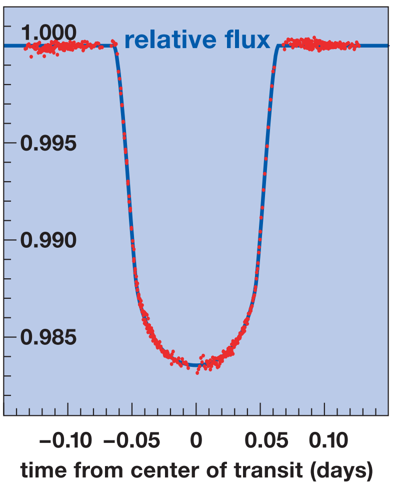


Figure 12
Light curve of HD 209458 during transmit as measured by HST (D. Charbonneau).

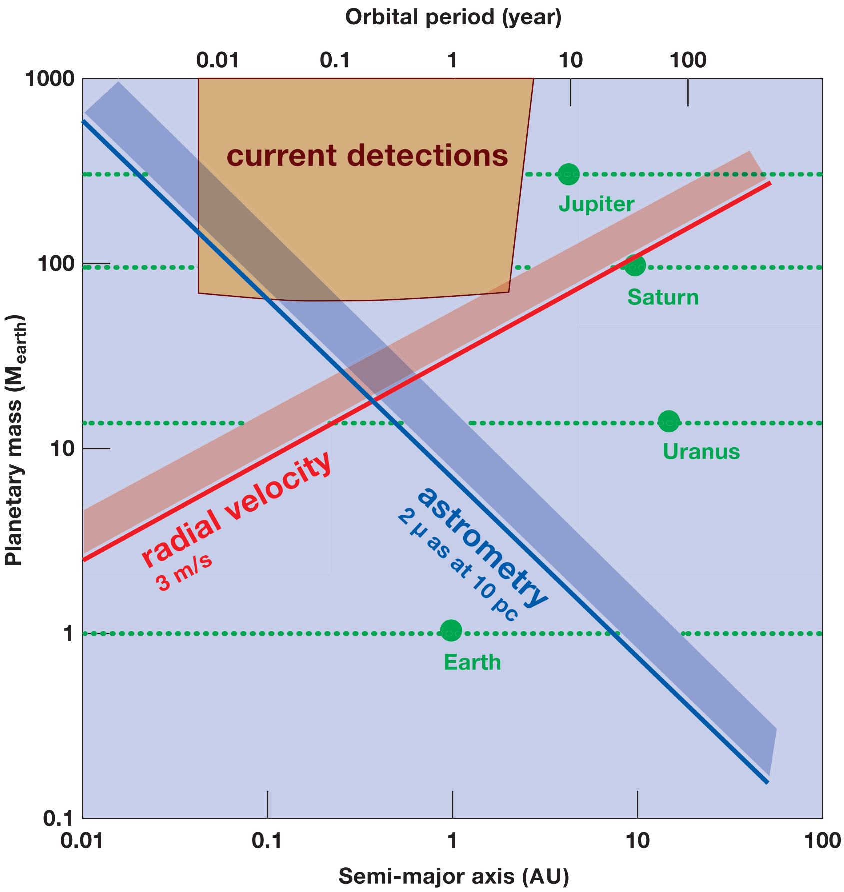


Figure 13
Detection limits of extra-solar planets for radial velocity searches (red line) and astrometric searches (blue line). The limit has been computed for the instrumental precision indicated. For reference, a few planets of the solar system are indicated. The brown box delineates the area of current discoveries.

Today, the most successful planet detection method remains the detection of radial velocity variations. The best measurements to date have an accuracy of approx 4 m/s (the speed of a casual biker). HARPS, a new instrument developed for ESO by a consortium including the Physics Institute of the University of Bern and led by Geneva Observatory, has been designed to reach an accuracy of 1 m/s (the speed of a pedestrian). Improving accuracy further is thought to be pointless because stellar surface motions induce an intrinsic noise in velocity measurements of this order. For comparison, the sun's reflex motion due to the presence of the Earth is about 8 cm/s, or more than an order of magnitude smaller.

Radial velocity searches are most sensitive to massive planets close to their parent star. To detect more distant and/or smaller planets, other methods are required. One of the most promising consists at measuring not the radial wobble of the star but its periodic displacements in the plane of the sky. In other words, high precision astrometry is used to detect again not the planet itself but the motion of the star. Since for a given measurement precision, one method is most sensitive to objects close to stars while the other to distant objects, both methods are actually complementary **(Figure 13)**.

To measure the difficulties involved in making these astrometric measurements, one has to realize that angles as small as 1 micro-arsecond have to be measured. This represents an angle smaller than the one sustained by a hair more than 100 km away! While just a dream a few years ago, such measurements will soon be possible using long baseline optical interferometers. Astrometry with an interferometer consists at combining the light of two or more separate telescopes and measuring angles (by adjusting optical delay lines) between the target star and a nearby reference star. To actually build such instruments is a very challenging technological task. For example, to get an efficient beam combination requires active distance control at a few tenths of nanometer precision level over more than 100 m of light path!

Despite these difficulties, long baseline optical interferometers, are scheduled to become standard facilities at the largest observatories in Europe as well as in the US. The European Southern Observatory (ESO) Very Large Telescope Interferometer (VLTI) located in Chile will be within the next three to five years the most powerful instrument of this sort **(Figure 14)**.

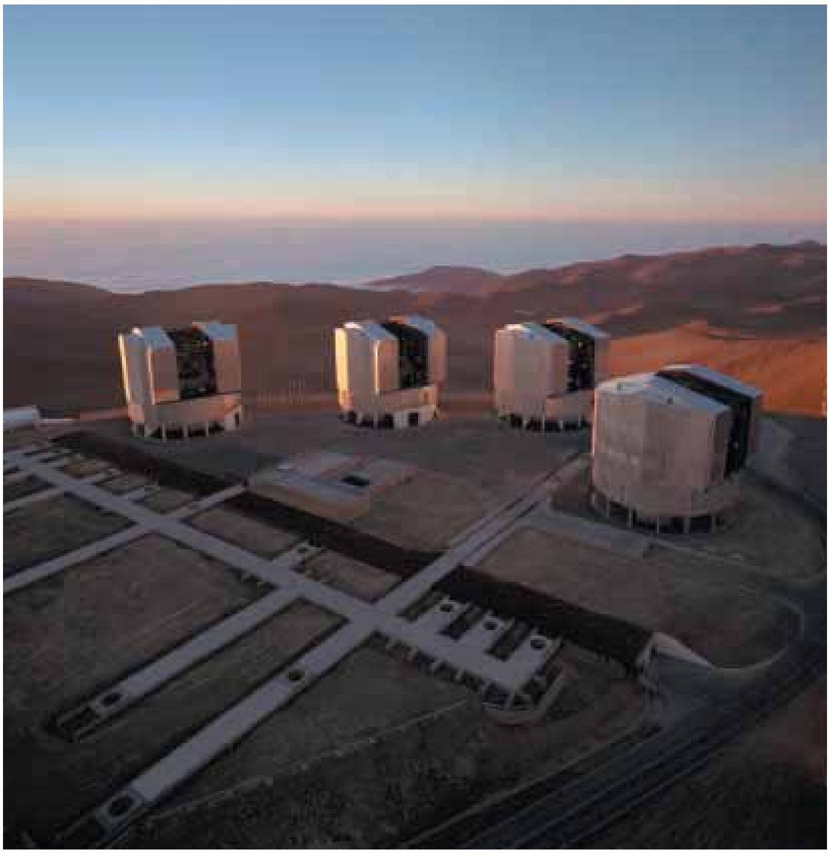


Figure 14
ESO's Very Large Telescope (VLT) array on the Paranal mountain in Chile consists of four 8.5 m diameter mirror telescopes. These telescopes can be coherently combined together with auxiliary telescopes (not on the picture but the rails are visible) to provide an interferometer with a baseline of up to 200 m).

Direct detection

While indirect methods will certainly yield a wealth of data about extra-solar planetary systems, the direct detection of photons originating from the planet itself would enable much more detailed physical studies. Examples include determining the chemical composition and temperature of the planet's atmosphere through spectroscopy, and studying surface structure and rotation by analyzing the light-curve. Even the presence of life could be tested by searching, for example, for oxygen lines in the spectrum. Indeed, the ozone absorption feature near 10 mm in the Earth's spectrum but absent in Venus' or Mars' only exists because of the photo-synthetic activities taking place on Earth.

Figure 15
In this colourful galaxy NGC 4214 star formation has taken place since billions of years (J. MacKenty/STScI et al. & the Hubble Heritage Team/AURA, STScI, NASA).

The major observational challenge in direct detections resides in observing a very faint object extremely close to a very bright one. Indeed, the most favorable brightness ratio (in the infra-red) between a planet and a sun is of order 10^7 and both objects are typically separated by 0.01 to 0.1 arcsec. While there are no direct detection of planets as of today, **Figure 16** illustrates the importance of resolution when it comes to detect faint objects close to bright ones. Such direct imaging of a faint object close to a bright one is again possible using optical interferometry. By combining the light of two or more arms of the interferometer in such a way that destructive interferences occur on-axis, the star is literally eclipsed thus revealing the fainter objects nearby.

To combine the light in such a way to "null" the star's light requires the exact knowledge of the phase of each light beam. Unfortunately, atmospheric turbulence renders this task extremely difficult and thus these so-called "nulling interferometer" will have to be flown in space. Such space-based instruments are on ESA's **(Figure 18)** as well as NASA's list of possible missions of the next decade albeit none has been definitively selected yet.

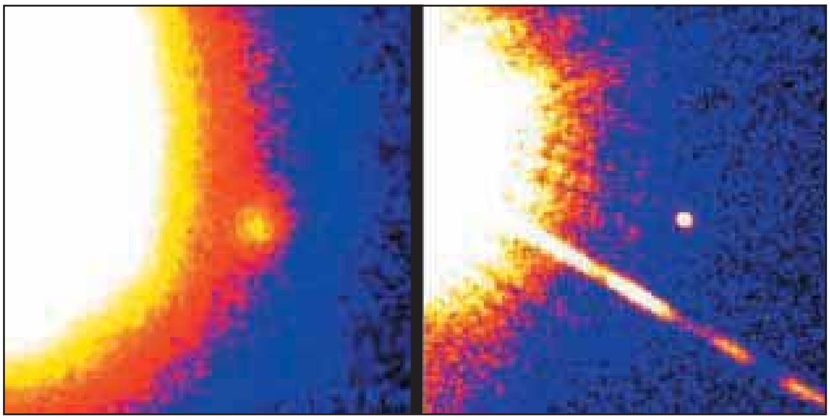


Figure 16
The direct detection of brown dwarf Gliese 229 from the ground (left) and from space using HST (right). This example illustrates how important high resolution is when it comes to detect faint objects close to bright ones. Detecting planets is orders of magnitudes more difficult and will be only possible from space.

Conclusions

The detection and study of earth-like planets outside our solar system will be one of the great scientific, technological, and philo-sophical undertakings of our time. Considered yesterday by most as a wild dream, the search for and studies of planets outside the solar system will become reality within the next decade or two. From the ground and later from space incredibly powerful instruments will search for and analyze the light of distant planets in an unprecedented world-wide effort to understand the origin and evolution of planets and maybe most importantly to check for the presence of even primitive life elsewhere in the universe.

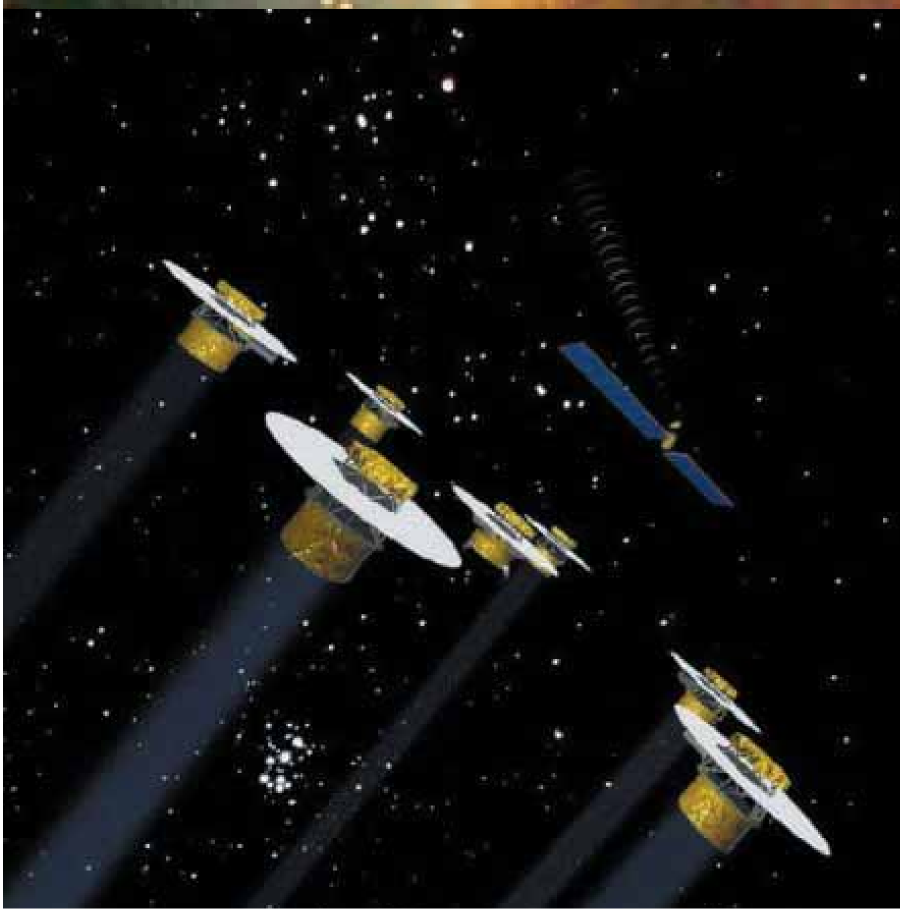


Figure 17
Star formation in the Omega Nebula (European Southern Observatory).

Figure 18
DARWIN: The InfraRed Space Interferometer. Six free-flying telescopes are connected interferometrically to detect Earth-like planets orbiting nearby solar-type stars.

The author

Willy Benz was born in Neuchâtel, Switzerland in 1955. Fascinated by natural sciences, he went to the University of Neuchâtel, where he acquired the diploma in physics. He carried his studies further as a research assistant at Geneva Observatory working in the group of Prof. Mayor (the well- known Swiss scientist having discovered the first planet orbiting a star other than the sun). It is during this period that his interests lead him from observations (he has spent many nights observing from the South of France and from the European Southern Observatory at La Silla, Chile) to a more theoretically oriented research. A doctoral thesis on the subject of star formation and rotation concluded his Geneva years.

After his doctorate, he went for two years to the Los Alamos, National Laboratory, USA, to hold a postdoctoral position in the Theoretical Division working with S. Colgate with whom he is still collaborating. There, he became interested in collision theories which are at the heart of our understanding of planet formation. A visit by A. Cameron triggered another collaboration that lasted many years and that led to his move to Harvard University where he was an assistant and later an associate professor for 5 years. During these years, his research focused on giant impacts and stellar explosions. He explored the possibility of forming the earth moon from the ejecta thrown into orbit following the impact on the young earth of a mars-sized body.

In 1991, he joined the rank of the faculty of the University of Arizona in Tucson were he became a professor of astronomy and planetary sciences. It was the time when comet Shoemaker-Levy 9 fell into Jupiter and thus a great time for someone interested in the physics of collisions!

After 13 years in the USA, he received a call from the University of Bern in 1997, where he was offered the position of physics professor in the Group of Space Research and Planetology at the Physikalisches Institut.

Willy Benz is currently living in Neuchâtel on the foothills of the Jura mountain range. He is married and has three daughters. Mountain biking in the Jura and long walks together with his dog are among his favourite activities allowing him to dive deeply into the loneliness of a wonderful nature.

A recognized teacher, Willy Benz aims at conveying his fascination of nature and the intense pleasure

provided by the understanding of some aspects of it to his students and to the general public as well. Space research has always benefited from the interest of the later and conveying to him the importance of the latest discoveries is seen by Willy Benz as an important task of a scientist. His sparkling lectures never fail the audience.

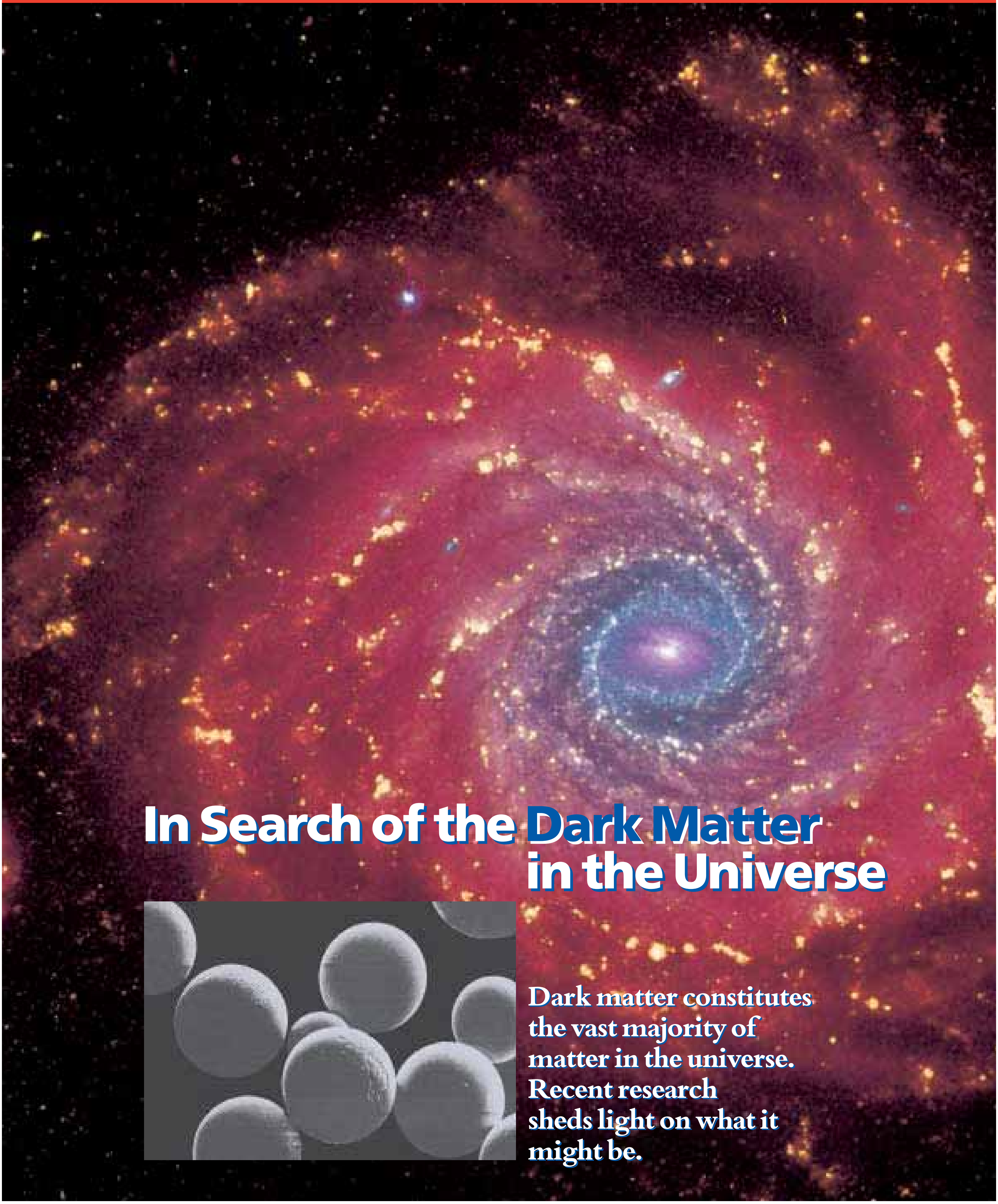
For him, despite its small size Switzerland plays an important role in the global scientific community. In many areas of space research Swiss scientists are at the forefront, as for instance in the discovery of planets outside the solar system. However, there is always the danger that Switzerland may loose this excellent position if the political will to support our stars dwindles, if we somehow fear to be ahead.

INTERNATIONAL SPACE SCIENCE INSTITUTE

No 7, May 2001

SPATIUM

Published by the Association Pro ISSI



In Search of the Dark Matter in the Universe

Dark matter constitutes the vast majority of matter in the universe. Recent research sheds light on what it might be.

Editorial

Dark matter. Dark energy.
Dark is beautiful!

But it is not so easy to find the dark. It requires charting the unknown as was done some six hundred years ago with the terra incognita. Vague contours, however, are known. Dark matter manifests itself by gravitational effects on visible matter, giving a general direction for those who want to move on in the dark. But much more remains to be found out.

Dark matter is a playground for creative scientists. Imagination is required to define the unknown and formulate hypotheses, which will most likely be thrown out once sufficient scientific progress has been made. Karl Popper, the great Austrian philosopher, showed that rejection – not confirmation – of a theory is the creative step when it is followed by the formulation of a more comprehensive theory.

Dark matter is full of surprises. Dark matter is everywhere. Dark matter may even be found in the dark underworld of Bern. The author of this issue of Spatium, Professor Klaus Pretzl, head of the Laboratory for High Energy Physics of the University of Bern, knows where to find it. He is in charge of the ORPHEUS experiment, now in its final stage of implementation some thirty meters below the University of Bern. This complex set-up seeks to investigate a specific class of dark matter particles, the weakly interacting massive particles.

In March 2000 Professor Pretzl reported on his institute's fascinating search for dark matter to an interested Pro-ISSI audience. We are very grateful to Klaus Pretzl for his kind permission to publish herewith his lecture.

Dark is beautiful.

Hansjörg Schlaepfer
Bern, May 2001

Impressum

SPATIUM
Published by the
Association Pro ISSI
twice a year

INTERNATIONAL SPACE SCIENCE INSTITUTE

Association Pro ISSI
Hallerstrasse 6, CH-3012 Bern
Phone ++41 31 631 48 96
Fax ++41 31 631 48 97

President
Prof. Hermann Debrunner,
University of Bern
Publisher
Dr. Hansjörg Schlaepfer,
legenda schläpfer wort & bild,
Winkel
Layout
Marcel Künzi, marketing · kommunikation, CH-8483 Kollbrunn
Printing
Druckerei Peter + Co dpc
CH-8037 Zurich

Front Cover
Spiral Galaxy NGC 1232
Copyright: European Southern Observatory ESO, PR Photo 37/e/98 (23 September 1998)

In Search of the Dark Matter in the Universe *)

Klaus Pretzl, Laboratory for High Energy Physics, University of Bern

Introduction

The story of the dark matter started in 1933, when the Swiss astronomer Fritz Zwicky published astonishing results from his studies of galactic clusters. In his paper he concluded that most of the matter in clusters is totally invisible. Zwicky argued that gravity must keep the galaxies in the cluster together, since otherwise they would move apart from each other due to their own motion. By determining the speed of motion of many galaxies within a cluster from the measurement of the Doppler-shift of the spectral lines he could infer the required gravitational pull and thereby the total mass in the cluster. Much to his surprise the required mass by far exceeded the visible mass in the cluster. His results were received with great scepticism by most astronomers at that time. It took another 60 years until he was proven to be right.

How much of this mysterious dark mass is there in the universe? Where do we find it? What is its real nature? Will we be able to directly detect it, since it does not radiate light or particles? It demonstrates its presence only by its gravitational pull on visible matter. These are the questions which intrigue astronomers, cosmologists, particle and nuclear physicists alike. In this talk I will try to give a short account of what we have learned about this mysterious kind of matter which dominates our universe.

How was matter created?

The very first instants

Our knowledge about the beginning of the universe is rather obscure, since it was born out of a state which is not describable by any physical law we know of. We simply call its indescribable moments of birth the Big Bang. After the Big Bang the universe rapidly expanded starting from an incredibly small region with dimensions of the order of 10^{-33} cm and an unthinkably high energy density of 10^{94} g/cm^3. Grand Unified Theories (GUT), which aim to describe the physical laws at this young age of the universe, tell us that physics was much simpler at that time, since there was only one force ruling everything. Today, however, we distinguish four fundamental forces: the force of gravity, which attracts us to the earth and the planets to the sun, the electromagnetic force, which keeps the negatively charged electrons in the outer shell of an atom attracted to the positively charged nucleus in the center, the weak force, which is responsible for the radioactivity we observe when unstable nuclei decay, and the strong force, which holds the protons and neutrons together in the nucleus. All these forces were unified in a single force at this early time. Physicists call this a state of symmetry, since the forces all have the same strength and are therefore indistinguishable from each other. However, while the universe was expanding rapidly this symmetry was quickly broken and gravity, the strong and the electro-weak forces appear today as separate forces with vastly different strength and reach.

The enormous isotropy and homogeneity of the universe, which we encounter when looking at the distribution of galaxies throughout space, and which we obtain from measurements of the cosmic microwave background radiation (CMB), is very puzzling. This is because there are regions in the expanding universe which have never been in causal contact with each other, i.e. light never had sufficient time since the Big Bang to travel far enough to transmit information from one to the other of these regions. To solve this puzzle the astrophysicists A. Guth and A. Linde introduced a scenario according to which the universe experienced an exponentially rapid expansion during a very short period of time between 10^{-36} to 10^{-34} seconds after the Big Bang. During this period the universe was rapidly growing in size by about a factor of 3×10^{43}. Within this scenario of inflation it becomes possible to causally connect regions of the universe which seem to be otherwise disconnected from each other. Nevertheless, after that short period of inflation the universe continued its expansion with retarded speed. Inflation is very much favoured by most cosmologists and strongly supported by the recent observations of the cosmic microwave background

*) Pro ISSI lecture, Bern, March 15th 2000

radiation by the Boomerang and Maxima experiments (see also below).

According to this scenario the inflationary phase of the universe was abruptly ended by the creation of matter and radiation about 10^{-34} seconds after the Big Bang. At that time the universe contained all the basic building blocks of matter, the quarks, the gluons and the electrons, and their antiparticles. Nevertheless, most of the energy of the universe resided in radiation, mainly photons and neutrinos, etc… However, as the universe cooled by expansion, radiation lost its energy faster than matter and when the universe became about ten thousand years old the energy balance shifted in favour of matter.

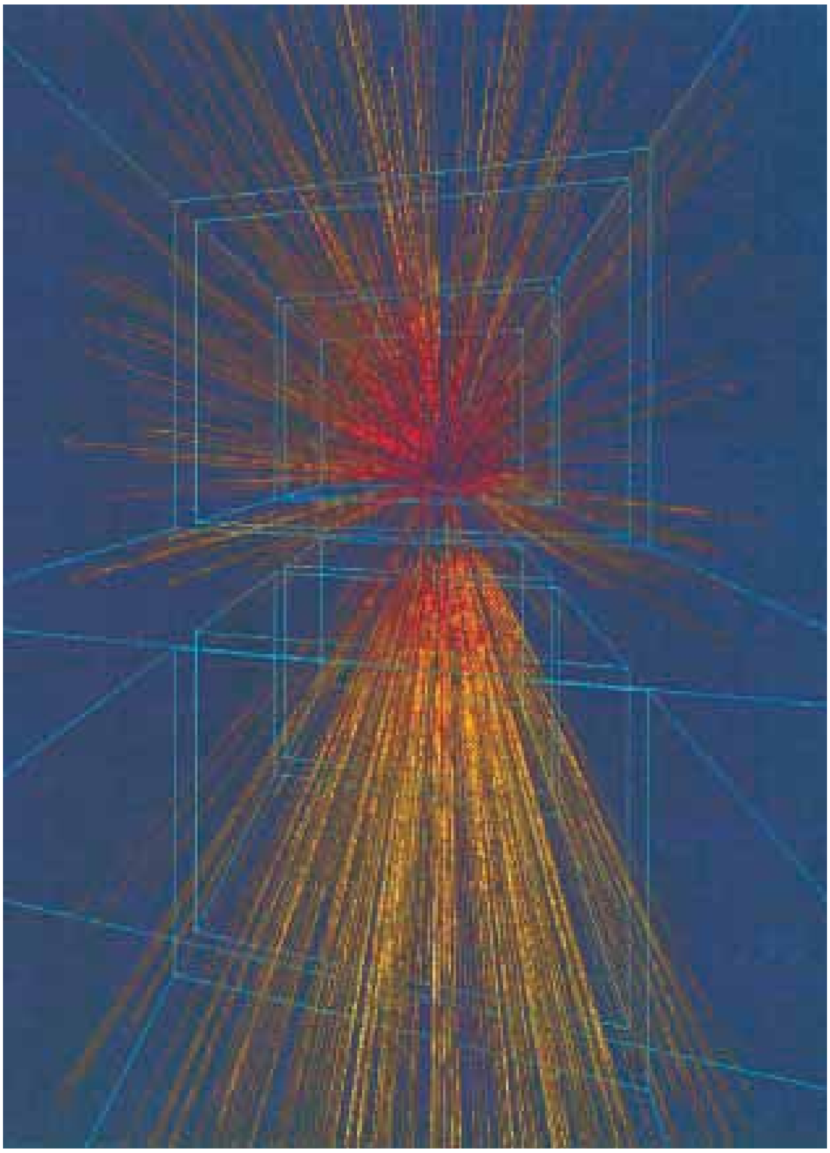


Figure 1
Temperatures of about 2×10^{12} were reached in this collision experiment at CERN. A lead target was bombarded with highly relativistic lead nuclei. At that temperature, the quark gluon plasma occurs. The picture shows the outburst of many particles which looks typical when transitions from the quark gluon to the nucleonic phase takes place. (NA49 experiment at CERN).

The quark gluon phase of matter ended about 10^{-6} seconds after the Big Bang, when the universe cooled to a temperature of 2×10^{12} Kelvin. At that temperature a phase transition from a quark gluon plasma to a nucleonic phase of matter took place **(Figure 1)**, where the protons and neutrons were formed. In this process three quarks of different flavour (so-called up-quarks and down-quarks) combine together to form a proton (two up-quarks and one down-quark) or a neutron (two down-quarks and one up-quark) and similarly antiprotons (two antiup-quarks and one antidown-quark) and antineutrons (two antidown-quarks and one antiup-quark). The gluons were given their name because they provide the glue for holding the quarks together in the nucleons. After this phase transition one would expect to end up with the same number of nucleons and antinucleons, which annihilate each other after creation, leaving us not a chance to exist. Fortunately this was not the case. The reason that we live in a world of matter with no antimatter is due to a very subtle effect, which treats matter and antimatter in a different way during

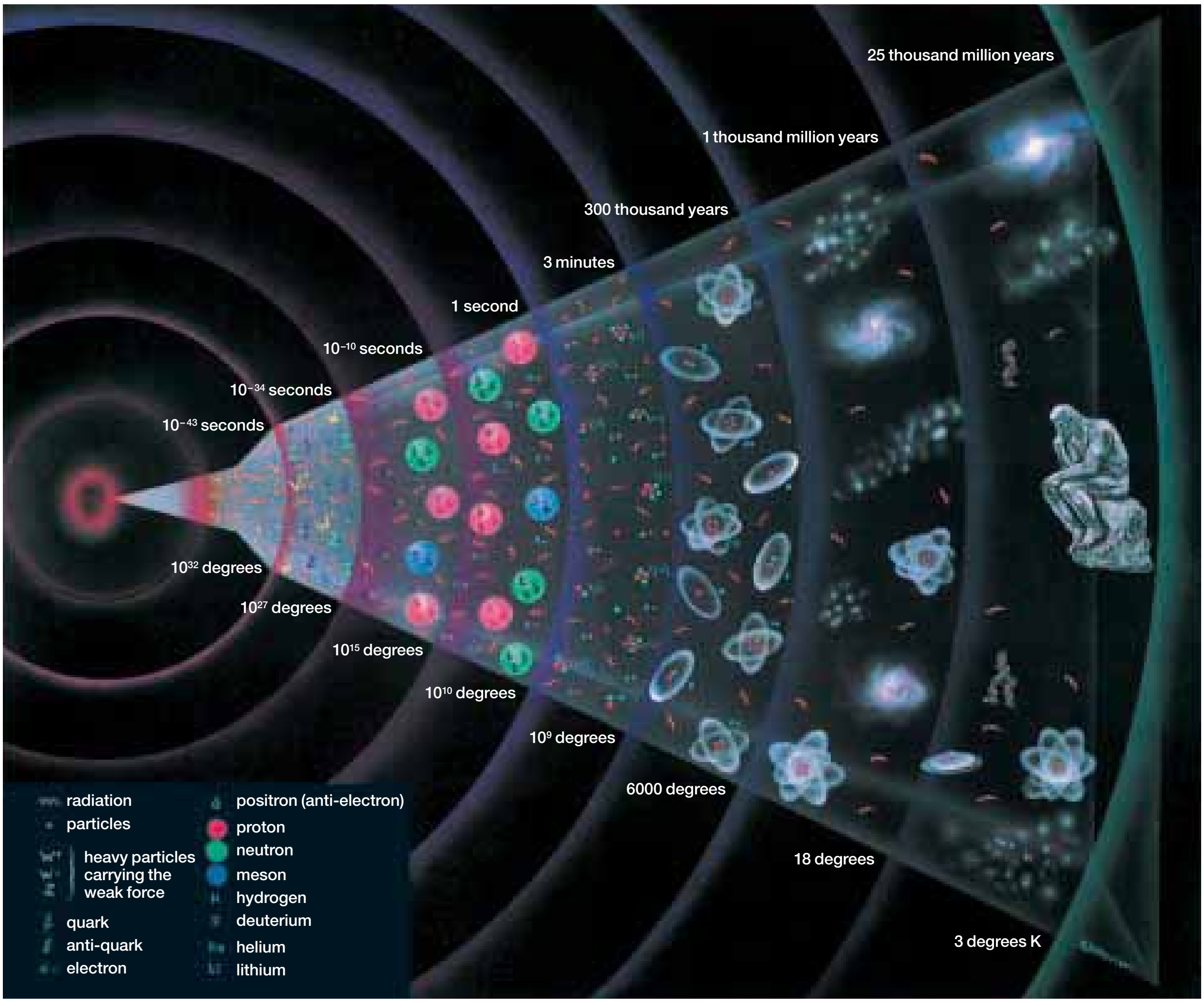


Figure 2
The evolution of the universe. Modern physics and experimental observations document the history of the universe from an incremental fraction of time after the Big Bang 15 billions of years ago up to its present state. Dark matter is seen today as having played a key role in the formation of stars and galaxies. (© CERN Publications, July 1991)

the phase of creation. This effect, which is known as CP-violation (Charge conjugation and Parity violation), was first discovered in an accelerator experiment by V. Fitch and J. Cronin in 1964, for which they got the Nobel prize in 1980, and was used by A. Sakharov to explain the matter antimatter asymmetry in the universe.

After further expansion the universe cooled to a temperature of 10^9 Kelvin, when protons and neutrons started to hang on to each other to form the light elements like helium, deuterium, lithium and beryllium. This phase of nucleosynthesis began a few seconds after the Big Bang. The heavier elements were only formed many million years later, mainly during star formation and supernova explosions. After their formation the light nuclei had hundred thousand years of time to catch electrons to build atoms.

The cosmic background radiation

About three hundred thousand years after the Big Bang radiation had not enough energy left to interact with matter and the universe became transparent for electromagnetic radiation. This radiation from the early universe was first discovered by R. Wilson and A. Penzias in 1965. They received the Nobel prize for this finding in 1978. Their discovery was made by chance, since they were on a mission from Bell Laboratories to test new microwave receivers to relay telephone calls to earth-orbiting communication satellites. No matter in what direction they pointed their antenna, they always measured the same noise. At first this was rather disappointing to them. But they happened to learn of the work of the astronomers R. Dicke and P. Peebles and they realised that the noise they were measuring was finally not the noise of the receiver, but rather the cooled down cosmic microwave background radiation (CMB) from the Big Bang. From the frequency spectrum and Planck's law of black body radiation the temperature of CMB was derived to be 2.7 Kelvin. Regardless which direction the cosmic radiation was received from, the temperature came out to be the same everywhere, demonstrating the enormous homogeneity of the universe. The most accurate CMB measurements come from the Cosmic Background Explorer satellite (COBE), which was sent into orbit in 1989. They found

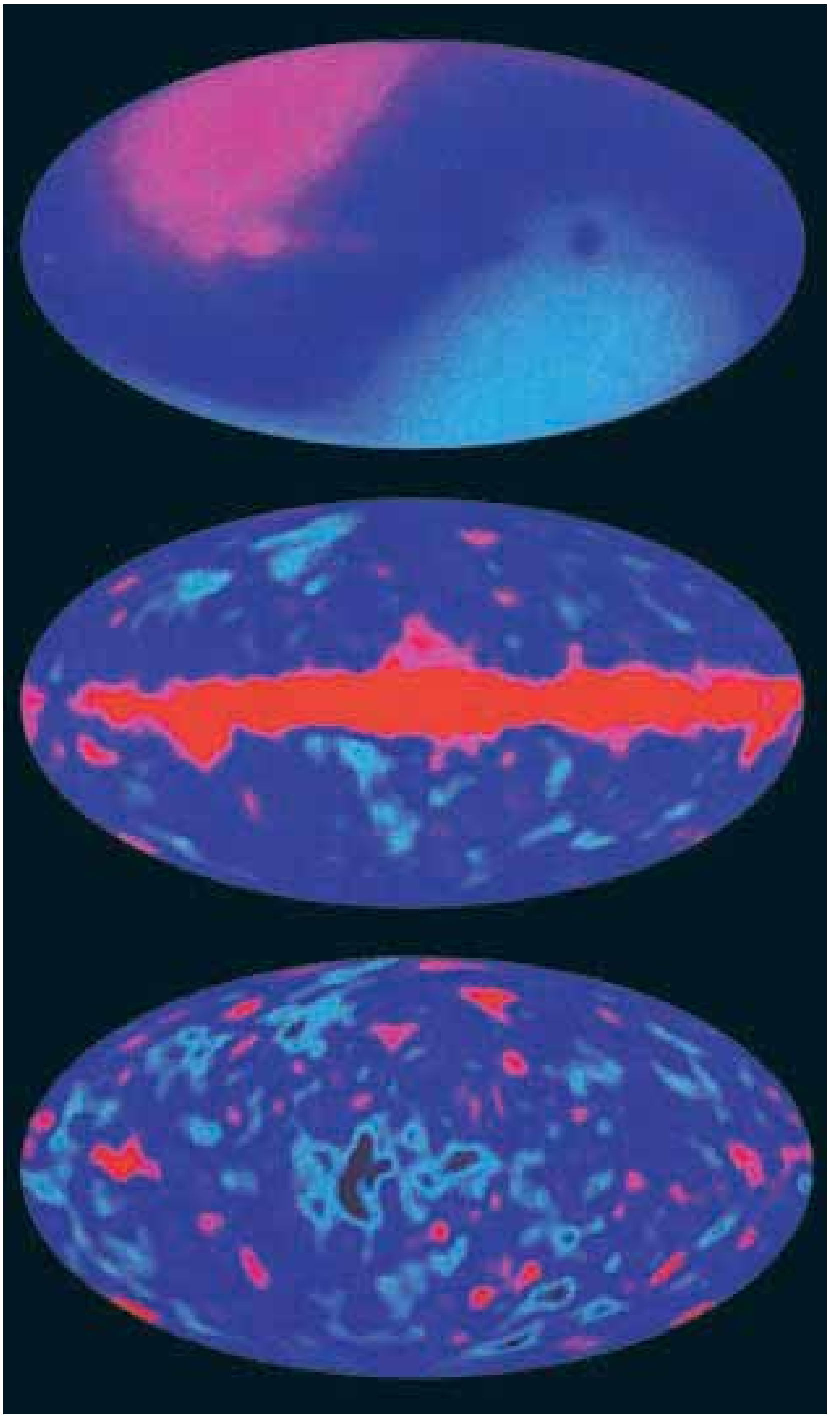


Figure 3
Full sky map of the cosmic background radiation as seen by the COBE mission. After subtraction of the dipol anisotropy (top) and our own galaxy's emission (center) temperature variation of 0.01% unveil matter density fluctuations in the very early universe (bottom). (© Physics Today)

temperature variations only at a level of one part in a hundred thousand.

Where does dark matter come from?

After this short description of the history of the universe and the creation of matter we ask ourselves: What about the dark matter? When and how was it created? A partial answer to this question is given to us by the COBE cosmic microwave background radiation measurements. They show islands of lower and higher temperatures appearing on the map of the universe which are due to density fluctuations of matter **(Figure 3)**. They were already present at the time radiation decoupled from matter, three hundred thousand years after the Big Bang, long before matter was clumping to form galaxies and clusters of galaxies. We have reasons to believe that these density fluctuations are due to the dark matter, which was probably created from quantum fluctuations during the inflationary phase of the universe. These tiny fluctuations expanded first through inflation and then retarded their expansion due to gravitational binding forces. They then formed the gravitational potential wells into which ordinary matter fell to form billion years later galaxies and stars. All galaxies and clusters of galaxies seem to be embedded into halos of dark matter.

How much matter is in the universe?

At first this question seems to be highly academic. It is not. The fate of our universe depends on its mass and its expansion velocity.

The destiny of the universe

In the 1920s the famous astronomer Edwin Hubble demonstrated that all galaxies are moving away from us and from each other. His discovery was the foundation stone of modern cosmology, which claims that the universe originated about 15 billion years ago in an unthinkably small volume with an unthinkably high energy density, the so-called Big Bang, and is expanding ever since. However, this expansion is counteracted by the gravitational pull of the matter in the universe. Depending on how much matter there is, the expansion will continue forever or come to a halt, which subsequently could lead to a collapse of the universe ending in a Big Crunch, the opposite of the Big Bang. The matter density needed to bring the expansion of the universe to a halt is called the critical mass density, which today would be roughly the equivalent of 10 hydrogen atoms per cubic meter. This seems incredibly small, like a vacuum, when compared to the density of our earth and planets, but seen on a cosmic scale it represents a lot of matter.

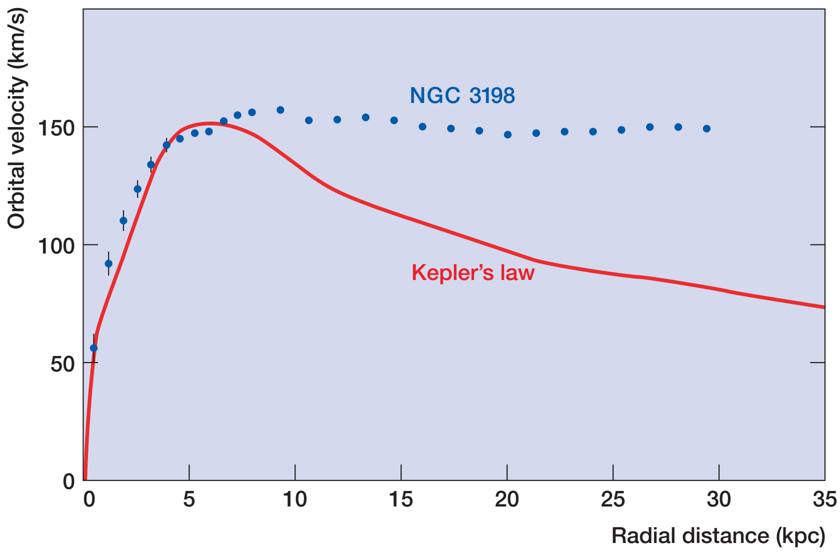


Figure 4
The observation of constant orbital velocities of stars around the galactic center (here the spiral galaxy NGC 3198) as a function of the radial distance provides convincing evidence for the presence of an extended halo of dark matter surrounding the galaxy. The expected curve from Kepler's law if there would be no dark matter is also shown.

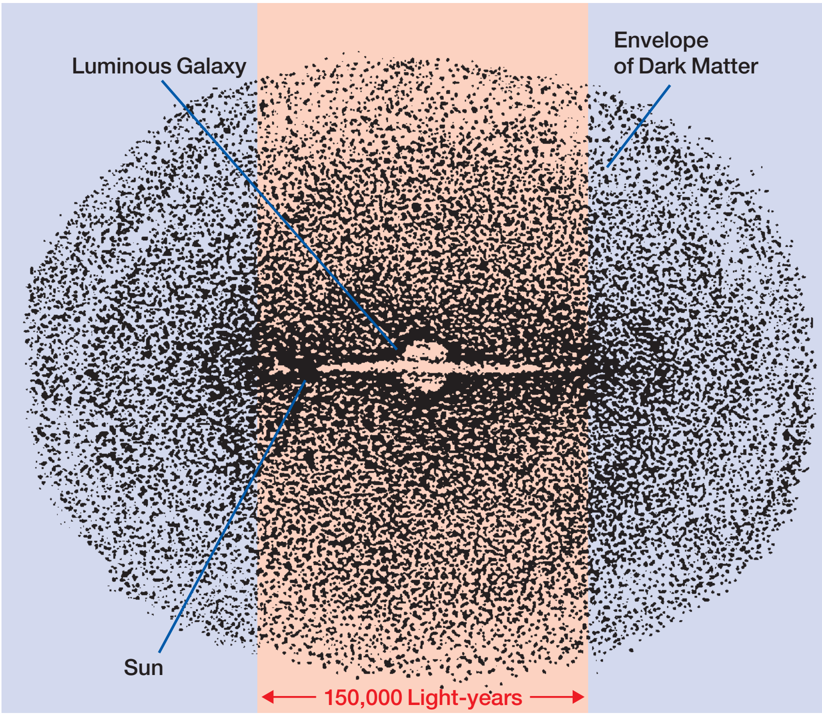


Figure 5
A galaxy as seen schematically from a distant point in the galactic plane. Dark matter forms a large halo extending far outside the outer edges of the galaxy.

How can we find out how much matter there is? When estimating the visible matter in the universe, astronomers look in a very wide and very deep region in space and count the number of galaxies. Typical galaxies containing hundreds of billions of luminous stars have a brightness which is proportional to their mass. Thus, by simply counting galaxies over a large volume in space and by assuming that galaxies are evenly distributed over the entire universe, one can estimate the total mass they contribute in form of visible mass to the universe. However, it turns out to be only 1% of the critical mass of the universe. Therefore, if the visible matter in form of stars and galaxies were the only matter in the universe, the universe would expand forever. We neglected here the amount of matter in form of planets, since they contribute not more than a few percent of the mass of a star. However, it came as a surprise when Vera Rubin and her team found out in the 1970s that the visible stars are not the only objects making up the mass of the galaxies. They measured the orbital speeds of stars around the center of spiral galaxies and found that they move with a constant velocity independent of their radial distance from the center **(Figure 4)**. This is in apparent disagreement with Kepler's law, which says that the velocity should decrease as the distance of the star from the galactic center increases, provided that all mass is concentrated at the center of the galaxy, which seems to

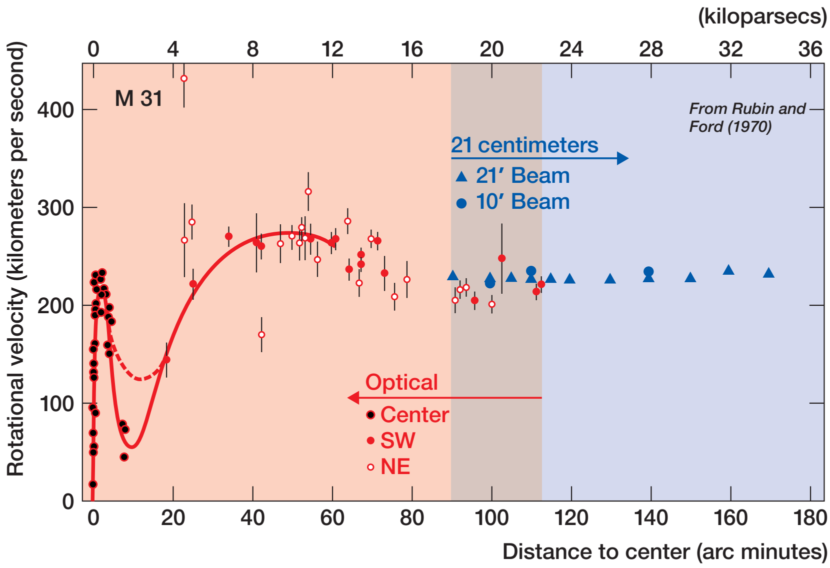


Figure 6
Experimental analysis of the rotational velocities in the Andromeda Galaxy M31 from optical observations (V. Rubin et al.) and radio observations at 21 cm wavelength (B. Roberts et al.).

be the case if only the luminous matter is considered. If however Kepler's law, which describes the orbital motion of the planets in our solar system very correctly, is valid everywhere in the universe, then the rotational velocities of the stars can only be explained if the mass of the galaxy is increasing with the radial distance from its center. Numerical calculations show that there must be at least an order of magnitude more matter in the galaxies than is visible. From their measurements, which they repeated on hundreds of different galaxies, they conjectured that each galaxy must be embedded in an enormous halo of dark matter, which reaches out even beyond the visible diameter of the galaxy **(Figure 5)**. Spiral galaxies are surrounded also by clouds of neutral hydrogen, which themselves do not contribute considerably to the mass of the galaxy, but which serve as tracers of the orbital motion beyond the optical limits of the galaxies. The hydrogen atoms in the clouds are emitting a characteristic radiation with a wavelength of 21 cm, which is due to a hyperfine interaction between the electron and the proton in the hydrogen atom and which can be detected. These measurements show that the dark matter halo extends far beyond the optical limits of the galaxies **(Figure 6)**. But, how far does it really reach out? Very recently gravitational lensing observations seem to indicate that the dark matter halo of galaxies may have dimensions larger than ten times the optical diameter. It is quite possible that the dark halos have dimensions which are already typical for distances between neighbouring galaxies within galactic clusters.

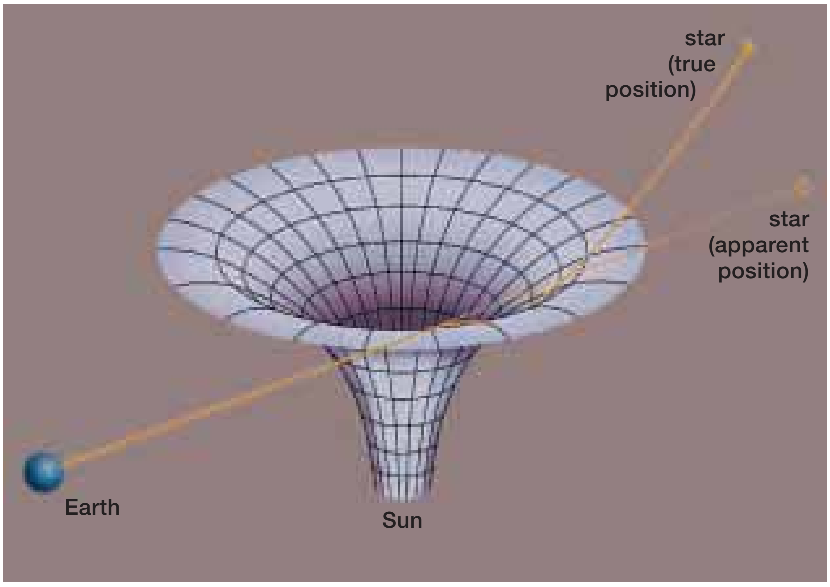


Figure 7
Space is curved by gravity. The light rays from a distant star are bent by the gravity field of the sun. The distant star therefore appears at a different position. (Spektrum der Wissenschaft)

Determining the mass in the universe

The effect of gravitational lensing is a consequence of Einstein's general relativity. It was first observed in 1919, when an apparent angular shift of the planet Mercury close to the solar limb was measured during a total solar eclipse. This was the first, important proof for the validity of Einstein's theory, according to which light coming from a distant star is bent when grazing a massive object due to the space curvature caused by the gravity of the object **(Figure 7)**. It was Fritz Zwicky in 1937 who realized that the effect of gravitational lensing would provide the means for the most direct determination of the mass of very large galactic clusters, including dark matter. But it took more than 50 years until his suggestion was finally realized and his early determination of the mass of the COMA cluster, in 1933, was confirmed. With the Hubble telescope in space and the Very Large Telescopes (VLT) at the Southern Observatory in Chile astronomers now have very powerful tools, which allow them not only to explore the visible, but also the dark side of the universe with gravitational lensing.

An observer sees a distorted multiple image of a light source in the far background, when the deflecting massive object in the foreground is close to the line of sight. The light source appears to be a ring, the so-called Einstein-Ring,

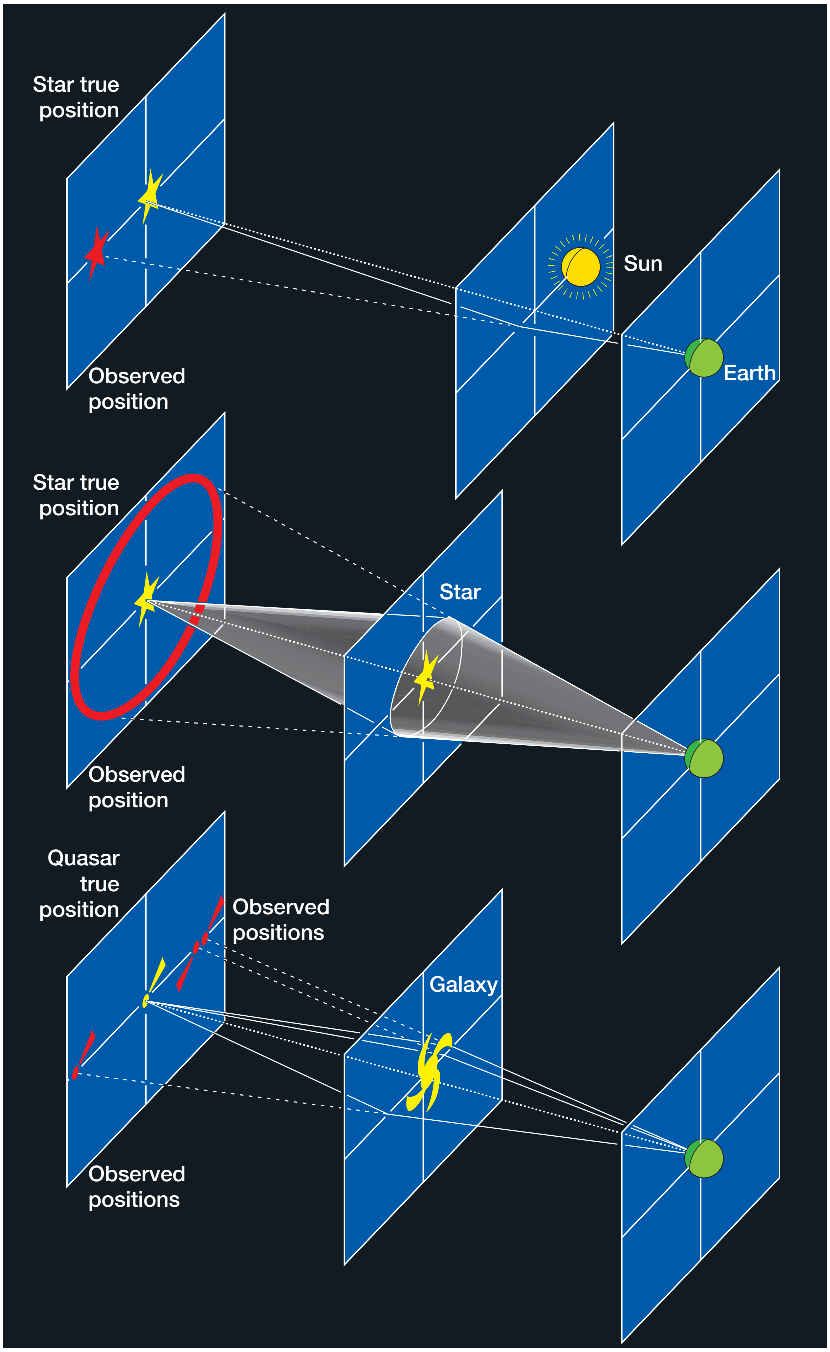


Figure 8
Gravitational lensing occurs when the gravity field of a massive celestial object bends the path of light emitted by a distant source. Einstein predicted the deflection of starlight by the sun (top) and the ring that would appear if the star and the celestial body were aligned perfectly (center). Lens systems found to date result from the alignment of extra-galactic quasars and galaxies (bottom). (© Scientific American, July 1988)

when the object is exactly in the line of sight **(Figure 8)**. If one knows the distance of the light source and the object to the observer one is able to infer the mass of the object from the lensing image. With this method it was possible to determine the mass of galactic clusters, which turned out to be much larger than the luminous matter. It seems that the gravitational pull of huge amounts of dark matter is preventing individual galaxies from moving away from each other and is keeping them bound together in large clusters, like for example the famous Coma cluster. By adding the total matter (dark and luminous matter) in galaxies and clusters of galaxies one ends up with a total mass which corresponds to about 30% of the critical mass of the universe. With only 1% luminous mass this would mean that there is 30 times more dark mass in the universe. In addition, the universe would be growing forever, since its total mass is subcritical to bring the expansion to a halt. But as we will see, this seems not to be the full story.

The discovery of dark energy

The big surprise came in 1998 from a supernovae type1a survey performed by the Super Cosmology Project (SCP) and the High z-Supernova Search (HZS) groups. Supernovae type1a are a hundred thousand times brighter than ordinary stars. They are still visible at very great distances for which their light needed several million years to travel until it reached us. In principle we experience now super-

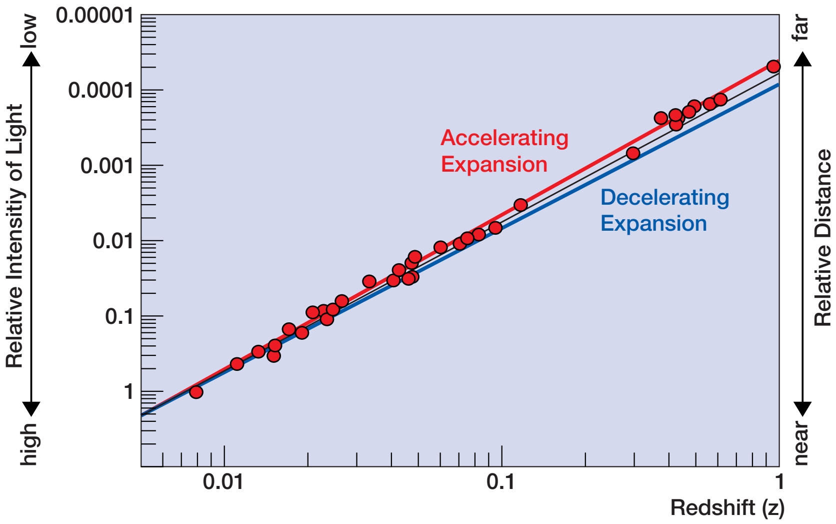


Figure 9
Recent supernovae distance measurements show that the expansion of the universe is accelerating rather than decelerating as assumed before. This observation suggests the presence of dark energy. (© Scientific American, January 1999)

novae explosions which happened several million years ago. Since in every supernova type1a explosion there is always the same total amount of energy released, they all have the same brightness and therefore they qualify as standard candles in the cosmos. Their distance from us can then be inferred from the measurement of their apparent brightness. By probing space and its expansion with supernovae distance measurements astrophysicists learned that the universe has not been decelerating, as assumed so far, but has rather been expanding with acceleration **(Figure 9)**. More measurements are still needed to corroborate these astonishing findings of the supernovae survey. But it already presents a surprising new feature of our universe, which revolutionizes our previous views and leaves us with a new puzzle. In order to speed up the expansion of the universe a negative pressure is needed, which may be provided by some unidentified form of dark energy.

This ubiquitous dark energy amounts to 70% of the critical mass of the universe and has the strange feature that its gravitational force does not attract, on the contrary it repels. This is hard to imagine since our everyday experience and Newton's law of gravity tell us that matter is gravitationally attractive. In Einstein's law of gravity, however, the strength of gravity depends not only on mass and other forms of energy, but also on pressure. From the Einstein equation, which describes the state of the universe, it follows that gravitation is repulsive if the pressure is sufficiently negative and it is attractive if the pressure is positive. In order to provide enough negative pressure to counterbalance the attractive force of gravity, Einstein originally introduced the so-called cosmological constant to keep the universe in a steady state. At that early time all observations seemed to favour a steady state universe with no evolution and no knowledge about its beginning and its end. When Einstein learned about the Hubble expansion of the universe in 1920 he discarded the cosmological constant by admitting that it was his biggest blunder. For a long time cosmologists assumed the cosmological constant to be negligibly small and set its value to zero, since it did not seem to be of any importance in describing the evolution of the universe. This has changed very recently, since we know about the accelerated expansion of the universe. However, there remain burning questions like why is the cosmological constant so constant over the lifetime of the universe and did not change similar to the matter density, and what fixes its value. Besides the cosmological constant, other forms of dark energy are also discussed by cosmologists, like for example vacuum energy, which consists of quantum fluctuations providing negative pressure, or quintessence, an energy source which, unlike vacuum energy and the cosmological constant, can vary in space and time.

In contrast to dark matter, which is gravitationally attractive, dark energy cannot clump. Therefore it is the dark matter which is responsible for the structure formation in the universe. Although the true nature of the dark energy and the dark matter is not known, the latter can eventually be directly detected, while the former cannot.

What is the nature of the dark matter?

Baryonic dark matter

Before speculating with exotic matter, the obvious thing is to look for non-luminous or very faint ordinary matter in form of planetary objects like jupiters or brown dwarfs for example. If these objects represent the dark matter, our galactic halo must be abundantly populated by them. Since they may not be visible even if searched for with the best telescopes, B. Paczynski suggested to look for them by observing millions of individual stars in the Large and the Small Magellanic Cloud to see whether their brightness changes with time due to gravitational lensing when a massive dark object is moving through their line of sight **(Figure 10)**. Several research groups, so-called MACHO, EROS and OGLE followed Paczynski's idea – Paczynski himself was member of the OGLE group – to look for these so-called Massive Astrophysical Compact Halo Objects (MACHOs) using gravitational lensing. They found some of these dark objects with masses smaller than the solar mass, but by far not enough to explain the dark matter in the halo of our galaxy. Other objects like black holes or neutron stars could also have been detected by this method, but we knew

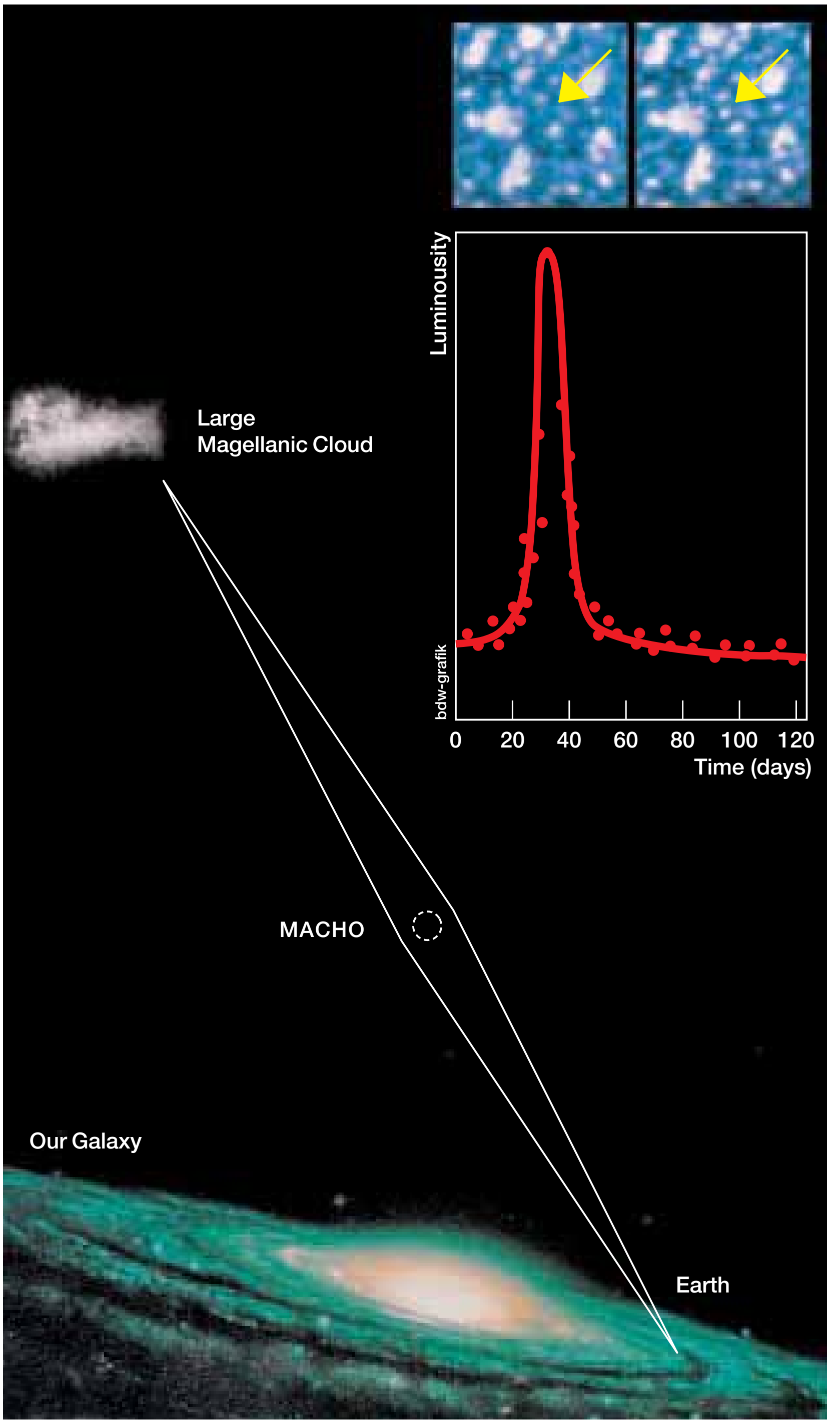


Figure 10
Massive dark objects (Massive Astrophysical Compact Halo Objects MACHOs) moving through the line of sight between the observer and a distant star in the Large Magellanic Cloud cause the apparent luminosity to change. (© Bild der Wissenschaft, 2/1997)

already that there are not many of them in the galactic halo.

Do we know how much ordinary matter exists in the universe? Under ordinary matter or so-called baryonic matter (barys meaning strong or heavy in ancient Greek) we understand matter in form of chemical elements consisting of protons, neutrons and electrons. About 3 minutes after the Big Bang the light elements, like hydrogen, deuterium and helium, were produced via nucleosynthesis. From the measurement of their present abundances we can estimate the total amount of the baryonic matter density in the universe. This amounts to not more than 6% of the critical mass density of the universe. It shows that most of the baryonic matter is invisible and most of the dark matter must be of non-baryonic nature.

Non-baryonic dark matter

The most obvious candidates for non-baryonic matter would be the neutrinos, if they had a mass. Neutrinos come in three flavours. If the heaviest neutrino had a mass of approximately 10^{-9} times the mass of a hydrogen atom, it would qualify to explain the dark matter. This looks like an incredibly small mass, but the neutrinos belong to the most abundant particles in the universe and outnumber the baryons by a factor of 10^{10}. For a long time it was assumed that neutrinos have no mass. The standard model of particle physics includes this assumption. All experimental attempts to determine the mass of the neutrinos ended in providing only upper limits. However, in 1998 an underground detector with the name SUPER-Kamiokande in Japan observed anomalies in the atmospheric neutrino flux which is highly suggestive that neutrinos have indeed a mass. These observations will have to be corroborated by planned accelerator experiments, like K2K in Japan, MINOS in the USA and OPERA in Europe. The OPERA experiment will be constructed in the underground Gran Sasso laboratory, which is located about 100 km north east of Rome. For this experiment, a neutrino beam will be sent from CERN to the Gran Sasso laboratory. If neutrinos have a mass they would change their flavour during their journey over the 735 km distance from CERN to the Gran Sasso. They would start as muon-neutrinos at CERN and would arrive as tau-neutrinos at the Gran Sasso. This change of flavour can be detected by the OPERA experiment, in which our group in Bern is also participating. The future will show whether neutrinos will be able to contribute significantly to the missing mass in the universe. Massive neutrinos may also provide the solution to the puzzle of the missing neutrinos from our sun.

A cocktail of non-baryonic dark matter

Computer models allow us to study the development of small and large scale structures under the hypothesis of various non-baryonic dark matter candidates. Two main categories are distinguished, namely the so-called hot and cold dark matter. Neutrinos would qualify under the category hot dark matter, since their velocities were very large when they decoupled from matter, a few milliseconds after the Big Bang. Because of their speed they were not able to clump on small, typical galactic scales, but their gravitational force would still allow for clustering on very large, typical supercluster scales. Thus in a hot dark matter dominated universe only the formation of large scale superclusters would be favoured. In such a model superclusters would fragment into smaller clusters at a later time. Hence galaxy formation would be a relatively recent phenomenon, which however is in contrast to observation. Cold dark matter candidates, on the other hand, would have small velocities at early phases and therefore would be able to aggregate into bound systems at all scales. A cold dark matter dominated universe would therefore allow for an early formation of galaxies in good agreement with observations, but it would overpopulate the universe with small scale structures, which does not fit our observations. Questions like, how much hot and how much cold or only cold dark matter, are still not answered. Some computer models yield results which come closest to observations when using a cocktail of 30% hot and 70% cold dark matter.

Neutralinos

Exotic particles like neutralinos are among the most favoured cold dark matter candidates. Neutralinos are stable elementary particles which are predicted to exist by Super Symmetry (SUSY), a theory which is an extension of the Standard Model of elementary particles. Thus if they exist, they would solve two problems at the same time: namely the dark matter as well as SUSY, which is a prerequisite for the unification of all forces in nature, the so called grand unification theory (GUT). Experiments at the Large Hadron Collider (LHC) at CERN, which is under construction and will be operational in 2005, will also search for these particles.

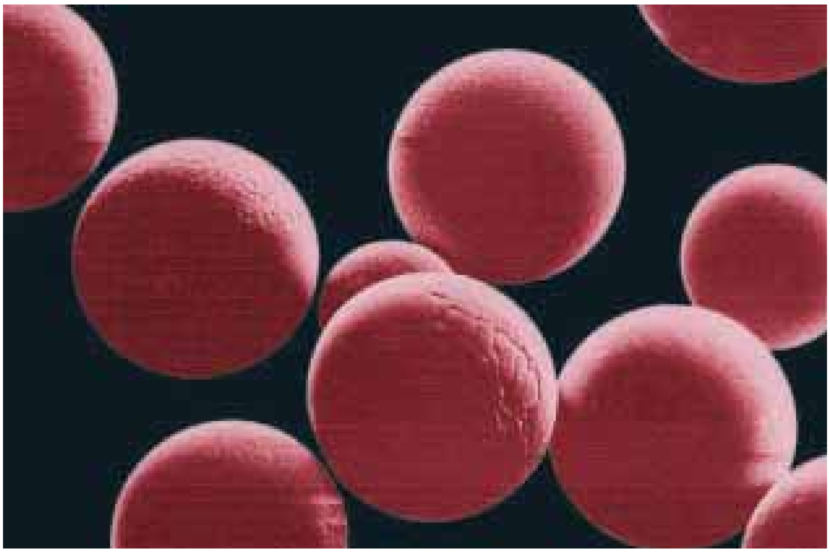


Figure 11
The ORPHEUS detector contains billions of small tin granules with a diameter of 30 micrometers. They are cooled down to –273 °C, where they are in a superconducting state. An interacting WIMP may generate enough heat in a granule to cause a phase transition from the superconducting to the normal conducting state. This phase transition of the granule can be measured with a sensitive read-out system.

Weakly interacting particles

If the dark matter consisted of neutralinos, which would have been produced together with other particles in the early universe and which would have escaped recognition because they only weakly interact with ordinary matter, special devices would have to be built for their detection. These detectors would have to be able to measure very tiny energies which these particles transfer in elastic scattering processes with the detector material. Because of the very weak coupling to ordinary matter these particles are also called WIMPs, for Weakly Interacting Massive Particles. They would abundantly populate the halo of our galaxy and would have a local density in our solar system which would be equivalent to one hydrogen atom in 3 cm^3. Since they would be bound to our galaxy allowing for an average velocity of 270 km/s, their flux (density times velocity) would be very large. However, because they only weakly interact with matter, the predicted rates are typically less than one event per day per kilogram detector material.

WIMPs can be detected by measuring the nuclear recoil energy in the rare events when one of these particles interacts with a nucleus of the detector material. It is like measuring the speed of a billiard ball sitting on a pool table after it has been hit by another ball. Because of the background coming from the cosmic rays and the radioactivity of the material surrounding the detector, which yield similar signals in the detector as the WIMPs, the experiment must be carried out deep underground, where cosmic rays cannot penetrate, and must be shielded locally against the rest radioactivity of materials and the radioactivity in the rock.

The race for WIMPs

Several groups in the USA, in Europe and in Japan are searching for WIMPs employing different techniques. However, in order to be sensitive to very small nuclear recoil energies, innovative new cryogenic detectors have been developed. At temperatures near the absolute zero point, at –273 °C, where the heat capacity is very small, a very tiny recoil energy can be transformed into a measurable signal. One of these innovative detection systems has been developed by our group at the University of Bern. The detector is called ORPHEUS in analogy to Greek mythology. However, we the experimentalists spend a lot more time underground than Orpheus ever did, because our detector is situated in an underground laboratory 30 meters below the University of Bern. The detector consists of billions of small superconducting tin grains with a diameter of a few micrometers, which are cooled down to –273 °C where they are in a superconducting state **(Figure 11)**. When an incoming WIMP hits a granule it can lead to a minute temperature increase of the granule, which can be just sufficient to cause a phase transition from the superconducting to the normal state. This phase transition of an individual granule can be detected by a pick-up loop which measures the magnetic flux change due to the disappearance of the Meissner effect. The ORPHEUS detector, consisting of about 0.5 kg of superconducting tin granules, just started to be operational **(Figure 12)**. It will however take some time before the background is sufficiently understood and true WIMP signals can be detected. Nevertheless, there is still the possibility that the dark matter does not consist of WIMPs, in which case we would not find any signals and other strategies would have to be developed in order to disclose the mystery of the dark matter in our universe.

Conclusions

Despite many interesting candidates, we still do not know what the dark matter is made of. Intensive experimental work is going on by trying to directly detect this abundant matter in our own galaxy or by learning about it through indirect methods. Recent studies of the Cosmic Background Radiation lead us to believe that the dark matter was present long before the chemical elements were created and was responsible for the fascinating architecture of our observable universe, providing also the necessary conditions for life to develop. Without the attractive pull of the dark matter, the universe would be extremely dull today. There would be no galaxies, no stars, no planets and no life. The total amount of dark matter and dark energy will determine the destiny of our universe. From all we have learned so far, it looks as if the universe will not end in a Big Crunch, instead it will expand forever.

In the coming years we are eagerly awaiting new insights into the dark side of the universe. I would like to close with a quote from Shakespeare: There is no darkness, only ignorance.

Acknowledgements

I would like to thank Dr. Hansjörg Schlaepfer for his valuable contributions and for enjoyable discussions. I am very grateful to Prof. Gerhard Czapek, Prof. Peter Minkowski and Irene Neeser for the critical reading of the manuscript.

Figure 12
Cross sectional view of the ORPHEUS dark matter detector. The detector is horizontally connected via a side axis with a dilution refrigerator, which cools it down to a temperature of –273 °C. It is surrounded by shielding material and scintillation counters to protect against background from particles other than WIMPs.

The author

Klaus Pretzl was born in Munich, Germany, in 1940. From 1941 to 1956 his family lived in Schliersee, south of Munich in the Bavarian Alps, where he also went to primary school. In 1956 he returned to Munich, where he made his Matura at the Wilhelms-Gymnasium.

His fascination for physics was inspired by public lectures about quantum physics given by Professor Auer, a former director of the Deutsches Museum in Munich. As a result he studied physics at the Technische Hochschule in Munich and wrote his diploma thesis in nuclear physics under Prof. H. Meyer Leibnitz and Prof. P. Kienle. He also gained a scholarship from EURATOM to work at the Centre des Recherches Nucléaires in Strasbourg.

His interest turned from nuclear physics to elementary particles, the fundamental building blocks of matter. For this reason he went to CERN in Geneva, where he met Prof. W. Paul (Nobel Prize winner 1989), who offered him to work on an experiment to study meson nucleon backward scattering at high energies. This work was the subject of his PhD thesis, which he submitted to the faculty of the University of Munich in 1968.

After receiving his PhD, Klaus Pretzl went to the USA for 4 years, where he worked at Fermi-Lab, near Chicago, under Prof. R. Wilson on the construction of the world's largest particle accelerator at that time. During the construction time of the accelerator, Klaus Pretzl conducted several experiments studying strong interaction processes at the nearby Argonne National Laboratory. When the construction of the Fermi-Lab accelerator was completed, he was involved in the first experiments to explore physics at the very highest energies.

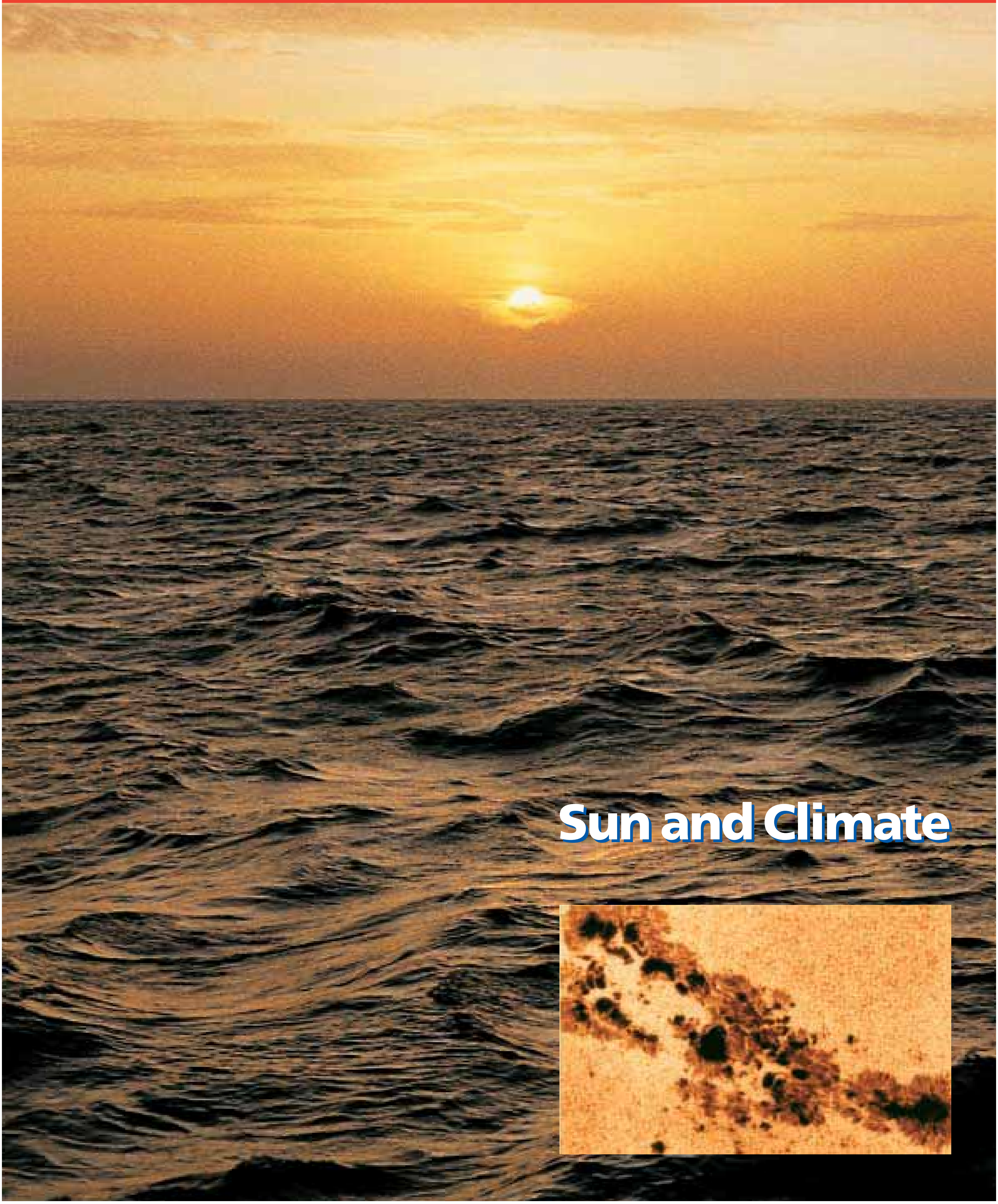
In 1973 Klaus Pretzl returned to Europe and joined the Max Planck Institute for Physics in Munich. There he first worked at the e^+ - e^- - storage ring DORIS at DESY (Deutsches Elektron Synchrotron), in Hamburg, with the so-called DASP (Double Arm Spectrometer) experiment, which lead to the discovery of the charmonium states. Later he performed several experiments at CERN studying QCD (Quantum Chromo Dynamic) processes. In 1982 he became interested in the topic of dark matter and started to study new concepts for its direct detection.

In 1988 he became professor of physics and head of the Laboratory for High Energy Physics at the Bern University. There he started a new group with the aim to develop cryogenic detectors for a dark matter experiment, which is now in operation in the Bern Underground Laboratory and is called ORPHEUS. Klaus Pretzl and his group are also involved in the ATLAS experiment at the Large Hadron Collider (LHC) at CERN and in the long baseline neutrino oscillation experiment OPERA at the Gran Sasso Laboratory in Italy.

INTERNATIONAL SPACE SCIENCE INSTITUTE
No 8, November 2001

SPATIUM

Published by the Association Pro ISSI



Sun and Climate

Editorial

Man's early ancestors were wise enough not to call themselves *homo sapiens.* They were also wise enough not to exploit their habitat to an extent which could cause the climate to change. But the change occurred all the same: droughts entailed the jungles to retreat, leaving wide-open space for savannas. In order to survive in such radically different surroundings, the hominids had to adapt their behaviour appropriately. They were successful: they were able to propagate all over the Earth.

Much later, their descendents coined the term *homo sapiens,* the wise man, for themselves. They were wise enough to invent machines that burn huge amounts of fossil fuel. They did so to such an extent that the Earth's atmosphere increasingly became a greenhouse: letting the warmth of the Sun enter but not radiate back to space. The wise man has got now the chance to prove his wisdom adapting his behaviour appropriately.

Scientific research must have the goal of augmenting collective wisdom. Climate research is needed to increase our knowledge of the properties of that thin layer around Earth where life can exist. But the term *wisdom* is appropriate only if knowledge turns into action, with a full awareness of responsibility.

The present issue of *Spatium* is devoted to a general overview of the Earth's climate system, with special emphasis on the role of the Sun, because it is by far the most important forcing factor of the climate system. Since the Industrial Revolution some 150 years ago, however, the anthropogenic influence can no longer be neglected. Climate research provides the knowledge required for mankind to act in awareness of responsibility towards its future generations.

We are very grateful to Dr. Jürg Beer, head of the Radioactive Tracers Group at the Swiss Federal Institute for Environmental Science and Technology in Dübendorf (ZH) for his kind permission to publish herewith a summary of his fascinating lecture on climate research given at our association's meeting in late autumn 2000.

Hansjörg Schlaepfer
Bern, November 2001

Impressum

SPATIUM
Published by the
Association Pro ISSI
twice a year

INTERNATIONAL SPACE SCIENCE INSTITUTE

Association Pro ISSI
Hallerstrasse 6, CH-3012 Bern
Phone ++41 31 631 48 96
Fax ++41 31 631 48 97

President
Prof. Hermann Debrunner,
University of Bern
Publisher
Dr. Hansjörg Schlaepfer,
legenda schläpfer wort & bild,
Winkel
Layout
Marcel Künzi, marketing · kommunikation, CH-8483 Kollbrunn
Printing
Druckerei Peter + Co dpc
CH-8037 Zurich

Front Cover
Sun-set over the Thyrrenian Sea.
(Photo: H. Schlaepfer)

Sun and Climate *)

Jürg Beer, Swiss Federal Institute for Environmental Science and Technology (EAWAG)

Introduction: Setting the Scene

The Sun is by far the most important driving force of the Earth's climate system. However, only little is known how variably this force acts on time scales ranging from minutes to millennia and how the climate system reacts to changes in this forcing.

In the core of the Sun nuclear fusion processes generate energy, which is then transported towards the solar surface. The total solar radiation that arrives at the top of the Earth's atmosphere depends on several factors: the energy production in the core of the Sun, the energy transport through the Sun's radiative and convective zone, the emission of radiation from the photosphere and the distance between Sun and Earth.

Satellite-based measurements over the past two decades have revealed a clear correlation between the solar irradiance and the 11-year sunspot cycle. The observed changes in total solar irradiance between solar maximum and solar minimum of an 11-year cycle amount to about 0.1% (annual average). This is too small a change to significantly affect the climate. However, there is growing evidence that, on longer time scales, solar variability is much more pronounced and the climate system is much more sensitive to solar forcing due to feedback mechanisms.

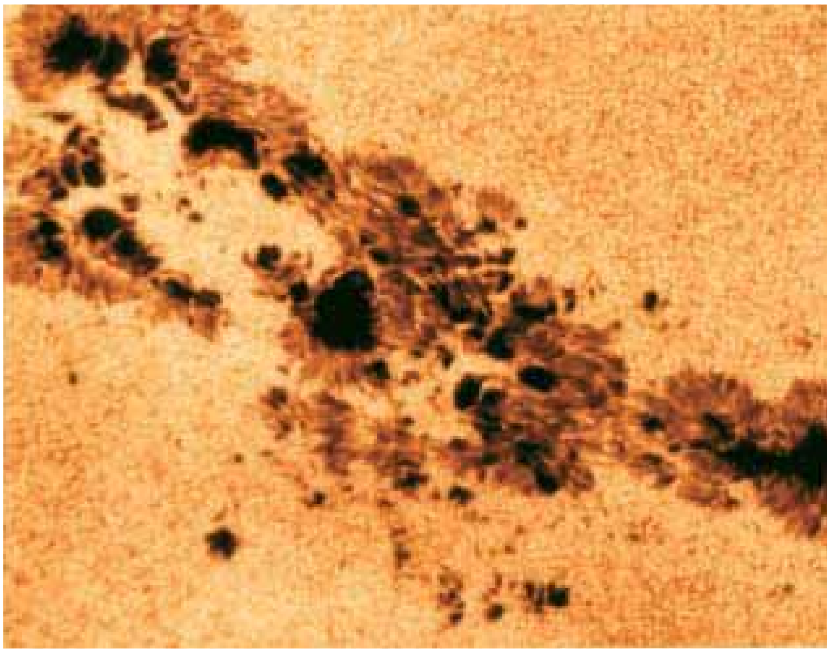


Figure 1
Sunspots represent areas where dense magnetic flux tubes cross the solar surface. As a consequence of the intense magnetic fields, the local convective heat transport is reduced and the surface temperature lowered by 1,000 to 2,000 K. In spite of the still very high temperatures of ca. 4,000 K, a sunspot looks dark compared to the temperature of 6,000 K of the surrounding area. The number of spots waxes and wanes with a periodicity of about 11 years (Schwabe cycle).

The response of the climate system to changes in solar forcing depends not only on the intensity of the radiation, but also on its spectral composition, seasonal distribution over the globe and on feedback mechanisms connected with clouds, atmospheric water vapour, ice cover, atmospheric and oceanic transport of heat and other terrestrial processes. It is therefore impossible to establish a simple quantitative relationship between reconstructed climate changes in the past and reconstructed solar variability. There is, however, growing evidence that periods of low solar activity coincide with the advance of glaciers, changes in lake levels and other significant environmental changes. These findings indicate that the Sun played an active role in past climate changes in concert with other geophysical climate forcing factors such as volcanic eruptions, greenhouse gases and internal variability of the climate system.

To study the climate in the past one has to rely on archives. The last few hundred years are quite well documented. The sunspots recorded by many observers tell us about solar activity and meteorological stations provide information about local air temperatures.

*) Pro ISSI lecture, Bern, October 31st, 2000

Unfortunately, these stations are not equally distributed around the globe. This instrumental time span, however, is too short to document slow climate changes and does not allow us to study the underlying long-term processes. We need archives that record climate changes over much longer periods of time. Glaciers and polar ice caps are especially promising types of archives which were formed from snowfalls over very long periods of time (up to several hundred thousand years). Snow scavenges a variety of solid and liquid constituents from the atmosphere, which are subsequently stored in the ice in chronological order and can later be retrieved and analysed by use of sophisticated techniques.

The present issue of Spatium is devoted to the topics of solar forcing, the Earth's climate and to some aspects of how information is retrieved from ice cores. However, a complete overview on all aspects of climate change would by far exceed the scope of this small booklet. The interested reader is referred to the following sources:

- PAGES (Past Global Changes) Programme, jointly sponsored by the United States and the Swiss National Science Foundation (http://www.pages-igbp.org.)

- R. von Steiger, Das neue Bild der Sonne (1998), Spatium Nr. 2.

- Hoyt, D. V. and K. H. Schatten (1997): The Role of the Sun in Climate Change. New York, Oxford University Press.

The Sun: Our Daytime Star

On the one hand, our Sun is simply an ordinary star, one of more than 100 billion in our galaxy. On the other hand, it is a very special star for us because it is by far the closest (8 light minutes compared to 4 light years). In comparison to the planet Earth, several of the Sun's properties are quite impressive. Its diameter of 1,390,000 km is more than 100 times the Earth's diameter, its mass equals more than 300,000 times the Earth's mass and its surface temperature equals 5,800 K. The conditions in the Sun's core (the inner 25% of its radius) are extreme compared to the conditions on Earth. The temperature is 16 million Kelvin and the pressure is 250 billion atmospheres. These are the conditions required to maintain the present energy production, which makes the Sun shine.

The Sun is by far the largest object in the solar system. It contains more than 99.8% of the total mass of the Solar System, and Jupiter contains most of the rest. At present, the Sun consists of about 92.1% hydrogen and 7.8% helium by number of atoms; other heavier elements such as metals only make up for 0.1%. This composition changes slowly over time as the Sun converts hydrogen to helium in a giant nuclear power plant in its core. Every second, about 700 million tons of hydrogen are converted to about 696 million tons of helium. This means that 4.3 million tons of mass are converted into energy every second, generating a power of $3.86 \cdot 10^{26}$ Watt. This power travels from the core towards the surface of the Sun in the form of radiation (radiative zone). On its way it is continuously absorbed and re-emitted at increasingly lower temperatures. Having reached about $^2/_3$ of the solar radius the gas is too cool and hence too opaque for radiation to pass, so convection sets in. In the adjacent convection zone, energy is transported by hot material rising to the surface and cool material sinking to the bottom, forming huge convection currents.

Overall, it is a long way. It takes the radiation several million years to work its way out from the core to the solar surface (photosphere). From the surface, the power is radiated into space. If this total power reached the Earth it would take only 10 seconds to evaporate all the water stored in the oceans and the ice caps on Earth! In reality, only about 1 part per billion reaches the Earth. All the same, the amount of solar energy that arrives on the surface of the Earth every hour is greater than the total amount of energy that the world's population consumes in a whole year. The radiation emitted is primarily visible light. This is no coincidence: during the hundreds of millions of years of the evolution the spectral sensitivity of our eyes has been perfectly matched to that part of the sunlight's spectrum which contains most of the energy.

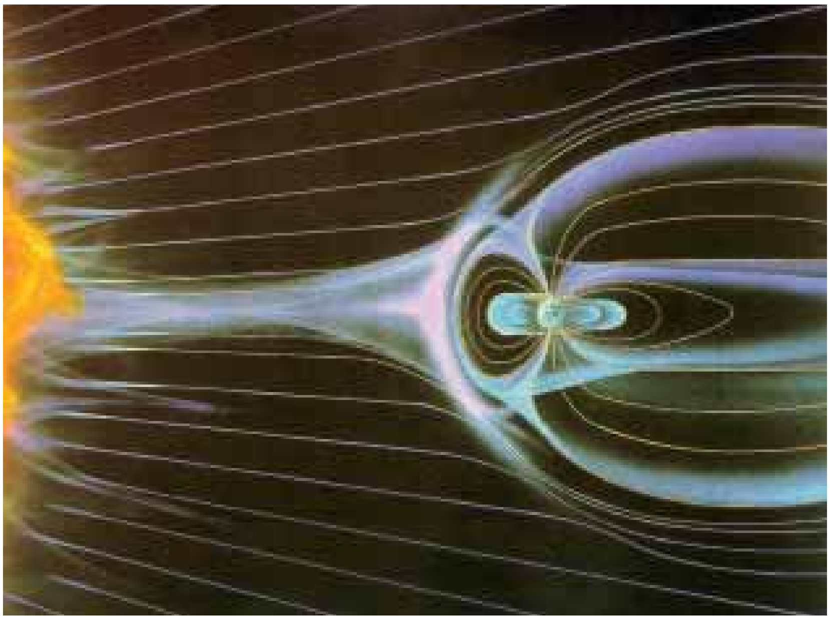


Figure 2
In this cartoon the Sun's atmosphere continually expands from the left into interplanetary space. The flow of ionised gas, called solar wind, carries frozen-in magnetic fields from the Sun into space. When sweeping by the Earth, it deforms the geomagnetic field. Explosive eruptions lead to sudden "space storms" which can seriously affect the Earth. The solar wind fills a space larger than 100 times the distance from Sun to Earth. The magnetic field deflects the charged cosmic ray particles. Therefore, the cosmic ray flux is reduced considerably within the solar system during periods of high solar activity. (Credit: ESA).

The radiating surface of the Sun, called photosphere, is a very thin layer of only a few hundred kilometres with a temperature of about 5,800 K. In a more or less regular rhythm of 11 years, the golden face of the Sun is disgraced by dark spots: the famous sunspots **(Figure 1)**. Sunspots are caused by dense magnetic flux tubes that cross the solar surface and hinder the convective heat transport to the surface. Therefore, sunspots are "cool" regions, with temperatures of only 4,000 K. They look dark in comparison with the surrounding areas. Sunspots can grow very large, up to 100,000 km in diameter.

The outermost region of the solar atmosphere, the corona, extends into space for millions of kilometres. However, the beautiful crown of the Sun is visible to the naked eye of the observer only during an eclipse.

The Sun's magnetic field is very strong compared to the geomagnetic field and also very complicated. The magnetosphere extends well beyond the outermost planet Pluto and plays an important role in shielding the Earth's surface from high-energy cosmic rays, which, otherwise, would endanger life on Earth **(Figure 2)**.

Apart from heat and light, the Sun also emits a low-density stream of charged particles (mostly electrons and protons) called solar wind, which spreads throughout the solar system with a speed of several hundred kilometres per second. The solar wind and the much higher energy particles ejected by solar flares can have dramatic effects on the Earth ranging from power line surges to the beautiful aurora phenomenon.

The Sun has a life cycle like a living creature. It was born out of cosmic dust about 4.5 billion years ago. Since then it has used up about half of the hydrogen in its core. It will continue to burn hydrogen for another 5 billion years or so (although its luminosity will approximately double in that time, see **Figure 3**). Eventually, it will run out of hydrogen fuel in the core and its life cycle will come to an end. This will set off a number of dramatic changes: The Sun will turn into a red giant by increasing its diameter beyond the Earth's orbit, swallowing all the inner planets and then creating a planetary nebula by throwing out part of its gaseous shells. This will be the termination of all life on Earth. Finally, it will collapse to a white dwarf and its light will extinguish.

Besides these large but very slow and steady changes over billions of years, the Sun also exhibits variability on much shorter time scales. After a period of large variability in its young age the Sun calmed down and became a relatively stable star, an important condition for the evolution of life on Earth. Nevertheless, the Sun is a variable star. Today, on time scales of millennia and shorter, fluctuations in the energy production in the core are extremely small (10^{-8}), and very small (10^{-6}) in the energy transport through the radiative zone (1⁄3 to 2⁄3 of the Sun's radius). In the convection zone, however, turbulence gives rise to fluctuations of up to percents on various time scales. Finally, the emission of radiation from the solar surface is given by the temperature of the photosphere, which is not homogeneous. Active magnetic regions (faculae) are warmer and sunspots are colder than the surrounding solar surface. On average, the faculae overcompensate the cooling effect of the sunspots leading to a slightly brighter Sun during periods of high activity. Periods of higher solar activities with many sunspots are followed by minima in an 11-year cycle. In addition, there are other variations with periodicities of approximately 90, 200 and possibly 2,000 years.

Sometimes, the Sun relaxes and becomes quiet for several decades or even a few centuries. The famous Maunder Minimum (1645–1715 AD) was such an event, a period during which hardly any sunspots were detected. It is interesting to note that the Maunder Minimum falls into a period characterised by generally lower temperatures and glacial advances known as the Little Ice Age.

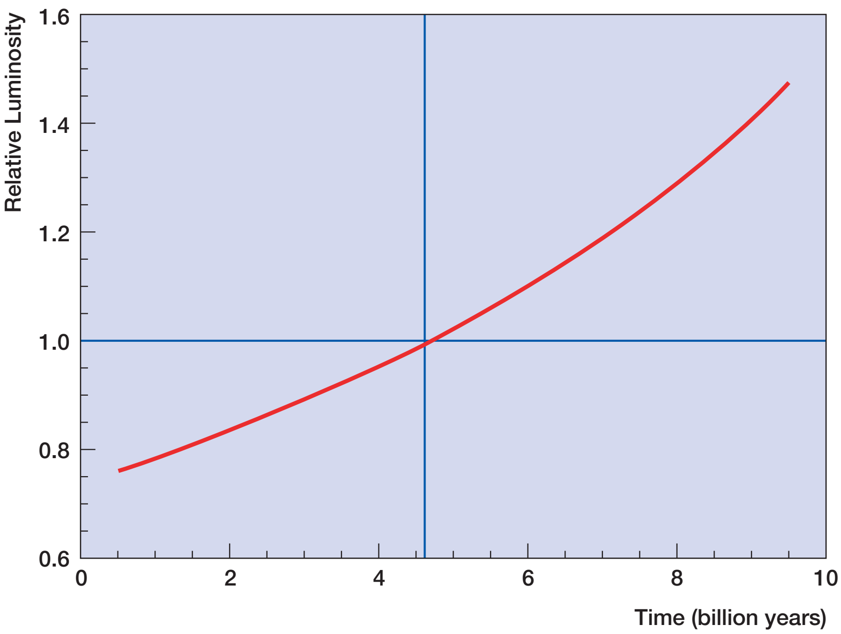


Figure 3
The standard solar model reveals that the solar luminosity (total solar energetic output) has continuously increased since the formation of the solar system 4.55 billion years ago and will continue to do so for another 5 billion years until most of the hydrogen in the core is burned to helium. This change is too slow to be relevant on time-scales of millennia. (Credit: W. Mende).

The Climate System: The Distributor and Equalizer of Energy on Earth

Following the laws of thermodynamics, the Earth's climate system seeks to redistribute the energy input from space equally: On top of the atmosphere, every square meter perpendicular to the axis Sun–Earth receives approx. 1,365 W of power from the Sun. This value is called the solar constant by tradition although we know that it is not constant over time. There are other sources of energy on Earth such as the tidal forces related to the Moon orbiting around the Earth or geothermal heat resulting from radioactive decay processes in the Earth's interior. However, their contribution is negligible in comparison to the solar input.

A very thin layer of only approximately 1⁄1,000 of the Earth's diameter forms the stage on which the climate system acts. Basically, it consists of the atmosphere, the hydrosphere, the biosphere and the continents. For simple geometrical reasons the mean annual radiative input of the Sun is strongest in the equatorial regions, where the Sun is near the zenith and becomes lower at higher latitudes. This unequally distributed energy input leads to thermal gradients, which the climate system tries to level out by transporting energy from lower to higher latitudes. The energy is transported by wind, ocean circulation and water vapour. Instead of following a direct path to the poles, the transport follows a rather complex regime. This is due to several reasons: firstly, warm air masses are less dense and move upwards forming convection cells. Secondly the distribution of land and water and the physical topography affect the energy flow. Thirdly, the Coriolis force deflects air and water masses moving towards the poles to the right on the northern hemisphere and to the left on the southern hemisphere.

The first redistributor is the lower part of the *atmosphere,* the troposphere, where large wind systems are generated as a result of the different factors influencing the air masses moving north or south. The trade winds in the tropical zones, which played a key role at the time of the first *conquistadores* are probably the best known of these. The wind systems are the domain of meteorology dealing with cyclones and anticyclones, with air humidity and temperature eventually causing clouds and precipitation. The short term behaviour of all these quantities determines the daily weather. The long-term development of weather patterns defines the climate.

The second redistributor is the *hydrosphere,* which consists of the water in oceans, rivers and lakes, the ice in glaciers and polar ice caps, the vapour in the air, the droplets in clouds and the ground water. The thermal gradients induced by solar insolation cause the water to flow. Warm tropical waters flow pole-wards over the cooler deep-sea water. Evaporation increases the salinity of the water and thus its density. Cooling further increases the density of the water until it begins to sink to the bottom of the ocean from where it flows back, closing the loop. These processes generate the great streams on Earth such as the Gulf Stream in the northern Atlantic or the Humboldt Stream west of South America. The amount of water involved in the circulation system of the North Atlantic, which is an important part of the global oceanic conveyer belt, is equal to about 100 Amazon Rivers. Interruption of this conveyer belt would change the climate conditions in northern Europe dramatically. London would experience winters like Irkutsk in Siberia.

The third redistributor is the radiation, which is reflected, absorbed and emitted by the Earth's surface, by greenhouse gases and clouds in the atmosphere.

An astronaut travelling through the solar system will inevitably have his attention drawn to the blue planet, the Earth, which looks completely different from all the other planets. The main reason for this special appearance are the oceans and the clouds. The *biosphere* also plays an important role. In spite of the fact that it uses only a very small part of the solar energy input, the biosphere has e.g. produced all the oxygen present in the atmosphere. All the fossil fuel we are combusting within a few centuries is solar energy which was accumulated by the biosphere over millions of years. Presently, mankind lives primarily on solar energy delivered to Earth in the past.

Although, at first glance, the task of redistributing energy seems rather simple, this is not the case. In fact, the climate system is extremely complex. Everybody knows that forecasting the weather for more than a day is a challenging task. Even professional meteorologists who can rely on a dense global network of weather stations, satellites and most sophisticated computer models are unable to make reliable predictions for more than a few days. The reason for this is that the weather conditions are the result of a very large number of processes and their complex interaction on various time and length scales. Because the behaviour of the weather system is to some part chaotic even the most powerful computers will not be able to solve this problem in the future.

In the case of climate change, which we will discuss later on, the situation is less chaotic, but still very complex. This is illustrated by the fact that the best answer at present to the question of the response of the climate system to a doubling of the atmospheric CO_2 concentration is not a precise figure but a temperature range (2–5 °C). This as yet unsatisfactory answer can only be improved by a

Figure 4
The glacier near Grindelwald photographed in 1858 and 1974. (Credit: F. Martens, H.J. Zumbühl in Hydrologischer Atlas der Schweiz).

better understanding of the climate system. The clue to this lies in the past. Careful study of the climate changes in the past will allow us to identify the key processes and determine the response to changes in the forcing. Nature is generous enough to provide us with a variety of archives, which contain all the necessary information to reconstruct the history of the Earth's climate and the forcing. However, reading and deciphering this information is not always an easy task.

Climate Archives: The Diaries of Nature

For the last several centuries, reconstruction of the climate history is comparatively simple. Beside direct physical measurements, there is a wealth of information in written historical documents. Even paintings can be a useful source of climate information. Paintings and more recently, photographs of a glacier from the same site document very nicely how its size changes with time and how dramatic the present global warming is **(Figure 4)**. Another example is depicted in **Figure 5**, showing a market on the frozen River Thames in the winter 1813/1814. It is reported that such markets and other events took place on the Thames during winter quite regularly as from the 15th century, but never again after 1814.

For earlier times, ice itself serves as an archive. Glaciers in high mountains and ice sheets in the

Figure 5
Market on the River Thames in London in the year 1813/1814. Between 1400 and 1814 such markets were quite common. In the winter of 1683/1684 the ice reached a thickness of 26 cm and the river was completely frozen for 2 months.
Painting by Luke Clement. (Credit: Wetterbuch, Christian Verlag, 1982, ISBN 3-88472-080-5).

Figure 6
The extent of ice cover on Mt. Kilimanjaro (5,895 m) decreased by 81% between 1912 and 2000. In 1889, when Hans Meyer first climbed Kibo, the crater rim was nearly encircled by ice. Today only a small fraction of this ice remains. Clearly visible is the banding of the ice sheet associated with the annual cycles. In the background Mt. Meru. (Credit: Andri Schlaepfer).

polar zones store climate information for centuries to hundreds of millennia. They truthfully record the average climate conditions, levelling out the capricious moods of the weather. One example of such an archive is the glacier on Mt. Kilimanjaro **(Figure 6)**, which is of special interest since it stores climate information on the tropical zone. Unfortunately, this irreplaceable archive is melting away quickly due to the increasing mean temperature.

Precipitation in form of snow scavenges various constituents of the atmosphere and stores them chronologically, layer by layer. Increasing pressure from additional layers causes the snowflakes to compact and become ice. The consequence of this formation process is that the ice not only preserves all the atmospheric constituents such as aerosols and dust, it also contains air bubbles that enable scientists to determine the atmospheric composition and in particular the concentration of greenhouse gases in the past. This unique property makes ice the only archive that virtually stores all the climate forcing factors such as greenhouse gases, aerosols and volcanic dust and cosmogenic isotopes reflecting solar variability. Ice cores also contain information on the corresponding climate responses: temperature, precipitation rate, wind speed and atmospheric circulation. **(Figures 7** and **8)**.

Reading and deciphering the information nature stored in archives is, however, not a simple task. For example, the temperature signal is stored in the form of the oxygen isotope ratio $^{18}O/^{16}O$ in ice. The corresponding temperature can only be determined if this ratio can be measured with high precision and if the relationship between the $^{18}O/^{16}O$ ratio and the temperature is known. Solar variability, on the other hand, is recorded in the ice through solar wind induced modulation of the cosmic ray flux. Some high energetic cosmic ray particles travelling through our galaxy penetrate into the Earth's atmosphere where most of them collide with nitrogen or oxygen atoms. If these atoms are hit by fast travelling

Figure 7
View of a typical drill-site for shallow ice-cores up to 300 m length in Greenland. The building in the background is the former American radar station of Dye 3.

cosmic ray particles they can break up into smaller atoms, so called cosmogenic radioisotopes such as ^{10}Be. Opposite to stable isotopes, radioisotopes disappear by radioactive decay and are therefore very rare. Within a few half-lives they reach an equilibrium between production and decay. On Earth there is a total amount of only about 100 tons of ^{10}Be.

Isotopes are atoms which are chemically identical but differ in mass due to a different number of neutrons in the nucleus.

During periods of high solar activity there is a stronger solar wind. Magnetic fields carried by the solar wind deflect the cosmic ray particles more efficiently and hence reduce the production rate of cosmogenic isotopes in the atmosphere. These isotopes are removed from the atmosphere mainly by rain and snow and some of them are stored in ice caps and glaciers preserving information on the prevailing space weather conditions. Therefore, if we measure a reduced ^{10}Be concentration in an ice core we can conclude that less ^{10}Be was produced by cosmic rays. Such lower production rate indicates a stronger shielding of the cosmic rays induced by more intense solar wind, which, in itself, is the result of increased solar activity.

This relationship between the concentration of the Beryllium isotope ^{10}Be in a Greenland ice core and the sunspot number as a proxy of solar activity can be seen in **Figure 9**. More specifically, the analysis of the cosmogenic isotopes ^{10}Be and ^{36}Cl in ice and ^{14}C

Figure 8
Deep-drilling (2,000 to 3,000 metres) is performed underground and takes about 2 to 3 years (Summer seasons). The drill is lowered into the borehole, drills a core of about 2,5 m length and is then brought up to the surface again. The picture shows how the ice-core is taken out of the drill. The liquid is used to prevent the borehole from closing.

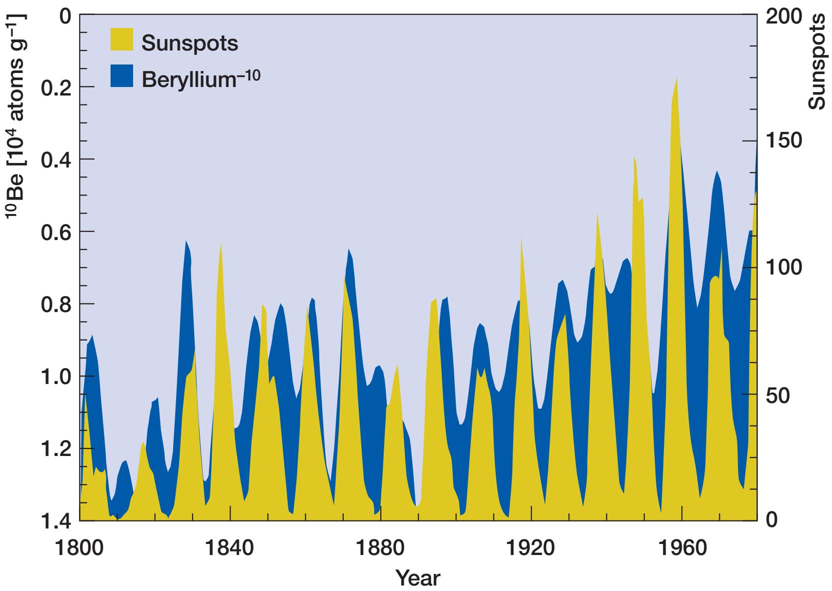


Figure 9
Comparison between the observed mean annual number of sunspots with the concentration of ^{10}Be atoms in the ice-core from Dye 3 (Figure 7). A large number of sunspots correspond to higher solar activity, more solar wind, fewer cosmic rays, lower production rates of ^{10}Be and therefore a lower ^{10}Be concentration. Both records show a clear 11-year Schwabe cycle and a long-term trend interrupted by short periods of reduced solar activity around 1815 and 1900.

in tree rings reveals not only that solar variability is partly cyclic with periodicities around 11, 90, 200 years and possibly longer ones, but also that the beginning of extended periods of low solar activity, so called grand minima, often coincide with rapid climate changes. This points to a relationship between solar activity and climate.

Combining the information from ice with additional information from other archives such as lake and sea sediments, loess deposits, peat bogs, tree rings, corals, speleothems and others provides a much more complete picture of how the climate system behaved in the past and which were the driving forces.

Accelerator Mass Spectrometry: Unravelling the Secrets of Cosmogenic Isotopes in Ice Cores

Since the late seventies, there has been great progress in the analytical techniques used to read the information stored in archives. An increasing number of additional parameters can be measured and the analytical sensitivity has improved significantly. This is especially true in the case of cosmogenic isotopes, which can now be measured by so-called Accelerator Mass Spectrometers (AMS). The Swiss Federal Institute of Technology of Zurich (ETHZ) was among the first laboratories in Europe to operate such an instrument.

Mass spectrometers are complex technical systems designed to separate and measure atoms with different masses. In a conventional mass spectrometer the material is ionised and then accelerated by a high voltage. The accelerated ions are formed to a beam by electrical and/or magnetic lenses. A magnet separates the ions into different beams depending on their masses. In the case of cosmogenic isotopes (^{14}C, ^{10}Be, ^{36}Cl) we are faced with extremely small numbers of atoms, which even further decrease with time due to radioactive decay. For example, the $^{14}C/^{12}C$ ratio in our body is about $12 \cdot 10^{-13}$, in the body of the ice man it is about $6 \cdot 10^{-13}$ and in the body of a Neanderthal man it is as low as 10^{-15}. Measuring such small ratios exceeds the capability of conventional mass spectrometers due to background problems caused by isobars (^{14}N in the case of ^{14}C) and molecules (e.g. ^{13}CH). A great step forward was the development of the Accelerator Mass Spectrometer (see **Figure 10**), which works along the same principle as a conventional mass spectrometer, but accelerates the ions to approximately 1,000 times higher energies. At these energies the molecules are destroyed and every single ion reaching the detector can be identified and counted individually. The power of this analytical technique is best illustrated by a simple comparison. The task of determining the ^{10}Be concentration in 1 kg of ice is comparable to measuring the ink concentration in 1 litre of water from lake Constance after putting a small drop (3 microliters) of ink into the lake and allowing it to mix completely!

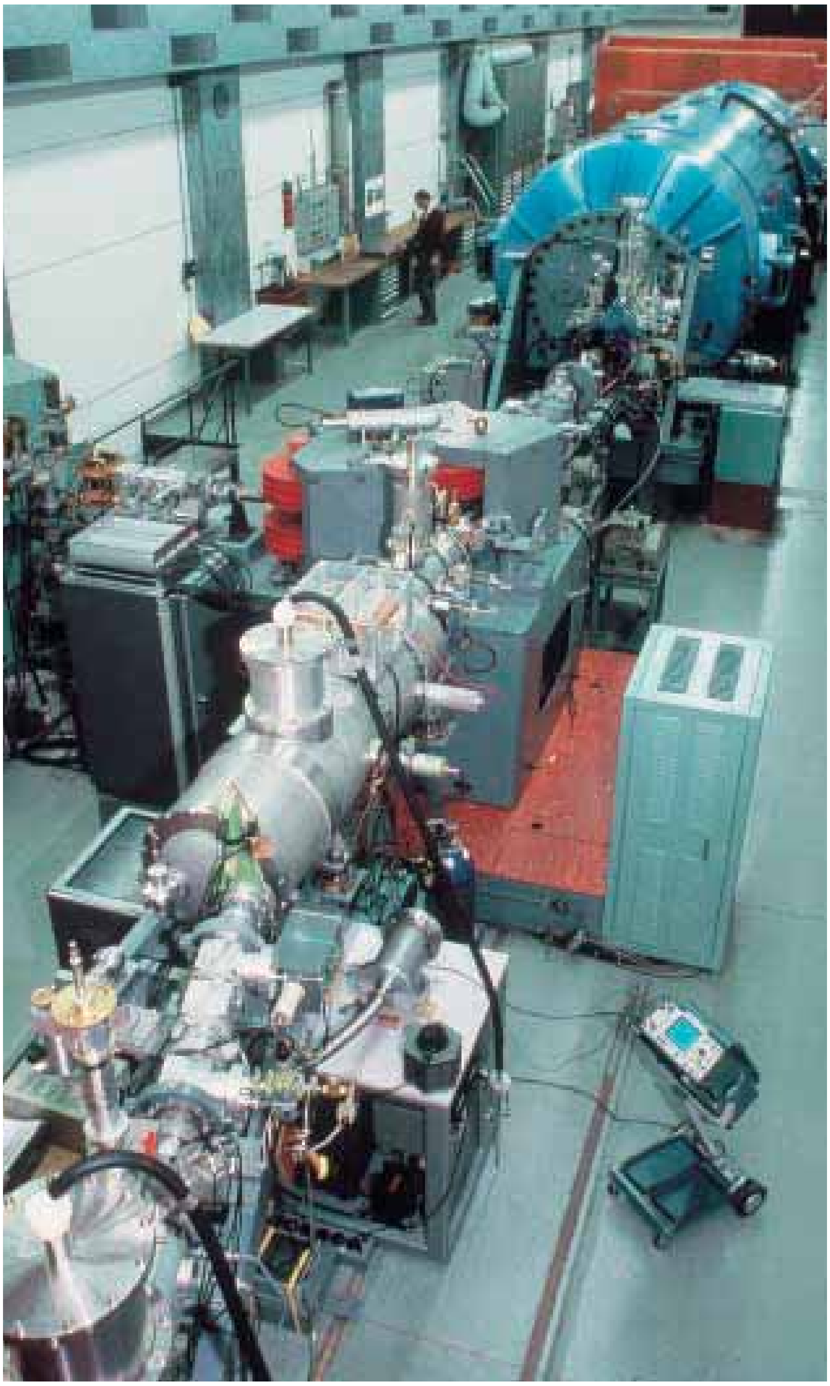


Figure 10
View of the accelerator mass spectrometer of ETH/PSI, which is used to measure the ^{10}Be concentration in ice. The central part of this instrument is the tandem accelerator (blue tank in upper right corner), which accelerates the Beryllium ions to very high energies (20 MeV). This is necessary to analyse ice samples of 1 kg containing as few as 10 million ^{10}Be atoms.

Climate Change: The Story of Perpetual Change

Newspapers and scientific journals tell us the story of the present climate change, the global warming. Ice cores and other archives tell us the story of the climate in the past. It is a story of perpetual change:

20,000 years ago: The world looked completely different. Sea levels were lower by some 100 m. Large parts of Switzerland were covered by ice. Glaciers carved the landscape and transported huge rocks over hundreds of kilometres found today as erratic blocks.

5,000 years ago: A young man tried to cross the mountains from Italy to Austria. He died at an altitude of about 3,200 m and was buried under snow and ice for more than 5000 years. In the year 1991, his tomb of ice was opened due to global warming and his body was discovered by hikers.

2,000 years ago: The Romans expanded their empire and conquered part of England. Thanks to favourable climatic conditions they were even able to produce wine on the British isle.

1000–1300: The Vikings took advantage of the favourable climate conditions to explore the sea in the north and west. On a big island in the north, Erik the Red

settled down with a group of people and lived well on agriculture and keeping cattle. They called this island Greenland.
Gothic churches were constructed all over Europe. Culture flourished on a global scale.

1400–1850: The weather became worse. Sea ice severed the connection to Greenland. The cattle died. Greenland turned from a green to an icy island.
In Switzerland, the glaciers advanced considerably. They covered areas where there had been forest **(Figure 4)**. Since then, they are generally retreating.

Today: Mankind is concerned about the ongoing rapid global warming.

These are just a few examples to illustrate that the climate system never reached stable conditions for long periods of time. In other words: the climate has always been variable. What are the causes of this variability? We can distinguish between external and internal factors with respect to the planet Earth.

Potential External Factors: Forcing the Climate

Since the Sun is the dominant source of energy on Earth it is also the main candidate for introducing variability in the climate system. First, we will consider sources of variability of the Sun itself. Then we will discuss how the Earth's orbit around the Sun is perturbed by the gravitational forces of the other planets and how this affects insolation.

Climate Change and Solar Forcing: The Solar Connection

On time scales of solar evolution, reliable models show that, compared to today, solar luminosity was lower by approximately 30% shortly after the formation of the solar system and has increased slowly but steadily over the past 4.5 billion years **(Figure 3)**. As mentioned earlier, there may be fluctuations in the transport of heat from the core of the Sun towards its surface, in particular in the convection zone where, like in boiling water, hot gas bubbles rise quickly to higher levels while cooler gas sinks down to be reheated. This turbulent boiling creates a characteristic pattern of granulation on the surface. Fluctuations in the heat transport cause changes in the solar luminosity.

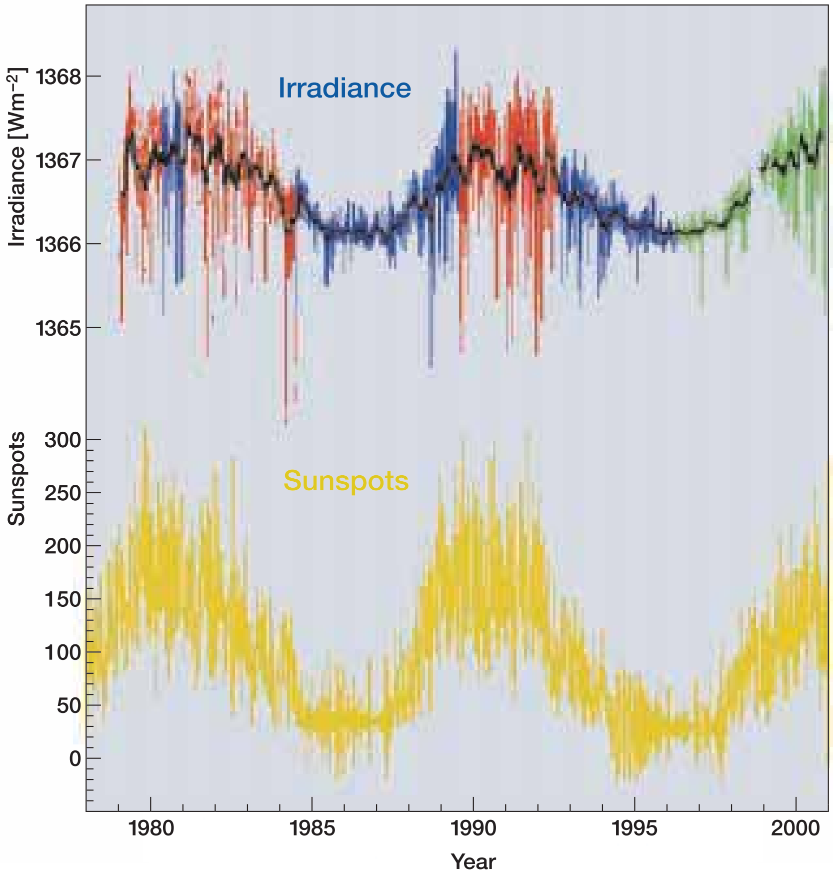


Figure 11
Comparison of the solar irradiance with the sunspot numbers for the last two Schwabe cycles. The irradiance record is a compilation of data from different satellites. During periods of high solar activity there are more sunspots, darkening a small part of the solar disk (visible in the negative excursions of the irradiance). However, the brightness of the Sun is increased at the same time, overcompensating the darkening effect of the sunspots. (Irradiance data: credit: C. Fröhlich, PMOD).

On very short time scales, information on the inconstancy of solar forcing comes from different sources:

Since 1978, direct measurements of the total solar irradiance are made by satellites equipped with radiometers. They show a relatively small amplitude of total solar irradiance of about 0.1% within one 11-y cycle **(Figure 11)**. Unfortunately, at present, these measurements cannot tell us how much the solar constant changes over decades to centuries. However, from the climate point of view, centuries and millennia are the more important time scales.

Luminosity: Total power emitted by the Sun.

Irradiance: Total power arriving at the top of the atmosphere in the distance of 1 astronomical unit (average distance between Sun and Earth)

Since, for obvious reasons, nobody wants to wait for decades to millennia to learn more about the long-term behaviour of the Sun, a complementary project was initiated to simultaneously study a large number of sun-like stars (stars with properties very similar to the Sun). A monitoring program of this kind began 3 decades ago. The results obtained to date from sun-like stars indicate a potential for much larger solar variations than those observed by satellites over the short time period of 20 years.

Theoretical considerations suggest that, as a general rule, slow changes in solar variability tend to have larger amplitudes than fast changes. Variability on time scales up to 20,000 years seems possible. Moreover, the climate system generally reacts more sensitively to slow changes than to fast changes. Building up an ice sheet, for example, takes a long time.

The so-called Little Ice Age (approx. 1400–1850) is characterised by generally lower solar activity compared to today. During this time there were several periods with very low activity. During the Maunder Minimum (1645–1715) hardly any sunspots were observed in spite of the continuous efforts of several researchers directing their newly invented telescopes towards the Sun. Since approx. 1700, the solar activity has steadily increased and was only interrupted by some weak minima around 1815 and 1890 **(Figure 9)**. From **Figure 11** we know that an increase in solar activity results in an increase in solar irradiance. If this relationship, which was only established for the last two decades, also holds true for longer periods of time, we have a tool in hand to estimate solar irradiance in the past. Using sunspot numbers or ^{10}Be data from ice cores we can reconstruct the history of solar activity, which in turn reflects the solar irradiance. This postulated relationship can be tested by comparing the solar forcing signal (^{10}Be) with the climate response (temperature). In fact, **Figure 12** shows that high ^{10}Be concentrations (low solar activity) in an ice core from Dye 3, Greenland, **(Figure 7)** coincide with cold temperatures (reduced irradiance). It has been estimated that the solar constant was reduced during the Maunder Minimum by about 2.5‰ compared to today. This could account for most of the observed cooling. ^{14}C data from tree rings and ^{10}Be data from ice cores show that during the past 10,000 years a number of such minima in solar activity occurred. Many of them can be related to climate deteriorations, which suggests a solar connection. A good example is the market on River Thames mentioned earlier, which took place in the winter 1813/1814, precisely at the onset of such a solar minimum **(Figure 12)**.

When it comes to the question of the mechanisms responsible for the observed climate changes, no final answers are yet available. Beside the most important changes in the total irradiance, which induce a variety of feedback and amplification processes there are other mechanisms which may also play an important role. For example, the relative change of ultraviolet radiation during a solar cycle is much larger than the change of total radiation. Since the ultraviolet component of the spectrum is responsible for the production of ozone in the stratosphere, considerable effects on the stratospheric chemistry and dynamics can be expected.

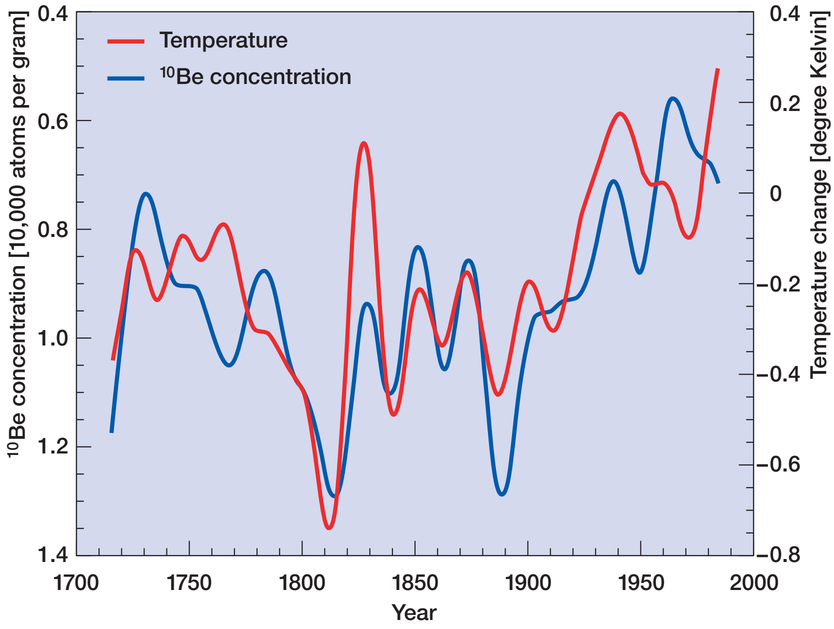


Figure 12
Comparison of the ^{10}Be concentration measured in an ice core from Dye 3 (Figure 7) reflecting the solar activity with a reconstruction of the northern hemispheric temperature. Both records were filtered to only show changes on time scales longer than 20 years. Note the solar minima (high ^{10}Be concentrations) around 1815 and 1890 which coincide with cold periods (Figure 5) and the rapid temperature increase during the first half of the 20th century.

Model results point to changes in tropospheric winds and storm tracks resulting from realistic variations in stratospheric ozone and solar ultraviolet radiation.

Finally, several authors have postulated indirect ways of how the Sun can affect the climate. For example, the interaction of cosmic ray particles with the atmosphere induces strong electrical currents, which may to some extent affect climatically important processes such as the condensation of water vapour and the formation of clouds. Although very speculative and controversial, these examples illustrate that we are far from understanding all the processes relevant for climate changes.

Climate Change and Variations in the Earth's Orbit: The Astronomical Connection

The Earth travels once a year around the Sun in an orbit, which, although elliptical in shape, is actually almost circular. At present, the deviation between the elliptical orbit and a perfect circle (the so-called eccentricity) is small, and the resulting maximum difference in the Sun-Earth distance between July 4th, when the Earth is farthest from the Sun (aphelion, 151.2 million kilometres) and January 4th, when it is closest (perihelion, 146.2 million kilometres) amounts to only 3%. However, averaged over a full year, the total radiation received by the Earth travelling along its present elliptic orbit differs only slightly from what it would receive if its orbit were circular (0.05 W m^{-2}). The effect of the small change in distance is too small to cause the seasons. But if one compares the total amount of solar radiation received by one hemisphere during summer and winter, it becomes obvious that this small discrepancy does exert a remarkable influence on the severity of the seasons. At present, the Northern Hemisphere is experiencing slightly cooler summers and warmer winters, while in the Southern Hemisphere, seasonality is, in contrast, somewhat amplified.

The seasons are the result of the tilt of the Earth's axis (the obliquity) relative to its plane of travel around the Sun (23°). Summer prevails on the hemisphere facing towards the Sun, whereas on the hemisphere facing away from the Sun, the days are generally shorter and colder. If the tilt were zero there would be no seasons. A larger tilt angle, on the other hand, would cause more extreme seasonality, with warmer summers and colder winters.

But the situation described above is not stable. The eccentricity, obliquity and the date of the perihelion all change slowly with time. These secular changes in the Earth's orbit are caused by the gravitational forces of the other bodies in the solar system. The relative positions of Jupiter and Saturn, especially, affect the eccentricity and the obliquity of the Earth's orbit. The effects on the orbital parameters are quasi-periodic because they consist of several components.

Eccentricity: This parameter changes with periodicities of 100,000 and 400,000 years. This effect influences the smallest and largest Sun–Earth distances. The greater the eccentricity, the larger the hemispheric asymmetry. In **Figure 13** the eccentricity is plotted for the past 400,000 and the coming 100,000 years. It was at its maximum 200,000 years ago, and is now declining towards a 400,000-year minimum, which will be reached in about 30,000 years. The eccentricity is the only orbital effect, which affects the annual global average of incident solar radiation. This global effect, however, is very small (a maximum of 0.6 Wm^{-2} in the daily mean or 2.4 Wm^{-2} for the solar constant), and cannot therefore be responsible for the observed 100,000-year periodicity in many paleoclimate records.

Obliquity: Gravitational forces cause the obliquity of the Earth's orbit to oscillate between 22° and 25°, with a periodicity of 41,000 years. The larger the obliquity, the greater the seasonal temperature variation at high latitudes.

Precession: The Earth's axis does not point in a fixed direction. Like a spinning top, which starts to gyrate around the vertical axis if you give it a nudge, the Earth's axis spins around, describing a large circle in the sky. The 'Polar Star' to which the axis now points will therefore be quite distant from the pole in a few thousand years. The cause of this precession, as it is called, is the equatorial bulge of the Earth. This bulge is formed by the Earth's rotation, which changes its shape from a perfect sphere to a slightly flattened one. The gravitational attraction of the Moon, the Sun and the planets on the bulge acts as the "nudge" which causes the Earth to precess like a spinning top.

As a consequence of the precession of the Earth, the perihelion and aphelion do not always occur at the same date each year, but instead move slowly through the seasons with a periodicity of about 23,000 years. This means that 11,500 years from now the perihelion will occur on July 4th, and seasonal temperature variations in the Northern Hemisphere will be larger than at present (colder winters, warmer summers), because the Earth will then be closest to the Sun during the Northern Hemisphere summer.

$\delta^{18}O$

relative deviation of the $\frac{^{18}O}{^{16}O}$ ratio of a sample from the $\frac{^{18}O}{^{16}O}$ ratio measured in present ocean water.

The last two orbital effects do not affect the total amount of the solar radiation received by the Earth, they only affect its geographical distribution. The variation of the obliquity spreads the radiation symmetrically polewards, while the precession controls its asymmetric distribution over the hemispheres.

The $\delta^{18}O$ data measured in deep sea sediments show a very distinct 100,000 year cycle of glacial and interglacial periods in phase with the eccentricity **(Figure 13).** This seems to be in conflict with the weakness of the direct forcing. However, the importance of the eccentricity lies not in the change in the amount of the total radiation, but in its coupling with the precession. As matter of fact, the larger the eccentricity of the Earth's orbit, the greater the amplitude of the precessional variation. That means that the precessional amplitudes of the insolation at any given location on the Earth depend very much on the instantaneous degree of eccentricity. Hence, the eccentricity period is reflected in the amplitudes of the precessional effects (amplitude modulation of the precession by the eccentricity).

For example in Berne, the difference in insolation between July 4th and January 4th has changed considerably with time. 80,000 years ago it was 425 W m^{-2}. Then it decreased by 85 Wm^{-2} to a minimum of 340 Wm^{-2} 65,000 years ago. After having attained a maximum of 410 Wm^{-2} about 60,000 years ago, it then decreased by 55 Wm^{-2} to reach another minimum of 355 Wm^{-2} 23,000 years ago,

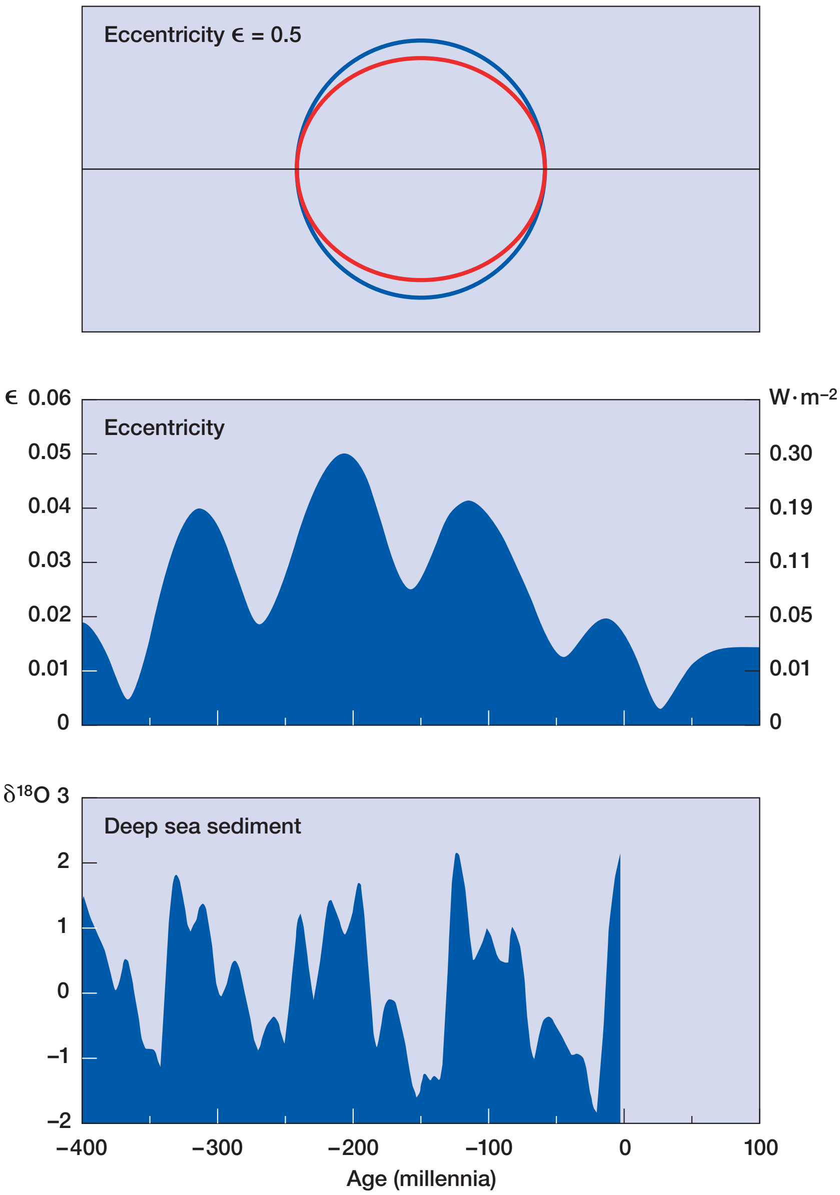


Figure 13
The analysis of the isotopic $^{16}O/^{18}O$ ratio in foraminifera (skeleton of plankton) stored in ocean sediments provides a means of reconstructing the sequence of glacial and interglacial periods (Bottom panel). Large values correspond to warm periods and vice versa. This sequence is dominated by a 100,000-year cycle but also shows shorter periodicities. It can be explained by the effects of the near-by planets on the orbit of the Earth (Milankovich theory). Small deviations from a circular orbit (eccentricity) (top panel) occur regularly with a dominant 100,000-year cycle. The middle panel shows the calculated changes of the eccentricity for the past 400,000 and the next 100,000 years. Note that the dramatic climate effects observed on Earth are due to very small deviations from the average annual forcing. (Lower panel: Credit: SPECMAP curve).

which coincides with the maximum of the last glaciation. This was followed by a rapid increase of 68 Wm^{-2} at the transition from the last Glacial to the Holocene, 11,000 years ago. Since then, the seasonal difference has been decreasing to its present value of about 365 Wm^{-2}. As already discussed, it will rise again to a new maximum within the next 11,500 years. However, for the next 50,000 years the seasonal difference will never again rise by more than 30 Wm^{-2} due to the declining eccentricity. Also, the two other orbital forcings reach a minimum in the next 30,000 years. This means that we must expect only small changes in the global climate due to orbital forcing during this time.

The uneven distribution of sea and land between the two hemispheres is important for the climatic response to changes in solar radiation. Therefore, a change in the hemispheric distribution of radiation can have a net effect on the global mean temperature. Processes, such as the melting and ablation of glaciers and the growth of plants, that are more sensitive to summer than to winter temperatures can accumulate the seasonal effects over thousands of years.

It was Milutin Milankovich who, 60 years ago, postulated that periodic variations in the Earth's orbital parameters affect the local insolation and could be the clue for some of the most prominent climate changes during the past million years.

Climate Change and Internal Forcing: The Terrestrial Connection

A great variety of internal factors affect the Earth's climate system and its reaction to the solar input. It is beyond the scope of this short overview to mention all of them. Instead, a limited selection of some important parameters are presented below:

For many years, climatologists have observed a connection between large explosive *volcanic eruptions* and short-term climatic change. For example, one of the coldest years in the last two centuries was the year following the Tambora volcanic eruption in 1815. Explosive volcanic eruptions have been shown to have a short-term cooling effect on the atmosphere if they eject large quantities of sulphur dioxide and fine-grained dust into the stratosphere forming an opaque compound with water vapour. The coarse-grained dust returns to the Earth's surface within a few months and hence has little impact on the climate.

Continents slowly move over time. Their drift not only changes the direction of the large ocean currents but also the position of Antarctica. This large continent collects enormous amounts of water in form of ice, thereby changing the sea levels on a global scale. This in itself influences the amount of water transported by the winds. Higher humidity in the air masses tends to increase the cloud cover and the Earth albedo, thereby contributing to an increased scatter of sunlight back into space and hence causing a reduction of atmospheric temperature.

Greenhouse gases such as CO_2 allow the visible sunlight to enter the atmosphere and be absorbed by the land and sea surface, but they hinder the thermal (infrared) radiation that is emitted by the Earth's surface from escaping into space. Increasing density of greenhouse gases cause the climate to become warmer. The tropical rain forests bind huge amounts of carbon to build up their biomass. When it is burnt in the process of deforestation this carbon is released in form of CO_2, which contributes to the greenhouse effect **(Figure 14)**. As a result of industrialisation, a large part of the solar energy which had been accumulated as biomass over millions of years is now released within a few centuries by combustion of oil, coal and gas.

"…whereas all experiences are of the past, all decisions are about the future… it is the great task of human knowledge to bridge this gap and find those patterns in the past which can be projected into the future as realistic images…"

K. Boulding 1973

Epilogue

4.5 billion years ago, when the solar system was formed, Earth and possibly also Mars were the only planets with an appropriate distance to the Sun to become habitable. Since then, the Earth's climate has varied all the time. In spite of this dramatic variability at times, it always remained within the boundaries of habitability and never turned into a hot-house like Venus nor into a snowball like Mars. Most of this variability is related to the main source of energy, the Sun. Even in ancient times, people intuitively recognised the importance of the Sun and worshipped it. Stable and favourable climate conditions brought wealth and prosperity, unstable and unfavourable conditions such as droughts and floods caused famines, forced people to migrate to other places and sometimes even caused wars. Understanding the climate and forecasting changes was always an important issue so that people who were able to make reliable predictions were highly respected. Today our knowledge of how the climate system works is much advanced compared to those times. Our observation systems and computing capabilities allow us to collect huge amounts of data and to model the complex processes in the climate system. However, we are still far from understanding it in detail. The most promising strategy to improve our understanding is to study past and present climate changes in parallel. Obviously, studying past changes is more difficult.

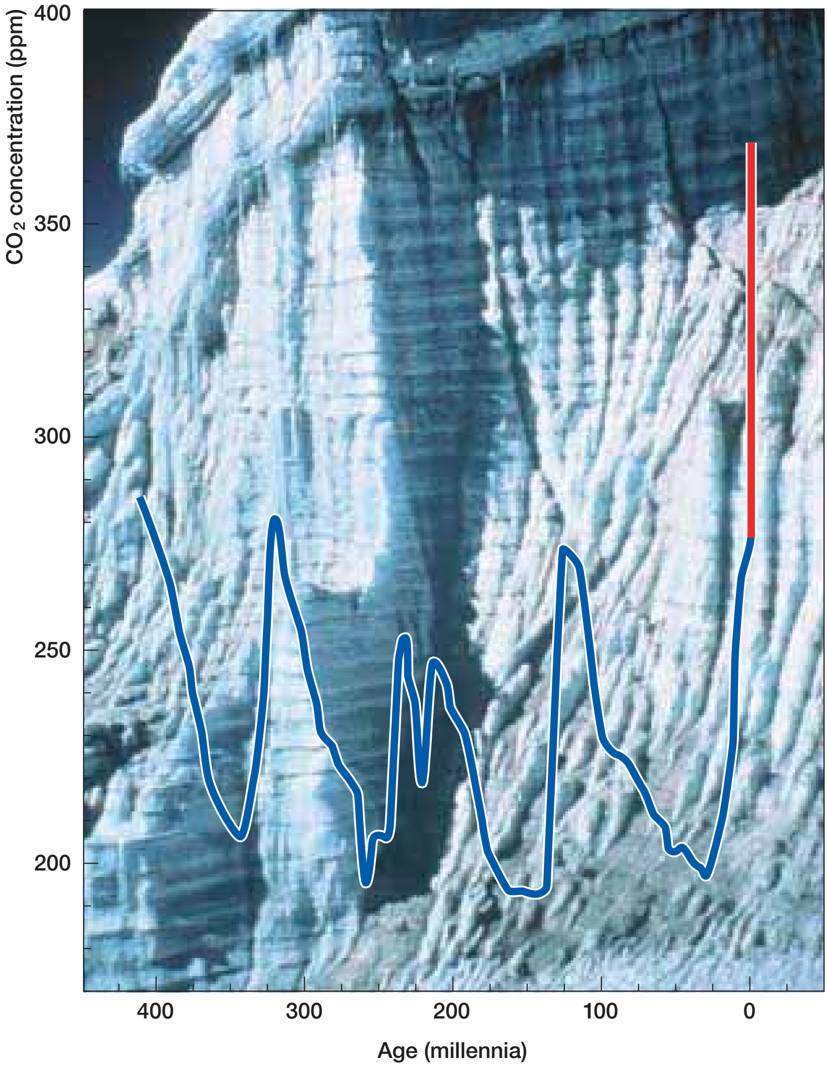


Figure 14
Global CO_2 concentration in parts per million (ppm) measured in air bubbles occluded in the Vostok ice core from Antarctica (blue line, credit: Nature 399: 429–436). The red line shows the man made increase in CO_2 since 1750 based on an ice core (Siple station, Antarctica, 1750–1953, credit: Nature 315, 45–47) and direct atmospheric measurements from Mauna Loa, Hawaii, 1959 to present, credit: http://cdiac.esd.ornl.gov). As a result of the global warming the glacier in the background (Quelccaya, Andes) no longer exists. (Photo: L. Thompson, Science 203: 1240–43).

Fortunately, nature built its own archives and kept track of most changes. We just have to find these archives, read and interpret the stored information and put together the individual pieces to have a complete picture of the temporal and spatial changes in the past. Without looking back into the deep past it will be impossible to estimate the anthropogenic and solar contribution to the present climate change and make reliable predictions about the future.

Looking back into the past is particularly important in the context of the controversial debate about the role man plays in the present global warming.

It is only about a century ago that mankind began to significantly interact with nature and perform a global experiment by increasing the atmospheric greenhouse gas concentrations to levels never experienced in the past half million years **(Figure 14)**. In view of our still very limited understanding of how nature works and our responsibility for future generations we should think twice before performing such blind and uncontrolled experiments. Many species have been extinguished before. Up to today, Homo Sapiens is the first with the potential to cause his own extinction.

Acknowledgments

This article was written in close contact with several colleagues. The numerous scientific discussions with W. Mende who contributed many ideas were particularly useful and enjoyable. The conversations with H.J. Schlaepfer who handled all the editorial aspects with great experience and patience were very stimulating. The comments of R. Muscheler and M. Vonmoos and the careful linguistic corrections by C. Wedema improved the manuscript considerably.

The author

Jürg Beer grew up in Gerlafingen in the Swiss midlands. He finished school with a "type A Matura" including Latin and Greek in Solothurn. In spite of this classical education, Jürg Beer was always interested in natural sciences, mainly in experimental and theoretical physics. In 1968 he registered at the University of Berne to study physics, mathematics and astronomy. He joined the group of Prof. Oeschger where he became familiar with low level counting techniques used to study environmental processes and reconstruct climate changes in the past. The development of a new and extremely sensitive analytical technique in 1977, the Accelerator Mass Spectrometry (AMS), opened up a large variety of new applications. Cosmogenic nuclides could now be measured virtually everywhere in the environment. This was the beginning of a very fruitful collaboration with the group of Prof. Woelfli at the nuclear physics department of the ETH in Zürich. Supported by the Swiss National Science Foundation, ETH and PSI and following the design of M. Suter and G. Bonani, the Tandem accelerator of the ETH developed into one of the world's leading AMS facilities. Jürg Beer took advantage of this unique opportunity and started to explore potential applications of cosmogenic radionuclides (mainly ^{10}Be and ^{36}Cl) in various natural archives such as polar ice cores and Chinese loess deposits.

In 1987, Prof. Imboden set up the department of environmental physics located at the Swiss Federal Institute of Environmental Science and Technology (EAWAG) in Dübendorf. Jürg Beer became the leader of the group "Radioactive Tracers" and extended his research activities. Since then, his group concentrates mainly on the analysis of the GRIP ice core from central Greenland and works on the reconstruction of the geomagnetic field, the solar activity and its influence on climate change. From 1994 to 1995 he spent a year at the High Altitude Observatory of the National Centre of Atmospheric Research (NCAR) in Boulder, USA to expand his knowledge of solar physics.

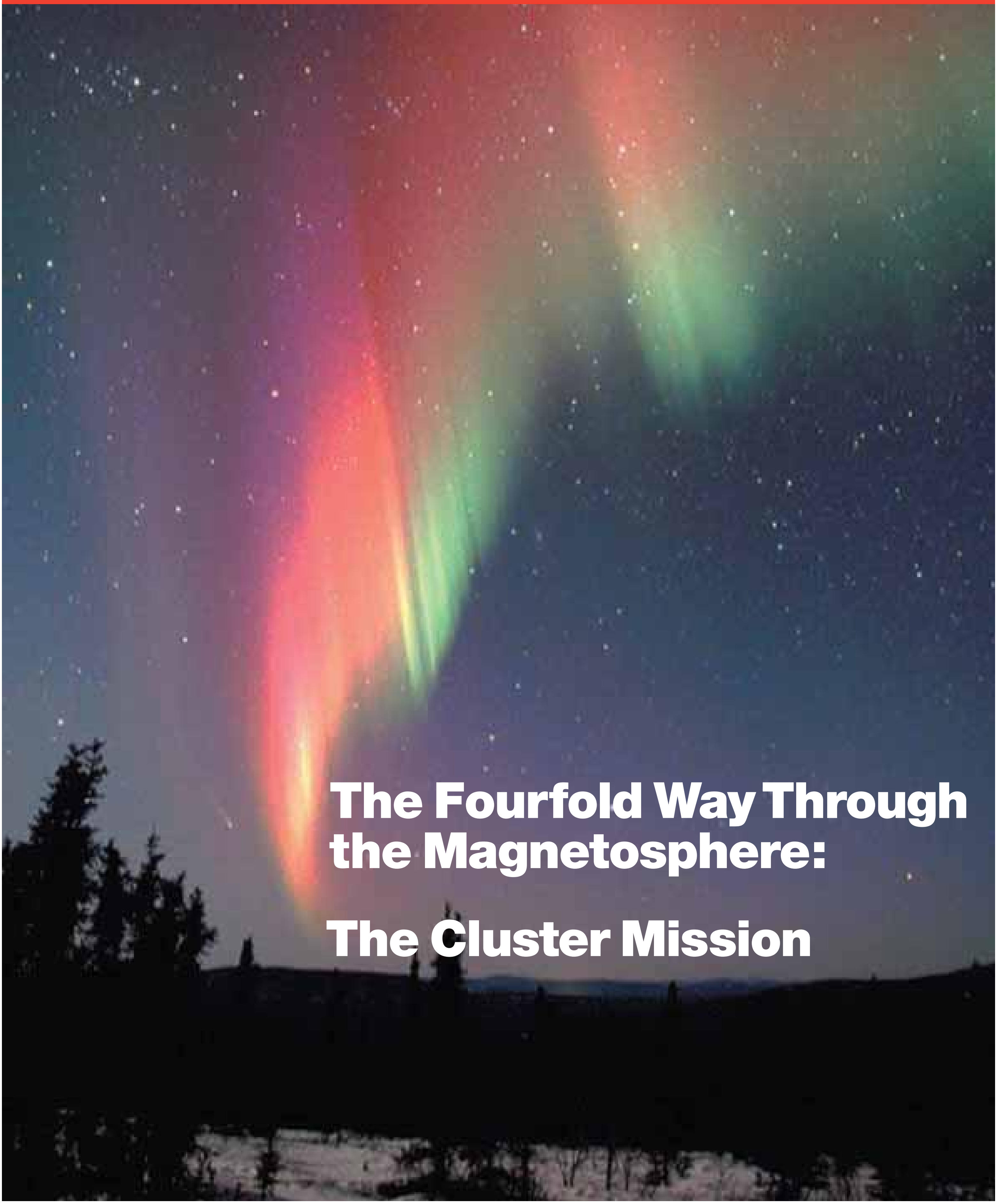
Jürg Beer is convinced that the key to understanding present and future climate changes lies in the past. He often sees himself in the role of a detective seeking for natural archives to collect all available traces of the mechanisms responsible for environmental changes. These traces are often partly erased and written in hieroglyphs, which can only be deciphered by the joint efforts of scientists from different fields.

Jürg Beer lives with his family near Zürich. His interest in nature is not restricted to his professional work. During his free time he enjoys hiking, biking and watching the beauties of nature.

INTERNATIONAL SPACE SCIENCE INSTITUTE
No 9, June 2002

SPATIUM

Published by the Association Pro ISSI



The Fourfold Way Through the Magnetosphere:

The Cluster Mission

Editorial

Imagine a lonely spacecraft travelling through space. Its mission is to probe the value of, say, the strength of the local magnetic field. Its instruments continuously monitor the field strength and deliver the data to the on-board telecommunication system, which forwards them to the ground control station on Earth. There, scientists try to interpret the data.

But now a basic dilemma arises: due to the continuous motion and temporal evolution of the ambient medium and its boundaries, the recorded data are a complicated mix. The observer cannot distinguish whether the magnetic field experienced a change in time or it varied spatially or both happened simultaneously.

That was the problem faced by the fathers of the European Space Agency's Cluster mission in 1982 intended to exploring the Earth's magnetosphere. Their idea was to solve the problem by flying a quartet of closely spaced identical spacecraft, since time and space need a set of three co-ordinates to be defined completely. The instruments aboard the four spacecraft may now take measurements at identical times but at different places in space allowing the scientists to work out the relevant temporal and spatial characteristics of the magnetic field and the other quantities that characterize the ambient medium, such as the electric field, and plasma density, velocity and temperature.

The author of the present issue of Spatium, Dr. Götz Paschmann, currently Director of the International Space Science Institute in Bern and Senior Scientists at the Max Planck Institute for extraterrestrial Physics in Garching (Germany) is one of the fathers of Cluster. He is also the Principle Investigator of the Electron Drift Instrument designed to determine the strength of the local electric field around the spacecraft.

In November 2001, very shortly after the commissioning of the Cluster spacecraft G. Paschmann presented the first results to our members. In the mean time, this mission has delivered a wealth of additional data, which partially could be considered in the present revised version of his lecture. We are very indebted to G. Paschmann for his kind permission to publish the present text on the fascinating world of magnetic fields.

Hansjörg Schlaepfer
Bern, June 2002

Impressum

SPATIUM
Published by the
Association Pro ISSI
twice a year

INTERNATIONAL SPACE SCIENCE INSTITUTE

Association Pro ISSI
Hallerstrasse 6, CH-3012 Bern
Phone ++41 31 631 48 96
Fax ++41 31 631 48 97

President
Prof. Hermann Debrunner,
University of Bern
Publisher
Dr. Hansjörg Schlaepfer,
legenda schläpfer wort & bild,
Winkel
Layout
Marcel Künzi, marketing · kommunikation, CH-8483 Kollbrunn
Printing
Druckerei Peter + Co dpc
CH-8037 Zurich

Front Cover: The aurora is the only visible proof of the Earth's magnetosphere. This wonderful picture shows the glow from outbursts of ionic particles from the Sun in the ionosphere. (Credit & Copyright: Dennis Mammana, Skyscapes)

The Fourfold Way Through the Magnetosphere: The Cluster Mission *)

Götz Paschmann, International Space Science Institute, Bern and Max Planck Institut für extraterrestrische Physik, Garching

Introduction

The Earth's magnetosphere has been investigated by direct in-situ measurements since the beginning of the space age and has turned out to be the site of many fascinating phenomena and processes that must be studied in-situ because they are not accessible by remote sensing techniques. Examples are shock formation, magnetic reconnection, particle acceleration, wave-particle interaction and turbulence. While interesting in their own right, as part of the exploration of the Earth's environment, they can also guide our understanding of similar processes on the Sun or the distant universe, places where one can never make in-situ measurements. **Figure 1** shows a view of the magnetosphere that emphasizes its main features, except for the bow shock that has been left out for clarity.

While much has been learned about the magnetosphere through measurements from single satellites, a limitation has been the fundamental inability of a single observer to unambiguously distinguish spatial from temporal changes. The magnetosphere is constantly changing its shape and size, and many of the processes governing it are known to vary on short spatial and temporal scales. A solution to this dilemma is a fleet of mobile stations, very much like what one does to study weather patterns and their temporal evolution. Cluster is the first attempt to do this in space, admittedly with only the minimum number that can be arranged in three dimensions, namely four. There have previous attempts with two spacecraft (ISEE and AMPTE), but with two stations one can only infer variations along the line between the spacecraft, and three stations always form a plane. Thus Cluster is a giant step forward. One should not forget though that because of the complex structure of the magnetosphere and the strong interrelationship between its parts, there is the danger that spacecraft flying in close formation will miss the large-scale context. Fortunately, observations from the ground and from other spacecraft are available that serve to fill in this global context.

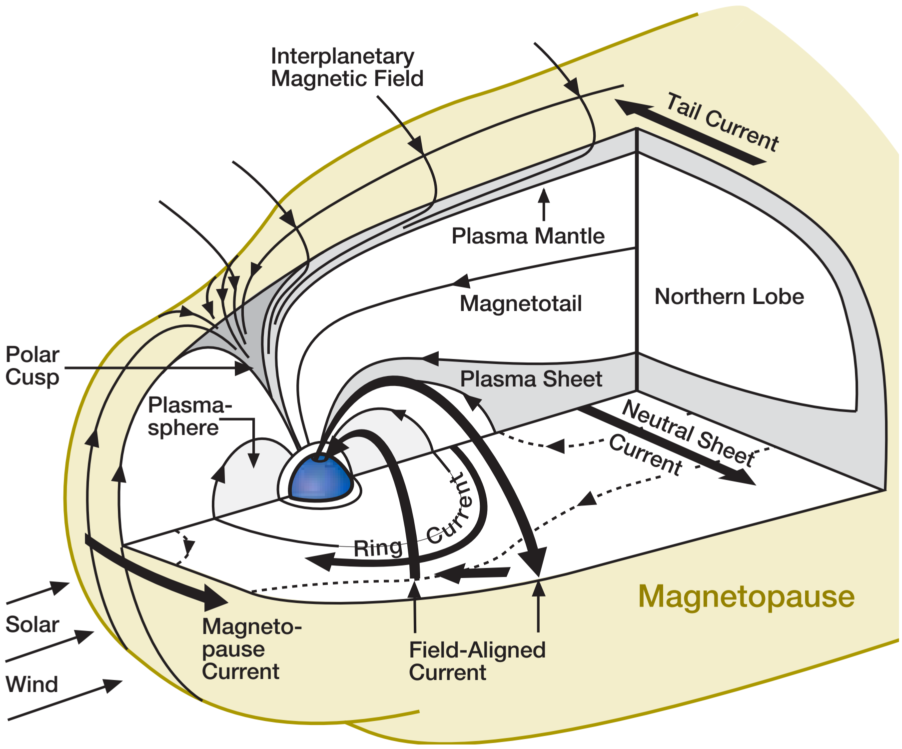


Figure 1
Cutaway view of the Earth's magnetosphere that identifies the key regions encountered by Cluster: the solar wind approaching from the left; the magnetopause (yellow surface) that is the outer boundary of the magnetosphere; the polar cusp, a funnel-shaped indentation in the magnetopause; the magnetotail, with the hot plasma sheet at the centre and the lobes on either side. Thin black lines with arrows are the magnetic field lines, dashed lines show plasma flow. The black arrows mark electric current flow. The bow shock that stands in the solar wind ahead of the magnetopause, and the region in between, the magnetosheath, have been omitted for clarity.

*) Pro ISSI lecture, Bern, 8th November 2001

The Cluster Mission

History

Cluster's history goes back to November 1982, when a group of European scientists proposed the mission. In February 1986 ESA chose Cluster and the Solar and Heliospheric Observatory (SOHO) as the first "Cornerstone" in its Horizon 2000 Science Programme. SOHO was launched in December 1995 and has sent back a stream of exciting data about the Sun, as vividly described in Spatium Nr. 2. Cluster was not so fortunate. Ready for launch in 1995, the four spacecraft were destroyed when the Ariane 5 rocket exploded during its maiden flight on 4 June 1996. At first it seemed that the only remaining option was to put together a single spacecraft from the leftover parts, appropriately named Phoenix after the mythical bird that rose from the ashes. But recognizing the unique importance of the mission, the ESA Science Programme Committee (SPC) on 3 April 1997 agreed that three new Cluster spacecraft should be built alongside Phoenix, and thus Cluster was reborn. Through a tremendous effort by all involved, the fully instrumented satellites were rebuilt in only three years and successfully launched on two Soyuz-Fregate launchers from Baikonur on 16 July and 12 August 2000, respectively. After extensive commissioning of the payload, the science phase officially began on 1 February 2001.

This being a Pro ISSI lecture, I should point out that several people associated with ISSI were deeply involved in Cluster. Two of its directors, Bengt Hultqvist and myself, and a long-time member of its Science Committee, Gerhard Haerendel, were among the seven who proposed the mission in 1982. Roger Bonnet, then Director of Scientific Programmes at ESA, was the key force behind the recovery initiative, and Hans Balsiger was chairman of ESA's

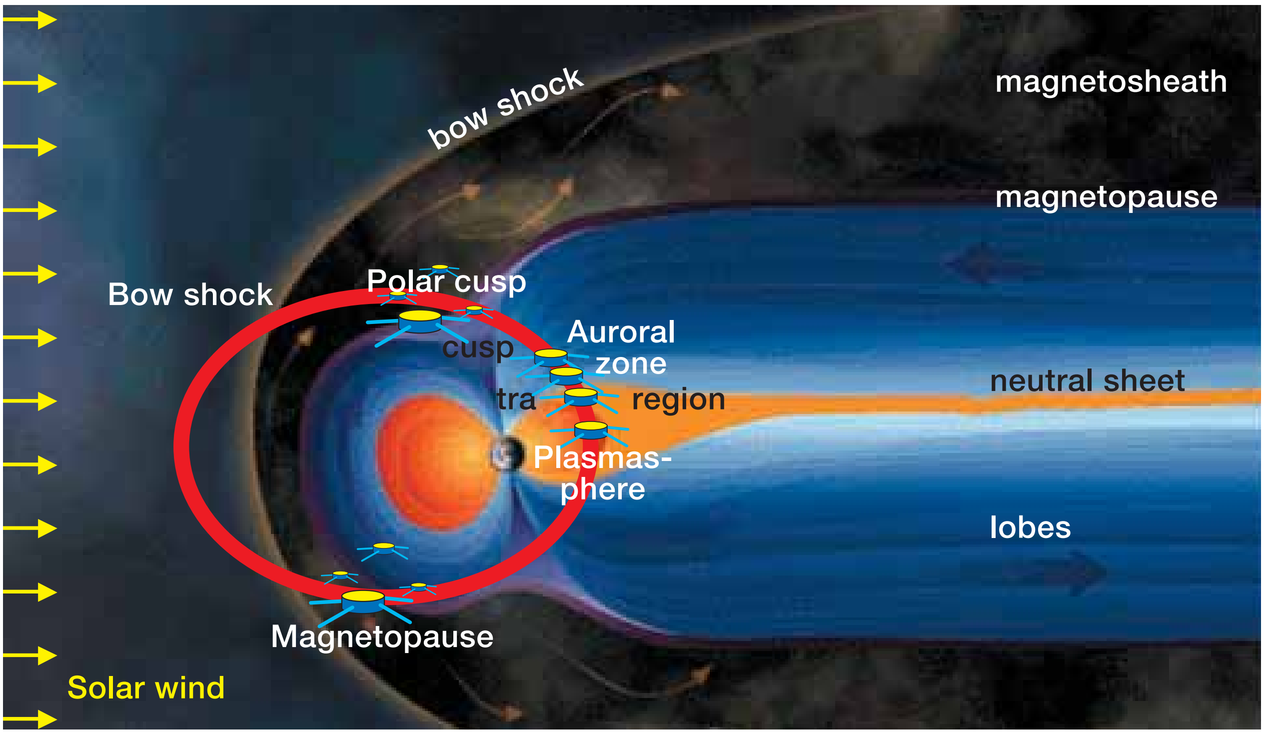


Figure 2
Cluster orbit in February, as it cuts through the key regions of the magnetosphere.

SPC, when the decision to repeat Cluster was finally made in April 1997. Last but not least, Hans Peter Schneiter and Contraves Space had a leading role in the revival efforts by the industry that had built Cluster. It is therefore only natural that the interpretation of the Cluster measurements will be one of ISSI's future workshop activities.

Orbits and Separations

After their successful launch from Baikonur, the Cluster spacecraft were placed in nearly identical, highly eccentric polar orbits, with an apogee of 19.6 Earth, a perigee of 4 Earth radii, and a period of 57 hours. **Figure 2** shows the orbit superimposed on a cut of the magnetosphere, and illustrates that Cluster moves outbound over the northern polar cap, crosses the magnetopause and bow shock into the solar wind, before recrossing those boundaries in reverse order and moving over the southern polar cap back towards perigee. **Figure 2** applies to February of each year. As the orbit is inertially fixed, it rotates around the Earth once a year, as the latter revolves around the Sun. As a result, the apogee of the orbit will be located in the geomagnetic tail half a year later, as shown in **Figure 3**. In the course of the mission Cluster will thus encounter the regions of interest several times. The orbits are tuned so that the four spacecraft are located at the vertices of a nearly

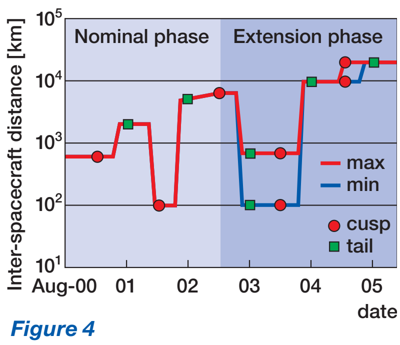


Figure 4
Spacecraft separation distances as planned for the entire mission.

regular tetrahedron when crossing the major boundaries. The separation distances between the spacecraft are being changed between 100 km and tens of thousands of km during the mission. This allows studying the phenomena at sufficiently different scales. **Figure 4** shows the sequence

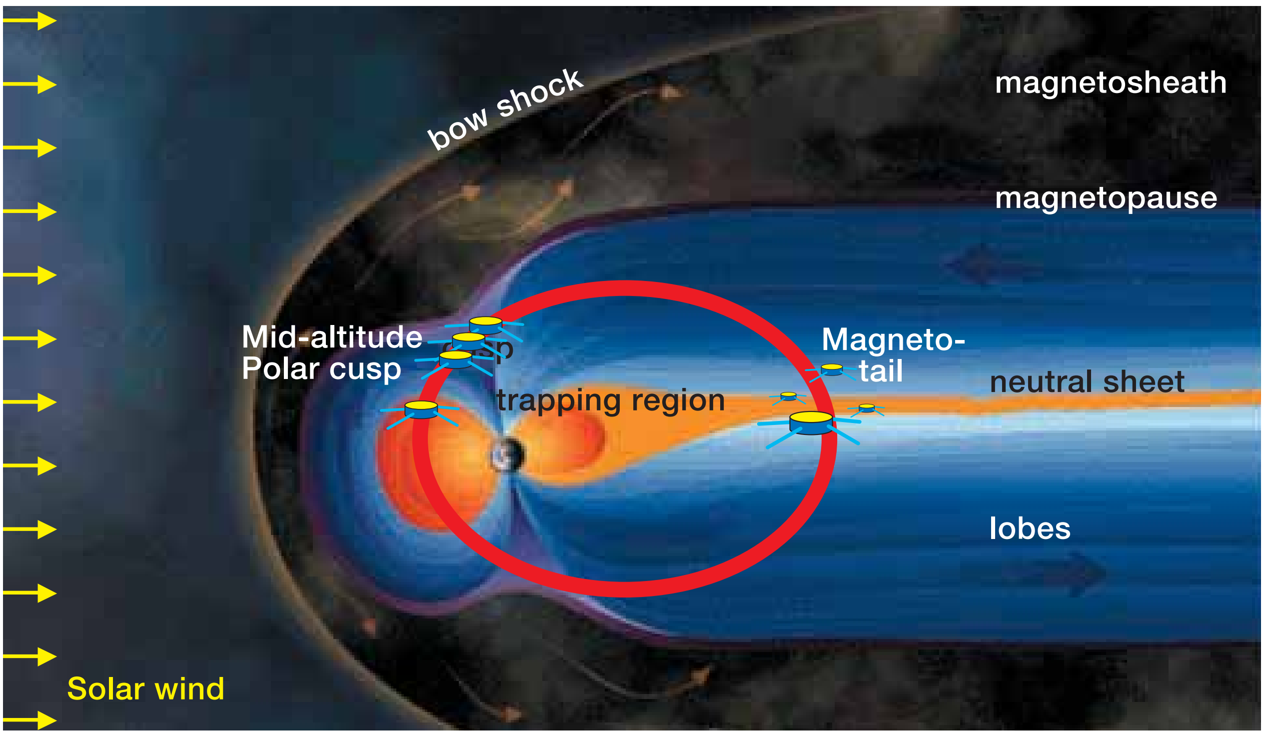
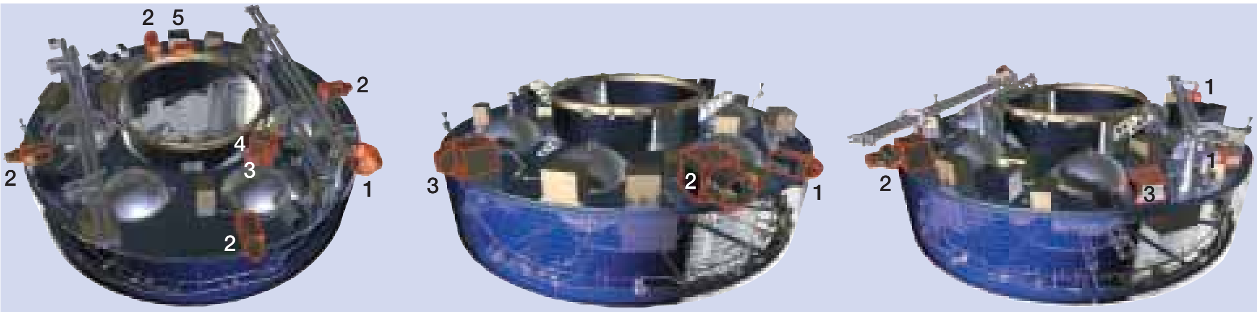


Figure 3
Same as Figure 2, but for the month of August.



1 **STAFF** (N. Cornilleau-Wehrlin, F)
Magentic and electric fluctuations
2 **EFW** (M. Andre, S)
Electric fields and waves
3 **DWP** (H. Alleyne, UK)
Wave processor
4 **WHISPER** (P. Decreau, F)
Electron density and plasma waves
5 **WBD** (D. Gurnett, USA)
Electric field wave-forms

1 **PEACE** (A. Fazakerley, UK)
Electrons (E < 30 [keV])
2 **CIS** (H. Reme, F)
Ion composition (E < 40 [keV])
3 **RAPID** (P. Daly, D)
High energy electrons and ions
($2 < E_{ions} < 1500$ [keV/nuc],
$20 < E_e < 400$ [keV])

1 **FGM** (A. Balogh, UK)
Magnetic field
2 **EDI** (G. Paschmann, D)
Plasma drift velocity
(0.1 < E < 10 [mV/m],
5 < B < 1000 [nT])
3 **ASPOC** (W. Riedler, A)
Spacecraft potential control

Figure 5
The science payload carried by each of the 4 Cluster spacecraft. The group listed on the left comprises the Wave Experiment Consortium, the three in the middle deal with the particle component, and the top two on the right address measurements of the magnetic field and the plasma drift velocity, respectively, while the last controls the spacecraft potential.

of separation distances planned until the end of the mission. This separation strategy is possible because each Cluster spacecraft carried 600 kg of fuel at launch, of which about half was expended in order to reach the operational orbit.

Science Payload

The four Cluster spacecraft carry identical sets of 11 scientific instruments, designed to measure the ambient electromagnetic fields and particle populations over a wide range of frequencies and energies, respectively (see **Figure 5**). Each instrument was built by a team under the leadership of a Principal Investigator (PI). There are more than 200 Co-Investigators from ESA member states, the United States, Canada, China, the Czech Republic, Hungary, India, Israel, Japan and Russia.

Figure 5 also gives an impression of the large size of the spacecraft. Its diameter is almost 3 m and its mass at launch 1200 kg, of which half was in form of fuel needed to reach the final orbit and to carry out the separation strategy. The mechanical structure of the spacecraft was developed and built by Contraves Space, while Dornier was ESA's Prime Contractor for Cluster, which included integration and test of the entire system, including instruments.

Almost unnoticed by even the teams involved, one of the spacecraft is carrying an extra piece of hardware, namely an engraved 10-by-10 cm Titanium plate that is part of a "kinetic sculpture" called Moving Plates by their creators, Viennese artists Eva Wohlgemuth and Andreas Baumann. Other such plates are mounted on stationary as well as moving objects all over the world.

Excursion: The Electron Drift Instrument

The Cluster payload contains one instrument, the Electron Drift Instrument (EDI), which is unique because it actively emits beams of electrons to sense the electric fields in the ambient plasma. This instrument was developed by a consortium consisting of the University of New Hampshire, the University of California at San Diego, Lockheed Palo Alto Research Laboratory, and the

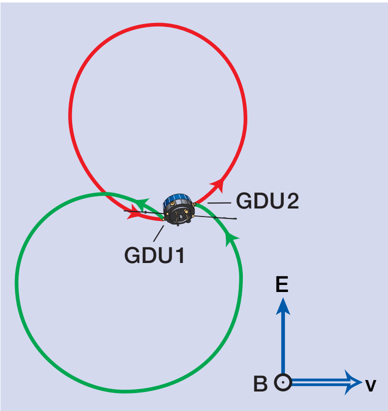


Figure 6
EDI's principle of operation. For any combination of magnetic field B and drift velocity V (induced by an electric field E), only a single electron trajectory exists that connects each electron gun with the detector located on the opposite side of the spacecraft. The two trajectories have different path lengths and thus different times-of-flight.

Max-Planck-Institut für extraterrestrische Physik in Garching. Because I happen to be the PI for the EDI instrument, I will dwell on it in some detail.

When injected perpendicular to a magnetic field, electrons perform a circular orbit (with a radius proportional to their velocity) that returns them to their origin after a complete gyration. Gyro radius and gyro time both are inversely proportional to the strength of the magnetic field. For electrons with 1 keV energy and a magnetic field of 100 nT, the gyro radius is 1000 m and the gyro time 0.35 ms.

In the presence of an electric field, the electrons execute a drift motion in addition to their gyration. As a result, the circular orbits are distorted into cycloid orbits, such that the electrons are displaced from their origin after one gyration. The displacement (the "drift step") is directly proportional to the electric field and inversely proportional to the square of the magnetic field, and ranges from a few centimetres to hundreds of meters. Under these conditions the beam will return to the spacecraft only when emitted in either of two precisely defined directions. By employing two beams and two detectors, the two unique directions can be monitored continuously and the displacement obtained by a triangulation procedure. **Figure 6** illustrates the principle, and also shows that the two beams travel different distances, and thus have different travel times. When the magnetic field becomes small, the electric field is obtained from these time-of-flight differences.

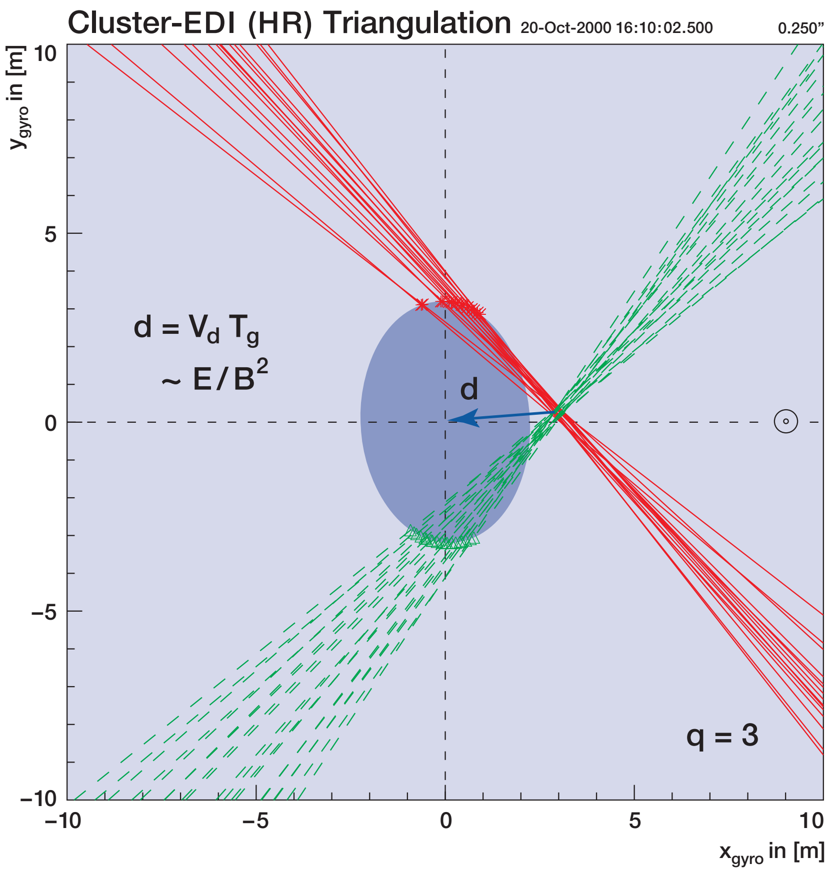


Figure 7
Sample EDI triangulation plot for an interval of 0.5 s duration. Red and green lines are the firing lines of the two electron guns for those times when the beams were properly aimed to return to the detectors. The gun positions, projected into the plane perpendicular to the magnetic field, move on an ellipse because the spacecraft spin axis and the magnetic field direction are not aligned. The 0.5 s interval corresponds to 1/8 of a full spin. The measured "drift step" is the vector from the beam intersection to the centre of the figure, and is proportional to E/B^2.

While exceedingly simple in principle, it is the implementation that has made the Electron Drift Instrument a formidable challenge. The first difficulty lies in the need to keep the beams precisely perpendicular to the ambient magnetic field, which can have arbitrary direction and can vary rapidly. This is overcome by a unique electron gun design that can emit a narrow beam in any direction within more than a hemisphere, and by having a direct on-board link to the Cluster magnetometer instruments (FGM and STAFF) that allows reconstruction of the instantaneous magnetic field direction. The next difficulty is to find the direction that returns each beam to its associated detector. This is achieved by sweeping the beam in the plane perpendicular to the magnetic field and waiting until the detectors record a signal. While this is going on, the detector look direction is steered in synchronism with the beam firing directions. Once signal has been acquired, the beam is swept back and forth in order to continuously track the target direction.

Because there usually are plenty of natural electrons at the same energy as the beams, care must be taken to distinguish the beam electrons from the natural electrons. This is achieved by coding the beams with a pseudo-noise code, much the same way as in a GPS system. By correlating the received electrons with the code of the emitted electrons, one not only discriminates against the natural electrons, but one also measures the times-of-flight of the electrons. As a by-product, one also obtains a very precise measure of the magnetic field this way (from the average of the two times-of-flight, which is equal to the gyro time). Further difficulties arise because the beams diverge. Thus fewer electrons return to the detectors the larger the distance they have to travel, i.e., the smaller the magnetic field strength is. This is partially overcome by continuously adjusting the beam current (between 0.1 and 300 nA) and the detector sensitivity.

The onboard software runs a 4-ms servo loop to execute the beam acquisition and tracking tasks and to navigate through the various control parameters. More than 50 patches of that software were needed before it worked satisfactorily under the variable conditions experienced along the Cluster orbit. **Figure 7** shows one of the first successful measurements, using the triangulation technique. The vector, d, from the beam intersection to the centre of the spacecraft is the "drift step", which is proportional to the ratio E/B^2. Using the magnetic field, B, measured by the FGM instrument, the electric field, E, is readily obtained. Results obtained with EDI are discussed in the next section.

Following Cluster from the Solar Wind to the Ionosphere

We now will follow Cluster through key regions of the magnetosphere and demonstrate its capabilities with illustrative examples, starting from the solar wind and ending with the aurora in the polar ionosphere.

Solar Wind

The solar wind is a plasma stream that approaches the Earth with speeds between 300 and 800 km/s. It carries with it the magnetic field of solar origin, the interplanetary magnetic field (IMF). Solar wind and IMF are both highly variable, which is one of the reasons the magnetosphere is so dynamic.

As an example, let us look at a sudden change in IMF direction caused by the passage of the heliospheric current sheet (HCS) over the Cluster spacecraft. The HCS is a warped surface dividing the heliosphere into two hemispheres of oppositely directed magnetic fields. Cluster encountered the HCS on 13 February 2001 at 10:48 universal time (UT), and the results are presented in **Figure 8.** The figure shows the magnetic field measurements and their analysis by the FGM team that highlights the unique

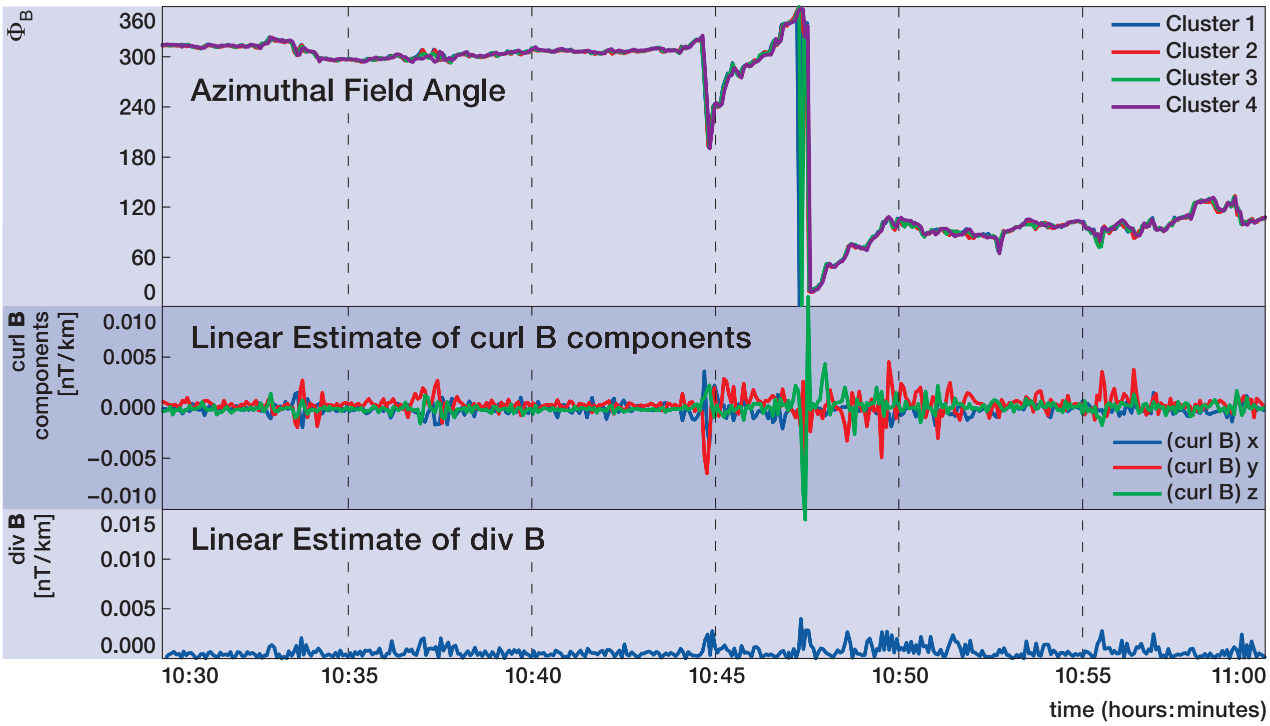


Figure 8
An encounter of the heliospheric current sheet (HCS). The top panel shows the magnetic field direction that showed a large rotation near 10:47 UT as the HCS crossed the Cluster formation. The middle panel shows the curl of the magnetic field, estimated from the four spacecraft magnetic field data: the large values during the passage of the HCS are caused by the current within the structure. The bottom panel shows an estimate of div B: this remains small throughout, providing a consistency check for the estimate of curl B (from Eastman et al., 2001).

capability of Cluster to derive spatial gradients in three dimensions. Assuming that the field varies linearly between the spacecraft positions, the density of electric current flowing in the volume between the spacecraft can be derived from the magnetic field measurements at the four points, using Ampère's law. The figure demonstrates the success of the technique: precisely when the HCS passes over the spacecraft, there is a spike in the computed current density. Note that as a consistency check, one also computes the divergence of B, which according to Maxwell's equations should be zero.

Bow Shock

The first sign of any interaction between the solar wind and Earth's magnetic field is a shock wave in space on the Sun-facing side of the Earth. This bow shock is rather like the sonic boom caused when an aircraft flies at supersonic speed. The role of the bow shock is to slow down the solar wind to subsonic speed such that it can be deflected around the obstacle formed by the Earth's magnetic shield, the magnetosphere.

Simply inspecting the raw magnetometer data already reveals the superiority of a four-spacecraft vantage point. **Figure 9** shows a crossing of the bow shock when the IMF had a direction that made it nearly aligned with the shock surface, a geometry referred to as (quasi-)perpendicular because the field is then perpendicular to the normal to its surface. In such a situation the shock has a very simple step-like structure and the profiles recorded by the 4 spacecraft are essentially just time-shifted copies of each other. From the time shift and the known distance between the spacecraft one obtains the speed at which the shock passed over the spacecraft, about 5 km/s in this case. Together with the measured duration of the shock

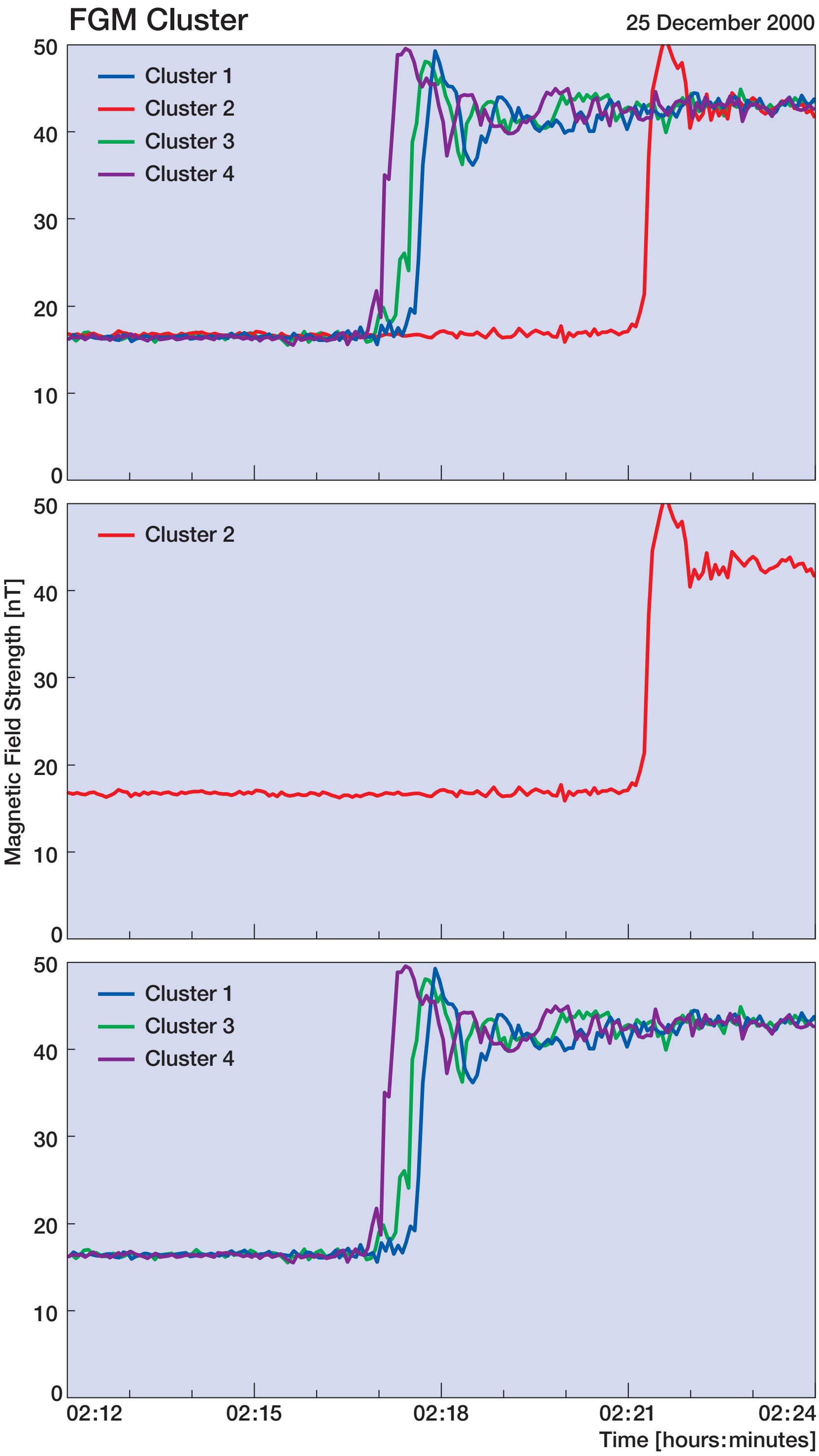


Figure 9
Magnetic field measured by the FGM instruments on the four spacecraft for a crossing of a quasi-perpendicular shock.

transition, this speed yields the thickness of the shock, a quantity that is an important for understanding shock formation in collisionless plasmas.

When, by contrast, the IMF is more nearly aligned with the shock normal, it becomes a (quasi-)parallel shock. Parallel shocks are characterized by oscillating transitions, and are much harder to analyse. They are, however, exceedingly important, because it is such parallel shocks (created by supernovae explosions) that are thought to be responsible for acceleration of galactic cosmic rays. As a consequence of the parallel geometry, particles are free to traverse the shock in both directions. This result is an extended "foreshock" region that is characterized by particle beams, waves, turbulence and first-order Fermi acceleration. Deep in this foreshock one finds Short Large Amplitude Magnetic Pulsations (SLAMP's), which are believed to play the dominant role in the thermalisation of the incident plasma (via reflection and subsequent mixing analogous to that at quasi-perpendicular shocks). **Figure 10** shows two SLAMPs observed by Cluster on 2 February 2001, when the spacecraft were separated by about 600 km. Despite both theoretical and previous experimental estimates of SLAMP sizes larger than 3000 km, the Cluster measurements reveal considerable spatial variability on scales less than 600 km. It will be interesting to see how the situation changes when Cluster separations have been reduced to 100 km (see **Figure 4**).

Magnetopause

Because charged particles are deflected by magnetic fields, the Earth's magnetic field is an obstacle that the solar wind cannot simply penetrate. Instead it flows around it, confining the terrestrial magnetic field to a cavity, the magnetosphere (see **Figure 1**). Its boundary, the magnetopause, is located where the plasma pressure exerted by the solar wind is in equilibrium with the magnetic pressure from the Earth's magnetic field. Across the magnetopause, the magnetic field usually undergoes a sharp change in magnitude and direction. The magnetopause is thus a current layer that separates the solar wind and its embedded interplanetary magnetic field from the Earth's magnetic field and plasma.

Density Gradients

The four spacecraft measurements allow the measurement of spatial gradients of scalar quantities, such as the gradient of the plasma density, as shown in **Figure 11**. The plasma density is deduced in this case from the plasma frequency measured by the WHISPER instrument. Using the values of the density measured simultaneously at the four spacecraft, the density gradient can be calculated, as has been done for a magnetopause crossing on 2 March 2002 shown in **Figure 11**. At the top, the figure shows the measured plasma densities on the four spacecraft. At the beginning of the interval, the spacecraft were in the magnetosphere, characterized by a tenuous plasma, then they crossed the magnetopause around 03:31 UT and went into the magnetosheath with a density about 20 cm^{-3} for less than a minute before re-entering the magnetosphere at about 03:32 UT. Note that at this time, the Cluster spacecraft were only about 100 km apart. As a result, they were all located in the thin magnetopause simultaneously, as demonstrated that the times of the four density traces in **Figure 11** clearly overlap. The same is true for the magnetic field traces from the FGM instrument (not shown). From the differences in the densities measured at a given time on the four spacecraft, the density gradients inside the current layer can be determined. Similarly, the current density can be determined from the magnetic field differences, using the method illustrated in **Figure 8**. The gradient vectors are shown at the bottom of **Figure 11**, projected onto one of three orthogonal planes (the other two projections not shown for clarity). Knowledge of the instantaneous density gradients and current densities are essential for understanding the formation of the magnetopause, and for distinguishing different plasma transport mechanisms, but neither information was available before Cluster.

Magnetopause Thickness and Motion

As a result of dynamical variations in the solar wind and interplanetary magnetic field, the magnetopause is in constant motion, which makes determination of its structure virtually impossible from single-spacecraft data.

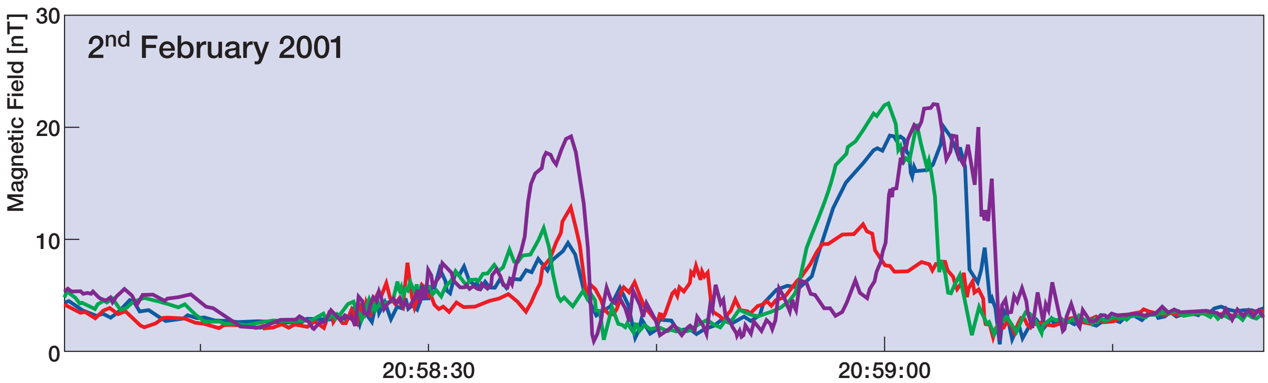


Figure 10
Shock Large Amplitude Pulsations observed by the 4 Cluster spacecraft in the vicinity of the quasi-parallel bow shock. Note the significant differences between the various spacecraft despite their modest (~600 km) separation (from Lucek et al., 2001).

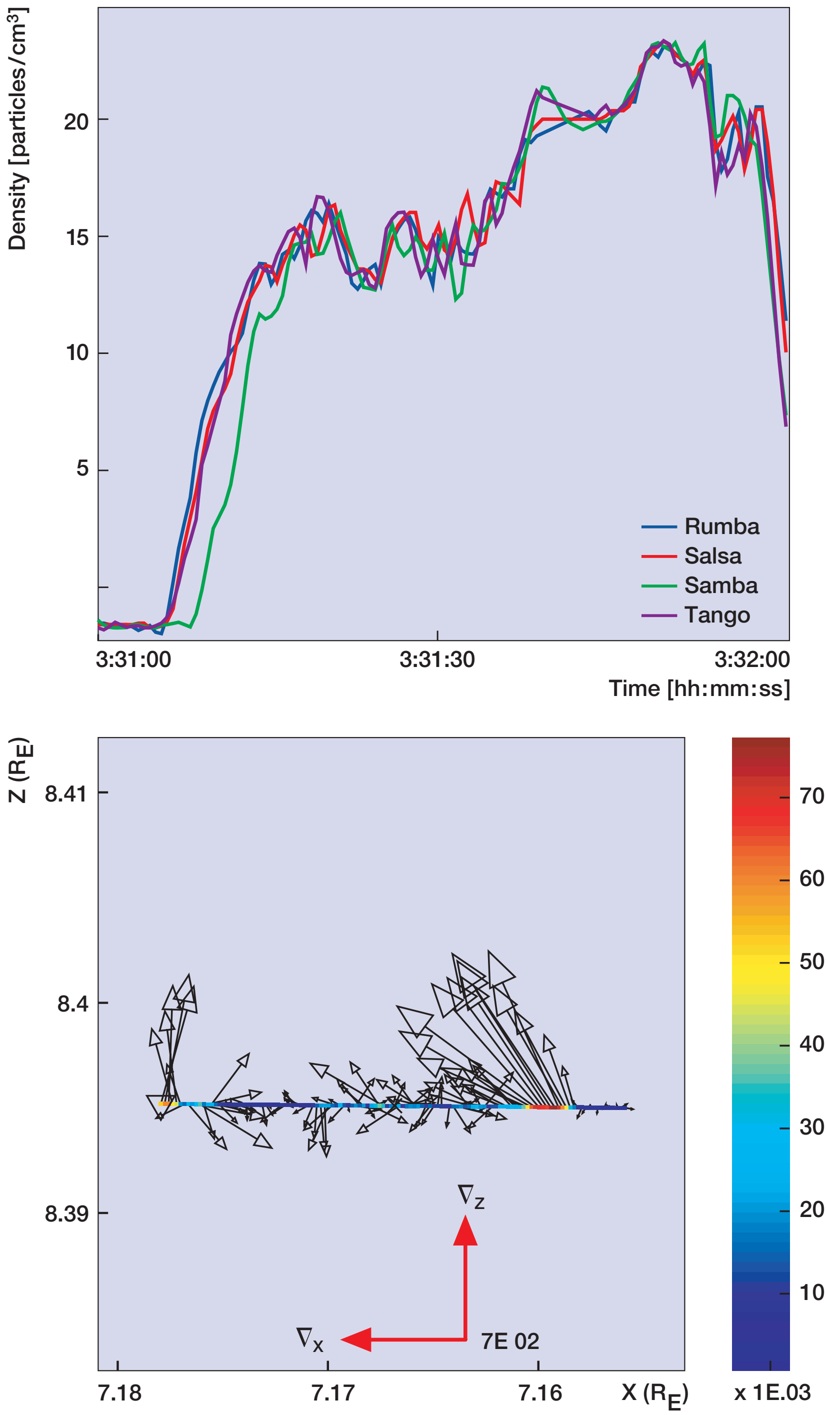


Figure 11
Plasma densities measured on the four Cluster spacecraft by the WHISPER instruments for a magnetopause crossing on 2 March, 2002 (top), when the spacecraft separation distances were only 100 km. At the bottom, the figure shows the gradient vectors inferred from the instantaneous differences in the densities (from Decreau et al., 2002).

Figure 12 (top panel) shows time series of the magnetic field measured by the FGM instruments on the four spacecraft for a magnetopause crossing on 4 July 2001. From the four crossing times and the known spacing between the spacecraft one can derive the magnetopause velocity, 30 km/s in this case. The bottom panel shows an expanded view of the magnetopause crossing by spacecraft C3. From **Figure 12** one infers a thickness of the magnetopause current layer of approximately 200 km, corresponding to about three ion Larmor radii, which is physically reasonable, although theory makes no definite prediction.

While determination of the current layer thickness has importance in its own right, the most striking aspect of the data in **Figure 12** is the sharp rise in plasma density (from 0.1 to 24 cm^3) across the current layer (shown in the bottom panel), because it means that at this time essentiality no solar wind plasma was able to enter the magnetosphere, neither at this location nor further upstream. This is an unexpected result, because it is generally thought that solar wind plasma can enter the magnetosphere by any number of processes.

The four-spacecraft timing provides the magnetopause velocity only at four discrete instances in time, which is often not enough to infer the boundary thickness and its evolution. But it turns out that the measurements of the plasma drift velocity with the Electron Drift Instrument (EDI), referred

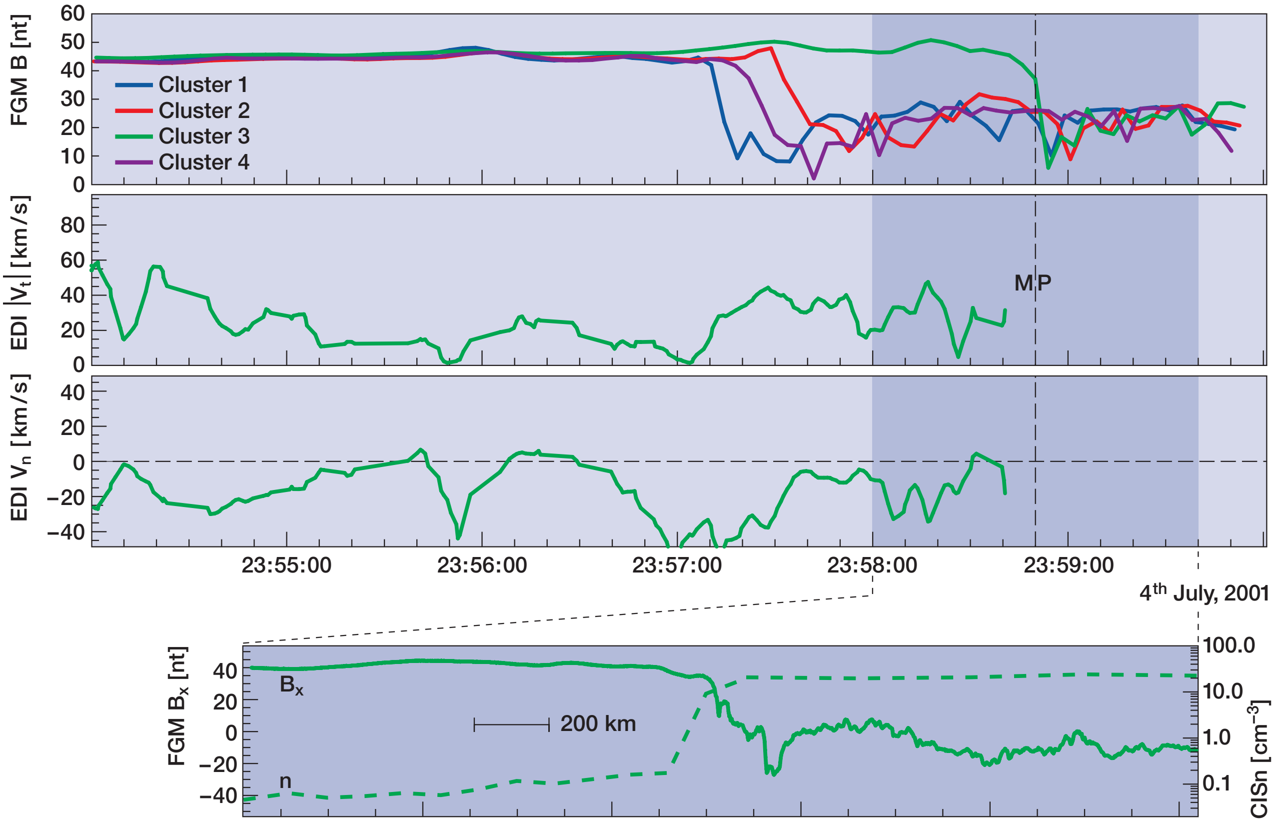


Figure 12

Magnetopause crossings by the four Cluster spacecraft on 4 July 2001, identified by the sudden drop in magnetic field strength measured by the FGM instruments (top panel). The second and third panels show, respectively, the components of the plasma convection velocity tangential (V_t) and normal (V_n) to the magnetopause, as measured by EDI on C3. The velocity V_n is interpreted as the magnetopause velocity. When integrated over time, it gives the instantaneous distance of spacecraft C3 from the magnetopause (not shown). The bottom panel shows, for an expanded time scale, the x-component of the magnetic field (from FGM) and the plasma density (from CIS), measured on C3 across the magnetopause. The horizontal bar shows a distance scale (200 km) obtained from the measured magnetopause velocity.

to earlier, and by the ion spectrometer CIS, can provide a continuous estimate of this motion, because, to first order, plasma and magnetopause move together. **Figure 12** shows, as the third panel, the projection, V_n, of the drift velocity measured by these two instruments on C3 onto the magnetopause normal direction. V_n varies, but is negative on average, corresponding to a net in-ward motion of the magnetopause that eventually causes the spacecraft to exit the magnetosphere. Integrating this velocity over time provides a continuous estimate of the magnetopause distance (not shown. This information permits reconstruction of the complete spatial profile, for example of the plasma density.

Magnetic Reconnection

To first order, the magnetopause is an impenetrable surface that completely separates the Earth's magnetic field and the solar wind. The example in **Figure 12** demonstrates that this situation can actually occur. More commonly, however, that separation is far from perfect, and the investigations of the processes that allow transfer across the magnetopause are at the heart of the Cluster mission. The one process known to play a key role for transfer of mass, momentum, and energy across the magnetopause, is magnetic reconnection. This process is the result of the breakdown of the "frozen-in" picture. Under ideal conditions, the interplanetary magnetic field is frozen into the solar wind plasma, and therefore slips along the magnetopause with the solar wind flow. But when this concept breaks down locally, by processes not yet identified, the IMF can diffuse relative to the plasma and connect with the terrestrial magnetic field across the magnetopause. This is sketched in **Figure 13**. Once reconnected, the magnetic field lines are dragged over the poles and into the magnetotail because they are embedded in the solar wind flow.

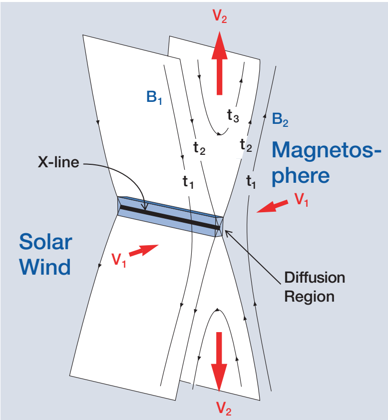


Figure 13
Schematic to illustrate reconnection across the magnetopause between magnetic field lines from the Sun and from Earth. A pair of field lines that at time t_1 are still entirely in the solar wind and magnetosphere, respectively, connect through the magnetopause at time t_2 and become two sharply kinked field lines at time t3 that now have one end on the Earth, the other on the sun. The reconnection site is a line that is termed X-line or reconnection line. The solar wind plasma approaches at speed V_1 and becomes accelerated to V_2 as a result of the tension in the sharply bent field lines. It is those high-speed plasma jets that are the most directly observable signature of magnetic reconnection at the magnetopause.

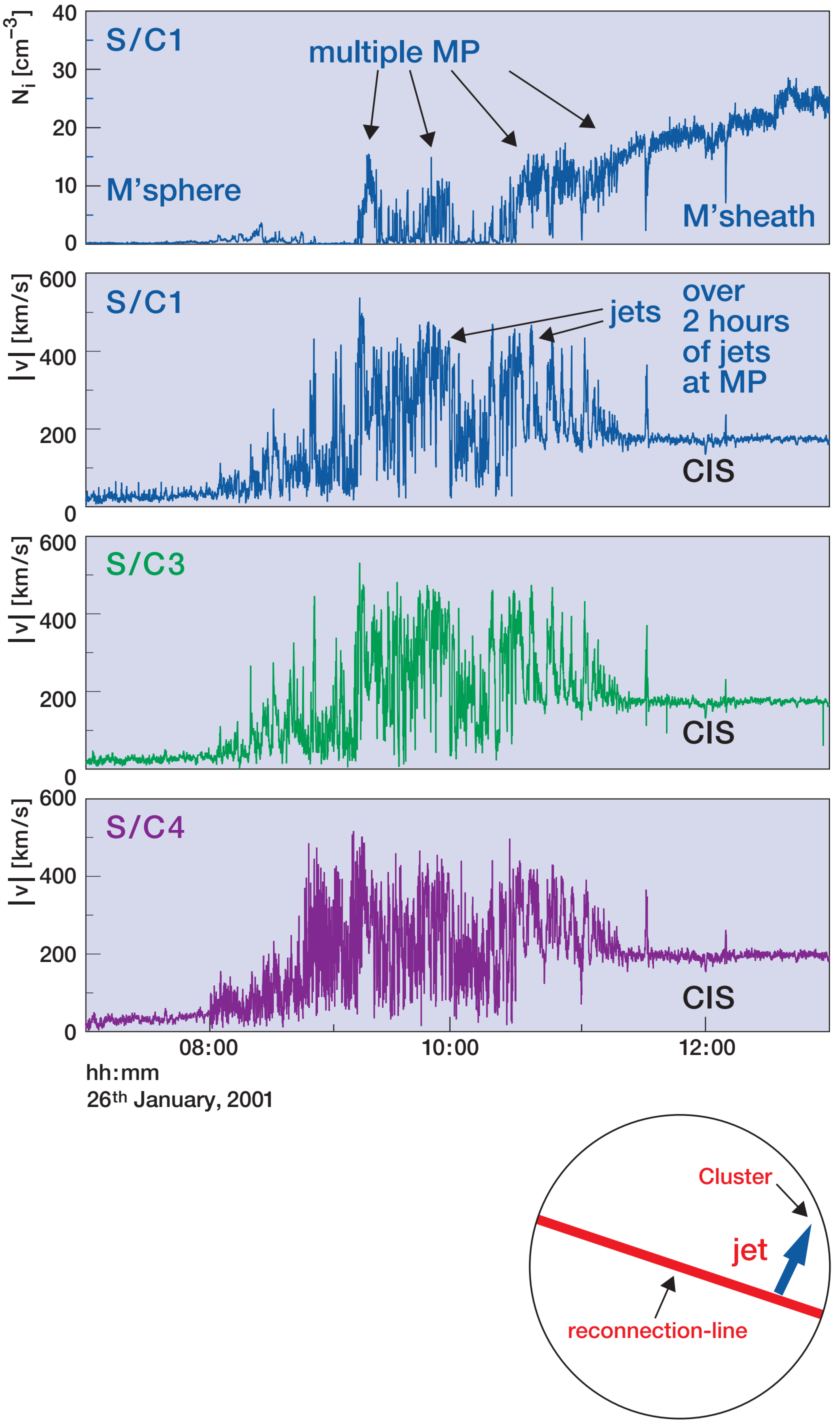


Figure 14
Plasma density (on C1) and velocities (on C1, C3, C4) measured by CIS-HIA for the magnetopause encounter on 26 January 2001. The high velocities indicate the encounter with plasma jets emanating from a reconnection-line located somewhere below the spacecraft (courtesy Phan et al.).

When reconnection occurs, solar wind plasma can enter the magnetosphere by fluid flow along reconnected field lines. Upon entry it is accelerated to large speeds by the magnetic tension forces that exist across the magnetopause because the reconnected field lines are sharply bent, as shown in **Figure 13**. We were the first to observe the resulting plasma jets with an instrument on ISEE in 1979. This discovery formed the "smoking gun" evidence for reconnection at the magnetopause.

One of the open questions is whether reconnection can happen in a quasi-stationary fashion, or whether it is necessarily intermittent. As the plasma jets are localized in thin layers, they can be observed only when this layer passes over the spacecraft. With a single spacecraft one could never be sure that reconnection had not actually stopped in between successive crossings. With Cluster one now has closely spaced crossings that allow to fill the gaps. **Figure 14** shows data from the CIS instrument that demonstrate that the plasma jets are observed whenever any of the spacecraft crosses the magnetopause. For this interval of almost 1-hour duration, reconnection must therefore have happened almost continuously, although not necessarily at a fixed rate.

Polar Cusps

The polar cusps are weak spots in the defence of the magnetosphere against the on-rush by the solar wind. This is due to their funnel shape and their small magnetic field. This unique feature was one of the drivers for the Cluster idea, as illustrated by **Figure 15**, taken from the 1982 proposal. As illustrated by **Figure 2**, the Cluster obit crosses the distant cusp region when its apogee is one sunward side. In fact it was this property that drove the selection of the Cluster orbit. By contrast, the orbit crosses cusp magnetic field lines at mid-altitudes when the apogee is on the night side (see **Figure 3**).

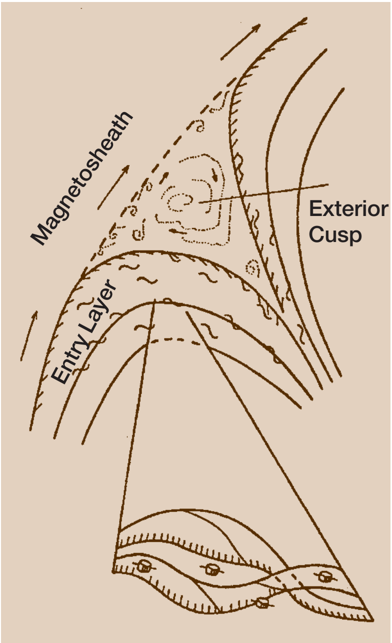


Figure 15
Expectations for the distant polar cusp region, as envisaged in the 1982 Cluster proposal.

Figure 16 shows what is observed at such a mid-altitude cusp crossing and illustrates the successive encounter with the particles entering the magnetosphere through the cusp. Close to perigee, when these measurements are made, the Cluster spacecraft have much larger separation distances than further out along their orbit. This explains the large time shift between the encounters, especially that by spacecraft C3. The difference in appearance is probably caused by a reconfiguration of the cusp before the encounter by C3.

Polar Cap

Moving tailward from the polar cusp, one enters the polar cap region that is permeated by magnetic field lines that extend deep into the geomagnetic tail where the form the tail lobes (see **Figure 1**). The polar cap is often void of any plasma, but CIS/CODIF data show the frequent presence of cold oxygen ions on polar cap magnetic field lines. **Figure 17** shows an energy-time spectrogram of O^+ ions for an hour on 4 March 2001. The feature to note is the narrow band of intense fluxes that proceeds to lower and lower energies while the spacecraft moves from the cusp over the polar cap towards the magnetotail. For three selected times, t_1–t_3, the upper diagram shows the velocity distribution of the oxygen ions, confirming that they have an extremely narrow spread in velocity, with the mean velocity becoming smaller with time.

The schematic diagram at the bottom illustrates the interpretation. Ions injected with a broad energy spectrum from a narrow source at low altitudes on the dayside, all will move up the magnetic field lines, but end up at different locations, depending on the ratio of the speed at which they move along the magnetic field and the speed with which those field lines move over the polar cap, as shown in the bottom panel. The black line is derived from the O^+ distributions, while the red symbols are from the measurements of the drift of artificially injected electrons by EDI. Noting the entirely different nature of the measurements, the agreement is simply remarkable.

Data shown in **Figure 17** are taken from one spacecraft only. But if one compares the measurements on the other spacecraft, one finds that the convection velocities are essentially identical, with zero time shifts. This means that the observed variations in the convection velocity, in particular the short bursts of sunward convection (positive V_x) are temporal in nature, probably caused by "dipolarizations" of the tail magnetic field that occur as a result of a magnetospheric substorm (see below).

Magnetotail

As a consequence of the solar wind flow past the Earth, magnetic field lines are being stretched in the anti-sunward direction, creating a long tail. This magnetotail consists of a northern and southern lobe containing very tenuous plasma and permeated by magnetic fields of opposite polarity that emanate from the northern and southern polar cap. At the centre there is a region of hot dense plasma, the plasma sheet (cf. **Figure 1**). The plasma sheet in turn straddles the so-called neutral sheet, where the magnetic field reverses direction, and which therefore is an ideal site for magnetic reconnection. Through magnetic reconnection in the tail, magnetic field lines that were opened by reconnection at the subsolar magnetopause (see **Figure 1**), and were carried into the magnetotail by the action of the solar wind flow, are being closed again, convect sunwards, and start

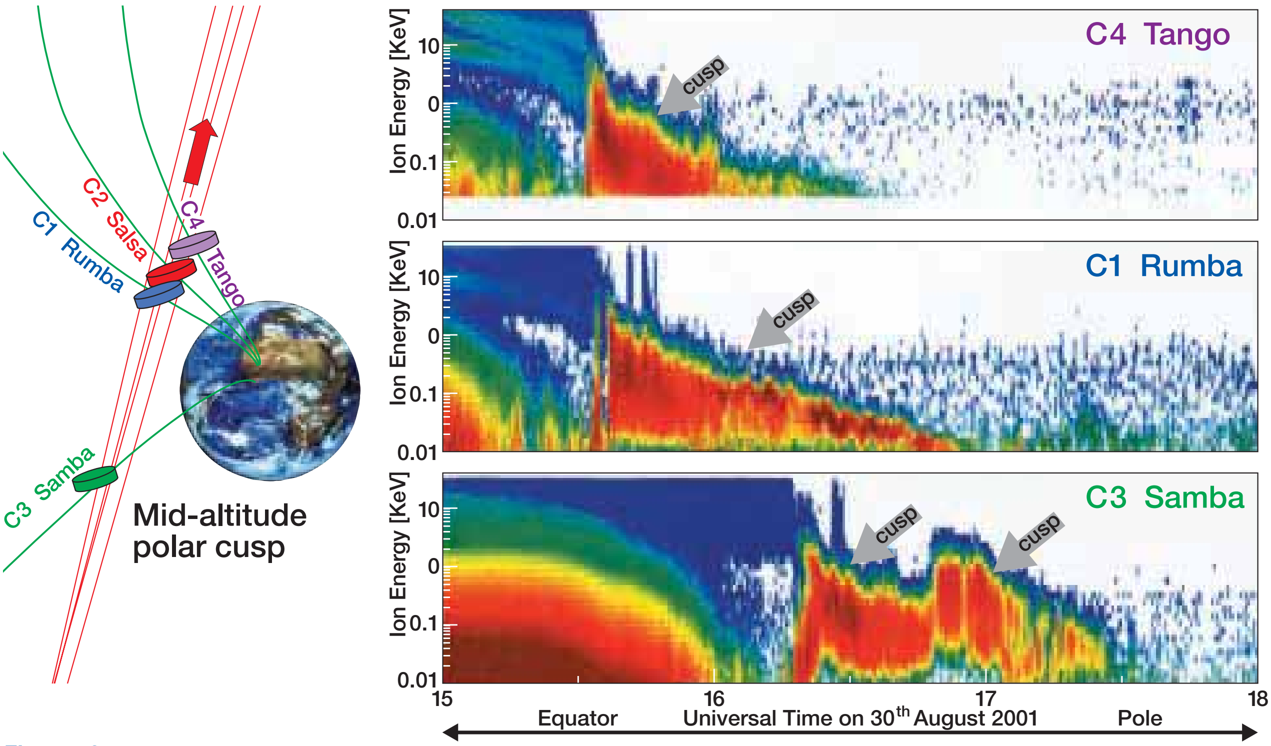


Figure 16
Crossing of the mid-altitude cusp (around 5 Re altitude) by the 4 spacecraft on 30 August 2001. The energy-time spectrograms from CIS on C1, C3 and C4 on the night show the encounter with the ions entering the magnetosphere through the cusp (from Bosqued et al., 2001).

the process all over again once they have reached the subsolar magnetopause. The combination of magnetic reconnection at the dayside magnetopause and in the magnetotail sets up a global circulation of the plasma and magnetic field in the magnetosphere.

The plasma sheet is the site of violent events referred to as magnetospheric substorms, which form one of the most intriguing phenomena in the magnetosphere. Substorms are manifestations of the sudden release of energy stored in the geomagnetic tail. For reasons not yet understood, the energy is not released at the same rate as it is stored, but quite suddenly, with enormous consequences for the entire magnetosphere, including an intensification and expansion of the aurora. **Figure 18** shows how, at the onset of such a substorm, the magnetic field measured by the Cluster spacecraft suddenly changes from a very stretched configuration, where the magnetic fled is essentially along the tail axis, indicated by a small elevation angle, λ, to an orientation characteristic of a magnetic dipole, i.e., having a large λ. This dipolarisation is not simultaneous, but propagates as a front passing over the four spacecraft in succession. This allows for the first time to precisely determine the speed and direction of such a front.

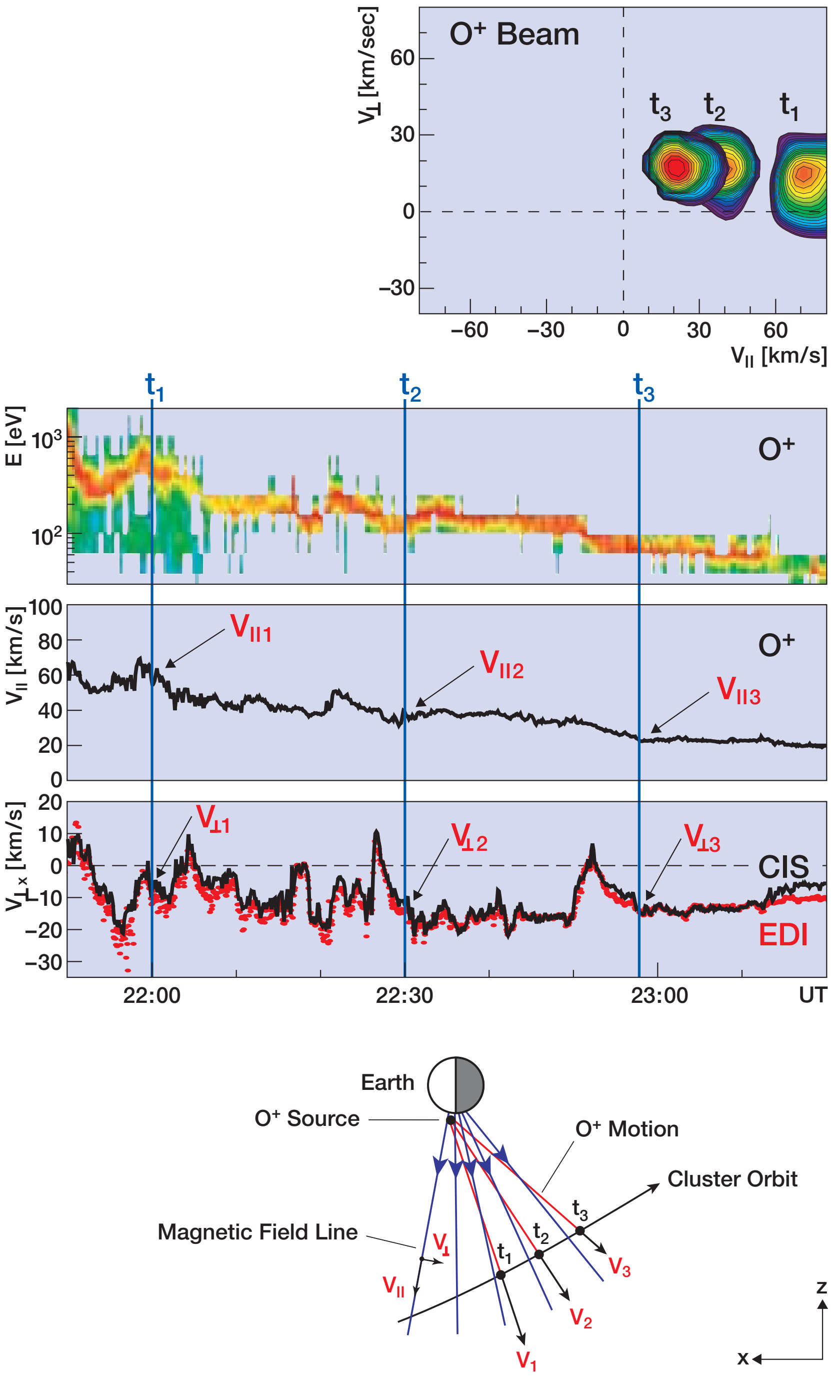


Figure 17
Oxygen ion observations and field line convection velocities for a polar cap pass on 4 March 2001. The top panel shows the distributions of O^+ ions in velocity-space, measured by CIS-CODIF at three different times; below are the energy time spectrograms of these ions, their inferred parallel velocity, and the convection velocity, V_x, inferred from the O^+ measurements (solid black line) and from EDI (red symbols). The interpretation of these observations is illustrated by the sketch at the bottom.

Plasmasphere

The upper atmosphere and ionosphere are important sources of plasma for the magnetosphere. In the region of closed magnetic field lines in the inner magnetosphere, outflow of ionospheric plasma produces the so-called plasmasphere, with a usually sharp outer boundary, the plasmapause.

Figure 19 shows a traversal of the plasmasphere, first inbound and then outbound, as illustrated by the insert in the lower left. In addition to the central plasmasphere, the figure shows some density structures further outward. The shape of these structures looks strikingly similar in the four profiles. Replotting the density measurements against the McIlwain L-parameter confirms that these structures are fixed in space and do not vary much over time-scales of about one hour. (The L-parameter is a measure of the equatorial distance of magnetic field lines in the inner magnetosphere, originally introduced to characterize the radiation belts)

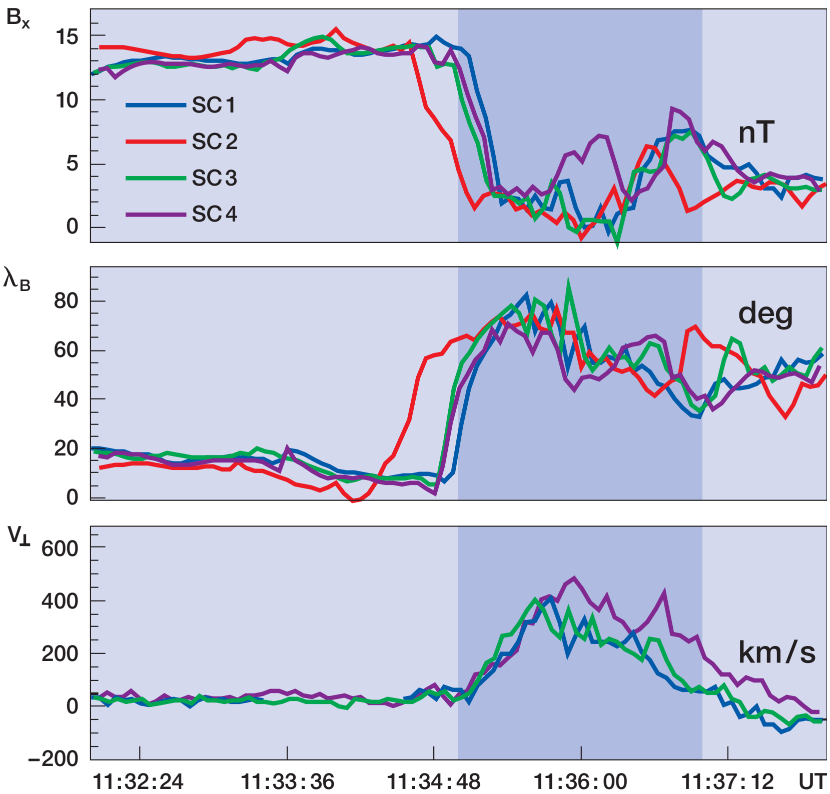


Figure 18
Magnetic field observations from the four Cluster spacecraft during the onset of a magnetospheric sub storm, when the stretched magnetic configuration, indicated by large values of the magnetic field along the tail axis, B_x, and small values of the field elevation angle λ, suddenly relaxes to a more dipolar configuration, characterized by small values of B_x and a large angle λ.

Aurora

The only visible consequence of the interaction between the solar wind and Earth's magnetic field is the aurora. It occurs in an oval around both magnetic poles and is generated by energetic particles that precipitate into the upper atmosphere and cause light emission as well as ionisation. The cusp and boundary layers on the dayside and the plasma sheet on the night side are the sources of this precipitation. On the other hand, ions escaping from the auroral oval contribute to the plasma in the magnetotail and in the magnetopause boundary layers.

Black Aurora

The visible aurora is associated with converging electric field structures producing an upward directed electric field that accelerates electrons towards Earth, generating the auroral light emission and upward field-aligned currents. The opposite happens in the auroral return current region: there the electric field structures are divergent and accelerate electrons upwards into space, causing a void of electrons and a lack of auroral light emissions. This is why this region is often referred to as "black aurora". **Figure 20** schematically illustrates the juxtaposition of visible and black aurora.

Figure 21 shows, schematically, the observations by the EFW and PEACE instruments from a crossing of a U-shaped potential structure around 04:30 UT on 14 January 2001. At the time of the first overpass, the electric field measurements by EFW showed a weak bipolar signal (signifying a sign reversal of the field, as required by **Figure 20**), which then became stronger during the second and third overpasses, but which had

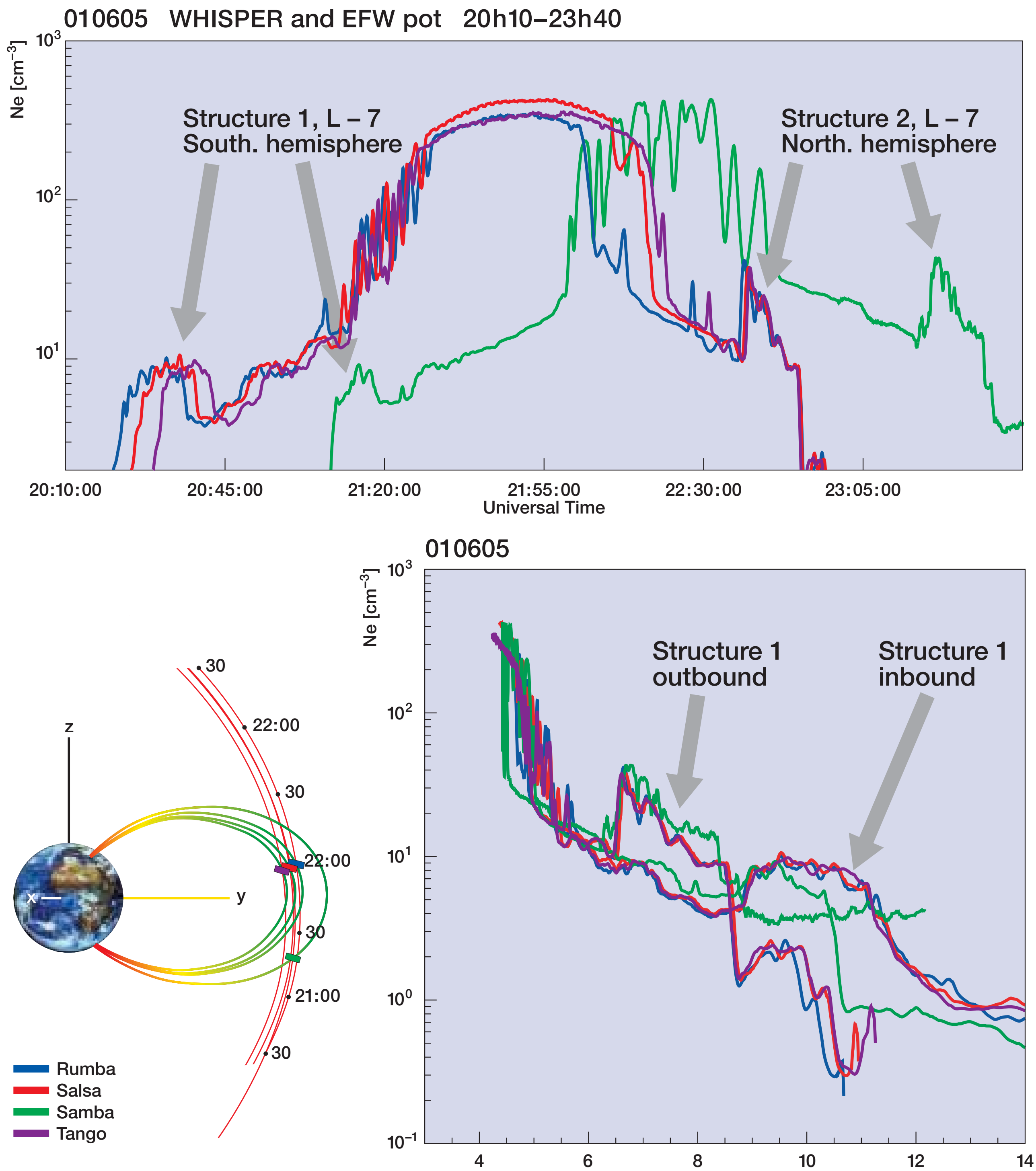


Figure 19

Plasma density deduced from the WHISPER and EFW instruments during the crossing of the plasmasphere on 5 June 2001 (top panel). The spacecraft orbit is shown on the lower left. When plotted against the McIlwain L-parameter (lower right), the density structures prove to be spatially fixed (from Decreau et al., 2001).

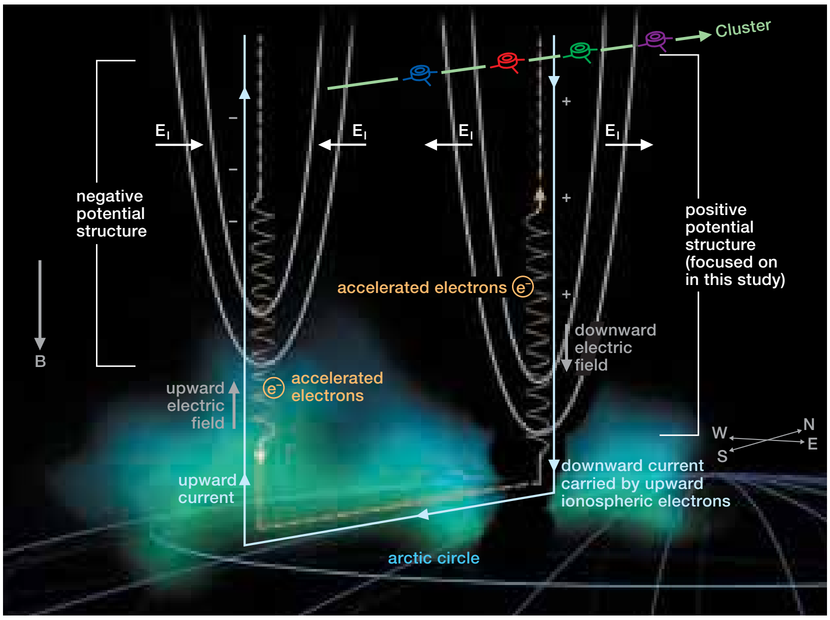


Figure 20
U-shaped electric potential configuration over an aurora (left) and its counterpart over a dark region referred to as black aurora (right). In the structure on the left, electrons are accelerated downward into the atmosphere to cause the characteristic auroral emissions, while the structure on the right accelerates electrons upward, causing a lack of aurora emissions, hence its name.

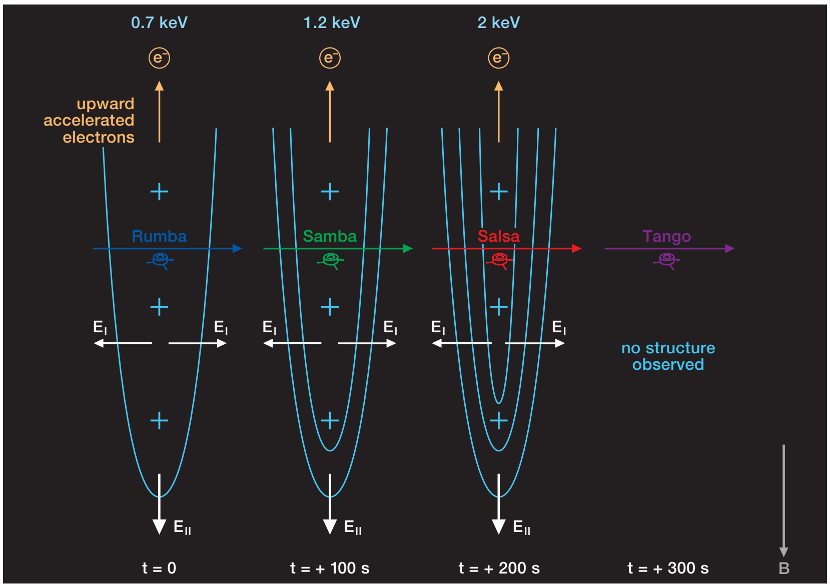
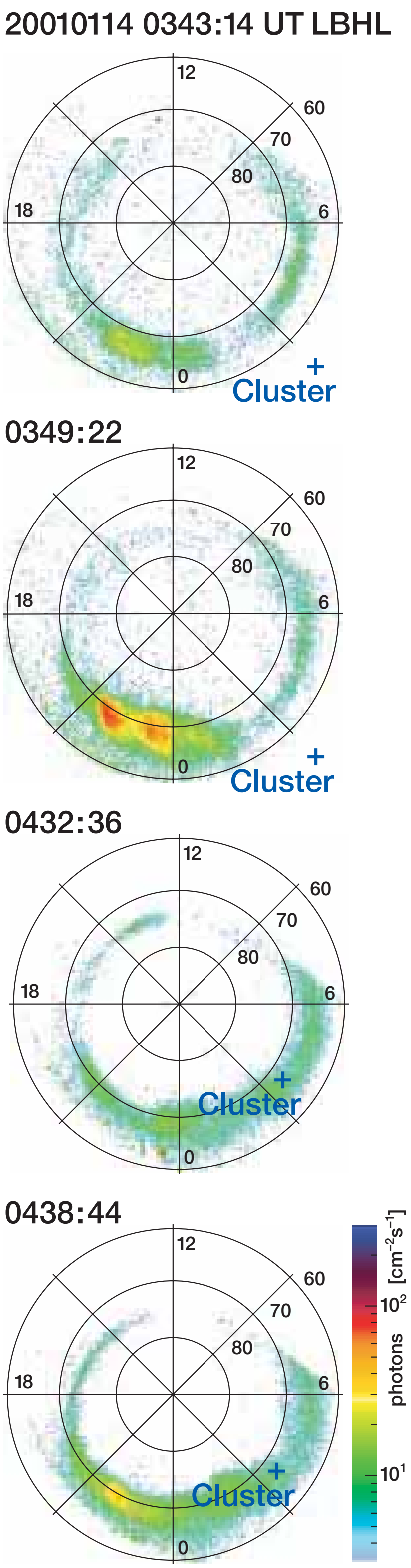


Figure 21
Schematic representation of the evolution of the electric potential pattern that Cluster observed at around 04:30 UT on 14 January 2001 when the four spacecraft sequentially flew over a black aurora, recording initially growing electric fields and electron energies, before electric field and electrons disappeared altogether (after Marklund et al., 2001).

essentially disappeared when the fourth spacecraft flew over the same region. Quite consistent with this, the PEACE instrument observed upward directed electron beams with increasing energy, before the electrons disappeared as well. The results show that this diverging electric field structure extends to altitudes greater than 20,000 km and that it grows in size, intensifies and decays over time scales of a few hundred seconds.

Ion Outflow and Energy Input

Electric fields directed such that they accelerate (negatively charged) electrons downward into the ionosphere to cause auroral displays will at the same time accelerate ions from the upper atmosphere/ionosphere in the opposite direction because of their positive electric charge. **Figure 22** shows a sequence of auroral images, together with Cluster ion measurements taken during a traversal of the plasma sheet, recognized as the intense flux (shown in red) of energetic ions in the spectrogram in the top panel. Inspection of the fourth panel on the right shows that oxygen ions are flowing up the magnetic field lines, as indicated by the predominance of large fluxes at pitch-angles of 180 degrees, and are detected on Cluster whenever Cluster is magnetically connected with the auroral region. In return, the plasma sheet is a source of electromagnetic energy for the aurora. At 04:31:30 UT, when Cluster crosses the edge of the plasma sheet, recognized as



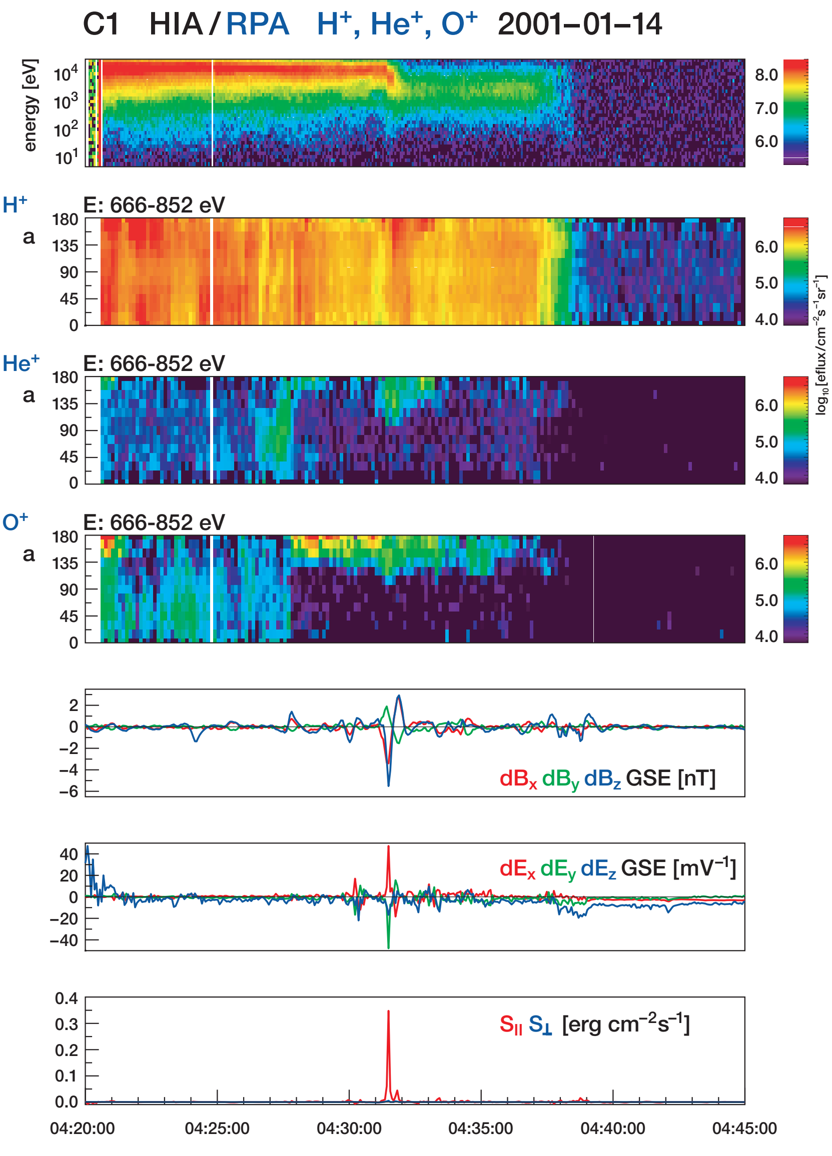


Figure 22

On the left, a sequence of four auroral images from the UVI instrument on NASA's polar spacecraft, with the mapped Cluster positions shown as blue crosses. On the right, Cluster particle and field measurements for the same time interval. The top panel shows an energy-time spectrogram of ions with energies between 5 eV and 27 keV from the CIS-HIA instrument. The next three panels show the pitch-angle distribution of hydrogen, helium and singly charged oxygen, respectively. The magnetic and electric field fluctuations measured by EFW and FGM are shown in the third and second panels from the bottom. The bottom panel, finally, shows the energy flux density, S, computed from these electric field and magnetic field fluctuations. The key features are the appearance of upflowing oxygen ions at pitch-angles near 180°, and the large earthward directed energy flux right at the edge of the hot plasma sheet, which serves to power the aurora (from Wilber et al., 2002).

the drop in flux of energetic ions in the spectrogram in the top panel, a high flow of earthward directed electromagnetic energy, S, is observed. This energy flux provides a way to power the aurora.

Auroral Kilometric Radiation

Since the mid 1960s it is known that the Earth is a powerful radio source in the frequency range from 50 to 500 kHz. Because electromagnetic radiation in this frequency range has wavelengths in the kilometer range, and because of the close association with the aurora, this radiation is now called auroral kilometric radiation, abbreviated AKR. One of the primary objectives of the WBD instrument on Cluster is to use the unique four-spacecraft configuration to conduct very-long-baseline interferometry (VLBI) investigations of AKR, because they have the potential of locating the AKR source better than can be achieved with any other technique. **Figure 23** illustrates the differential delay location determination method, simplified to only two spacecraft.

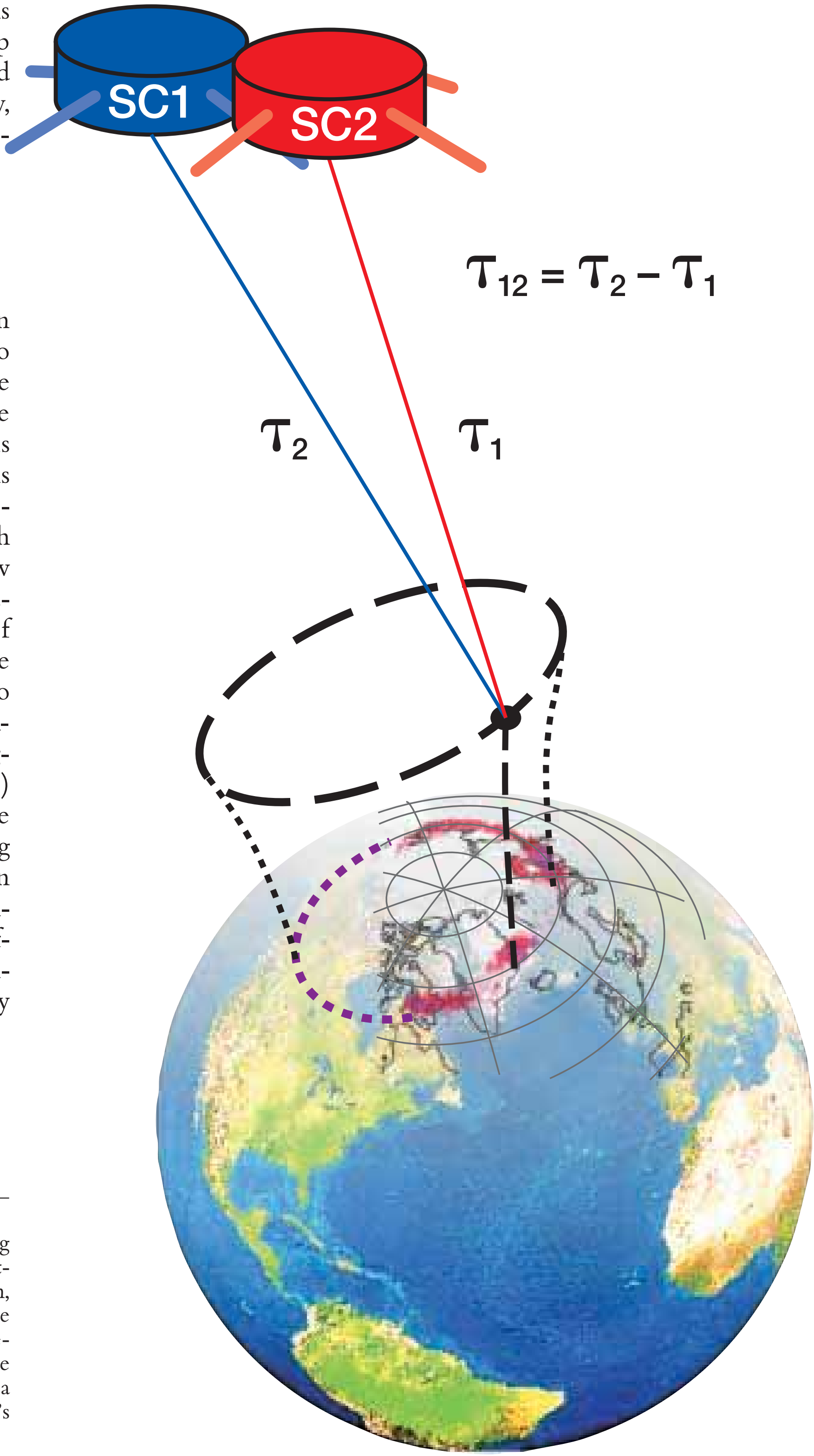


Figure 23
Illustration of the scheme for locating the source of the Auroral Kilometric Radiation (AKR) by triangulation, based on differential delays between the radiation recorded by pairs of spacecraft. With only two spacecraft, the source is constrained to locations in a ring at some distance from the Earth's centre (courtesy WBD team).

Future

Originally intended for a two-year lifetime, the Cluster mission has now been extended until late 2005. Although at the time of this writing, the Cluster teams are still busy with final calibrations and validation of the analysis methods, it is clear from even the cursory glance at the data provided in this lecture, that at the end we will have a highly improved understanding of the various physical processes and phenomena operating in collisionless plasmas. However, it is also clear that in view of the many scale sizes that are involved, four spacecraft are simply the minimum. Ideas to fly swarms of dozens of spacecraft are presently being debated. Such a fleet would allow coverage of multiple spatial scales simultaneously. Coupled with global measurements, this would improve our understanding of the complex linkage between the various processes and regions, an important goal in its own night, but also a key in understanding space weather.

Acknowledgment

The author is indebted to his colleagues in the Cluster community for their help in putting this lecture together.

The author

Götz Paschmann was born in 1939 in Schwelm/Germany. He studied physics, first at the Universität Tübingen, then at the Technische Universität München, where he got his Diploma in 1965. Fascinated by the emerging field of space science, he asked Prof. Lüst, then Director of the newly founded Max-Planck-Institut für extraterrestrische Physik (MPE) in Garching for a PhD thesis opportunity. The theme was related to Earth's Van Allen belts and depended on data to be obtained from a sounding rocket flight into these radiation belts. When this failed, he was sent to the Lockheed Palo Alto Research Laboratory, where he worked in R. G Johnson's group on auroral particle precipitation. Back at MPE, he then wrote his thesis on this subject, under the supervision of G. Haerendel and got his PhD from the Technische Universität München in 1971.

At MPE G. Paschmann worked on instrument development and on the analysis of data from ESA's Heos2 spacecraft on the polar regions of Earth's magnetosphere. He then joined forces with S. J. Bame and his group at the Los Alamos Scientific Laboratory to build plasma detectors for the NASA/ESA ISEE mission launched in 1977. The discovery of high-speed plasma jets at the magnetopause as the "smoking gun" evidence of magnetic reconnection is one of the highlights of this work. Later, G. Paschmann was involved in further magnetosphere research missions like the ill-fated Firewheel project and the highly successful AMPTE mission.

In 1982, G. Paschmann was among the group of European scientists that proposed to ESA a mission called Cluster, intended to explore the magnetosphere with a fleet of four spacecraft flying in close formation. The article in this issue of Spatium describes Cluster's long history, including its failure in 1996 and eventual recovery, ending in the successful launch in the summer of 2000. G. Paschmann is the Principal Investigator for the Electron Drift Instrument on Cluster .

At the time of the failure of the first Cluster attempt, G. Paschmann had already started his involvement with ISSI, initially as the leader of a team that wrote a book on Analysis Methods for Multi-Spacecraft Data that has become the reference for much of the Cluster data analysis. In July 1999 he became a Director at ISSI in Bern, where he now spends 50% of the time, and the other 50% still at MPE in Garching.

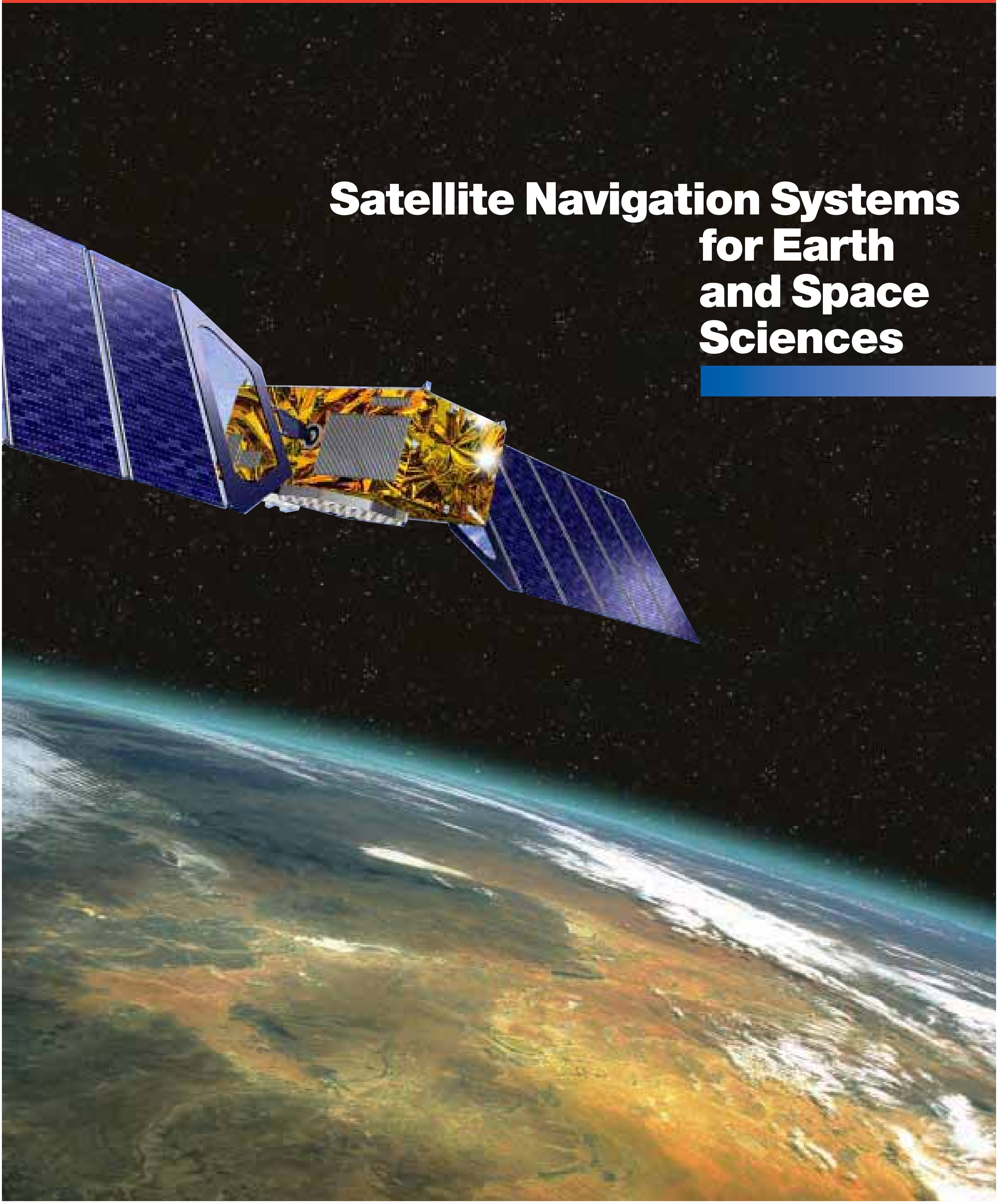
INTERNATIONAL SPACE SCIENCE INSTITUTE

No 10, June 2003

SPATIUM

Published by the Association Pro ISSI

Satellite Navigation Systems for Earth and Space Sciences



Editorial

It was in 1915 when the German scientist Alfred Wegener published the "The Origin of Continents and Oceans" outlining the concept of continental drift for the first time. According to this theory, all the continents on Earth had formed from one single mass about 300 million years ago, called Pangaea. Later on, Pangaea had split, and its pieces had been moving away from each other ever since. Where such continental plates collide, mountain ranges may be pushed up, like for example the Alps that are the collision product of the African and the Eurasian plate.

Initially, Wegener's theory was strongly contested not least because he did not provide convincing explanations for the mechanics allowing for the movement of the continents through the oceans (as it was seen at that time). Today we know that below the continental plates there is a hot liquid zone, the asthenosphere, on which the continents float. The plates are separated by ridges, where fresh magma spills out from the asthenosphere providing the thrust to the continental plates moving them apart thereby.

In the meantime such different scientific branches as geology, palaeontology, botany and zoology have provided ample proof in support of Wegener's theory. Even space technology has contributed by the direct measurement of the continental drift. As an example, the drift velocity between America and Europe has been determined to be in the order of 2.5 centimetres/year.

Global satellite navigation systems are based on a constellation of Earth orbiting spacecraft emitting signals with precise orbital and time data. Suitable receiver equipment combines the signals from at least four spacecraft yielding the time and the three space coordinates. The US Global Positioning System (GPS) for instance consists of 24 spacecraft orbiting the Earth in six different planes. The Russian Glonass system as well as the planned European Galileo system are based on the same concepts.

Of course, the GPS originally was not conceived as a scientific tool but much more as a sophisticated military infrastructure for the fulfilment of navigational tasks. It was only after its commissioning that its scientific value became apparent. The continental drift velocities measurement is just one of a number of fascinating scientific applications of global navigation systems to which the present issue of Spatium is devoted. We are greatly indebted to Prof. Gerhard Beutler, Astronomical Institute, University of Berne for his kind permission to publish herewith a revised version of his exciting lecture of November 12, 2002 for the members of our association.

Hansjörg Schlaepfer
Zürich, June 2003

Impressum

SPATIUM
Published by the
Association Pro ISSI
twice a year

INTERNATIONAL SPACE SCIENCE INSTITUTE

Association Pro ISSI
Hallerstrasse 6, CH-3012 Bern
Phone +41 (0)31 631 48 96
Fax +41 (0)31 631 48 97

President
Prof. Heinrich Leutwyler,
University of Bern
Publisher
Dr. Hansjörg Schlaepfer,
legenda schläpfer wort & bild,
Winkel
Layout
Marcel Künzi, marketing · kommunikation, CH-8483 Kollbrunn
Printing
Druckerei Peter + Co dpc
CH-8037 Zurich

Front Cover: The Galileo navigation satellite in an artist's view (credit: European Space Agency ESA)

Satellite Navigation Systems for Earth and Space Sciences *)

Gerhard Beutler, Astronomical Institute, University of Bern

Navigation and Science in the Past and Today

The introduction to Peter Apian's *Geographia* from 1533 in **Figure 1** nicely illustrates that *positioning* in the "good old times" in essence meant measuring angles – the scale was eventually introduced by one known distance between two sites (as indicated by the symbolic measurement rod in the centre of the wood-cut).

Figure 1
Geographia by Peter Apian, dated 1533.

Figure 1 also shows that relative local and absolute positioning was performed with the same instruments, the so-called cross-staffs, in Apian's days. Global positioning meant the determination of the observer's *geographical latitude* and *longitude* (relative to an arbitrarily selected site – first Paris, then Greenwich was used for this purpose). The latitude of an observing site was easily established by determining the elevation (at the observer's location) of the Earth's rotation axis, approximately represented by the polar star. In principle, longitude determination was simple, as well: One merely had to determine the *time difference* (derived either from the Sun [local solar time] or from the stars [sidereal time]) between the unknown site and Greenwich. The problem resided in the realisation of Greenwich time at the observing site in the pre-telecommunication era. One astronomical solution to this problem, illustrated in **Figure 1**, consisted of measuring the so-called lunar distances (angles between bright stars and the Moon). With increasing accuracy of the (prediction of the) lunar orbit the angular distances between the Moon and the stars could be accurately predicted and tabulated in astronomical and nautical almanacs with Greenwich local

*) Pro ISSI lecture, Bern, 12th November 2002

time as argument. Lunar distances were used for centuries for precise positioning. For navigation on sea the method became eventually obsolete with the development of marine chronometers, which were capable of transporting accurately Greenwich time in vessels over time spans of weeks. **Figure 2** shows the first chronometer developed by the ingenious British watchmaker J.Harrison (1693–1776).

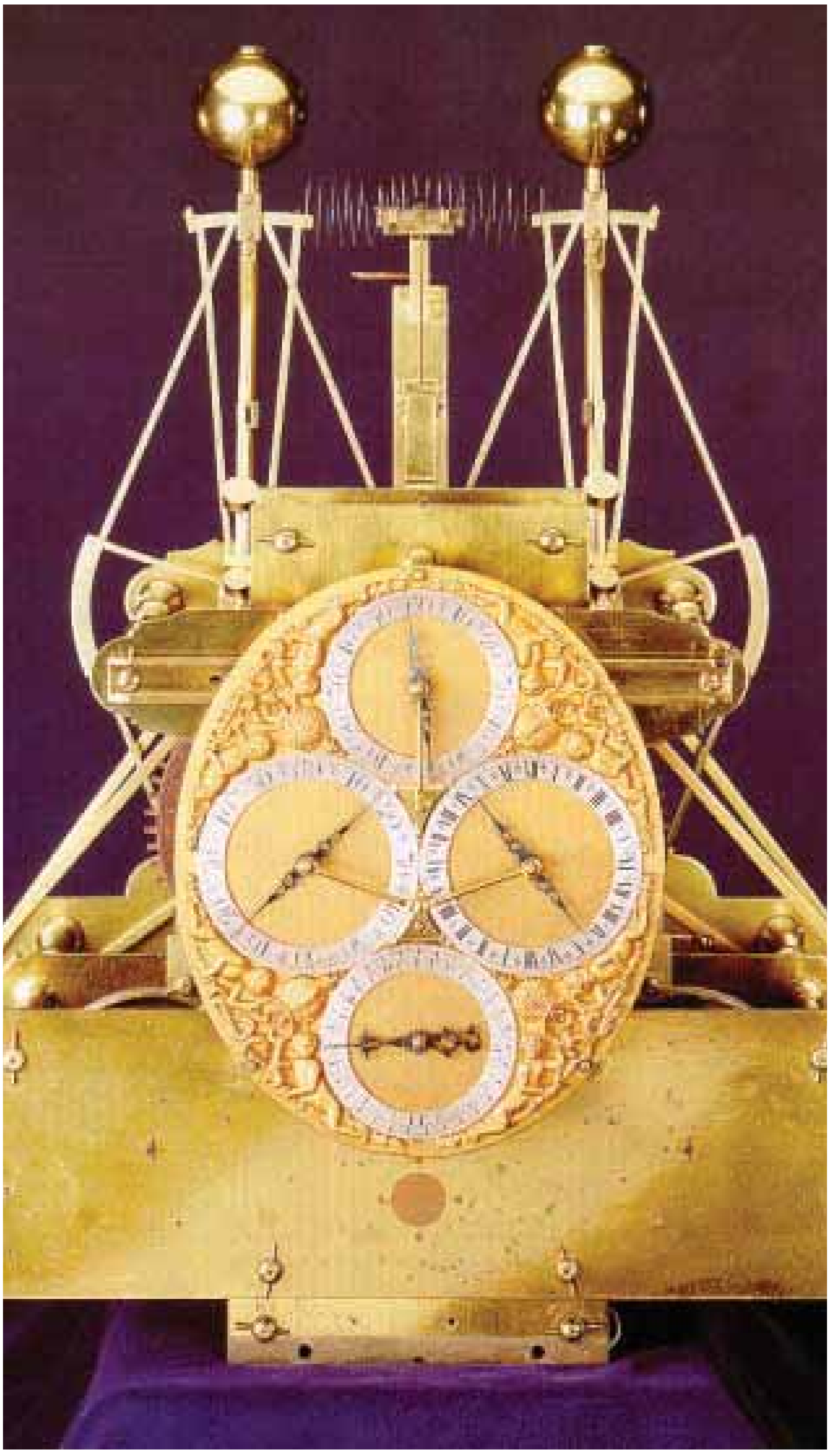


Figure 2
Harrison I, first marine Chronometer (credit: D. Sobel und W.J.H. Andrewes: Längengrad, Berlin Velag 1999).

The *principles* of precise global positioning and precise navigation remained in essence the same from Apian's times till well into the second half of the 20th century. The development of the accuracy was dramatic: The cross-staff was replaced by increasingly more sophisticated optical telescopes. More precise star catalogues (fundamental catalogues) were produced and the art of predicting the motion of planets was developed in analytical celestial mechanics. A long list of eminent astronomers, mathematicians, and physicists, from L. Euler (1707–1783), P. S. de Laplace (1749–1827), to S. Newcomb (1835–1909) were steadily improving the ephemerides. Highly precise pendulum clocks and marine chronometers allowed it eventually to time-tag the observations in the millisecond accuracy range.

The relationship between science on one hand and precise positioning and navigation on the other hand were truly remarkable: The discipline of *fundamental astronomy* emerged from this interaction between theory and application. In fundamental astronomy one defines and realises the global terrestrial and the celestial reference systems *including* the transformation between the systems. The terrestrial system was realised by the geographical coordinates of a network of astronomical observatories. Until quite recently the celestial reference system was realised through fundamental catalogues of stars. Celestial mechanics, so to speak the fine art of accurately describing the motion of celestial bodies in the celestial reference systems, is also part of fundamental astronomy.

The establishment of the transformation between the two systems implies the monitoring of Earth rotation in inertial space. **Figure 3** illustrates that the rotation axis of the Earth moves in inertial space.

It is well known that the rotation axis approximately moves on a straight cone inclined by 23.5° w.r.t. the pole of the ecliptic, an effect known as precession, which was already discovered in the Greek era (and usually attributed to the great Greek astronomer Hipparchos). This motion is not fully regular but shows short-period variations, which is why the astronomers make the distinction between precession and nutation. A study of ancient solar eclipses revealed eventually that the length of day was slowly (by about 2 msec per century) growing. The Earth axis also moves on the Earth's surface, an effect known as polar motion. This and other discoveries related to Earth rotation made in the era of optical astronomy are summarised in **Table 1**.

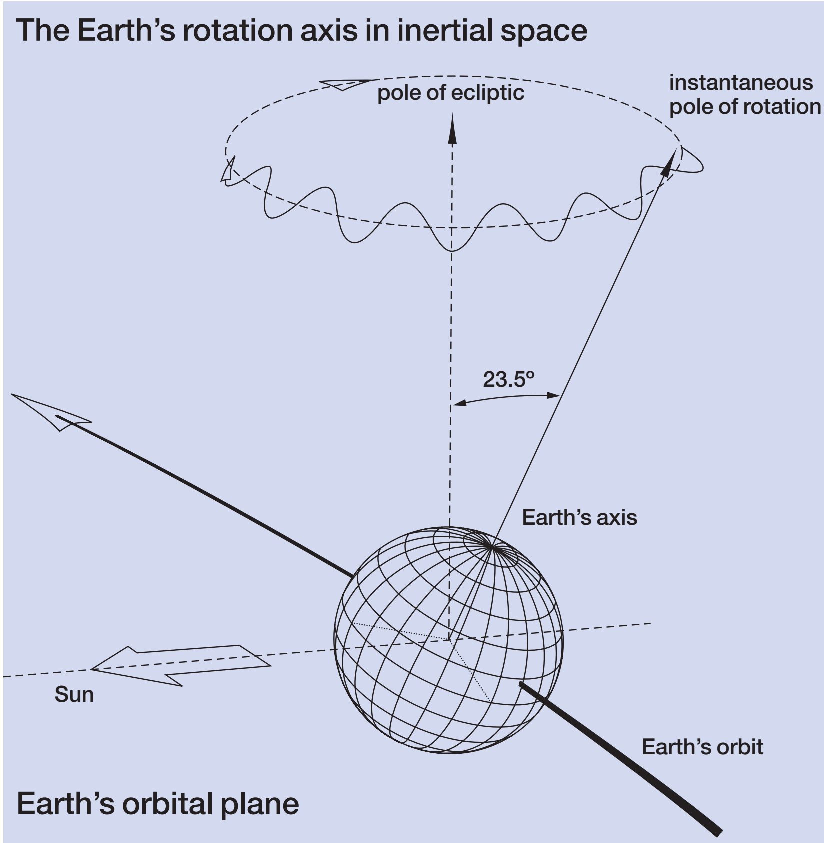


Figure 3
Precession and Nutation.

Year	*Discoverer*	*Effect*
300 B.C	*Hipparchos*	*Precession in longitude (50.4"/y)*
1728 A.D.	*J. Bradley*	*Nutation (18.6 years period, amplitudes of 17.2" and 9.2" in ecliptical longitude and obliquity, respectively)*
1765 A.D.	*L. Euler*	*Prediction of polar motion (with a period 300 days)*
1798 A.D.	*P.S. Laplace*	*Deceleration of Earth rotation (length of day)*
1891 A.D.	*S.C. Chandler*	*Polar motion, Chandler period of 430 days and Annual Period*

Table 1
Discoveries related to Earth rotation in the optical era of fundamental astronomy

The Global Positioning System (GPS)

Sputnik-I, launched on October 4 of the International Geophysical Year 1957, did broadcast a (relatively) stable radio-frequency f. The relative Doppler shift $\Delta f/f$ of such signals is proportional to the radial component v of the relative velocity between satellite and receiver, i.e., $\Delta f/f = v/c$, where c is the speed of light. Actually, this relationship is only the non-relativistic approximation of the more elaborate relativistic expression.

The Doppler shift of radio signals emitted by satellites may be easily measured by standard radio equipment. As the orbital velocity of a low Earth orbiter (LEO) is rather big (about 7 km/sec), the radial velocity is expected to vary roughly between the limits ±5 km/s. One may therefore expect to reconstruct rather accurately the orbit of the satellite using the Doppler shift of the signal, as measured by a few receivers located at known positions on the surface of the Earth. If, on the other hand, the orbit is assumed known, the measurement of the Doppler shifted signal recorded at an unknown location may be used to determine its position on the Earth. The first generation of navigation satellite systems, e.g., the NNSS (U.S. Navy Navigation Satellite System) was based on these simple principles. The advantages over optical navigation were considerable: the observations were weather-independent and available in digital form from the outset. The system worked remarkably well. Using the observations of one pass of one satellite over a receiver (duration about 5–10 minutes), the (two-dimensional) position of the observer could be established with an accuracy of a few ten meters. Apart from the limited precision the most serious disadvantage of the system had to be seen in the limitations of the system for high-speed navigation (the necessity to collect observations over several minutes ruled out the use of the NNSS for air- or space borne applications).

The second-generation satellite navigation systems cured this problem. They are based on

- the measurement of the propagation times of radio signals between the satellites and the observer and on
- the simultaneous observation of several satellites.

Neglecting propagation delays caused by the atmosphere and assuming that all clocks (those of satellites and receivers) are synchronised, the known numerical value of the speed of light c allows it to calculate (from the signal travelling time) the distance ρ between the satellite S at signal emission time and the receiver R at signal reception time. The receiver R thus must lie on a sphere with centre S and radius ρ. If three satellites are observed simultaneously, the receiver's position is obtained by intersecting three spheres with known centres and radii. (One may remember from school that in general two solutions emerge from the intersection process, but one easily recognises that one solution – lying far above the Earth's surface – may be ruled out). As one wishes to use rather cheap oscillators in the receivers, one cannot assume that the receiver clock error is zero (or known a priori), but that it has to be estimated together with the three coordinates of the receiver. The unknown position of R may thus be found by intersecting three hyperboloids with foci (S_1, S_2), (S_1, S_3), (S_1, S_4), and the corresponding differences of the semi-major axes $(\rho_1 - \rho_2)$, $(\rho_1, -\rho_3)$, $(\rho_1 - \rho_4)$. If a satellite system shall be used for positioning and navigation in this sense, one therefore has to make sure that at any time and for any location of the receiver R on the Earth's surface at least four navigation satellites are above the observer's horizon. This implies that the navigation satellites are rather high above the Earth's surface, that the inclinations of

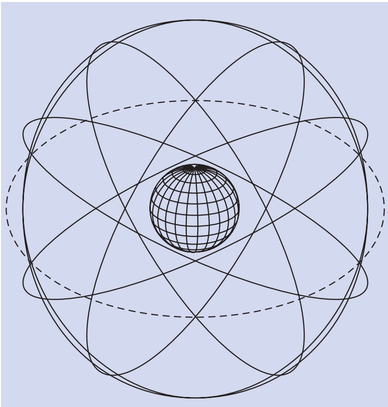


Figure 4
GPS constellation

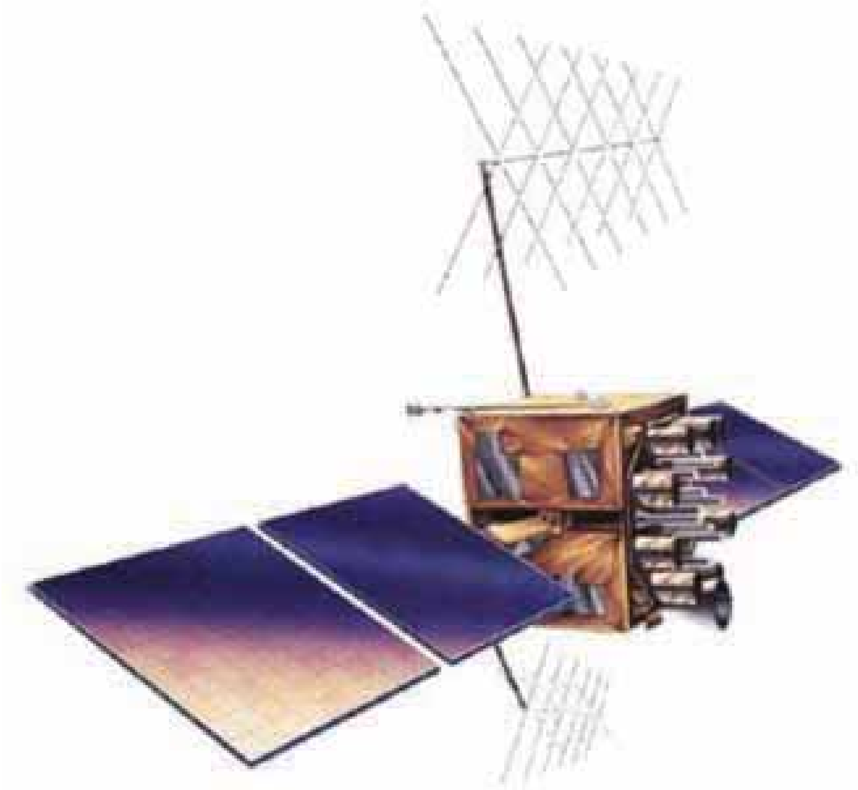


Figure 5
GPS IIR spacecraft (credit USAF).

their orbital planes w.r.t. to the equatorial plane are rather big, and that the satellites are well separated (equally spaced) in the orbital planes.

The best known of these second generation satellite systems is the fully deployed U.S. Global Positioning System (GPS). **Figure 4** illustrates the GPS configuration as seen from a latitude of 35° from outside the actual configuration. The constellation consists of six orbital planes, each inclined by 55° w.r.t. the equator and separated by 60o in it. The satellite orbits themselves are almost circular with radii of about 26.500 km – giving rise to revolution periods of half a sidereal day. There are (at least) four satellites, well separated from each other, in each orbital plane.
The satellites transmit information on two L-band wavelengths, L1 and L2, with wavelengths of $\lambda_1 = 19$ cm, $\lambda_2 = 24$ cm. **Figure 5** shows a GPS satellite. Obviously, the antenna array always has to point towards the centre of the Earth. The solar panels axes have to be perpendicular to the line Sun–satellite at any time. The attitude emerging from these requirements is actively maintained using momentum wheels within the satellite's body.

Alternative systems are the Russian GLONASS (Global Navigation Satellite System) or the European GALILEO system, to be deployed by the European Space Agency (ESA) in the first decade of the third millennium. All systems are based on the same principles.

In order to fully appreciate the results presented subsequently we have to inspect the actual observables of second-generation navigation satellite systems in more detail. The signal transmitted by satellite S at time t^S (satellite clock reading) contains this transmission epoch. This information may be used by the receiver R to compute the so-called pseudorange p using the reading of the receiver clock at signal reception time t_R:

$$p = c\,(t_R - t^S) \quad (1)$$

In the "best possible of fundamental-astronomical worlds" p would just be equal to the geometric distance ρ. Because neither the satellite nor the receiver clocks are exactly synchronised with atomic time, because the signal has to travel through the Earth's atmosphere, and because of the measurement error the actual relationship between the pseudorange and the distance ρ reads as

$$p = \rho + c\Delta t_R - c\Delta t^S + \Delta\rho_T + \Delta\rho_I(\lambda) + \varepsilon \quad (2)$$

where ρ is the distance between satellite S and receiver R, Δt_R the receiver clock error, Δt^S the satellite clock error, $\Delta\rho_T$ the propagation delay due to the electrically neutral atmosphere, $\Delta\rho_I$ the delay due to the ionosphere (caused by the free electrons in layers between 200 km and 1200 km height), and ε the measurement error. For the code measurements the error ε is of the order of a few decimetres allowing for real time positioning with about one meter accuracy. For scientific purposes the phase-derived pseudorange p' is extensively used. The receiver R generates this observation by counting the incoming carrier waves (integers plus fraction part) and by multiplying the result by the speed of light c. This observable is related to the geometrical and atmospheric quantities by

$$p' = \rho + c\Delta t_R - c\Delta t^S + \Delta\rho_T - \Delta\rho_I(\lambda) + N\lambda + \varepsilon' \quad \mathbf{3}$$

Obviously, the code- and phase-derived pseudoranges are intimately related. The key difference resides in the fact that the phase measurement error ε' is very small, of the order of few millimetres, whereas the code-derived measurement error ε is about a factor of 100 larger. This positive aspect is somewhat counterbalanced by the term $N\lambda$, the initial phase ambiguity term, which is unfortunately unknown. The parameter N, known to be integer, has to be estimated as a real-valued parameter. Under special conditions it is possible a posteriori to assign the correct integer number to the estimated real-valued unknown. Note that the ionospheric signal delay in **equation 2** is replaced by a phase advance in **equation 3.**

The term "best possible of fundamental-astronomical worlds" was coined to characterise a world, in which clock corrections and atmosphere delays do not exist. The following results will show that this label is not really justified when interested in other than geometrical aspects. The clock- and atmosphere-related "nuisance" terms are, as a matter of fact, exploited to synchronise clocks worldwide and to describe the Earth's atmosphere. The limited space allows us only to present one of these aspects (the derivation of ionosphere models from a worldwide net of GPS receivers). All the other results are related to the term ρ, the distance between satellite and receiver. The satellite orbits, the positions (and tectonic motions) of the observing sites, and the Earth rotation parameters are derived from ρ.

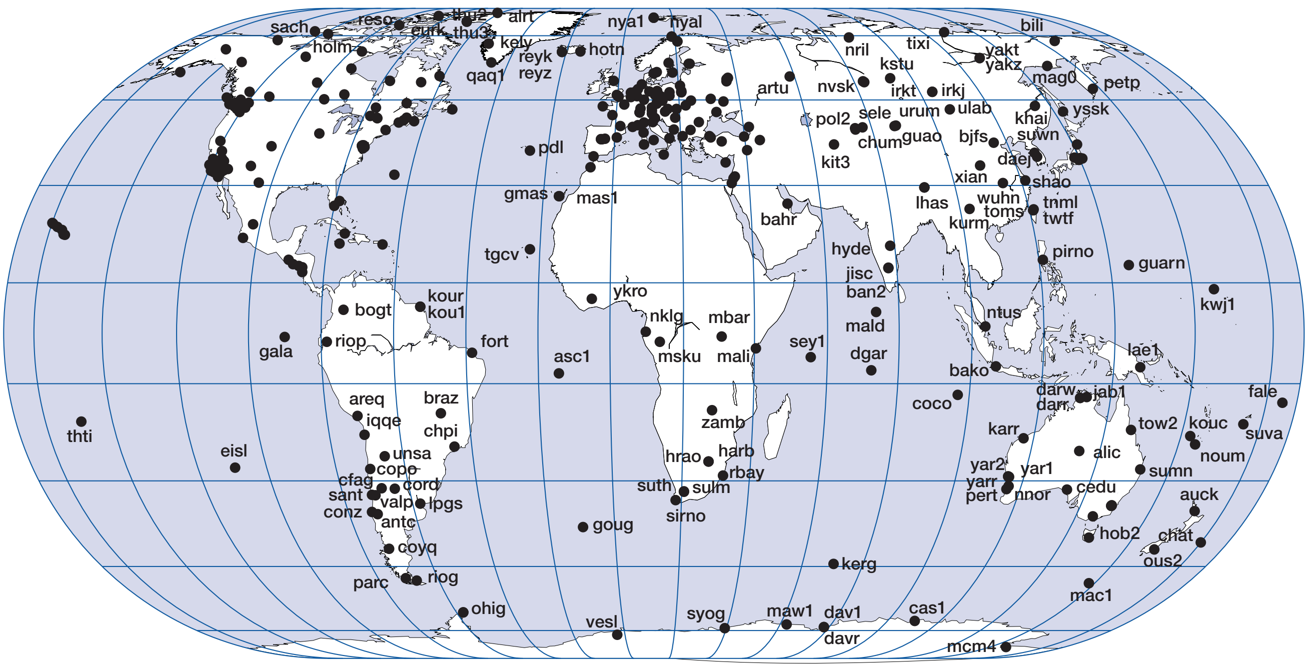


Figure 6
The IGS network in March 2003 (credit: IGS Central Bureau; JPL; Pasadena, California.

The International GPS Service (IGS) and the Code Analysis Centre

The International GPS Service (IGS) was created in 1991. It became fully operational, after a pilot phase of about two years, on January 1, 1994. The IGS is based on a voluntary collaboration of scientific and academic organisations in the fields of geodesy, space science, and fundamental astronomy. Big research organisations like NASA, JPL, ESA, etc. contribute as well as National Geodetic Survey institutions, and universities. The IGS deployed and maintains a network of more than 200 globally distributed permanent GPS sites. **Figure 6**, taken from the IGS homepage, gives an impression of the network.

The IGS network reflects to some extent the fact that the IGS is a voluntary collaboration: The station distribution is far from homogeneous, but one can also see that there are only few "empty" areas left today. Each IGS station tracks each GPS satellite in view. At least once per 30 seconds all available code and phase measurements on the two frequencies are stored – giving rise to 2 x 8640 measurement epochs per day. At least on a daily turnaround cycle the raw data are sent (via internet or telephone modems) to regional and global data centres, from where they can be retrieved by the wider scientific community, but also (perhaps more importantly) by the IGS Analysis Centres to generate the IGS products. There are currently seven IGS Analysis Centres (three in the USA, one in Canada, two in Germany (where the one at ESA should be considered as multi-national) and one, called CODE, situated at the University of Bern. CODE is a joint venture of the University of Bern's Astronomical Institute (AIUB) with the Swiss *Bundesamt für Landestopographie* (Swisstopo), the German *Bundesamt für Kartographie und Geodäsie* (BKG) and the French *Institut Géographique National* (IGN). The work of the IGS Analysis Centres (AC) is truly remarkable. Every day, a table of geocentric orbital positions is generated by each of the AC's for each active GPS satellite with a spacing allowing it to reconstruct the satellites' positions and velocities with cm-precision for any time argument within the day. These satellite orbits originally were the primary IGS product. They are a big relief for research and production organisations using the GPS for high-accuracy surveys. It is, e.g., worth mentioning that since 1991 the Swiss first order geodetic survey is uniquely based on GPS and on the IGS products. A similar development was observed in most other countries worldwide.

The IGS Analysis Coordinator is first comparing, then combining the orbits of all IGS Analysis Centres into one official IGS orbit. The work of the individual centres is rated by the root mean square error per satellite position (actually per geocentric coordinate) w.r.t. this official, so called final orbit, available about one week after real

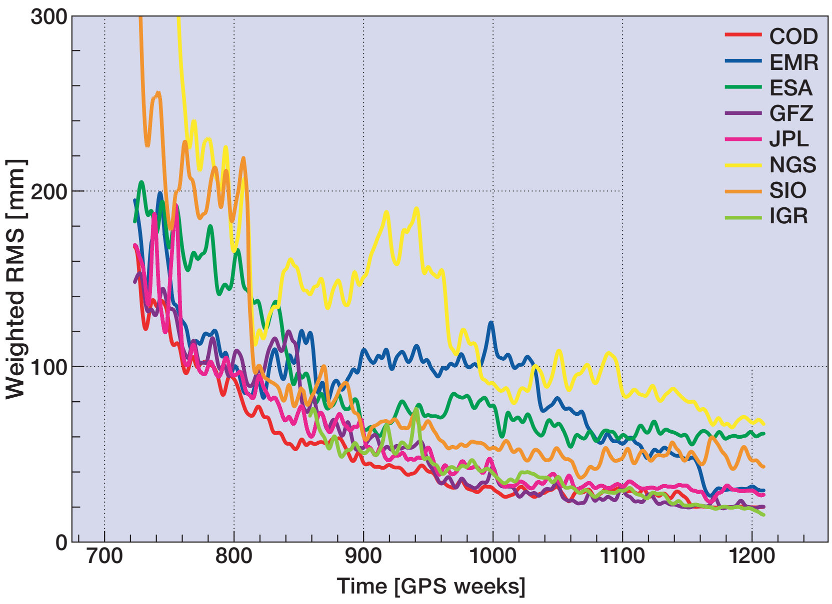


Figure 7
RMS-error per satellite coordinate for IGS-AC-orbits.

time. **Figure 7** shows these root mean square errors for each centre since 1994 to March 2003. The consistency of the best contributions is today of the order 2–3 cm. Not without pride we note that the CODE-contribution (red curve) always was among the best since the advent of the IGS. (Note that the green curve labelled "IGR" characterises the combined IGS rapid solution, which is already available within one day after "real time".) Elementary geometric considerations show that this orbit accuracy of few cm is sufficient for millimetre precision positioning on the Earth or in Earth-near space (e.g., for LEO's). It is a remarkable IGS achievement that the GPS orbits may be considered virtually "error-free" for most applications.

Scientific Applications

In 1991 it was the intention to base the IGS orbit determination on the Earth rotation parameters calculated by the (at that point in time) well established space techniques VLBI (Very Long Baseline Interferometry) and Satellite Laser Ranging (SLR). It turned out that the required information was not available in due time for IGS orbit production. This is why each IGS Analysis Centre has to solve for the daily position (and velocity) of the Earth rotation pole. **Figure 8** shows the daily estimates of polar motion by the CODE Analysis Centre since 1993. Note that the units are arc seconds, implying that the daily estimates of the Earth's rotation axis are accurate to about 2–3 mm.

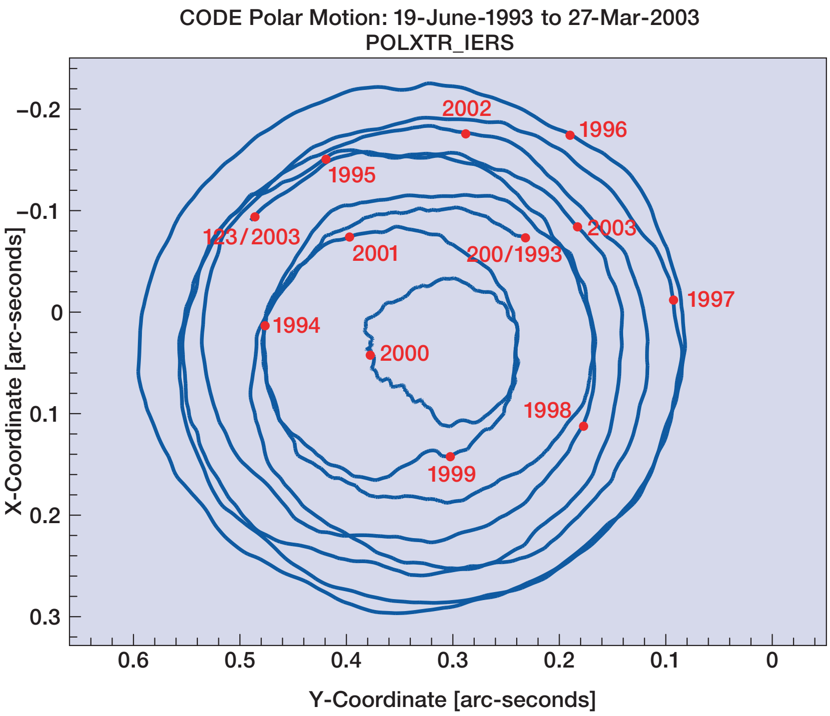


Figure 8
Daily estimates of polar motion by CODE since 1993.

Polar Motion

One arc second corresponds to about 30 m on the Earth's surface. The motion of the pole between 1993 and 2003 thus roughly takes place within a circle of about 7 m centred roughly at the endpoint of the Earth's axis of the maximum moment of inertia. Bigger excursions may occur occasionally. A spectral analysis of polar motion reveals two dominating periods, one of 430 days (amplitude of about 0.17″), the so-called Chandler period, and one of one year (amplitude of about 0.08″). The Chandler period is related to the motion of a solid (not completely rigid) body. The period may be derived from the numerical values of the Earth's principal moments of inertia and from the elastic properties of the Earth. The annual period originates from the interaction between the solid Earth and the atmosphere (and oceans).

The superposition of the Chandler and annual periods results in a beat period of about six years letting the pole move on a ("bad") circle with a radius varying between 2 m and 8 m. Many short-period variations show up in **Figure 8.** Most of them are related to the exchange of angular momentum between atmosphere and solid Earth.

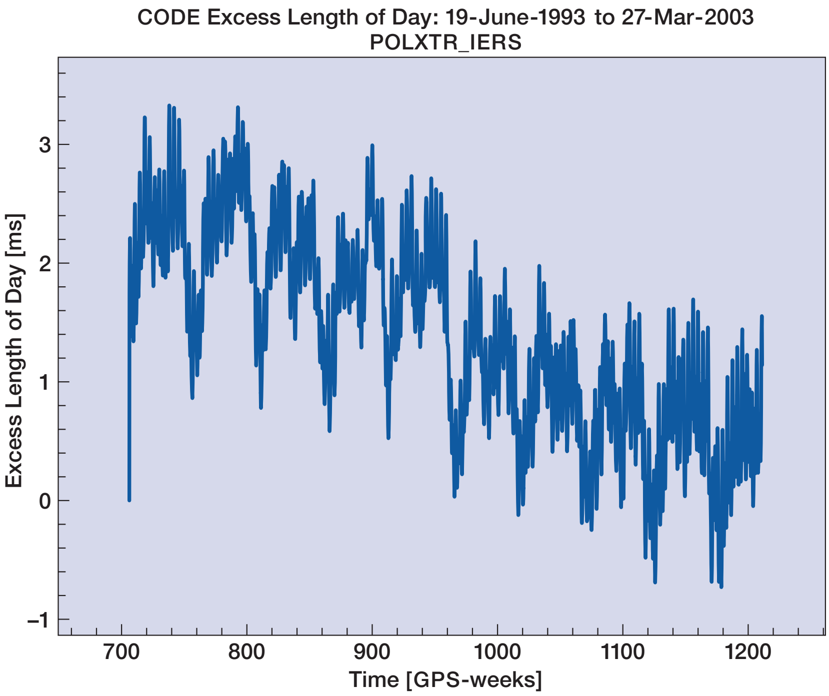


Figure 9
Length of day variations 1993–2003.

Length of Day

In inertial space (i.e., w.r.t. the stars) the Earth rotates once about its axis in a sideral day, corresponding to about 23^{h} 56^{m} of atomic time. **Figure 9**, emerging from the daily estimates of the CODE AC, shows that the actual length of day is "far" from constant. Short period variations (e.g., of two weeks) are easily explained by the tidal deformation of the Earth due to the Moon (the tidal displacement causes a varying polar moment of inertia, which in turn leads to a varying angular velocity of Earth rotation). Observe that the bi-monthly changes in the length of day (lod) are small (amplitude well below the millisecond), but that they are easily detected in this global analysis. The annual variations with amplitudes of about one millisecond are more interesting.

The phenomenon is explained by comparing the polar component of the solid Earth's angular momentum (which may be derived from the length of day variations) with the so-called atmospheric angular momentum (which is derived from the global meteorological networks, registering pressure, temperature, and wind profiles).

Figures 10a and **10b** compare the relative polar (axial) angular momenta of the solid Earth (red line) and of the Earth's atmosphere. The "true" angular momentum of the solid Earth was used in the case of **Figure 10a**, whereas a low-order polynomial was removed from this time series in **Figure 10b**. The

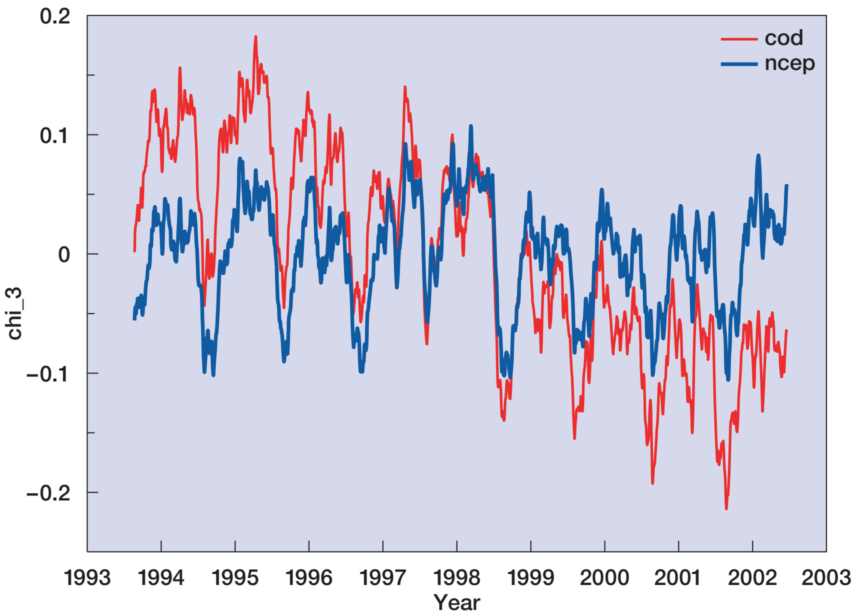


Figure 10a
Polar angular momenta solid Earth and athmosphere.

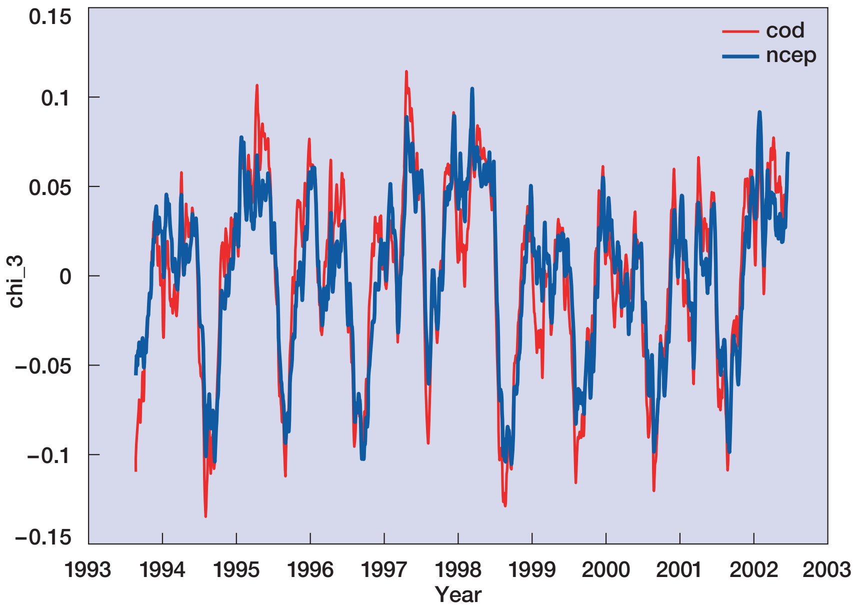


Figure 10b
Polar angular momenta solid Earth (trend removed) and athmosphere.

correlation of the two curves – which are completely independent – is striking in both cases. In **Figure 10b** the correlation coefficient is r = 0.98, implying that the polar component of the system (solid Earth plus atmosphere) is almost perfectly conserved – at least when considering variations with periods of one year or smaller.

Obviously there are variations of longer period ("decadal" variations) in the polar angular momentum **(Figures 10a, 9a)**, which do not occur in the polar component of the atmospheric angular momentum. They are undoubtedly real. These variations are most interesting from the global geodynamics point of view. They may be explained by the fact that the IGS observing sites are attached to the Earth's crust, which is, however, not rigidly attached to the Earth's inner shells (in particular to the fluid outer and the rigid inner core of the Earth). It would be very nice to present pictures corresponding to **Figures 10a,b** proving the exchange of angular momentum with the inner layers, namely mantle, inner and outer core. Unfortunately there are no direct measurements of the angular momenta of these layers. This may be considered as a disadvantage. On the other hand, one has to acknowledge that the study of the decadal lod-variations **(Figure 9)** or of the corresponding angular momentum (red curve in **Figure 10a)** contains very valuable information concerning the rotational behaviour of the Earth's interior. Geodynamical Earth models must explain the long-term development in **Figure 10a**. With the polar motion time series of **Figure 8** it is possible to calculate the other two components (the equatorial components) of the solid Earth's angular momentum and to compare them with the corresponding atmospheric quantities, as well. The correlation between the two series is rather pronounced, but with correlation coefficients around r = 0.7 not nearly as high as in the case of the polar component.

Plate Motion Velocities

The Chandler period of 430 days (instead of 300 days) clearly indicates that the Earth is not rigid. This fact is also supported by **Figure 11** showing that the observing sites are not fixed. The figure is derived from weekly estimates of the coordinates of the tracking stations of the IGS network. It turns out that in the accuracy range achieved by the IGS analyses ([sub]cm accuracy for the daily estimates of the coordinates of the tracking sites) it is not possible to assume that the polyhedron of observing sites is rigid. One has to take the relative motion of the stations into account. **Figure 11** shows the "velocity" estimates for the stations processed daily by the CODE analysis centre (from a long series solutions). Other space-geodetic techniques are of course also capable of determining station velocities. The unique aspect of GPS analyses is the comparatively high density of observing sites. The velocities seem small at first sight. One is inclined to consider velocities of the order of 1–10 cm/year as irrelevant. Nothing could be more wrong: A velocity of 1 cm/year gives rise to a posi-

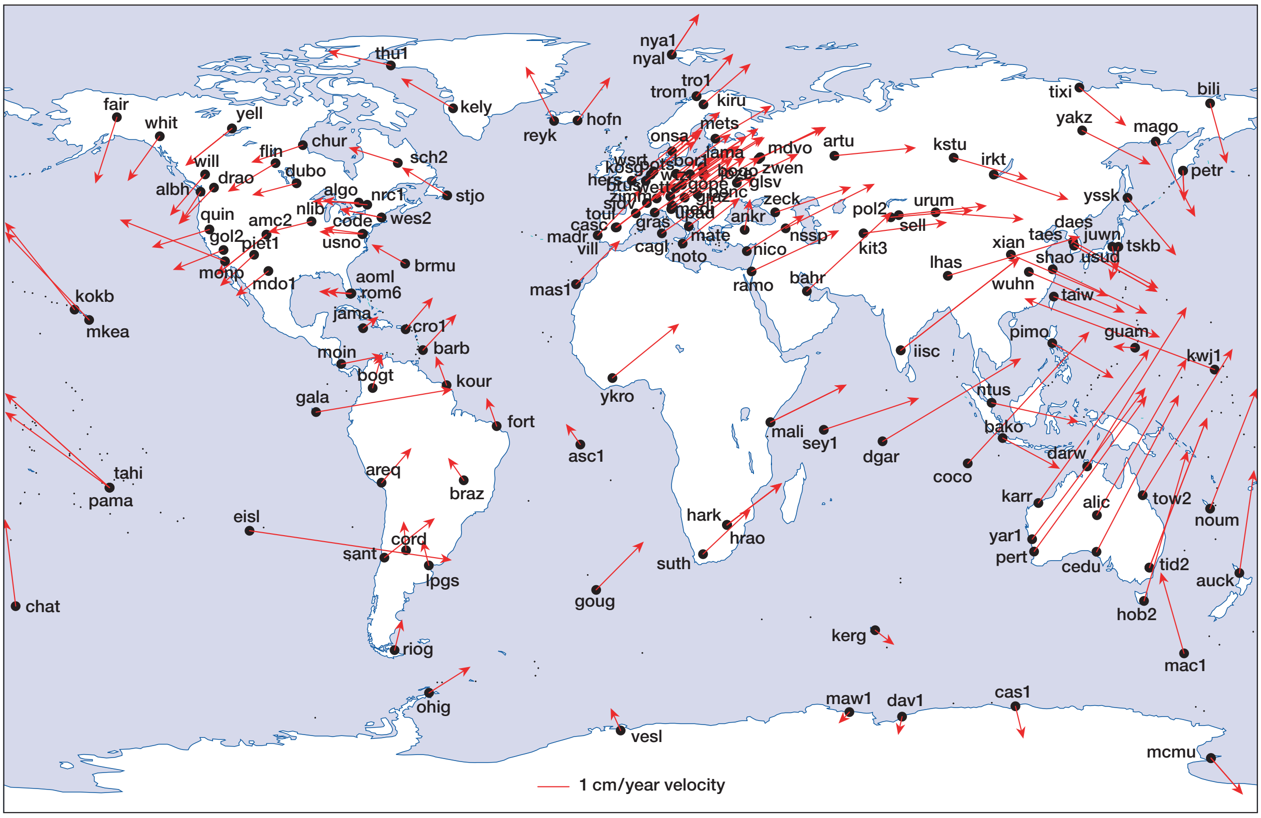


Figure 11
Station velocities estimated by CODE Analysis Center.

tion change of 1000 km in 100 million years! As a matter of fact, these velocities explain the "continental drift" postulated by Alfred Lothar Wegener (1880–1930) in 1915. At Wegener's epoch the continental drift, referred to more appropriately as plate motion, was pure speculation – today we are monitoring it in real time.

Electron Density in the Atmosphere

Let us continue our review of the scientific exploitation of the GPS by an example related to the atmosphere, the determination of the total number of electrons in the atmosphere. The information is extracted from the ionospheric refraction term $\Delta\rho_I(\lambda)$ in **equations 2, 3**. Whereas the results presented so far were obtained by analysing the so-called ionosphere-free linear combination of the L1- and L2-observables (which eliminate the wavelength-dependent ionospheric refraction term), the so-called geometry-free linear combination of the L1- and L2-observations is used for the purpose we have in mind now. This particular linear combination of observations is the plain difference of the **equations** of type **2** or **3** for the L1- and the L2-wavelengths. As all the other terms are wavelength *in*dependent, we obtain a direct observation of ionospheric refraction along the line of sight between observer and satellite by using this difference of observations. The result is proportional to the number of electrons contained in a straight cylinder between observer and satellite. Using the phase and code observations from the entire IGS network one may derive maps of the electron density (the simplest model assumes that all free electrons are contained in a layer in a given height H – usually set to H = 400 km). At CODE, such maps are produced with a time resolution of

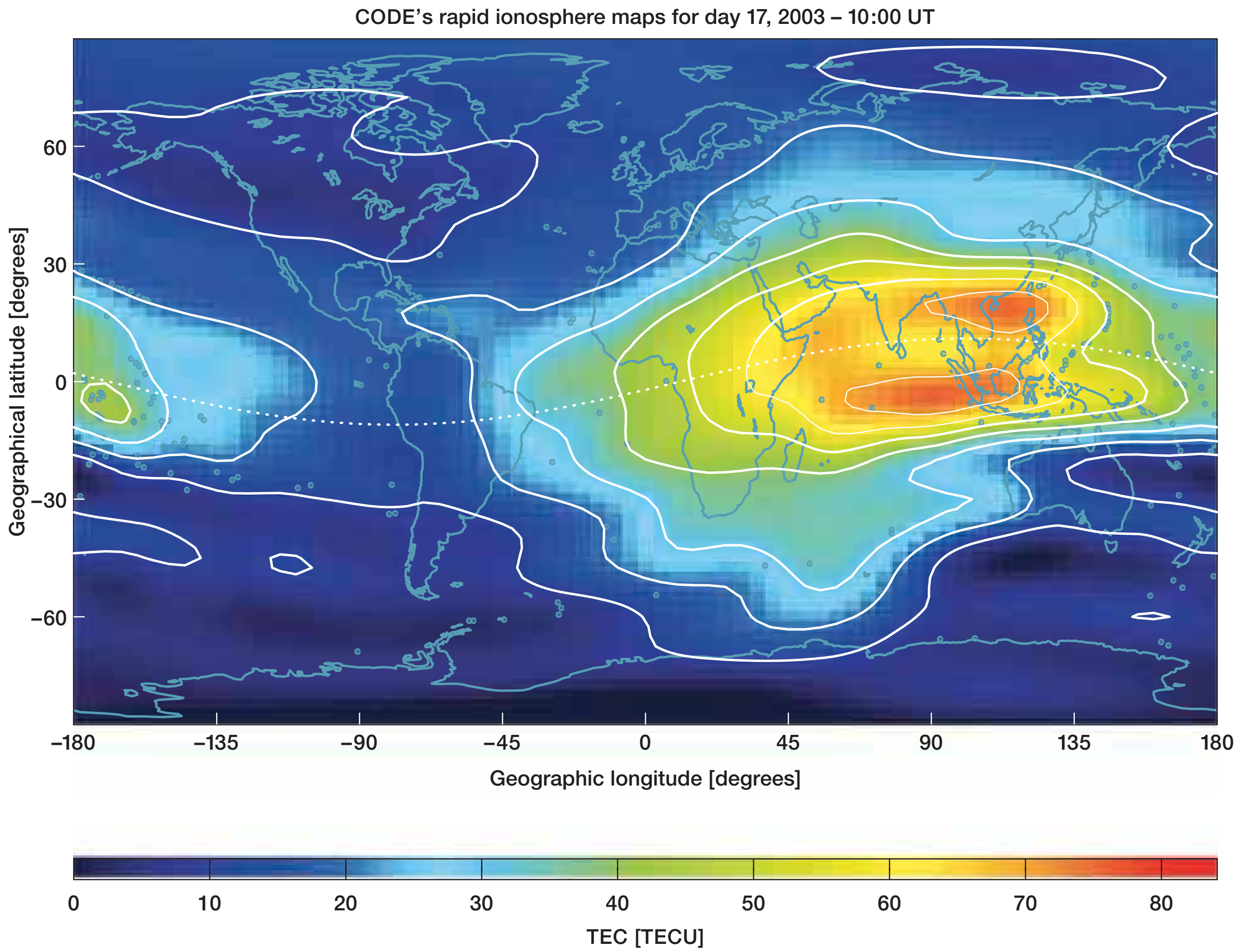


Figure 12
Map of total ionospheric content of electrons (1 TECU = 10^{16} electrons per m^3).

two hours. **Figure 12** shows one of these snapshots. The electron density is given in TECU's (Total Electron Content Units), 1 TECU corresponding to 10^{16} electrons per m^2.

Several mechanisms are responsible for producing free electrons in the Earth's atmosphere. The most important is related to the Sun's UV radiation ionising the molecules in the upper atmosphere. This is why the "spot" of maximum electron density closely follows the projection of the Sun onto the Earth's surface (subsolar point). The maximum is actually lagging behind the subsolar point by about two hours. In **Figure 12** we see a bifurcation of the ionosphere density along the geomagnetic equator (dotted line), an interesting aspect due to the Earth's magnetic field. Internally, the electron density is represented by a spherical harmonics series with about 250 terms. The zero-order term represents the mean electron density. **Figure 13** shows the development of this mean electron density, which may be inspected as a function of time.

This is a good example for solar-terrestrial relationships. A spectrum of the time series underlying **Figure 13** shows that the shortest period corresponds to the rotation

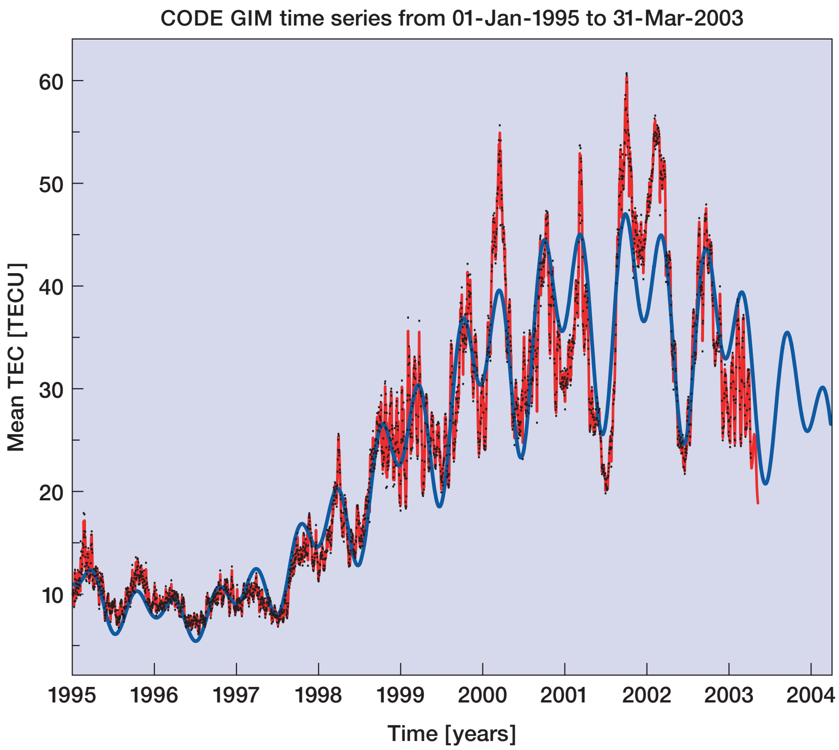


Figure 13
Mean TEC values 1995–2003.

period of the Sun (about 27 days), indicating that the UV radiation is related to the sunspots. Another prominent period is the annual period, which is due to the varying distance between Sun and Earth (because of the Earth's orbital eccentricity of e = 0.016). The longest period corresponds to the 11-year sunspot cycle. It should be pointed out that **Figure 13** probably represents the best history of the density of free electrons. Many more figures of this type might be generated by replacing the zero order term of the development of electron densities by the higher-order coefficients of the harmonic series underlying **Figure 12.**

Determination of the Earth's Gravity Field: Review and Outlook

It would have been fair and appropriate to discuss our knowledge of the Earth's gravity field in the introductory section together with the geometrical properties of the Earth. This was not done because our knowledge of the global aspects of the Earth's gravity field was very poor in the pre-space age: It was of course known that the Earth is, in good approximation, a sphere (meaning that the gravity field is that of a point mass). Thought experiments due to I. Newton and expeditions performed in the 18th century then revealed that the next better approximation for the shape of the Earth was that of a spheroid with a flattening of about $f={}^1/_{300}$ (giving rise to one second-order term of the gravity field, the so-called "dynamical flattening"). (Gravimeter measurements on the Earth's surface indicated that there were significant local and regional variations of the Earth's gravity field, but it was not possible to use these measurements to derive a consistent global gravity field of the Earth).

The analysis of the orbits of artificial Earth satellites, starting in the 1960s, greatly improved our knowledge of the Earth's gravity field. Instead of only two terms (the mass of the Earth and the dynamical flattening) a few thousand terms could be determined with reasonable accuracy in the first era of satellite geodesy. The two Lageos satellites (the acronym standing for Laser GEOdetic Satellite) were paramount for this task. The satellites were designed to minimise the impact of non-gravitational forces (the sphere of 70 cm diameter consists of Uranium) on the orbital motion and to optimise their observation by the Laser technique (Lageos II is shown in **Figure 14**).

Figure 15 shows University of Bern's 1 m telescope at Zimmerwald, which may be used as an optical telescope, but also as transmission and receiving device for Satellite Laser Ranging (SLR). The Laser technique is used to generate very short light pulses (few ten picoseconds). The measurement simply is the light travelling time of the Laser pulse from the observatory to the satellite and back. Due to the divergence of the Laser beam and due to the limited number of reflectors onboard the satellite **(Figure 14)** only few reflected photons may be detected in the observatory. The technique has outstanding properties: The accuracy is high (few picoseconds corresponding to about 1 cm in range) and the atmospheric effects may be modelled with highest accuracy (sub-cm) using only standard meteorological measurements at the sites. The technique has, however, also severe disadvantages: good weather is a precondition (the clouds in **Figure 15** would not allow it to track satellites). Moreover, a Laser observatory is rather bulky and expensive, which is why there exist only about thirty observatories today, which are globally coordinated by the International Laser Ranging Service (ILRS).

The sparseness of the Laser sites makes it therefore impossible to continuously track the orbit of a particular satellite. This situation is not at all ideal for the analysis of satellite orbits.

Figure 14
The Lageos satellite.

This circumstance was one of the key motivations to look for alternative space borne methods to determine the Earth's gravity field. All new methods rely on GPS to determine the orbit of the space vehicle(s) used to determine the Earth's gravity field and on the IGS products to model the GPS orbits and clocks. **Figure 16** shows an artist's view of the German research satellite CHAMP (ChAllenging Minisatellite Payload). Champ was launched in July 2000. It is, as the name implies, a multipurpose satellite, allowing atmospheric sounding, and the determination of the magnetic and the gravity fields. The gravity field is recovered from analysing the satellite's orbits using the GPS. This is possible because the satellite carries a space borne GPS receiver (JPL's blackjack receiver). CHAMP flies at rather low altitudes (from initially 450 km to 300 km at end of the mission in 2005). In view of the bulkiness of the satellite non-gravitational forces, residual air drag in particular, do seriously affect the orbit of the satellite – and therefore also the determination of the Earth's gravity field. CHAMP copes with this problem by so-called accelerometers, in essence a probe mass in the satellite's interior (shielded against non-gravitational forces). The displacement of this probe mass in the satellite-fixed reference frame may be used to "determine" the non-gravitational forces. This is, of course, only possible within certain limits given by the error spectrum of the accelerometers. It turns out that the separation of gravitational and

Figure 15
The Zimmerwald 1-m telescope.

non-gravitational forces is rather good for short periods and rather poor for the longer periods. It is in any case most encouraging to see that one month of CHAMP data gives better results for the short periodic part of the gravity field spectrum than about 40 years of Laser tracking! CHAMP marks the beginning of the new era of gravity field determination.

The two satellites GRACE A and GRACE B (see **Figure 17**) fly in the same orbit, separated by about 20 km. GRACE, launched in March 2002, stands for Gravity Recover And Climate Experiment. Obviously the two satellites are similar in shape to CHAMP. GRACE focuses on the time variability of the Earth's gravity field. For that purpose it determines the Earth's gravity field with a high time resolution allowing it, e.g., to study the seasonal water cycle (evaporation over oceans, precipitation over continents, ground water variability, flowing off into the oceans). It is amazing that such experiments can be performed by studying the Earth's gravity field! The orbits of GRACE A and B are reconstructed exactly like in the case of CHAMP by using the data of space borne GPS receivers. The orbit of GRACE A relative to GRACE B is, however, reconstructed with a so-called K-band link (distance measurements in the microwave range) between the two spacecraft. This K-band link establishes the distance between the two satellites with extreme precision. The measurements may be used to determine directly the second derivatives of the Earth's gravity potential. Both, the satellite orbit established by GPS and the K-band measurements, are used to reconstruct the Earth's gravity field. It is expected that the results will significantly improve our knowledge in a wide spectrum of Earth sciences.

The third of the new generation satellites to determine the Earth's gravity field is ESA's GOCE satellite, shown in **Figure 18** (GOCE standing for Gravity field and steady-state Ocean Circulation Experiment). As the name implies, GOCE is a combined mission to measure gravity and Ocean circulation. GOCE wants to establish the mean gravity field of the Earth with the highest possible accuracy and spatial resolution. When determining the gravity field associated with a mass distribution, one obtains, as a by-product, equipotential surfaces. The equipotential surface at sea level is called the geoide. The geoid to be determined by GOCE will be of greatest importance for the determination and monitoring of ocean circulation. For this purpose it is, in addition, necessary to determine the so-called sea surface topography using the measurements of altimeter satellites, reconstructing the sea surface using the radar echo technique. The difference between the geometrically established sea surface and the GOCE geoid gives the elevation of the sea surface over the geoid. Exactly as in the continents, the water in the oceans has to flow "mountains" to "valleys". It is indeed fascinating to see that sciences which were considered special disciplines for few experts are now developing into one multidisciplinary topic in Earth sciences.

Unfortunately we could only touch a few aspects related to this new era of gravity field determination. For more information we refer to the proceedings of the workshop *Earth Gravity Field from Space – From Sensors to Earth Sciences,* which was hosted by the ISSI in March 2002. The workshop brought together the leading experts in celestial mechanics, geodesy, Earth sciences, and in instrument manufacturing. The proceedings of the workshop underline the tremendous impact of this new sequence of space missions on a very broad field of science.

Figure 16
The CHAMP spacecraft launched in July 2000 (credit Astrium).

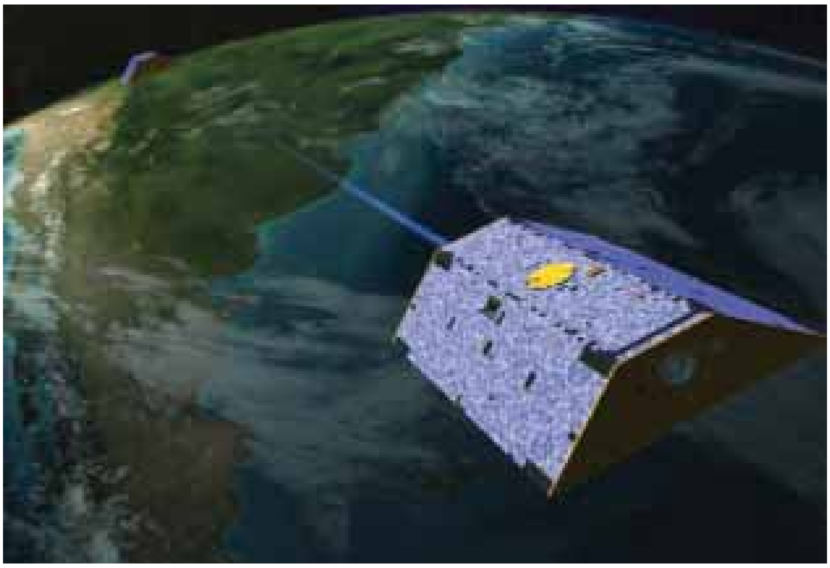


Figure 17
The pair of GRACE spacecraft launched in March 2002 (credit NASA).

The scientific aspects emerging from the new navigation systems are co-ordinated by the International GPS Service (IGS). Polar motion, length of day, plate tectonics, and the Earth's ionosphere are monitored with unprecedented accuracy. The IGS products are also paramount for the success of gravity field determination with the new generation of gravity missions. The new era of gravity field determination will lead to a unification of geometric and gravitational aspects, truly bringing together the three pillars of modern geodesy and fundamental astronomy, namely (1) positioning and navigation, (2) Earth rotation, (3) gravity field determination.

Summary

Navigation and fundamental astronomy are intimately related since hundreds of years. Until about thirty years ago global navigation relied on measuring angles. Scientific discoveries include the phenomena of precession, nutation, secular deceleration of Earth rotation, and polar motion.

With the advent of the space age the astrometric observation of stars and planets was replaced by the measurement of distances (or distance differences) between observers and artificial satellites. As weather-and daytime-independence is an important aspect of navigation, the observation techniques were moreover moved from the optical to the microwave band of the electromagnetic spectrum. In view of the fact that the "old" methods were used for centuries, one can truly speak of a revolution in navigation, geodesy, and fundamental astronomy.

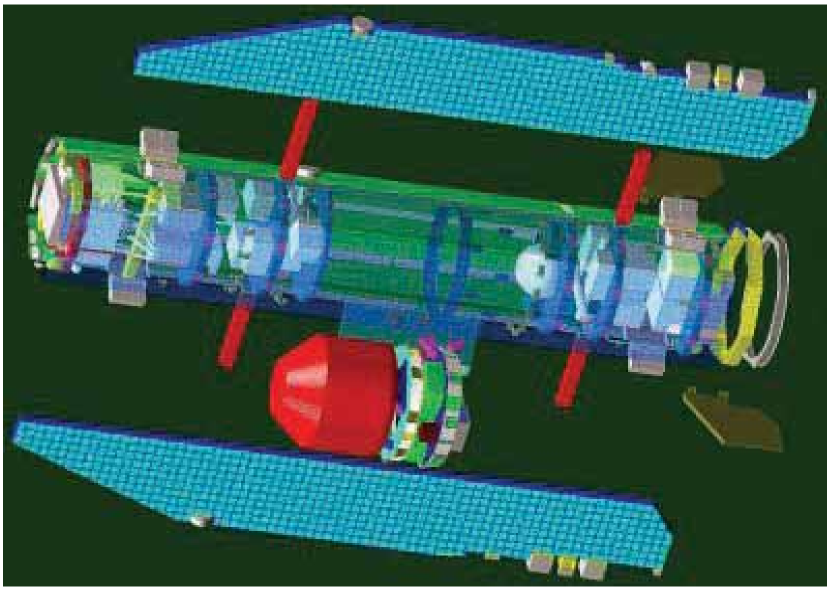


Figure 18
The GOCE spacecraft to be launched in 2006 (credit: ESA).

SPATIUM

The author

Gerhard Beutler was borne in Kirchberg in the Canton Berne in 1946, where he visited the elementary school. After receipt of the type C maturity in 1964 he enrolled at the University of Berne in astronomy, physics and mathematics. These subjects turned out to be the ideal theoretical foundation when it came to exploring the scientific potential of navigation satellites when the U.S. Global Positioning System became operational.

In 1976 G. Beutler received the Ph.D. degree with a thesis on Integral Evaluation of Satellite Observations. The second thesis was devoted to the Solution of Parameter Estimation Problems in Celestial Mechanics and Satellite Geodesy. Later on, he held positions as research associate from 1983 until 1984 at the University of New Brunswick, Fredericton, Canada and from 1984 until 1991 at the University of Berne. During the latter period he was engaged in the development of the Bernese Global Positioning System Software, which is used today at about 150 research institutions worldwide. In 1991 he was elected Professor of Astronomy and Director of the Astronomical Institute of the University of Berne. His lectures cover such fields as celestial mechanics, Earth rotation, stellar dynamics and statistics, parameter estimation theory and digital filtering theory.

The main research interests of Gerhard Beutler are devoted to fundamental astronomy, celestial mechanics, global geodynamics and satellite-based positioning und navigation. He is a member of numerous national and international committees and working groups, as for example the Schweizerische Geodätische Kommission, the American Geophysical Union and the European Space Agency's Scientific Advisory Board on the European Navigation Satellite System Galileo. As of July 2003 he will serve as President of the International Association of Geodesy (IGS).

Based on his initiative the CODE Processing Centre of the International Geodetic Society was created. CODE today is a joint venture of the Astronomical Institute of the University of Berne, the Swiss Federal Office of Topography, the Institute for Applied Geodesy in Germany, and the Institut Géographique National in France.

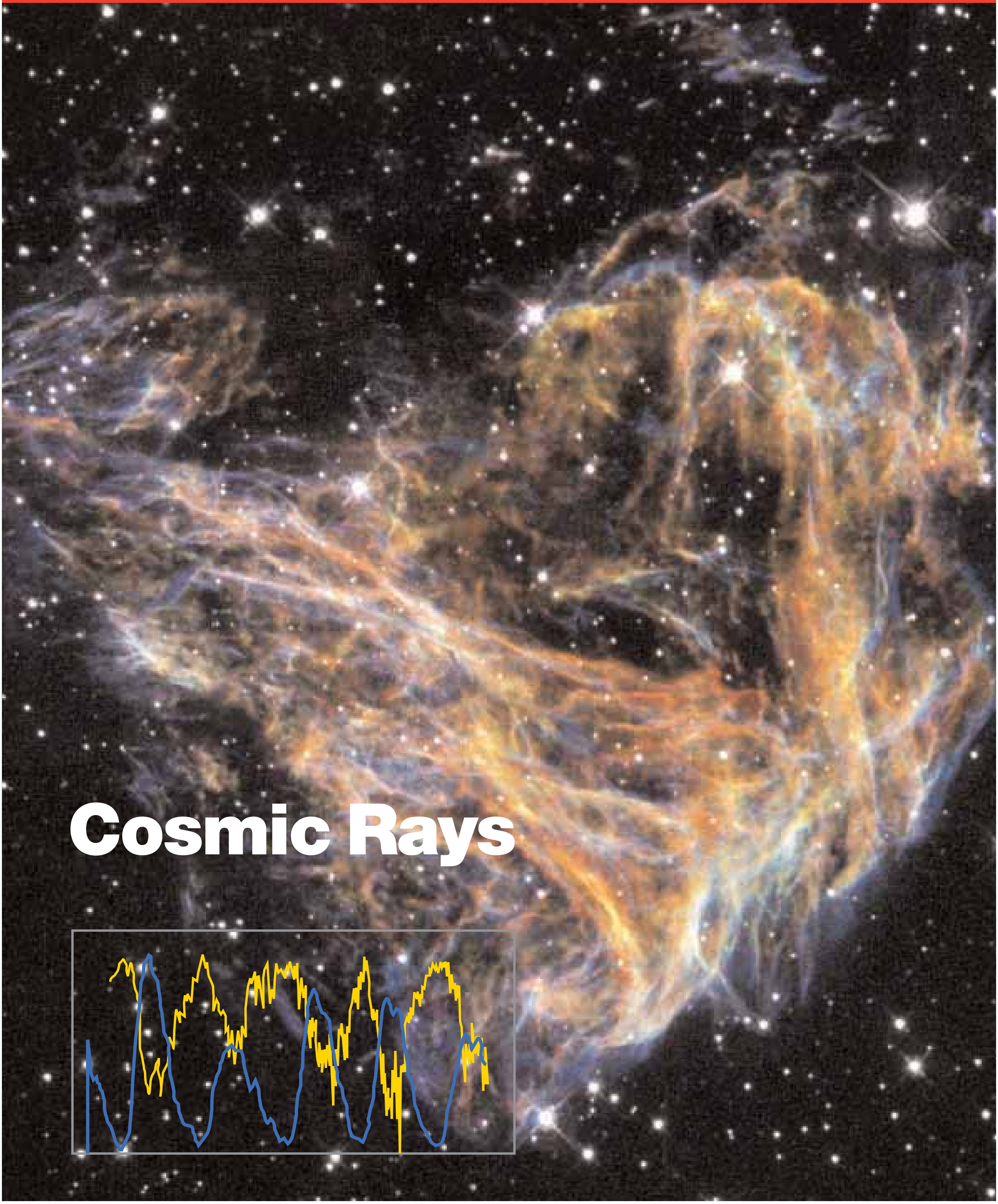
Gerhard Beutler is a fervent supporter of the planned European navigation and positioning system Galileo, not only because this system would make more satellites available but also because its different orbit characteristics as compared to the US GPS system would provide the scientific user with an increased potential of positioning accuracy.

Gerhard Beutler lives with his family in Schüpfen in the Canton Berne. In his free time he likes biking through the wonderful landscapes of the Canton Berne, as well as playing tennis. Not without pride he counts himself to the world top league of the celestial mechanics tennis players.

INTERNATIONAL SPACE SCIENCE INSTITUTE
No 11, November 2003

SPATIUM

Published by the Association Pro ISSI



Cosmic Rays

Editorial

In a deep cave some people have been caught since their early childhood. They are chained down in a way that they even cannot turn their heads around. They can only see shadows on the walls of their inconvenient shelter, which are cast by a fire blazing in the background. The shadows stem from objects of unknown form and material carried by some servants. During the years they have given the shadows names and they interpret them as the reality.

One of these cave dwellers is able to shake-off his chains and leave the cave. His eyes get dazzled by the light of the Sun at first, but after a while he becomes able to see all the wonderful objects that cast the shadows.

And again he names them all and calls them the reality. But, upon his return to his pitiable colleagues, he is far from being welcome: his view of reality has been revolutionised by his stunning experience outside and has nothing more to do with the prisoners' view of reality.

So far Plato's cave parable. It tells us about the relation between our ideas and the objects behind them. We are in a similar situation like the cave dwellers when it comes to exploring the universe. What we can see with the best of our sophisticated technical means are nothing more than shadows of the reality out there and, like the prisoners in the cave, we have to content ourselves with the images on the wall. But – in contrast to Plato's parable – an unexpected second fire lights up in the background casting additional shadows from the unknown objects onto the wall of our cave. This second fire are the cosmic ray particles that reach us from the depth of the universe. Apart from the electromagnetic spectrum, where astronomical observations have taken place since mankind started looking at the stars, the cosmic rays are independent and complementary messengers from violent processes in the universe. That is why cosmic rays are such a fascinating topic, which is still in our days rich of mysteries.

This issue of Spatium is a short summary of articles edited by the International Space Science Institute in several of its fascinating publications as outlined on the back-cover. We are convinced that our readers will enjoy these insights into the mysteries of cosmic rays.

Hansjörg Schlaepfer
Zürich, October 2003

Impressum

SPATIUM
Published by the
Association Pro ISSI
twice a year

INTERNATIONAL SPACE SCIENCE INSTITUTE

Association Pro ISSI
Hallerstrasse 6, CH-3012 Bern
Phone +41 (0)31 631 48 96
Fax +41 (0)31 631 48 97

President
Prof. Heinrich Leutwyler, University of Bern
Publisher
Dr. Hansjörg Schlaepfer, legenda schläpfer wort & bild, CH-8185 Winkel
Layout
Marcel Künzi, marketing · kommunikation, CH-8483 Kollbrunn
Printing
Druckerei Peter + Co dpc
CH-8037 Zurich

Front Cover: The remnants of the N49 supernova in the Large Magellanic Cloud, a source of cosmic ray particles (credit: Hubble Heritage Team [STScI / AURA], Y. Chu [UIUC] et al., NASA). The lower part shows data from the Climax Neutron Monitor, see Figure 6.

Cosmic Rays

The Early Years [1]

Cosmic ray studies now span an epoch of almost exactly 100 years. At the close of the nineteenth century, scientists using gold-leaf electroscopes to study the conductivity of gases discovered that no matter how carefully they isolated their electroscopes from possible sources of radiation they still discharged at a slow rate. In 1901 two groups investigated this phenomenon, J. Elster and H. Geitel in Germany, and C.T.R. Wilson in England. Both groups concluded that some unknown source of ionising radiation existed. Wilson even suggested that the ionisation might be "due to radiation from sources outside our atmosphere, possibly radiation like Roentgen rays or like cathode rays, but of enormously greater penetrating power." A year later two groups in Canada, Ernst Rutherford and H. Lester Cooke at McGill University, and J. C. McLennan and E.F. Burton, at the University of Toronto showed that 5 cm of lead reduced this mysterious radiation by 30%. An additional 5 tonnes of pig lead failed to reduce the radiation further.

In 1907 Father Theodore Wulf of the Institute of Physics of Ignatus College in Valkenburg, Holland, invented a new electroscope. Wulf's electroscope enabled scientists to carry the search for the origin of the mysterious radiation out of the laboratory, into the mountains, atop the Eiffel Tower and, ultimately, aloft in balloons. Assuming that the radiation came from the Earth, they expected to find a rapid decrease in the radiation as they moved away from the surface. They did not find the decrease they expected and in some cases there seemed to be evidence that the radiation actually increased. Intrigued by the conflicting results obtained by Wulf and his colleagues a young Austrian nuclear physicist, Viktor Hess, obtained support from the Austrian Imperial Academy of Sciences and the

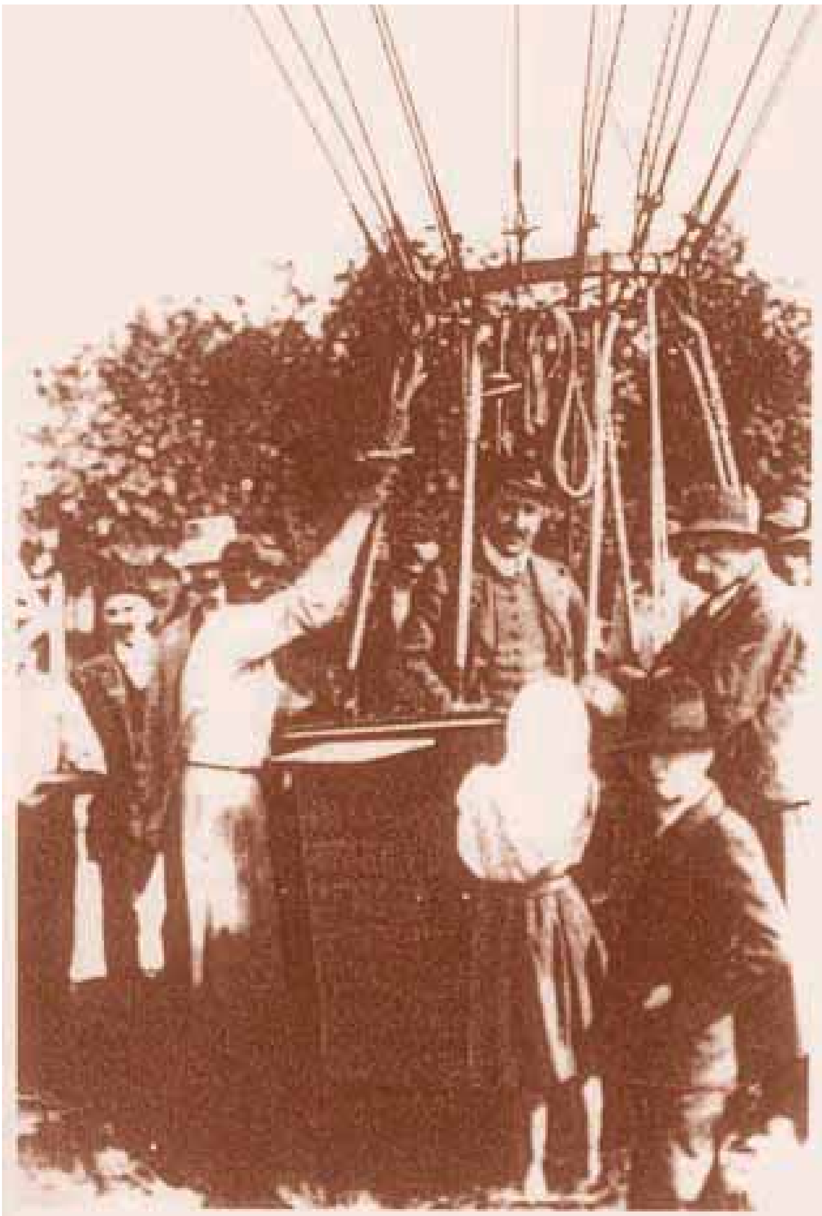


Figure 1
Viktor Hess after a balloon landing in 1912

[1]) *See reference list at the end of this article*

Royal Austrian Aero Club to conduct a series of balloon flights to study the radiation. Hess obtained a license to pilot balloons in order to reduce the size of the crew and thereby increase the altitude to which he could carry his electroscopes. On 12 August 1912, using the hydrogen-filled Böhmen, Hess reached an altitude of 5,350 m. Carrying two hermetically sealed ion chambers, he found that the ionisation rate initially decreased, but that at about 1500 m it began to rise, until at 5,000 m it was over twice the surface rate. Hess concluded that the results of these observations can best be explained by the assumption that radiation of a very high penetrating power from above enters into the atmosphere and partially causes, even at the lower atmospheric layers, ionisation in the enclosed instruments.

On a voyage from Amsterdam to Java, Clay observed in 1927 a variation in cosmic ray intensity with latitude with a lower intensity near the equator, thus establishing that before entering the Earth's magnetic field, the bulk of the primary cosmic rays were charged particles. In 1930 Bruno Rossi showed that if the cosmic rays were predominantly of one charge or the other there should be an east–west effect. In the spring of 1933 two American groups, Thomas H. Johnson of the Bartol Research Foundation and Luis Alvarez and Arthur H. Compton of the University of Chicago, simultaneously and independently measured the east–west effect. It showed the cosmic radiation to be predominantly positively charged.

In a series of balloon flights in the late 1930s, M. Schein and his co-workers used Geiger counter telescopes interspersed with lead absorbers to determine that most of the primary particles were not electrons, and hence protons were most plausibly the dominant constituent.

In 1948 research groups from the University of Minnesota and the University of Rochester flew nuclear emulsions and cloud chambers on the same high-altitude Skyhook balloon flight and discovered the presence of heavy nuclei in the primary cosmic radiation. Further studies by many other groups soon established that essentially all of the elements between H and Fe were present in the cosmic radiation near the top of the atmosphere – including an overabundance of the light elements Li, Be, and B. Then in 1950 it was found that a significant fraction of the cosmic radio emission was synchrotron radiation – indicating the presence of highly relativistic electrons throughout our Galaxy including some discrete sources as well as extragalactic sources. However, because of their small abundance (1% of the intensity of cosmic ray nuclei) electrons were not directly detected in the primary cosmic radiation until 1962. These discoveries made it possible to begin constructing realistic models of the origin and interstellar transport of galactic cosmic rays.

As early as 1934, Baade and Zwicky linked the appearance of supernovae with neutron star formation and cosmic ray generation. Fermi in 1949 regarded cosmic rays as a gas of relativistic charged particles moving in interstellar magnetic fields. His paper laid the groundwork for the modern theory of cosmic ray acceleration and transport. The close link between radio astronomy and cosmic rays was conclusively established at the time of the Paris Symposium on Radioastronomy in 1958. This marked the birth of cosmic ray astrophysics. The basic model of the origin of galactic cosmic rays was developed by Ginzburg and Syrovatskii in 1964.

The Nature of Cosmic Rays

Cosmic rays consist of electrons, neutrons and atomic nuclei, which have been accelerated to very high speed. Their elemental composition provides information on chemical fractionation in the source region as well as some insight into the nature of this region and of the propagation of cosmic rays in interstellar space. Cosmic ray isotopes probe more deeply the nature of the source region and the timescales of the injection and initial acceleration. Radioactive isotopes such as ^{10}Be, ^{26}Al, and ^{36}Cl reveal the temporal history of cosmic rays in the disk and halo regions. The variation of the charge and mass composition with energy – their energy spectra – can be related to the acceleration process and to particle transport in the Galaxy. When improved measurements are available at ultrahigh energies it should be possible to determine whether these particles are of galactic or extragalactic origin. At the highest energies the cosmic ray arrival direction may also indicate the approximate direction of the most powerful sources.

The most remarkable feature of cosmic rays is their energy spectrum (**Figure 2**). From $\approx 10^9$ eV to $>10^{20}$ eV these spectra, over some 10 orders of magnitude variation in intensity show a relatively featureless power-law distribution.

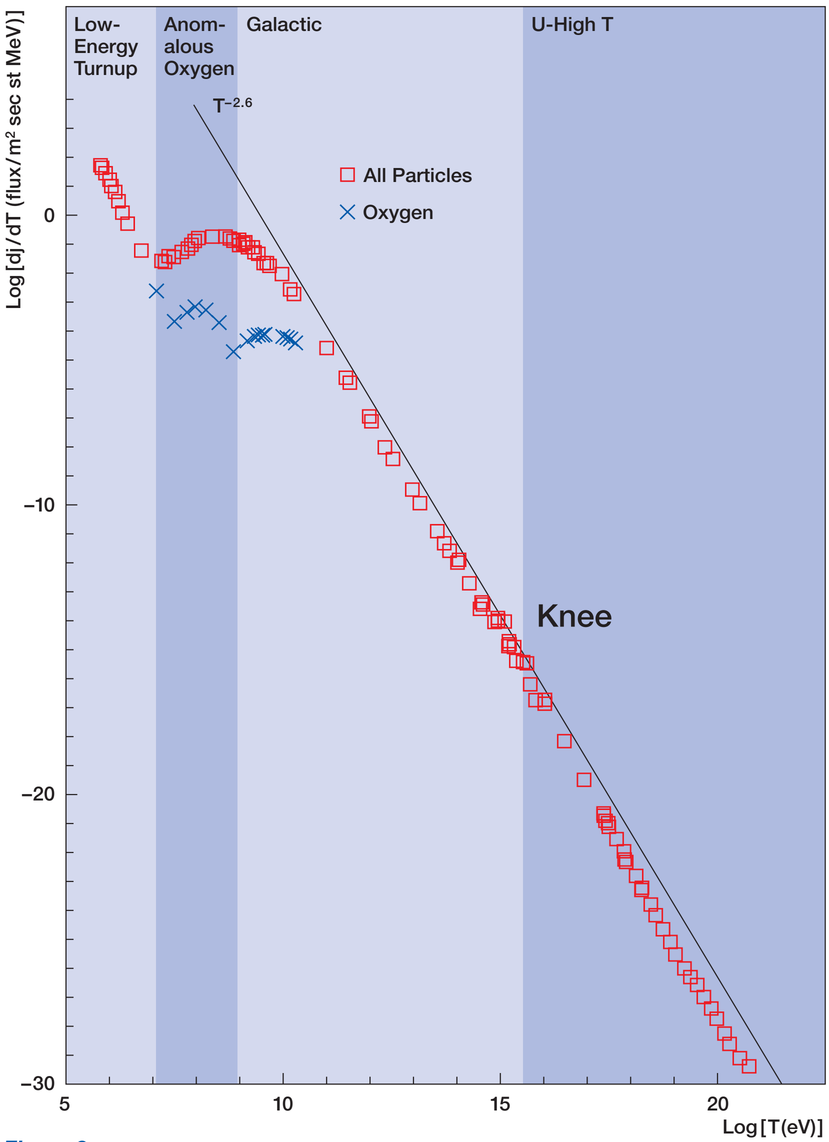


Figure 2
Energy spectrum of cosmic rays measured at the Earth.

At energies below a few GeV the influence of solar modulation becomes important with significant temporal variations at 1 AU related to the 11- and 22-year solar and heliomagnetic cycles. At energies of less than 40 MeV the oxygen spectra in **Figure 2** show the presence of so-called anomalous cosmic rays as discussed later. Those are partially ionised interstellar atoms accelerated at the solar wind termination shock. Near 10 MeV there is a highly variable turn-up in the ion spectrum produced by particles of solar/interplanetary origin although the acceleration up to energies more than tens of GeV was registered in some solar flare events.

The experimental limitation imposed by the size of current detector systems does not allow measurements of the cosmic-ray intensity at energies greater than 3 x 10^{20} eV. Above 10^{12} eV the composition is not well known; this is the region where acceleration by supernova shocks becomes difficult. It is generally assumed that the cosmic ray "knee" at ~10^{15} eV may reflect a gradual transition in the composition to particles of increasingly higher charge. At energies >10^{19} eV questions of galactic magnetic confinement lead to the assumption that these particles are of extragalactic origin. At energies above 4 x 10^{19} eV even protons should experience significant deceleration by the 3 K blackbody radiation (the Greisen–Zatsepin–Kuzmjn effect). At energies of 10^{12}–10^{14} eV there are small anisotropies of ≈ 0.1%, which are thought to be due to local effects. At this time there are no meaningful anisotropies observed at higher energies except the ultra-high energies ~10^{18} eV.

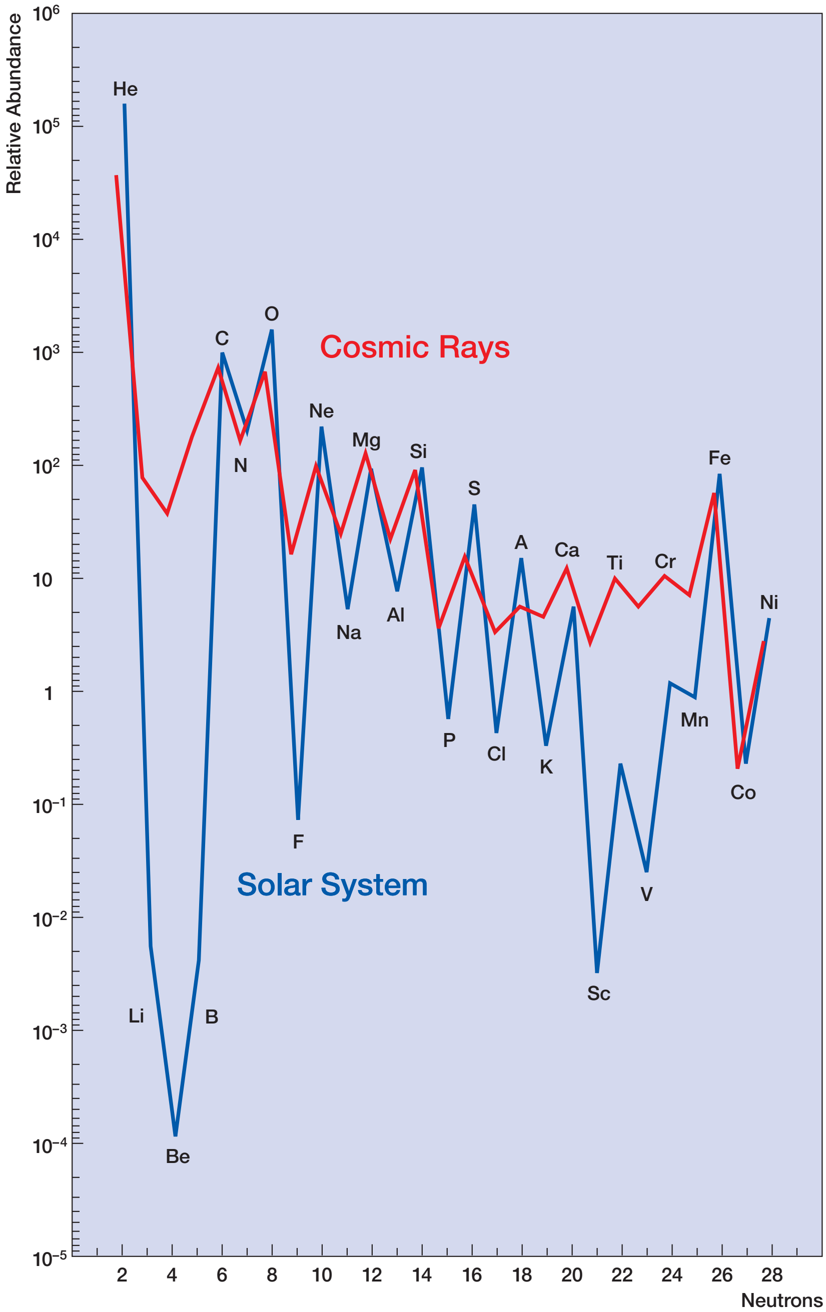


Figure 3
The relative abundance of He to Ni in cosmic rays (red line), and in the Solar System (blue line).

Cosmic ray particles are mostly hydrogen (87%), and some helium (12%) with diminishing amounts of carbon, oxygen, etc and of heavier elements **(Figure 3)**. All are fully ionised. Electrons account for approx. 1% of the cosmic rays. With a few exceptions, their chemical composition corresponds to the elemental abundances in our solar system. The exceptions (H, He are under-abundant, Li, Be, B are over-abundant) yield important information about the matter traversed by cosmic rays. Isotopic abundances of galactic cosmic rays are very similar to the isotopic composition of the interstellar gas. An important exception is the larger ratio of $^{22}Ne/^{20}Ne$. The isotopic composition provides key information about the origin, acceleration and transport mechanisms of cosmic rays in our galaxy.

The Origins of Cosmic Rays

Cosmic ray particles arrive evenly from all directions of the sky, but this does not necessarily mean that their sources are evenly spread around us. More likely, they are constantly deflected and scattered by magnetic fields in the galaxy, until any trace of their original motion is lost.

There are three categories of cosmic ray particles according to their energy content:

Energy level [eV]	Cosmic ray type; origin and acceleration mechanism
Below the knee $E < 3 \times 10^{15}$ eV	Galactic cosmic rays I (GCR I); galactic origin, diffuse shock acceleration in the shock waves of supernova remnants (SNR)
Above the "knee" $3 \times 10^{15} \leqslant E \leqslant \approx 10^{18}$ eV	Galactic cosmic rays II (GCR II); galactic origin, second stage acceleration of GCR I by shocks
Above $E \geqslant \approx 10^{18}$ eV	Extragalactic cosmic rays; acceleration in extragalactic shocks?

Galactic cosmic rays come from outside the solar system but generally from within our Milky Way. They have probably been accelerated within the past few million years, and have travelled many times across the galaxy, trapped by the galactic magnetic field; they have been accelerated to nearly the speed of light, probably by supernovae explosions. A supernova is a star of several times the mass of our Sun, that has run out of the "nuclear fuel" of light elements, especially hydrogen, needed to keep it shining. Its nuclear burning gradually converts the light elements into heavier ones, and the heat it produces keeps it from collapsing under its own immense weight. When the star can no longer produce nuclear heat, it suddenly collapses to a small volume, releasing enormous amounts of gravitational energy. Much of that energy is spent in a grand explosion as shown in **Figure 4**, blowing the star's outer layers out to space and creating thereby a huge expanding shock front. It is this shock front which is believed to accelerate the cosmic ray particles to nearly the speed of light.

Anomalous Cosmic Rays

The discovery of a totally new component of nuclear radiation in the heliosphere – now called anomalous cosmic rays (ACR) – has greatly expanded our understanding of pickup ions, particle acceleration and our knowledge of neutral atomic abundances in the local interstellar medium. The story begins with satellite measurements at 1 AU of the proton and helium spectra as the heliospheric modulation of galactic cosmic rays declined towards solar minimum condition in 1971–1972. A spectrum of helium nuclei with energies below ≈ 100 MeV per nucleon appeared which could not be accounted for by either the solar modulation of helium of galactic origin or its isotopic composition.

It soon became clear that this anomalous helium component was also accompanied by anomalous nitrogen and oxygen. More recently, there has appeared evidence for C, Ne, Ar and H. Garcia-Munoz and collaborators showed that the anomalous component of He is modulated ^{4}He with no spallation ^{3}He and, therefore, must have a "local" origin. Fisk, Koslovsky and Ramaty proposed the most successful model to account for the anomalous components. As sketched in **Figure 5**, interstellar neutral atoms with high ionisation potentials enter the heliosphere, undergo single ionisation by solar ultra-violet radiation in the inner solar system and–by charged particle pickup in the solar wind – are carried outward to the solar wind termination shock, where they are

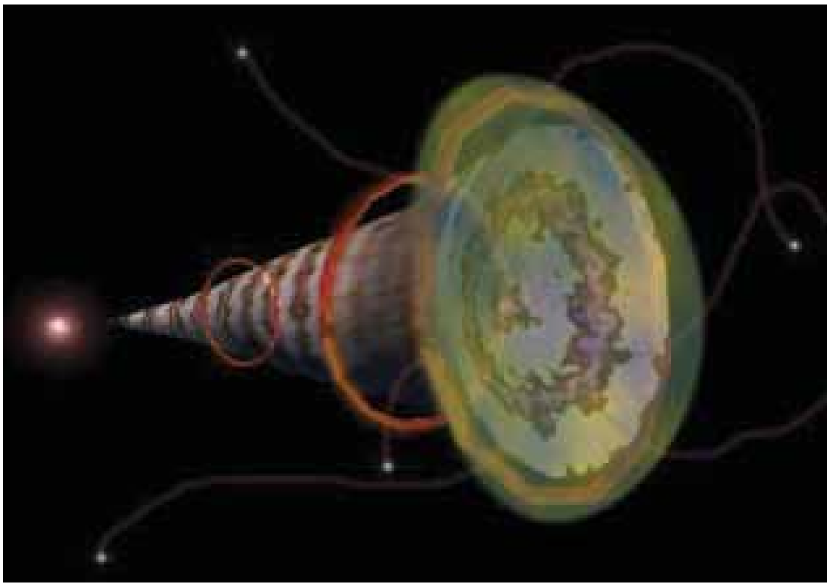


Figure 4
This artist's illustration shows a supernova explosion (at left) and a conical section of the expanding cloud of ejected material. Atoms are torn from the brownish bands of "dust" material by shock waves (represented by orange rings). The shocks in the expanding blast wave then accelerate the atoms to near light speeds firing them into interstellar space like cosmic bullets. (Credit: M. DeBord, R. Ramaty and B. Kozlovsky [GSFC], R. Lingenfelter [UCSD], NASA)

accelerated. Some of these accelerated nuclei – now possessing high magnetic rigidities since they are singly charged – propagate inward to undergo solar modulation, along with the galactic low energy cosmic ray nuclei.

Solar Cosmic Rays

Most of the time the cosmic rays arriving at Earth are of galactic or even extragalactic origin. However, from time to time, the Sun is also a source of cosmic rays. Evidence for the acceleration of energetic protons, ions and electrons in close association with solar flares and coronal mass ejections is provided by direct observations of the energetic particles in interplanetary space or at Earth, and by the detection of various neutral radiations produced in interactions of the accelerated particles with the solar atmosphere and the solar magnetic field. Solar protons, ions and electrons of energies extending into the 100 MeV region are directly measured by space borne instruments. Solar protons with energies above ~500 MeV/nucleon become increasingly rare and therefore must be detected by continuously operating ground-based cosmic ray detectors. The neutral radiations with an unambiguous link to interactions of accelerated particles in the solar atmosphere are radio- and microwaves, soft and hard X-rays, gamma rays and neutrons. On a long-term average, cosmic ray ground level intensity enhancements due to relativistic solar particles occur about once per year, while events with lower-energy particles are much more frequent. There is a distinct dependence on solar sunspot activity, with particle events being more frequent and having a larger fluence from around two years before to about four years after a solar maximum. However, there is evidence that the most energetic events do not occur right in the maximum phase of solar activity.

In the current paradigm, solar energetic particle events are generally classified as "impulsive events" or "gradual events". Impulsive events are characterized by the presence of type III radio bursts, that are radio emissions generated by streams of energetic electrons injected from the Sun into interplanetary space. Furthermore high energy protons, ions, and electrons are emitted by the Sun and accelerated in the deep corona in regions with temperatures $>10^7$ K. In gradual events, the solar cosmic rays are produced by shock acceleration in the high corona and in interplanetary space near the Sun. An expanded classification of solar cosmic ray events includes also the high-energy solar particles, which cause nuclear reactions in the solar atmosphere leading to the emission of gamma rays and neutrons.

The Sun is the most powerful particle accelerator in our neighbourhood and, therefore of fundamental physical and astrophysical interest. Ongoing research concentrates on the nature and the properties of the acceleration processes. One of the key instruments supporting these efforts is the Reuven Ramaty High Energy Solar Spectroscopic Imager (RHESSI), a NASA Small Explorer mission launched on February 5, 2002 with contributions from the ETH Zürich and the Paul Scherrer Institute, Würenlingen. Its primary mission is to explore the basic physics of particle acceleration and explosive energy release in solar flares.

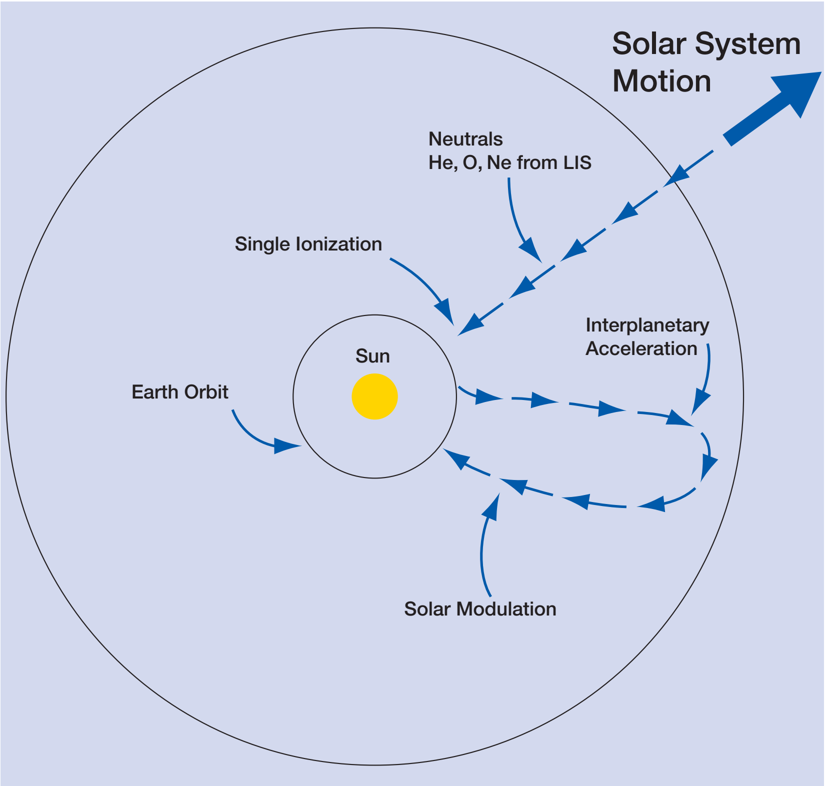


Figure 5
Sketch of the concept for the production and acceleration of the anomalous cosmic rays.

Detecting Cosmic Rays

The earliest cosmic radiation data acquired on a worldwide basis were obtained from Wilson chambers on ships. Routine monitoring of the cosmic radiation was initiated in January 1932 with the operation of ionising chamber at Hafelekar, Austria. Wilson chambers respond to muons generated by incident high-energy protons; however only nucleons greater than about 4 GeV have sufficient energy to generate a muon cascade capable of penetrating the atmosphere and reaching the Earth's surface. Thus it was desirable to develop a detector that would respond to lower energy nucleons as well as being relatively easy to maintain. In the 1950s John A. Simpson, at the University of Chicago, invented and developed the neutron monitor and found that the Earth's magnetic field could be used as a spectrometer to allow measurements of the cosmic ray spectrum down to low primary energies. The magnetic latitude of a particular neutron monitor determines the lowest magnetic rigidity of a primary that can reach the monitor, the so-called "cut-off rigidity". The station's altitude determines the amount of absorbing atmosphere above the station and hence the amount of absorption of the secondary cosmic rays (the higher the station, the higher the counting rate). By using a combination of lead (to produce local interactions), paraffin or polyethylene (to moderate or slow down the neutron component) and multiple slow-neutron counters, Simpson greatly increased the counting rate in his monitor design. After the invention of the standard neutron monitor this type of detector became widely used for cosmic ray monitoring all over the world. One of the longest uninterrupted observations is made by the Climax Neutron Monitor of the University of Chicago, see **Figure 6**.

As the counting rate of high energy particle counters increases with increasing altitude high elevation research stations became attractive locations for neutron monitors. In 1958 the first neutron monitor was installed on the Jungfraujoch where it is operated by the

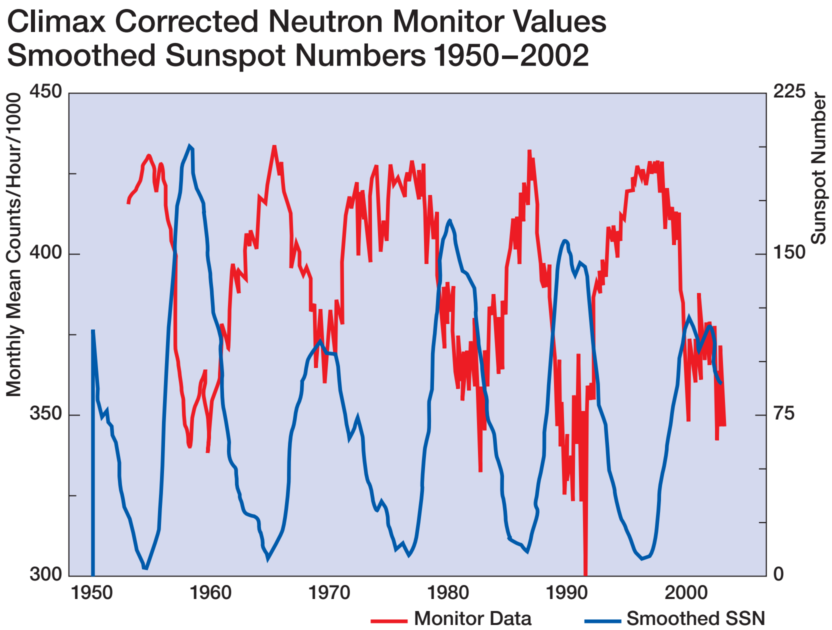


Figure 6
This plot shows data from the Climax Neutron Monitor operated by the University of Chicago. The cosmic rays show an inverse relationship to the sunspot cycle because Sun's magnetic field is stronger during sunspot maximum and shields the Earth from cosmic rays.

Figure 7
Two neutron monitors are installed at the Sphinx laboratory of the Jungfraujoch at an altitude of 3570 m. The 18-IGY-Detector is in continuous operation since 1958, while the Standard 3-NM64-neutron monitor operates since 1986.

International Foundation High Altitude Research Stations Jungfraujoch and Gornergrat (HFSJG) with headquarters in Bern, **Figure 7**.

A second detector was installed in 1986. The effective vertical cutoff rigidity at Jungfraujoch is 4.5 GeV. Hermann Debrunner, the former President of the Pro ISSI association, has started his career there as a student and served later as the foundation's director.

Propagation of Cosmic Rays in the Earth's Environment

The magnetic field and the atmosphere form two powerful protective layers against the cosmic radiation on the Earth's surface. The magnetic field acts both as a shield and as a giant natural spectrometer for cosmic ray particles. If the particles possess energy, which is greater than the magnetic cut-off energy, they will cross through the magnetosphere and reach the upper layers of the atmosphere. But if their energy is insufficient, they will have a tendency to follow the magnetic lines of force, with which they move "easily", due to their lack of energy, and succeed in reaching the poles. It is the reason why the areas located near the poles receive radiation in higher quantities than near the equator, which is better protected by the Earth's magnetic field, see **Figure 8**. The second protective layer is the Earth's atmosphere. Upon arriving in the upper parts of the atmosphere, the cosmic ray particles interact with the atoms, which they encounter. As shown in **Figure 9** these collisions create new cascades of particles that produce further successively lower energy nuclear disintegrations. This nucleonic cascade process caused by primary cosmic particles can be detected at the surface of the Earth by means of fast neutrons

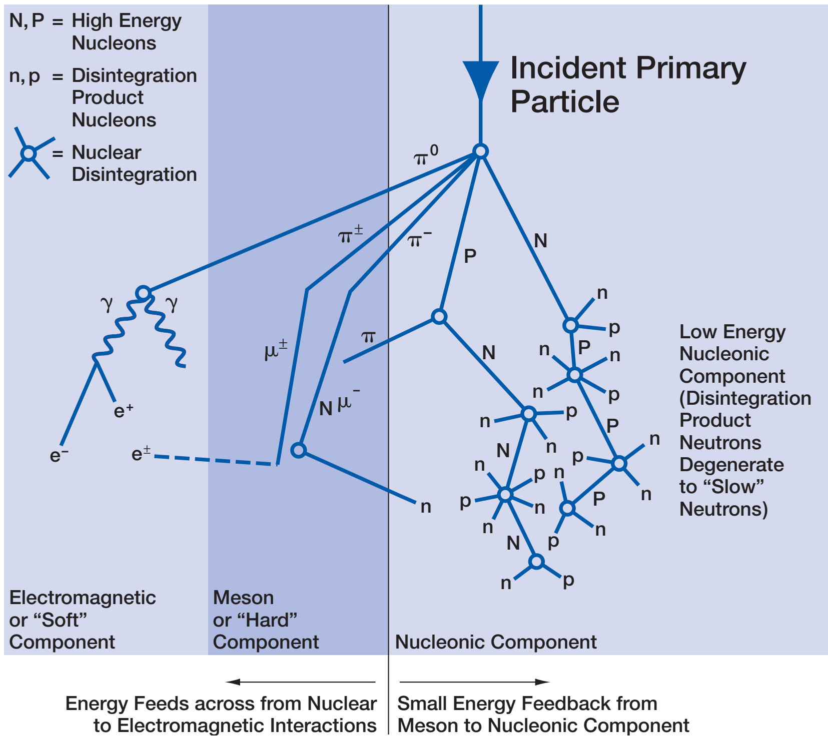


Figure 9
Schematic Diagram of a Cosmic Ray Shower. An incident cosmic ray particle interacts with the atoms at the top of the atmosphere. Due to its high energy it disintegrates the atoms producing a cascade of electromagnetic radiation, of muons and nucleons, of which the neutrons are detected by the neutron monitors.

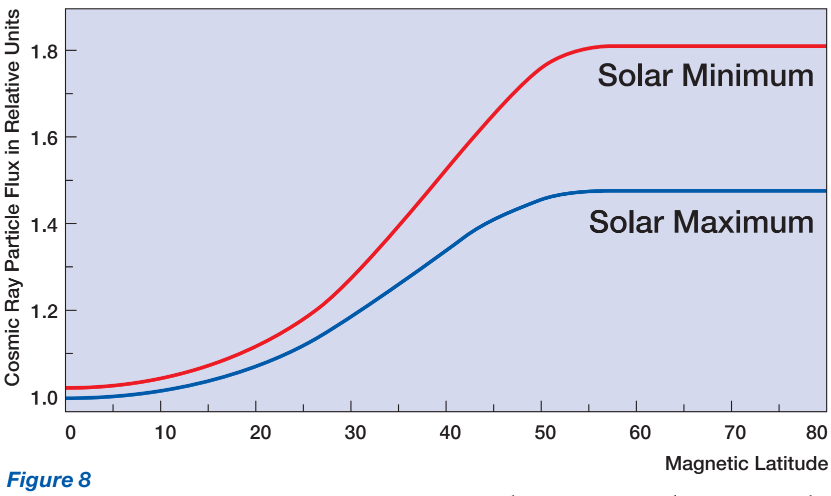


Figure 8
Illustration of the cosmic ray latitude curve. The minimum value occur at the equator and the maximum values at polar latitudes. The values are relative since the numbers vary with altitude and solar activity. At high latitudes the cosmic ray flux levels off, since the shielding effect of the Earth's atmosphere becomes larger than the cosmic ray cut-off by the magnetic field.

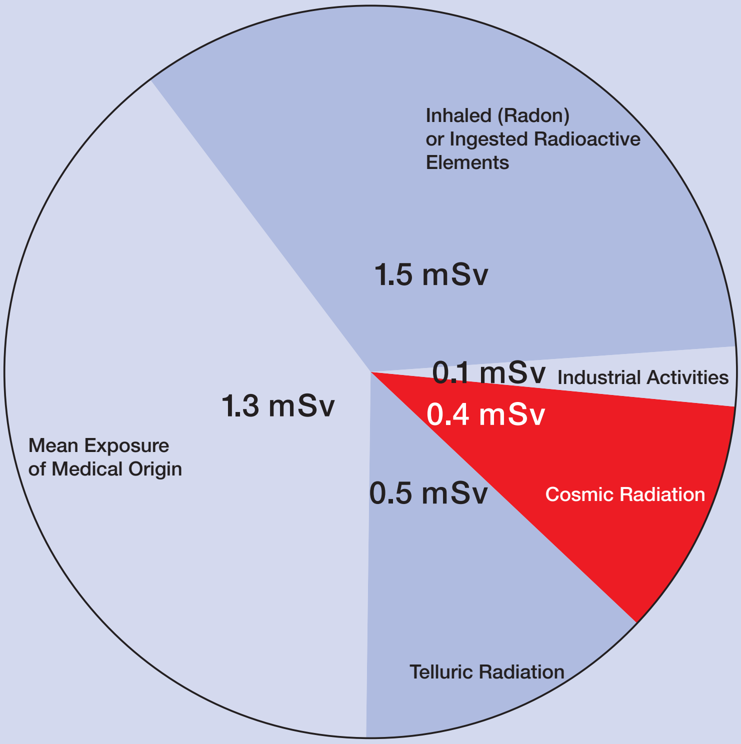


Figure 10
Sources of radioactivity as a percentage of the dose received by an average individual. (UNSCEAR, United Nations Scientific Committee on the Effects of Atomic Radiations)

produced in the atmosphere, which are counted by neutron monitors. As shown in **Figure 10** cosmic radiation represents at ground level only a small part (11%) of the ionising radiation to which an individual is commonly exposed. Natural land-based sources expose each of us to an average total dose of 2.4 mSv per year **(see box)**, though with significant variations according to regions. The larger part of the sources is a gaseous descendant of natural uranium, i.e. radon, which concentrates in enclosed areas such as houses. There is also soil-based radiation, coming from surface rocks, granite in particular, which contain radioactive elements such as uranium, dating from the formation of the planet. The water and foods, which we ingest also contain radioactive elements. Finally, there is also the internal radiation, i.e. coming from within our own bodies, namely from the potassium 40 which is naturally present in our tissues.

The Sievert

Assessing the biological risk of ionising radiation

When it comes into contact with matter, ionising radiation collides with the atoms comprising it. During these interactions, it releases a part or all of its energy. The absorbed dose (expressed in Gray) is defined by the ratio of this released energy over the mass of the matter. A Gray corresponds to one Joule of energy released in one kilogram of matter.

In order to have a single unit which expresses the risk of the occurrence of the stochastic effects associated with all possible exposure situations, physicists developed an indicator known as the "effective dose", a measurement using the Sievert (Sv), named after the Swedish physicist who was one of the pioneers in protection against ionising radiation.

The effective dose is calculated from the dose (expressed in Gy) absorbed by the various exposed tissues and organs, by applying weighting factors which take into account the radiation types (alpha, beta, gamma, X, neutrons), the means of exposure (external or internal) and the specific sensitivity of the organs or tissues.

Cosmic Rays and Climate

Changes in the Earth's climate are generally attributed to internal causes, e.g. volcanic dust in the atmosphere, atmospheric/ocean oscillations like the El Niño Southern Oscillation, and to external causes like variations in the Earth's orbital parameters and in the Sun's luminosity. Recently, it was suggested that the Earth's cloud cover is correlated with the intensity of galactic cosmic rays. The idea is that droplet formation is influenced by ionisation of atmospheric molecules by cosmic rays. The ionisation could potentially influence optical transparency of the atmosphere, by either a change in aerosol formation or an influence on the transition between different phases of water. As these hypotheses are still controversial and highly speculative they are currently the objects of intense world-wide research activities.

Cosmic Rays and Life

Since the cosmic radiation increases with increasing altitude, it could be expected that people living at high altitudes suffer more from cosmic rays than those at sea level. For example at Denver, Colorado at an altitude of 1'600 m the exposure to cosmic radiation is twice than for people living near sea level. Medical records for these populations exist for a significant number of years thus making it possible to search for health problems that might be correlated to differences in the exposure to background cosmic radiation. When these studies are done, a surprising paradox is found: the population living at mountain altitudes is generally healthier and has longer life span than the populations living at sea level. The conclusion is that other factors must be dominant and exposure to cosmic radiation at altitudes where people live does not appear to present a health hazard.

Biological effects of ionising radiation in living cells begin with the ionisation of atoms. Ionising radiation absorbed by living cells has enough energy to remove electrons from the atoms that make up molecules of the cell. When the electron that was shared by the two atoms to form a molecular bond is dislodged by ionising radiation, the bond is broken and thus, the molecule falls apart. This is a basic model for understanding radiation damage. When ionising radiation interacts with cells, it may or may not strike a critical part of the cell. We consider the chromosomes to be the most critical part of the cell since they contain the genetic information and instructions required for the cell to perform its function and to make copies of itself for reproduction purposes. Obviously, cosmic radiation has been an intrinsic boundary condition for the evolution of life on Earth since its very beginnings and nature, therefore, had to develop efficient repair mechanisms, which are able to repair cellular damage – including chromosome damage.

Public concern and legal regulations that became effective in Europe in May 2000 address the radiation risk and the individual dose assessment of airline crewmembers. On the average, at flight altitude radiation dosage is of the order of 10 microsievert / hour. Of course, this value varies with altitude, latitude, and solar activity, and it must be interpreted in comparison with the average natural dosage. Thus, approx. 400 hours per year at flight altitude would lead to the equivalent of the natural yearly radiation load of around 4 mSv. Interested flight passengers can evaluate the expected dosage using publicly available programs as e.g. SIEVERT (http://www.sievert-system.org/).

Above the Earth's first protective layer, the atmosphere, the radiation exposure increases strongly. The assessment of the radiation risk of astronauts for example during the ongoing construction of the International Space Station is essential to safeguard their health. The astronauts' suits provide some protection against cosmic rays especially against low energy particles and in areas, where human organs more susceptible to radiation hazards are, like for example the head or the heart. In any case, however, the layout of such a suit remains a compromise between its protective efficiency and the residual mobility required by the astronaut. The US Space Shuttle orbit the Earth with an inclination of 28° and receive between 42 to 62 μGy per day, while in the 51° inclination orbit of the International Space Station the radiation dose is higher due to the coverage of higher geomagnetic latitudes and amounts to 90 to 150 μGy per day.

Outside the Earth's second protective layer, the magnetosphere, the radiation loads become still more important. When it comes to planning a mission to the other planets like for example to Mars the radiation protection of the astronauts on their long-duration flights is a key issue. In order to systematically broaden the available radiation database the European Space Agency regularly equips its spacecraft with the Standard Radiation Environment Monitor (SREM). The SREM, a high-energy particle counter for electrons with up to 6 MeV and protons up to 300 MeV has been jointly developed by Contraves Space AG, of Zürich and the Paul Scherrer Institute of Würenlingen.

Cosmic Rays and Matter

The disintegrating effect of cosmic rays on molecular bonds not only effects living cells but of course also anorganic materials. Due to their large areas and their long exposure time the solar panels of spacecraft are especially endangered by cosmic rays. A continuous degradation of their power production efficiency or even their complete breakdown may be the result of cosmic rays. In addition, electronic micro-components are prone to radiation damages leading to malfunctioning of equipment or even to the complete loss of the spacecraft. The most widely known disturbances and failures are those of the Canadian ANIK satellite on January 20/21, 1994 and of the Telstar satellite on January 11, 1997, which both have been attributed to cosmic ray effects. The commercial losses amounted to more than 135 millions US$.

Reference List

[1]) Excerpt from Frank B. Mc. Donald and Vladimir S. Ptuskin, Galactic Cosmic Rays, The Century of Space Science, Kluwer Academic Publishers 2001; p. 677–697

[2]) Excerpt from John A. Simpson, The Cosmic Radiation, The Century of Space Science, Kluwer Academic Publishers 2001; p. 117–151

[3]) Extracts from M. A. Shea and D. F. Smart; Fifty Years of Cosmic Radiation Data; Cosmic Rays and Earth; Space Science Series of ISSI; Kluwer Academic Publishers 2000; p. 229–262

[4]) NOAA Satellite and Information Services, http://www.ngdc.noaa.gov/stp/SOLAR/COSMIC_RAYS/cosmic.html

[5]) The International Foundation High Altitude Research Stations Jungfraujoch and Gornergrat (HFSJG); http://www.ifjungo.ch

[6]) Extracts from M. A. Shea and D. F. Smart; Cosmic Ray Implications for Human Health; Cosmic Rays and Earth; Space Science Series of ISSI; Kluwer Academic Publishers 2000; p. 187–205

7) The Sievert System; http://www.sievert-system.org

Conclusions

Cosmic rays are the messengers from distant regions in our galaxy and beyond. Solar cosmic rays released occasionally in association with energetic processes at the Sun, provide first-hand information on astrophysical processes responsible for particle acceleration. Anomalous cosmic rays reveal new insight in the dynamic processes in the heliosphere and its interaction with the local interstellar medium. Apart from the electromagnetic spectrum, where the classical astronomical observations take place, cosmic rays open a second window to the universe providing an attractive complementary diagnostic tool for our understanding of the processes in the universe. Although the research of cosmic rays began nearly 100 years ago much is still unknown especially with regard to their origins and the mechanisms providing the particles nearly the speed of light. That is why the cosmic ray research continues to be one of the most fascinating adventures of modern space science. This article is based on publications of the International Space Science Institute on cosmic rays, which are strongly recommended for further reading. In addition I am very thankful to Erwin Flückiger, Institute for Physics, University of Bern and Rudolf von Steiger, International Space Science Institute for their his most valuable contributions.

Hansjörg Schlaepfer

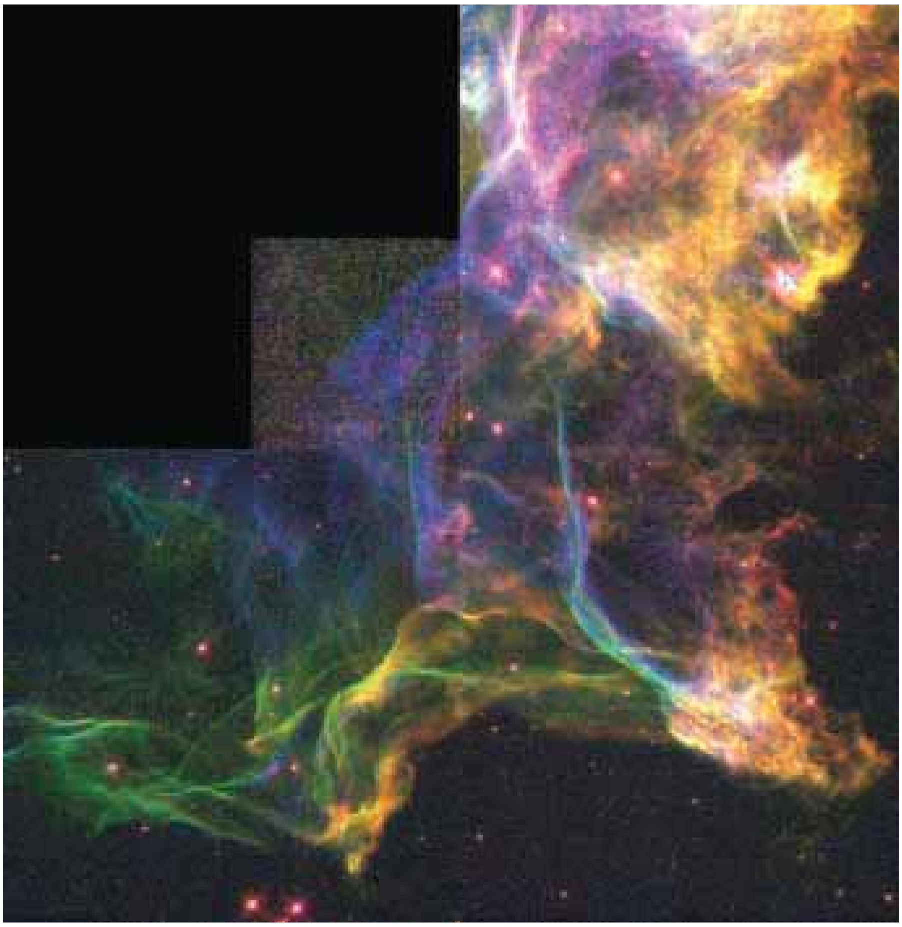


Figure 11
15,000 years ago a star in the constellation of Cygnus exploded – the shockwave from this supernova explosion is still expanding into interstellar space! Credit: J.Hester (ASU), NASA

SPATIUM

The Publications of ISSI as of October 2003

All books are published by Kluwer Academic Publishers, Dordrecht, Boston, London

Reports from Workshops

ISSI Scientific Reports

Volume 1	Analysis Methods for Multi-Spacecraft Data, 1998
Volume 2	The Radiometric Calibration of SOHO, 2002

Space Science Series of ISSI

Volume 1	The Heliosphere in the Local Interstellar Medium, 1996
Volume 2	Transport Across the Boundaries of the Magnetosphere, 1997
Volume 3	Cosmic Rays in the Heliosphere, 1998
Volume 4	Primordial Nuclei and Their Galactic Evolution, 1998
Volume 5	Solar Composition and its Evolution – from Core to Corona, 1998
Volume 6	Magnetospheric Plasma Sources and Losses, 1999
Volume 7	Corotating Interaction Regions, 2000
Volume 8	Composition and Origin of Cometary Materials, 1999
Volume 9	From Dust to Terrestrial Planets, 2000
Volume 10	Cosmic Rays and Earth, 2000
Volume 11	Solar Variability and Climate, 2000
Volume 12	Chronology and Evolution of Mars, 2001
Volume 13	The Astrophysics of Galactic Cosmic Rays, 2002
Volume 14	Matter in the Universe, 2002
Volume 15	Auroral Plasma Physics, 2003
Volume 16	Solar System History from Isotopic Signatures of Volatile Elements, 2003

Review of the Space Science Achievements in the 20th Century

The Century of Space Science

Two volumes, 1846 pages, 2002

INTERNATIONAL SPACE SCIENCE INSTITUTE

SPATIUM

Published by the Association Pro ISSI

No 12, June 2004

Ten Years Hubble Space Telescope

Editorial

Give me the material, and I will build a world out of it!

Immanuel Kant (1724–1804), the great German philosopher, began his scientific career on the roof of the Friedrich's College of Königsberg, where a telescope allowed him to take a glance at the Universe inspiring him to his first masterpiece, the "Universal Natural History and Theory of Heaven" (1755). Applying the Newtonian principles of mechanics, it is the result of systematic thinking, "rejecting with the greatest care all arbitrary fictions". In his later Critique of the Pure Reason (1781) Kant maintained that the human intellect does not receive the laws from nature, but rather dictates them upon it, since our mind requires a priori rules in space and time, in cause and in effect, in order to understand nature. And the result of his cosmological studies confirms the strength of his empirical thinking. Sapere aude: "dare use your mind," used he to say to his critics.

In his Universal Natural History he wrote: "Nature, on the immediate edge of creation, was as raw and undeveloped as possible. Only in the essential properties of the elements, which made up the chaos, can we perceive the sign of that perfection, which nature has from its origin, since its being is a consequence arising from the eternal idea of the Divine understanding". Can one better describe the dawn of this world?

The present issue of Spatium is dedicated to a scientist who lived a century after Kant and a telescope built another century later. The Hubble Space Telescope has revolutionised our understanding of the cosmos much the same as Kant's theoretical reflections did. Observing the heavenly processes, so far out of any human reach, gives men the feeling of the cosmos' overwhelming forces and beauties from which Immanuel Kant derived the order for a rational and moral human behaviour: "the starry heavens above me and the moral law within me…".

Who could be better qualified to rate the Hubble Space Telescope's impact on astrophysics and cosmology than Professor Roger M. Bonnet, the former Director of Science at the European Space Agency? It was under his guidance that in the frame of a joint NASA/ESA programme the Hubble Space Telescope was created, leading to one of the many marvellous achievements of the European Space Agency's science programme. R. M. Bonnet, now the International Space Science Institute's Executive Director, gave a fascinating lecture on the Hubble Space Telescope to the Pro ISSI audience on 6 November 2003. We are indebted to Professor Bonnet for his kind permission to publish herewith a slightly revised version of his lecture.

Sapere aude!

Hansjörg Schlaepfer
Zürich, June 2004

Impressum

SPATIUM
Published by the Association Pro ISSI twice a year

INTERNATIONAL SPACE SCIENCE INSTITUTE

Association Pro ISSI
Hallerstrasse 6, CH-3012 Bern
Phone +41 (0)31 631 48 96
Fax +41 (0)31 631 48 97

President
Prof. Heinrich Leutwyler, University of Bern
Publisher
Dr. Hansjörg Schlaepfer, legenda schläpfer wort & bild, CH-8185 Winkel
Layout
Marcel Künzi, marketing · kommunikation, CH-8483 Kollbrunn
Printing
Schellenberg Druck AG
CH-8330 Pfäffikon

Front cover: The beautiful Eskimo Nebula is an intricate structure of shells and streamers of gas around a dying Sun-like star 5000 light-years away. The disc of material is embellished with a ring of comet-shaped objects, their tails streaming away from the central, dying star. The planetary nebula began to form about 10,000 years ago, when the dying star started to expel an intense 'wind' of high-speed material out into space. (Credit: Space Telescope Science Institute, STScI)

Ten Years Hubble Space Telescope *)

Roger M. Bonnet, International Space Science Institute, Bern

The Man…

Edwin Powell Hubble was born in Marshfield, Missouri, USA, on November 29, 1889. In 1898, his family moved to Chicago, where he attended high school. Edwin Hubble was a fine student and an even better athlete, but he also found time to study and earn an undergraduate degree in mathematics and astronomy. He then went to Oxford University, where he did not continue his studies in astronomy, but instead studied law.

In 1913, Hubble returned from England and was called to the barrister, setting up a small practice in Louisville, Kentucky; but it didn't take long for him to realise that he was not happy as a lawyer, and that his real passion was astronomy. He therefore studied at the Yerkes Observatory and received in 1917 the doctorate in astronomy from the University of Chicago. Following a tour of duty in World War I, Hubble was employed at the Mount Wilson Observatory in California, where he devised a classification system for the various galaxies he observed, sorting them by content, distance, shape, and brightness. It was during these studies that he noticed the red shifts in the emission of light from galaxies, which he correctly interpreted as moving away from each other at a rate proportional to their distance. From these observations, he was able to formulate in 1929 the so-called Hubble's Law, allowing astronomers to determine the age of the universe, and proving that the universe was expanding.

It is interesting to note here that as early as 1917, Albert Einstein had already introduced his general theory of relativity, and produced a model of space based on that theory, claiming that space was curved by gravity, therefore that it must be able to expand or contract; but he found this assumption so far fetched, that he revised his theory, stating that the universe was static. Following Hubble's discoveries, he is quoted as having said that second guessing his original findings was the *biggest blunder* of his life, and he even visited Hubble to thank him in 1931.

After World War II, Edwin Hubble returned to the Mount Wilson and Mount Palomar observatories and continued his studies until his death on September 28, 1953. During his life, Hubble had tried to obtain the Nobel Prize, even hiring a publicity agent to promote his cause in the late 1940s, but all the effort was in vain as there was no category for astronomy. With or without a Nobel Prize he will forever be remembered as the father of observational cosmology and as a pioneer of the distant stars.

Figure 1
Edwin Hubble around 1920

*) Pro ISSI lecture, Bern, November 6, 2003

... and the Telescope

Its History

In the 1970s, NASA and ESA took up the idea of a space-based telescope. Funding began to flow in 1977. Later, it was decided to name the telescope after Edwin Hubble. Although the Hubble Space Telescope (HST) was downsized later to a 2.4 m primary mirror diameter from the initial 3 m, the project started to attract significant attention from astronomers.

The precision-ground mirror was finished in 1981 and the assembly of the entire spacecraft was completed in 1985. The plan called for a launch on NASA's Space Shuttle in 1986 but just months before the scheduled launch, the Challenger disaster caused a 2-year delay of the entire Shuttle programme. HST was finally launched on 24 April 1990. Soon after, the tension built up as astronomers examined the first images through the new telescope's eyes. It was soon realised that its mirror had a serious flaw: the mirror edge was too flat by a mere fiftieth of the width of a human hair, enough of focusing defect to prevent it from taking sharp images.

Fortunately, the HST was the first spacecraft ever to be conceived as serviceable. That made it possible for engineers and scientists at the Space Telescope Institute in Baltimore (USA) to come up with a cleverly designed corrective optics package that would restore the telescope's eyesight completely. A crew of astronauts including Claude Nicollier carried out the repairs necessary to restore the telescope to its intended level of performance during the first Hubble Servicing Mission (SM1) in December 1993. This mission captured the attention of both astronomers and the public at large to a very high degree: meticulously planned and brilliantly executed, the mission succeeded on all counts. It will go down in history as one of the great highlights of human space flight. Hubble was back in business!

The European contribution covered a nominal 15% stake in the mission including the Faint Object Camera, the first set of two solar panels that powered the spacecraft as well as a team of space scientists and engineers at the Space Telescope Science Institute (STScI) in Baltimore. Contraves Space of Zurich provided the Primary Deployment Mechanism, assuring the deployment of the solar panels either by a command from the onboard computer or manually by an astronaut.

The Spacecraft

The Hubble Space Telescope is a long-term, space-based observatory. Observations are carried out in the visible, infrared, and ultraviolet light. It circles the Earth in 96 minutes on a circular orbit 600 km above the ground, inclined 28 deg. to the equator. Hubble's orbit above the Earth and its excellent stability allow astronomers to make the high resolution observations that are essential to open new windows to planets, stars, and galaxies. Furthermore, access to the infrared and ultraviolet light window can only be achieved from space, because the atmosphere prevents it from reaching the ground.

At the heart of the HST are a 2.4 m primary mirror, fine guidance sensors and gyroscopes as well as a collection of four science instruments that work from the near infrared through the visible to ultraviolet light. There are two cameras and two combined camera / spectrographs. Power for the computers and the scientific instruments onboard is provided by large solar panels.

The telescope uses an elaborate system of attitude controls to secure its stability during observations. The spacecraft has a pointing stability of 0.007 arcsec, which is equivalent to a 1 € coin in Paris seen from Bern. A system of reaction wheels manoeuvres the telescope into place and its position in space is monitored by gyroscopes. The fine guidance sensors are used to lock onto guide stars to ensure the extremely high pointing accuracy needed to make very accurate observations.

The Hubble Space Telescope Servicing Concept

As stated above, the Hubble Space Telescope was built around a revolutionary servicing concept from the very beginnings. The concept aimed at allowing not only repair missions in case of faulty components, but also to replace subsystems once technologically more advanced instruments or computers have become available. In addition, this concept allowed the mission to be put on track after the repair of the initial flaw of the primary mirror.

Since SM1, three other Servicing Missions have been carried out: during SM2 in 1997 two new instruments were installed. In Service Mission SM3A (1999) many of the spacecraft's crucial technical systems were exchanged. During SM3B in 2002 the telescope was again equipped with new science instruments. In the wake of the Columbia accident, however, and as a first consequence of the strategic redirection of the US science programme, NASA has cancelled the next servicing mission, which was foreseen in 2003. At this moment in time it is not yet clear whether the telescope will be serviced again. This may lead to a much shorter service life as compared to the initial plans, which foresaw an end of mission in 2010.

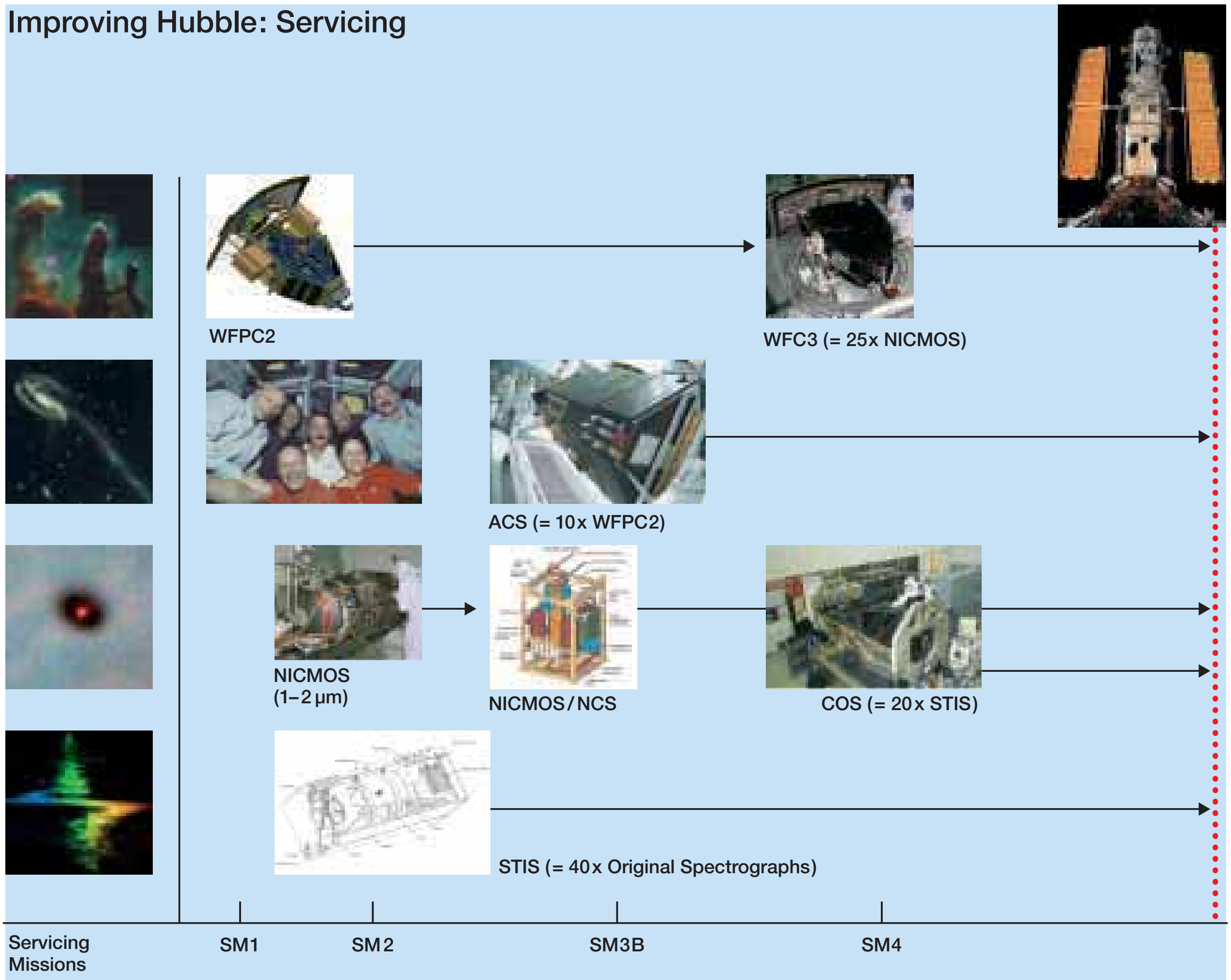


Figure 2 shows the servicing mission planning

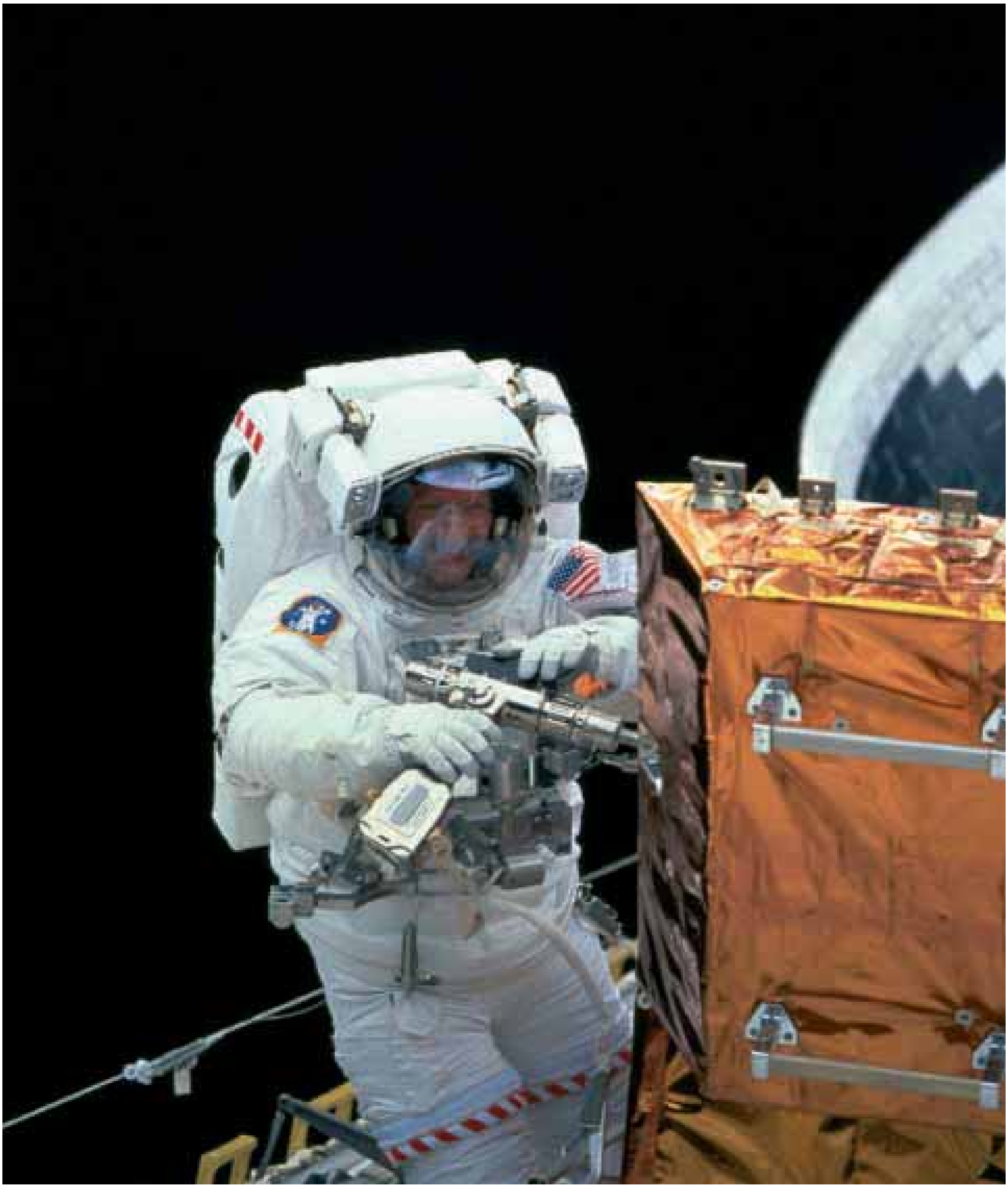


Figure 3: The famous photograph of Claude Nicollier as a happy space worker on the first Hubble Space Telescope Servicing Mission (Credit: NASA)

The Instruments

Hubble's science instruments currently include two cameras, two imaging spectrographs and a suite of fine guidance sensors:

The Wide-Field Planetary Camera 2 (WFPC-2)

The WFPC-2 is Hubble's workhorse camera. It records images through selection of 48 colour filters covering a spectral range from the far ultraviolet to visible and near-infrared wavelengths. It has produced most of the pictures that have been released as public outreach images over the years. Its resolution and excellent quality have made it the most used instrument in the first ten years of Hubble's life.

The Space Telescope Imaging Spectrograph (STIS)

The STIS is a versatile dual-purpose instrument consisting of a camera and a spectrograph operating over a wide range of wavelengths from near infrared to the ultraviolet.

The Near Infrared Camera and Multi-Object Spectrometer (NICMOS)

The NICMOS can take images and make spectroscopic observations of astronomical targets. It detects infrared light between 0.8 µm to 25 µm.

The Advanced Camera for Surveys

This camera replaced the initial Faint Object Camera built by the European Space Agency, which was returned to ESA after the Servicing Mission S3B.

The Fine Guidance Sensors (FGS)

The HST has three Fine Guidance Sensors on board. Two of them are used to point and lock the telescope onto the target and the third can be used for astrometry, making very precise position measurements to establish stellar distances and investigate stellar binary systems.

Figure 4 shows the Hubble Space Telescope after release from its first servicing mission.

Scientific Highlights

From Solar System Science…

Thanks to its unrivalled optical resolution power, HST provides visual information about the Solar System on a near continuous basis (as opposed to planetary fly-by's). The following examples show some of the most striking results in the field of Solar System research.

Figure 5 shows Jupiter's aurora with the footprints of its moons Io, Ganymede and Europa. Since Jupiter possesses a strong magnetic field, the footprints of its moons become visible. Io produces particles from volcanic eruptions. The ionised gas becomes trapped by Jupiter's magnetic field. As the particles spiral along the magnetic field lines, they hit Jupiter's atmosphere near the poles (the magnetic field threads the north and south poles similar to Earth's magnetic field). Generally, the particles follow the moon's field lines to the poles with some straggling particles providing the lagging lines. HST discovered this phenomenon because, unlike the earlier Voyager flybys, it can observe the planets continuously on timescales ranging from minutes to years. (Credit: STScI, John Clarke)

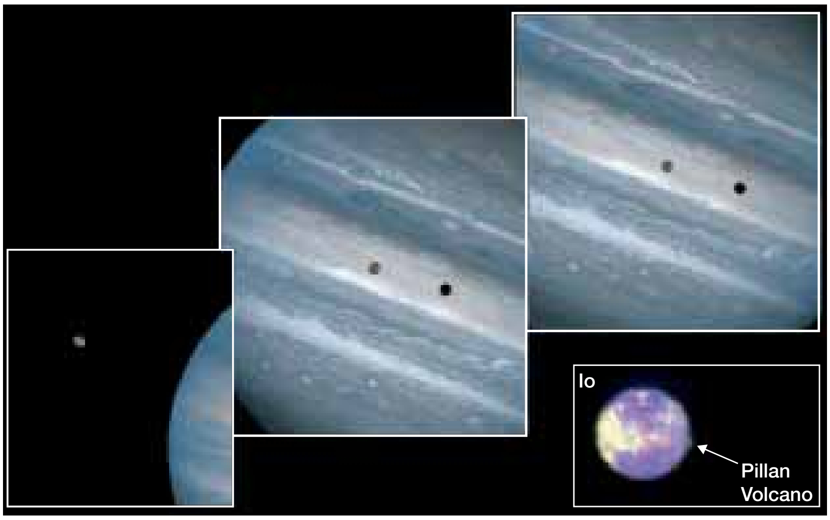


Figure 6: Transit of Io above Jupiter. HST is not only able to show the Jupiter moons, but also the shadows they cast onto the surface of Jupiter. On the left image, Jupiter's moon Io is seen on the left from Jupiter. In the centre and the right images Io, together with its shadow is seen before the planet. Even Io's volcanic activity can be identified with the 400 km high gas and dust plume emerging from its silhouette (right, bottom image). (Credit: STScI, J. Spencer)

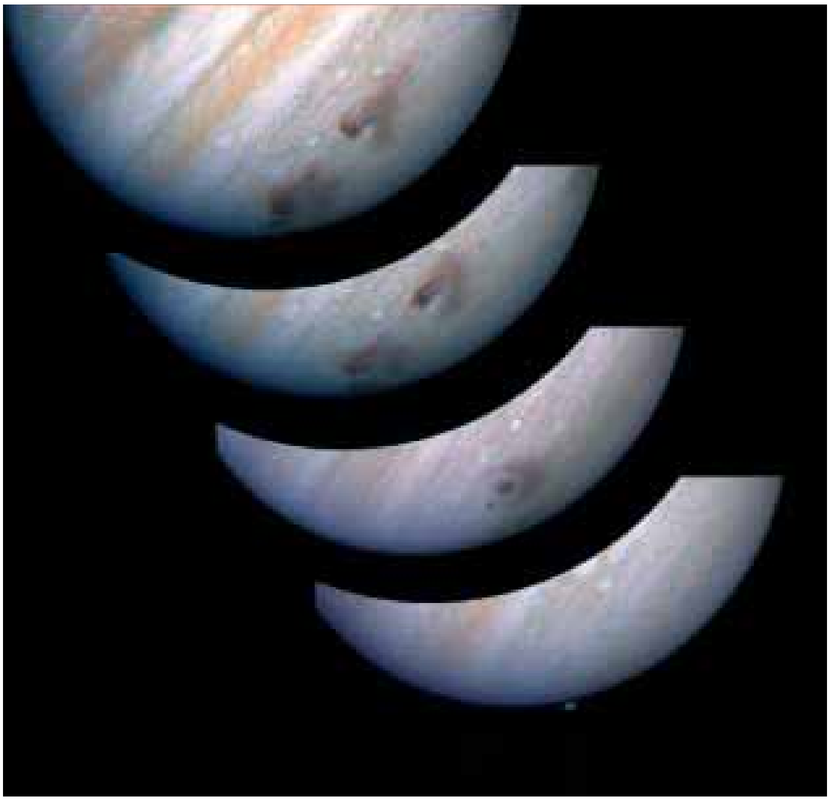


Figure 7 captures a cosmic catastrophe, when in 1994 the comet Levy-Shoemaker came near Jupiter, which by virtue of its giant gravitational field not only attracted the comet but also broke it into parts. This figure shows Jupiter being hit by two of Levy-Shoemaker's comet fragments and the subsequent disturbances in its atmosphere (from bottom to top). (Credit: STScI)

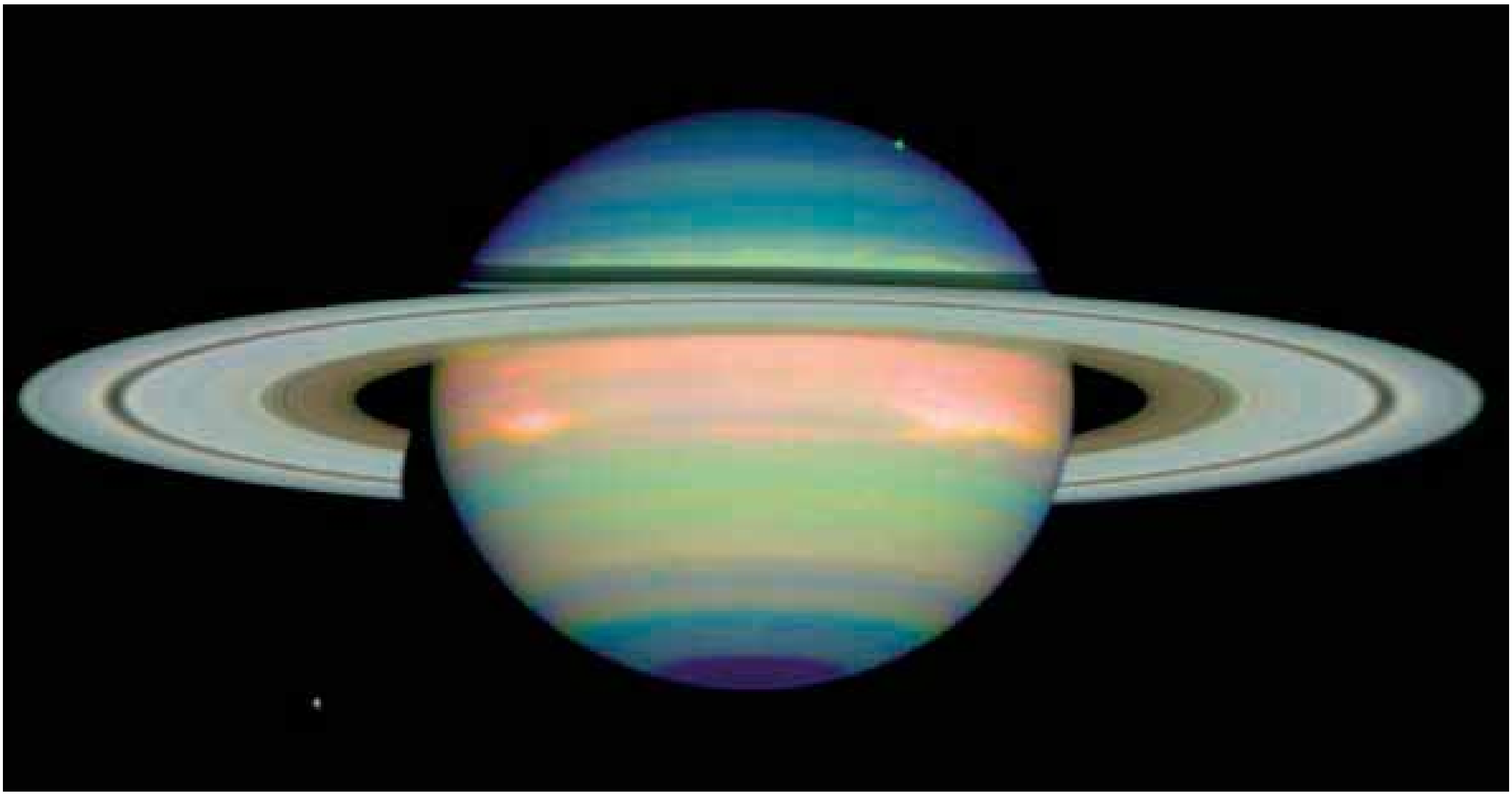
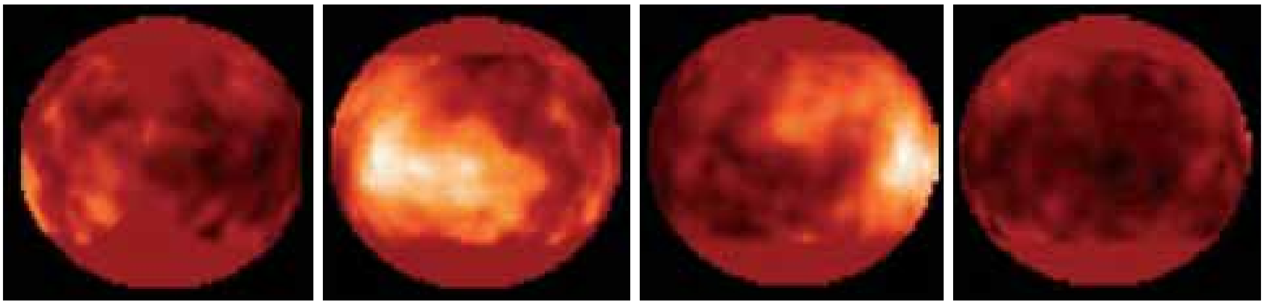


Figure 8 shows the planet Saturn in the infrared light. The picture is a combination of three images from Hubble's NICMOS instrument and shows the planet in reflected infrared sunlight. Different colours indicate varying heights and compositions of cloud layers generally thought to consist of ammonia ice crystals. The eye-catching rings cast a shadow on Saturn's upper hemisphere. The bright stripe seen within the left portion of the shadow is infrared sunlight streaming through the large gap in the rings known as the Cassini Division. Two of Saturn's many moons have also put in an appearance, Thetys just beyond the planet's disk at the upper right, and Dione at the lower left. (Credit STScI, E. Karkoschka)



Figures 9 (top) and 10 (bottom): HST renders invaluable services to mission planners when it comes to exploring planets and moons in the frame of future landing missions. In the case of the joint NASA / ESA mission Cassini / Huygens it was of primordial importance to get as much information on Titan and its atmosphere in order to design properly the details of the lander mission for Huygens. The Near Infrared Camera and Multi-Object Spectrometer (NICMOS) was best suited to penetrate Titan's opaque atmosphere containing nitrogen and carbon compounds and to provide information about its surface. Figure 9 above shows Titan seen by the NICMOS from four different directions in a distance of approx 1.5 billion km. This is approximately equivalent to observing a football in Zurich from Berne. The brighter areas are believed to be icy continents in oceans of methane and ethane, and of course scientists are eager to place Huygens on a beautiful site along one of Titan's beaches in January 2005, as sketched in Figure 10. (Credits: Figure 9: STScI, Figure 10: ESA)

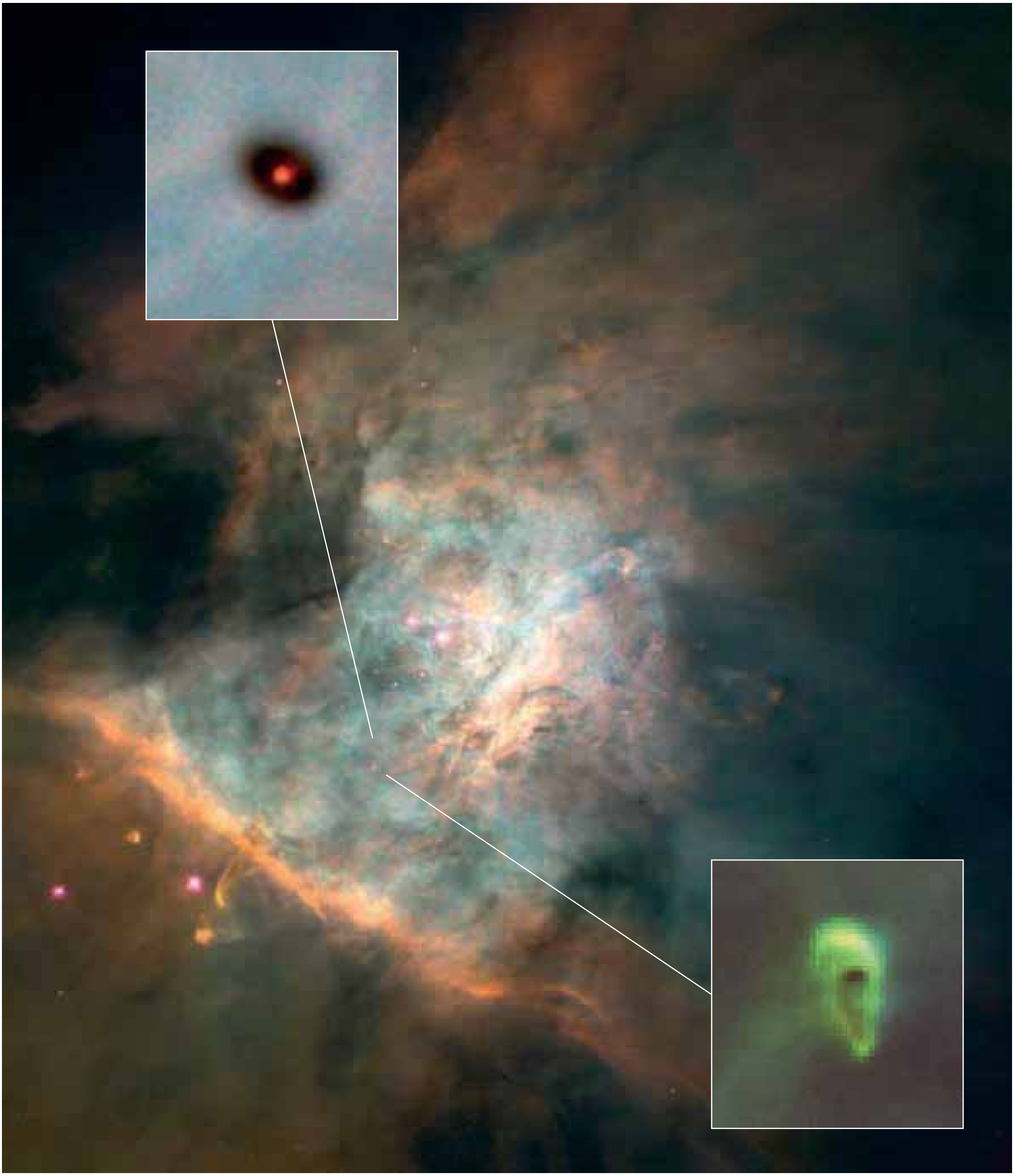


Figure 11 depicts one of the close-by cradles of formation of stellar systems in the Orion nebula. The two disk images show the two distinct disk types: a disk silhouetted against the background nebula (top left) and a disk inside the nebula whose outer parts are also ionised creating a glowing, teardrop-shaped bubble of ionised gas around it. These disks are the precursors of new planetary systems. Hubble Space Telescope's images of these disks were the first direct pictures confirming the general planet-creation scenario that has been stipulated by Immanuel Kant in 1755 and Pierre-Simon Laplace some years later. (Credit: STScI, C. Robert O'Dell)

...to Future Solar Systems...

Of course, the Hubble Space Telescope is instrumental not only in discovering the secrets of the Solar System but of other solar systems that exist around other stars. In the interstellar space, where dust and gas is extremely thinly distributed, irregularities of the density of matter eventually cause the higher densities to generate slightly increased gravitational fields, which in turn slowly attract additional matter from the ambient space, leading to a further increase of the gravitational forces. An avalanche process is initiated whereby more and more mass is concentrated in time spans of 10,000 of years. After around 1 million year a protoplanetary disk is formed, which, due to the angular momentum of the former dust cloud, rotates around an axis perpendicular to the future solar system's ecliptic plane. After a further 100 million years, when the mass of the new star has exceeded a certain critical threshold, the nuclear fusion process in its core is ignited, burning hydrogen into helium thereby releasing enormous amounts of energy, causing the new star to shine. From the debris remaining in the disk, planetesimals of increasing mass are formed and after another 1 billion years the planets have gathered the majority of matter circling around the central star. If there is a planet not too close to the central star and not too far away it may contain liquid water on its surface that may become a new shelter of life in the cosmic void. (See also Spatium 6 for further details).

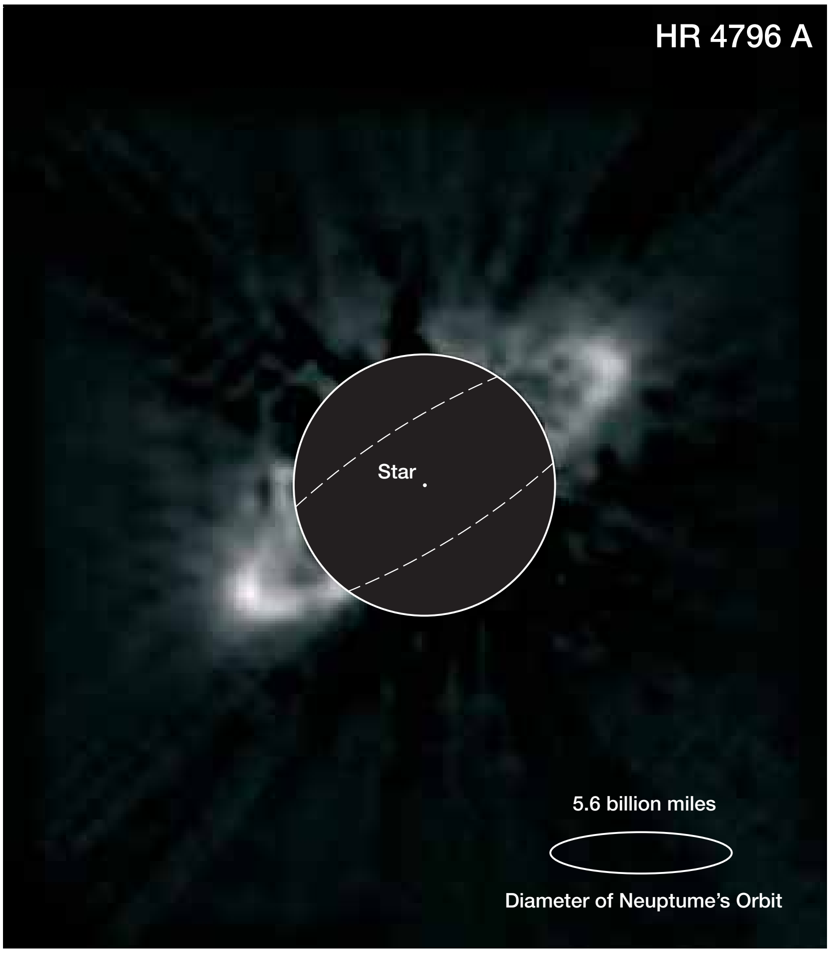
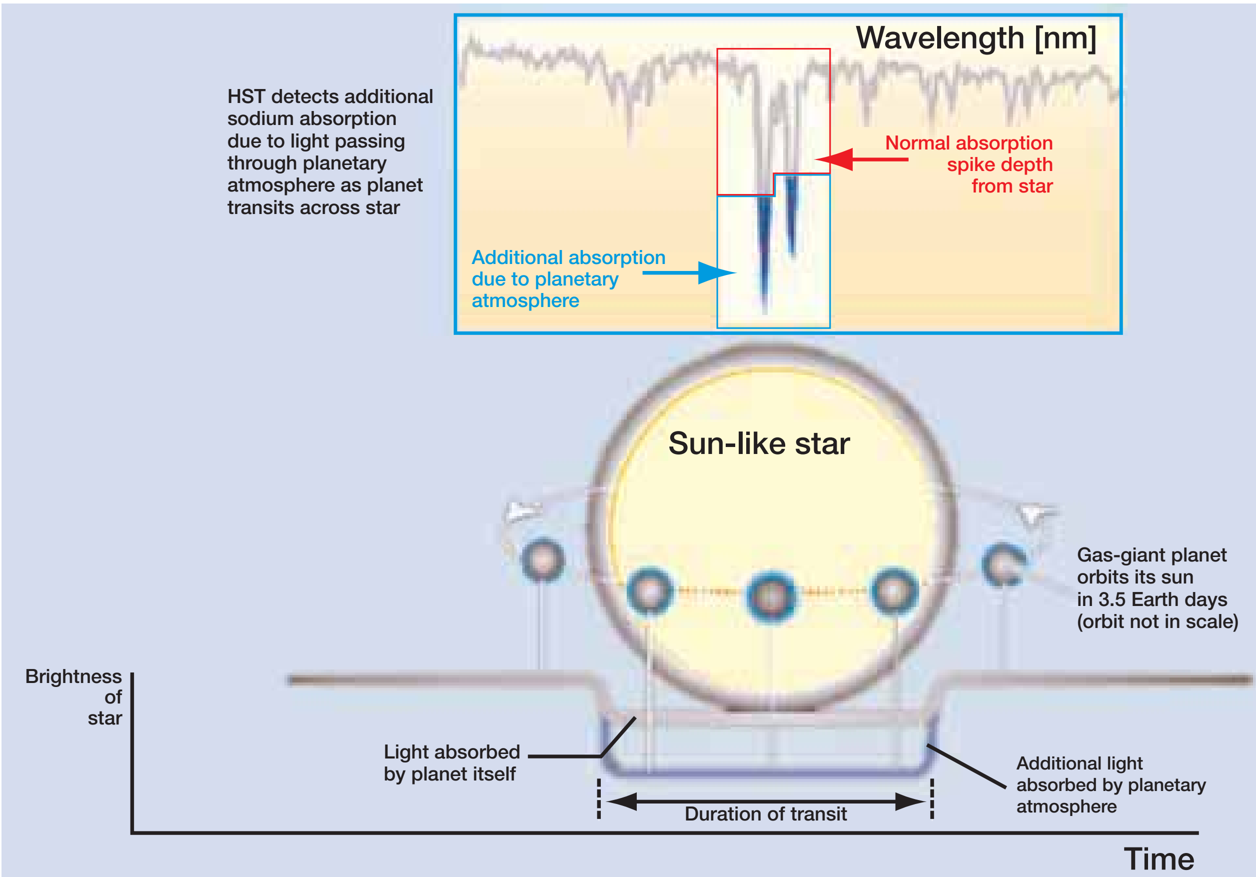


Figure 12 shows the capability of central obscuring of the NICMOS Instrument. By eclipsing the light of the central star like a solar eclipse but with an opaque mask, the faint light of rings of matter around HR 4796A becomes visible. This is the very place where planets will form. Note the dimension of this protoplanetary disk being in the order of 20 billion km, about twice the diameter of the Solar System. Rings seen in disks around main sequence stars are most easily interpreted as gravitational sweeping by small companions to the stars i.e. planets. The only way that we know how to produce rings is to have gravitational sweeping of the matter by companion bodies. (Credit STScI)



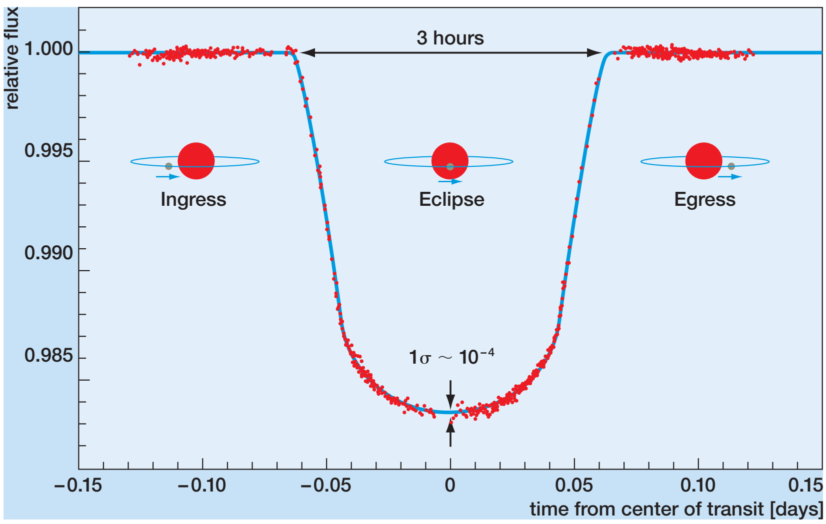


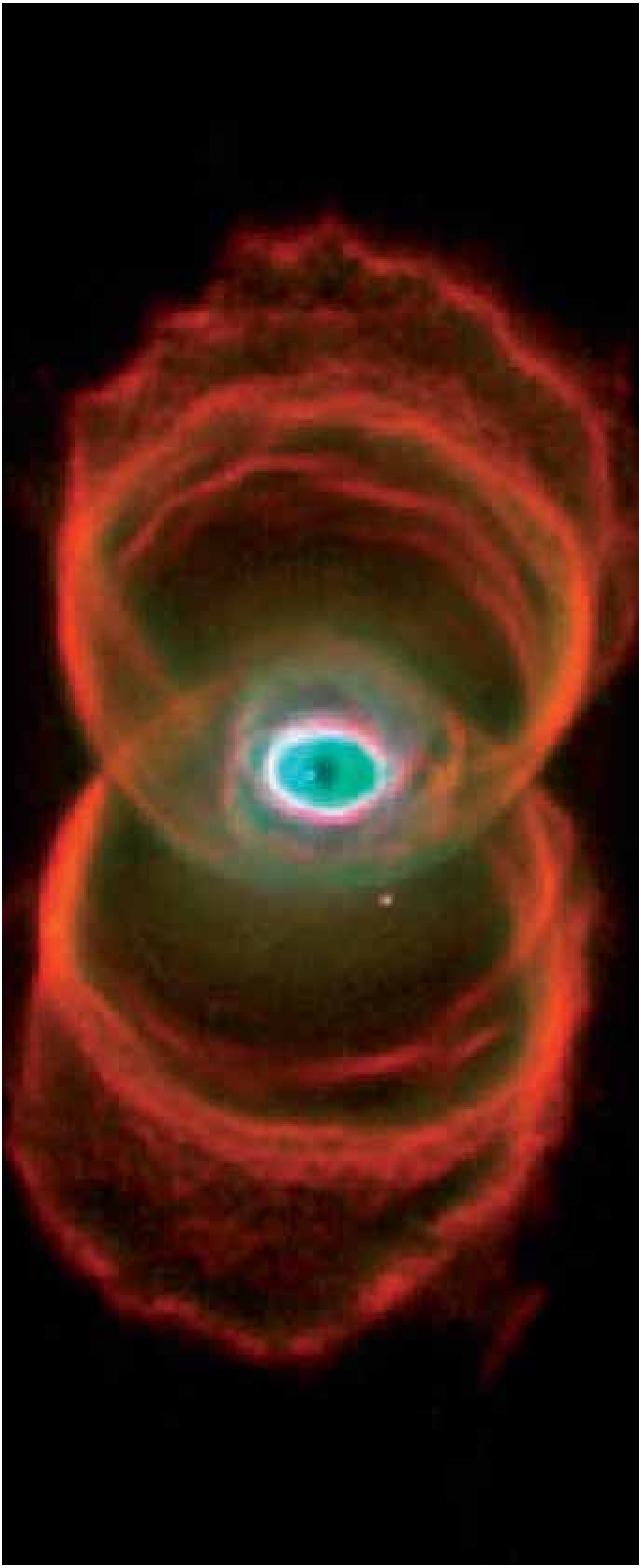
Figure 13: The STIS aboard HST is a powerful spectrograph. This figure shows schematically how the sodium abundance in the atmosphere of the planet HD 209458b was detected. The STIS was used to observe the star before, during, and after the planet partially eclipsed the star. Astronomers compared the depth of the eclipse at the wavelengths of two absorption lines of sodium with the depth at adjacent wavelengths free of absorption. Since the sodium in the atmosphere absorbed more starlight, the eclipse depth at those wavelengths was slightly greater than in the adjacent wavelengths. The versatility and power of HST made it possible to perform such pioneering measurements. (Credit: STScI)

... and the Death of Stars

While the birth of stars is a slow process their death is a spectacular show. A star with sufficient mass content shines thanks to the nuclear fusion of its light elements. When this fusion process comes to an end, no heat generating fusion process keeps the star shining anymore, it will, therefore, collapse under its own gravitational field and release thereby a lot of gravitational energy causing its relics partially to be merged to heavier elements, which are then scattered into space. Together with fresh hydrogen these heavier atoms may become the seeds of new generations of stars whose planets eventually may support life. When our Solar System formed around 4600 million years ago it also collected the relics of such former generations of stars, that at the end of their lives had produced heavy elements, which today are found on Earth and in our own bodies. (see also Spatium 3 for further information).

Figure 14: The shapes of planetary nebulae may become extremely complex like in the case of the Red Spider Nebula. It is still not understood how such complex structures may evolve. But in any case, nature is beautiful, when it explodes. (Credit: STScI)

Figure 15 shows the hourglass nebula through the light scattered by the gas expelled by the dying star at an earlier epoch. The shapes of dying stars show an astonishing variety of colours and forms depending on the properties of the former star. This HST snapshot of MyCn18, a young planetary nebula, reveals that the object has intricate patterns of "etchings" in its walls. A planetary nebula is the glowing relic of a dying, Sun-like star. The results are of great interest because they shed new light on the poorly understood ejection of stellar matter that accompanies the death of these stars. According to one theory on the formation of planetary nebulae, the hourglass shape is produced by the expansion of a fast stellar wind within a slowly expanding cloud, which is denser near its equator than near its poles. (Credits: Raghvendra Sahai and John Trauger, JPL)



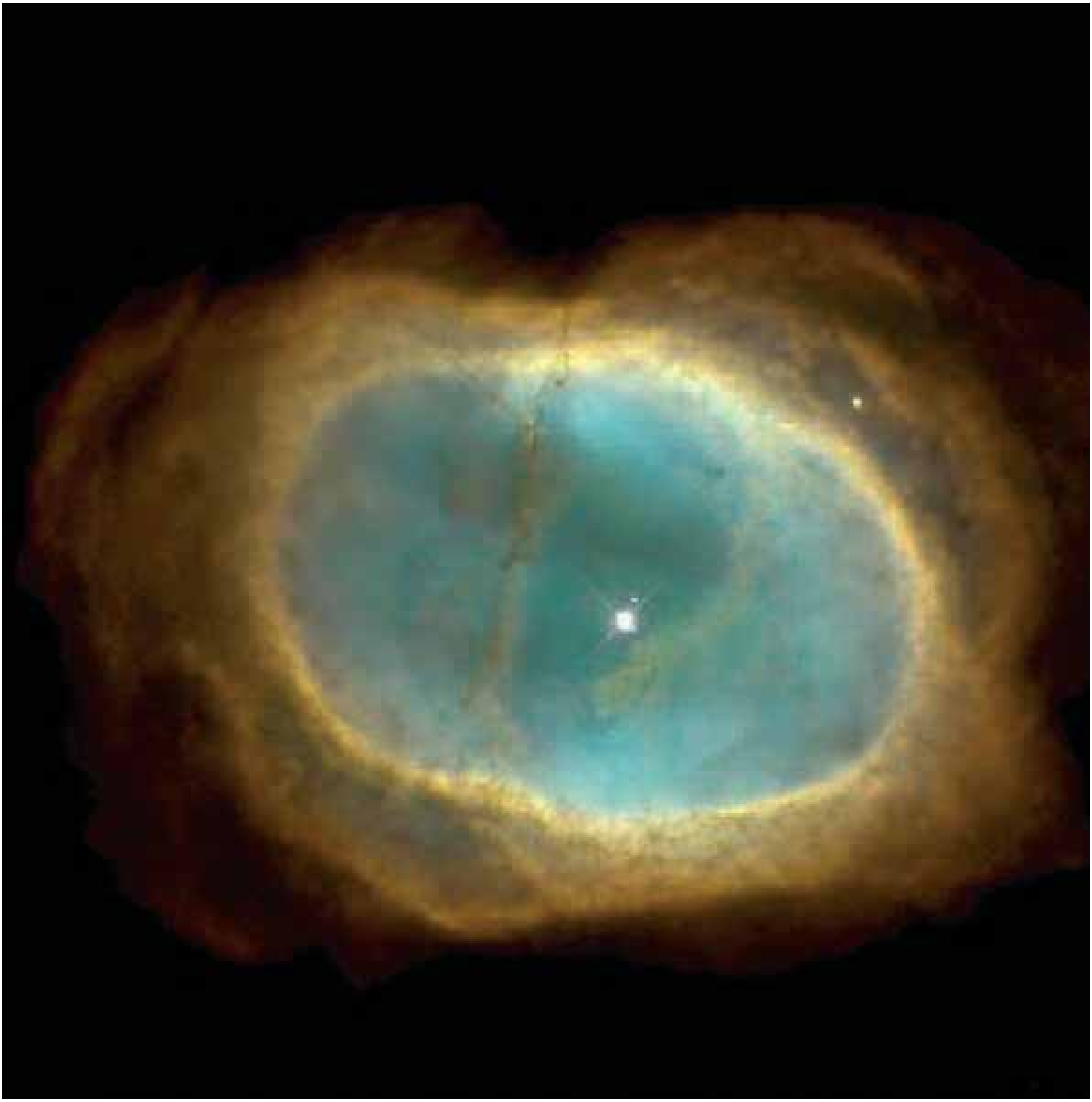


Figure 16: The Eight-Burst Nebula NGC 3132 is another striking example of a planetary nebula. This expanding cloud of gas, surrounding a dying star, is known to amateur astronomers in the southern hemisphere as the "Eight-Burst" or the "Southern Ring" Nebula. The name "planetary nebula" refers only to the round shape that many of these objects show when examined through a small visual telescope. In reality, these nebulae have little or nothing to do with planets, but are instead huge shells of gas ejected by stars as they near the ends of their lifetimes in fact probably ingesting any planet which they might have around them. NGC 3132 is nearly half a light year in diameter, and at a distance of about 2000 light years it is one of the nearer known planetary nebulae. The gases are expanding away from the central star at a speed of 15 km/s. This image clearly shows two stars near the centre of the nebula, a bright white one, and an adjacent, fainter companion to its upper right. (A third, unrelated star lies near the edge of the nebula.) The faint partner is actually the star that has ejected the nebula. This star is now smaller than our Sun, but extremely hot. The flow of ultraviolet radiation from its surface makes the surrounding gases glow through fluorescence. The brighter star is in an earlier stage of stellar evolution, but in the future it will probably eject its own planetary nebula. (Credit: STScI)

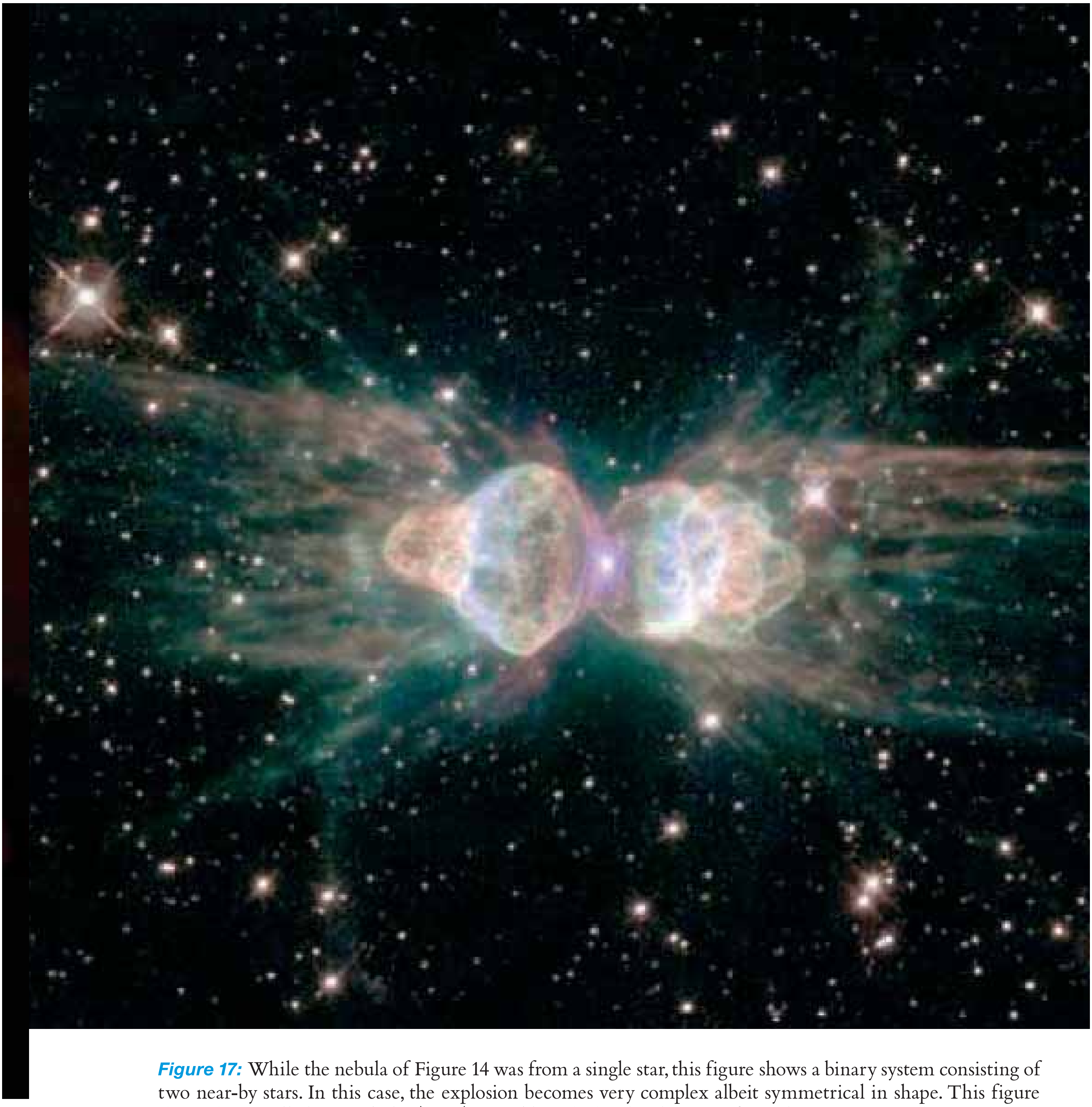


Figure 17: While the nebula of Figure 14 was from a single star, this figure shows a binary system consisting of two near-by stars. In this case, the explosion becomes very complex albeit symmetrical in shape. This figure shows the so-called "ant nebula" (Mz 3) resembling the head and thorax of a garden-variety ant. This HST image reveals the "ant's" body as a pair of fiery lobes protruding from the dying star. Scientists using Hubble would like to understand how a spherical star can produce such prominent, non-spherical symmetries in the gas that it ejects. One possibility is that the central star of Mz 3 has a closely orbiting companion that exerts strong gravitational tidal forces, which shape the outflowing gas. For this to work, the orbiting companion star would have to be close to the dying star, about the distance of the Earth from the Sun. At that distance, the orbiting companion star would not be far outside the hugely bloated hulk of the dying star. It is even possible that the dying star has consumed its companion, which now orbits inside it. A second possibility is that, as the dying star spins, its strong magnetic fields are wound up into complex shapes. Charged winds moving at speeds up to 1000 km per second from the star, much like those in our Sun's solar wind but millions of times denser, are able to follow the twisted field lines on their way out into space. These dense winds can be rendered visible by ultraviolet light from the hot central star or from highly supersonic collisions with the ambient gas that excites the material and make it to fluoresce. (Credit: STScI)

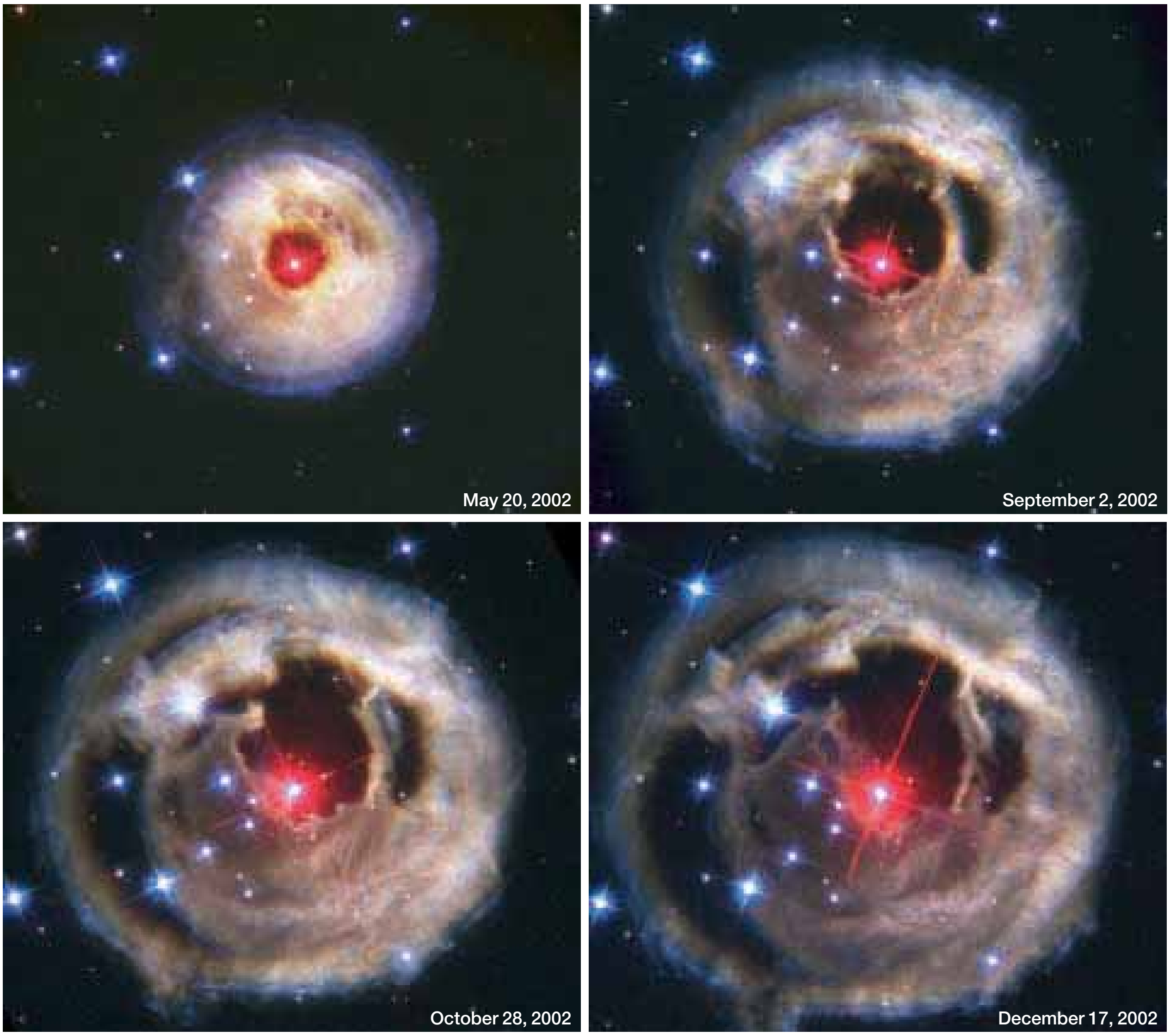


Figure 18: One of the most spectacular images obtained by the HST is this sequence of the V838 Monocerotis collected over a time span of about six months. Calculations of the speed of the shells showed that they move with the speed of light. This is obviously an illusion. The visible shells are in fact different parts of dust clouds released by the central dying star in the course of its life. They are illuminated by short bursts of light released by the star and spreading with the velocity of light. This effect is called a "light echo." The red star at the centre of the eyeball-like feature is an unusual erupting super giant called V838 Monocerotis, located about 20,000 light-years away in the winter constellation Monoceros (the Unicorn). During its outburst the star brightened to more than 600,000 times our Sun's luminosity. The outer circular feature is slightly larger than the angular size of Jupiter on the sky. For several more years its diameter will increase as reflected light is scattered from more distant portions of the nebula. Eventually, when light starts to be released from the backside of the nebula, the light echo will give the illusion of contraction, and finally it will disappear probably by the end of this decade. The black gaps around the red star are regions of space where there are holes in the dust cloud. (Credit: STScI)

Figure 19 shows the Whirlpool Galaxy perpendicular to its galactic plane. We clearly see the spiral arms, where billions of stars that build up this galaxy are concentrated. The red stars are young starting stars, while the blue dots are large massive stars. New pictures from Hubble are giving astronomers a detailed view of the Whirlpool galaxy's spiral arms and dust clouds, which are the birth sites of massive and luminous stars. This galaxy, also called M51 or NGC 5194, is having a close encounter with a nearby companion galaxy, NGC 5195, just off the upper edge of this image. The companion's gravitational influence is triggering star formation in the Whirlpool, as seen by the numerous clusters of bright, young stars, highlighted in red. (Credit: STScI)

Figure 20 shows Gomez's Hamburger, a Sun-like star nearing the end of its life. The hamburger buns are light scattered from the central star, which itself is obscured by a large band of dust in the middle. (Credit STScI)

Beyond the Milky Way

One of the major contributions of Edwin Hubble was the observation of galaxies beyond the Milky Way. Our galaxy is a flat disk with a diameter of approx. 100,000 light years. We find ourselves in one of several spiral arms far outside its centre. The large aperture of the HST as well as its (now) nearly perfect optical quality allows observations of distant objects out to regions, which have never been observed before. Since light travels with a limited speed of 300,000 km/s we observe distant objects not in their present state, but at an earlier state in the past. The further they are the earlier can we look into the past. In addition it allows us to observe galaxies like ours but from very different angles.

Figure 21 shows the ESO 510-G13 galaxy, whose galactic plane seems to be bent. HST has captured an image of this unusual edge-on galaxy, revealing remarkable details of its warped dusty disk. In contrast, the dust and spiral arms of normal spiral galaxies, like our own Milky Way, appear flat when viewed edge-on. (Credit: STScI)

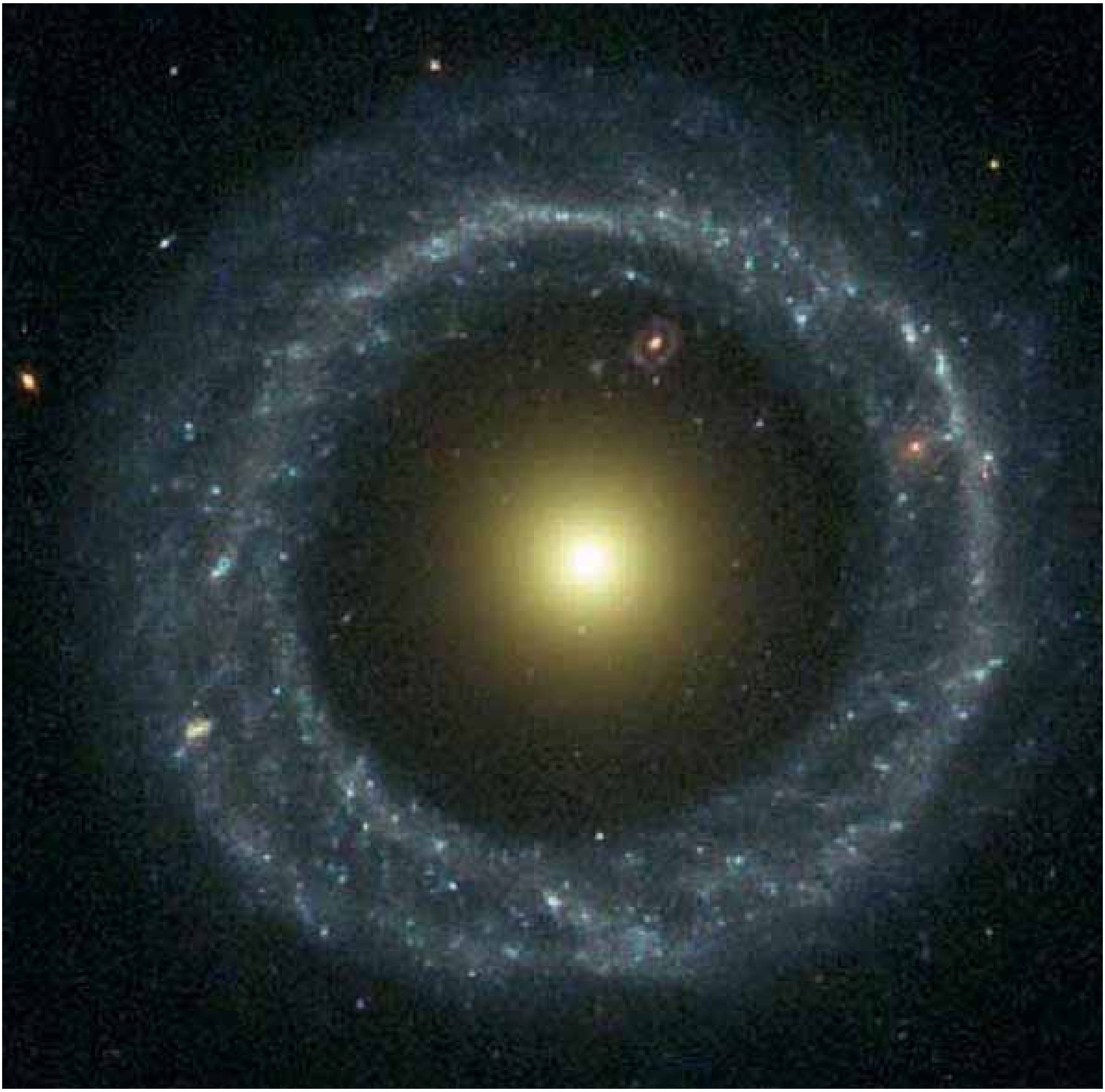


Figure 22 shows a ring galaxy called the Hoag's Object. It is a young galaxy, which has not yet developed its spiral arms. A nearly perfect ring of hot, blue stars pinwheels about the yellow nucleus. The entire galaxy is about 120,000 light-years wide, which is slightly larger than our Milky Way. The blue ring, which is dominated by clusters of young, massive stars, contrasts sharply with the yellow nucleus of mostly older stars. What appears to be a "gap" separating the two stellar populations may actually contain some star clusters that are almost too faint to see. Curiously, an object that bears an uncanny resemblance to Hoag's Object can be seen in the gap at the one o'clock position. The object is probably another ring galaxy.

Ring-shaped galaxies can form in several different ways. One possible scenario is through a collision with another galaxy. Sometimes the second galaxy speeds through the first, leaving a "splash" of star formation. But in Hoag's Object there is no sign of a second galaxy, which leads to the suspicion that the blue ring of stars may be the shredded remains of a galaxy that passed nearby. Some astronomers estimate that the encounter occurred about 2 to 3 billion years ago. The galaxy is 600 million light years away in the constellation Serpens. (Credit: STScI)

Figure 23 shows a cosmic catastrophe. This impressive image shows two colliding galaxies called NGC 3314, which lie about 140 million light years from Earth, in the direction of the southern hemisphere constellation Hydra. The bright blue stars forming a pinwheel shape near the centre of the front galaxy have formed recently from interstellar gas and dust. In many galaxies, interstellar dust lies only in the same regions as recently formed blue stars. However, in the foreground galaxy, NGC 3314a, there are numerous additional dark dust lanes that are not associated with any bright young stars. A small, red patch near the centre of the image is the bright nucleus of the background galaxy, NGC 3314b. It is reddened for the same reason the setting Sun looks red. When light passes through a volume containing small particles (molecules in the Earth's atmosphere or interstellar dust particles in galaxies), its colour becomes redder.

Collision does not mean that the single stars physically touch each other, since the distances between them are so large. But due to their gravitational fields collisions of galaxies may lead to collisions of stars caught by the gravitational field of large stars in one of the galaxies. Through an extraordinary chance alignment, a face-on spiral galaxy lies precisely in front of another larger spiral. This line-up provides us with the rare chance to visualise dark material within the front galaxy, seen only because it is silhouetted against the object behind it. Dust lying in the spiral arms of the foreground galaxy stands out where it absorbs light from the more distant galaxy. This silhouetting shows us where the interstellar dust clouds are located, and how much light they absorb. The outer spiral arms of the front galaxy appear to change from bright to dark, as they are projected first against deep space, and then against the bright background of the other galaxy. (Credit: STScI)

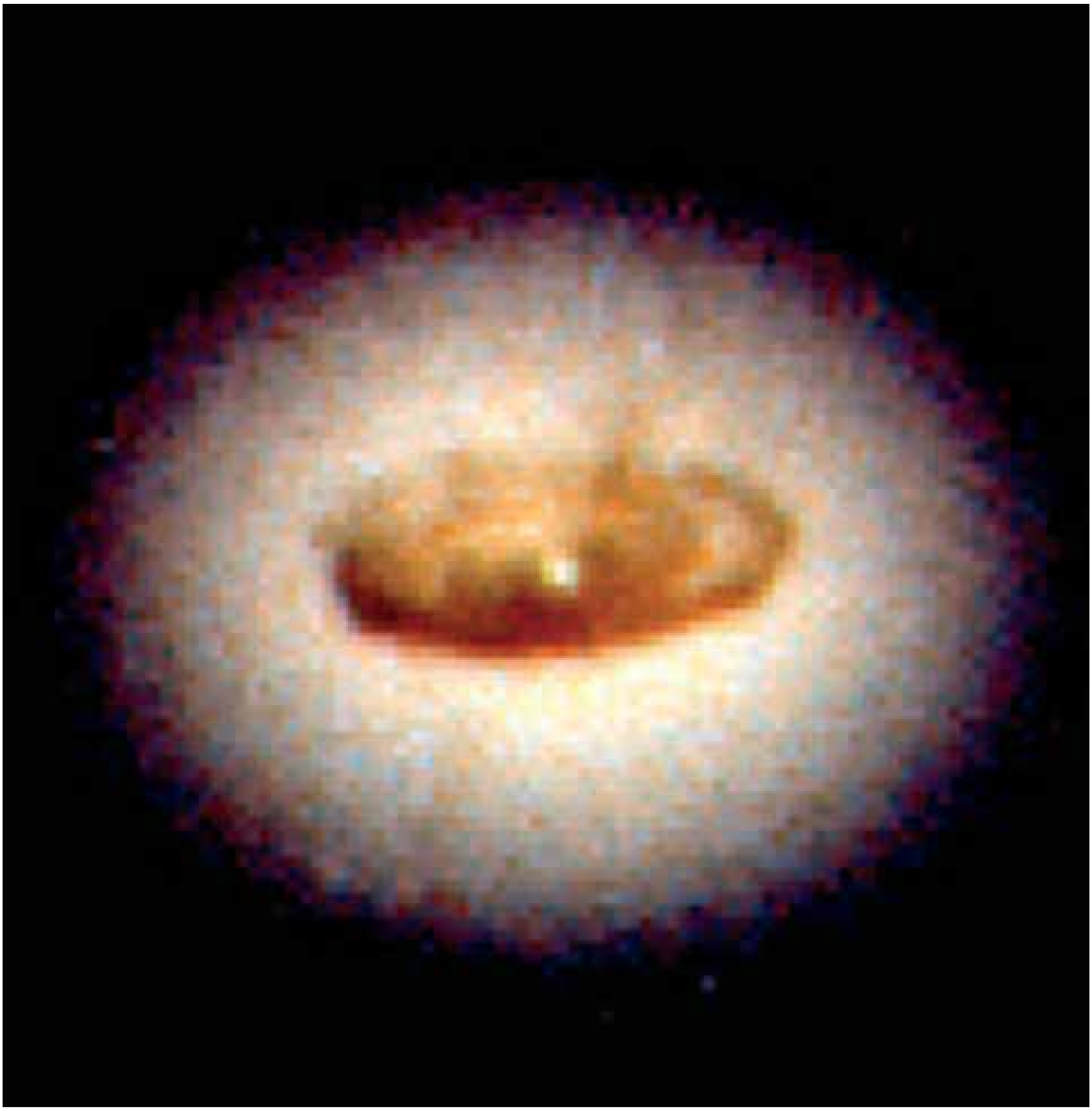


Figure 24 shows the Black Hole in the galaxy M84. HST has strongly contributed to the understanding of the mechanics of galaxies. Thanks to HST and new observations conducted in large ground based observatories such as the Very Large Telescope of the European Southern Observatory in Chile, we know that galaxies are powered by black holes in their centre. Their mass density is so high, that no matter, not even light can escape, hence their name. Due to their high mass they generate strong gravitational fields, the binding force of galaxies preventing the individual stars from escaping the common centre. In the case of M84, HST was able to provide pictures of the core of the galaxy. (Credit: STScI).

The Early Universe

HST's nearly perfect optical system combined with its excellent long-term stability allows to investigate regions of the universe which have never been reached before. Some 15 billion years ago, the universe started to expand and cooled over the next half million years. The glowing plasma of which it was composed, recombined into atoms of neutral gas, mostly hydrogen and some helium. The glow from this era of recombination has been observed as the cosmic microwave background radiation, and is used to study the large-scale geometry of the universe. Over the following half billion years or so, termed the Dark Age, the cold gas began to assemble into what we think are the first galaxies. The Dark Age ended when the light from the stars and the newly formed galaxies and quasars reionised the surrounding neutral gas.

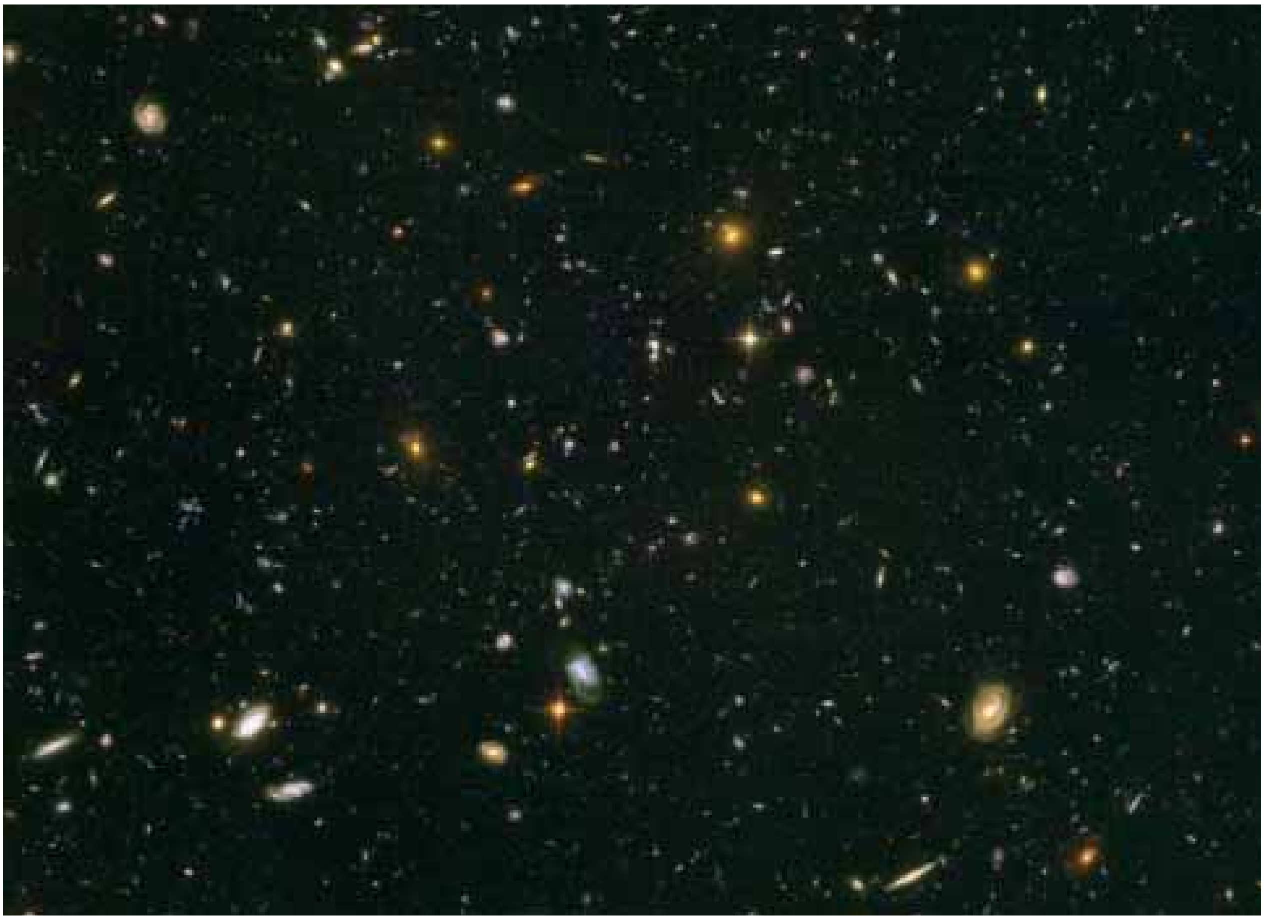


Figure 25 shows the deepest portrait of the visible universe ever achieved. This historic new view is actually based on a million-second exposure of two separate images taken by the Advanced Camera for Surveys and the Near Infrared Camera and Multi-Object Spectrometer. Both images reveal some galaxies emerging from the Dark Age, a mere 500 million years after the Big Bang. These galaxies are too faint to be seen in Hubble's previous faraway looks. The HUDF field contains an estimated 10,000 galaxies in a patch of sky just one-tenth the diameter of the full Moon. Besides the rich harvest of classic spiral and elliptical galaxies, there are many strange galaxies littering the field telling about a period when the Universe was more chaotic than today and when order and structure were just beginning to emerge.

Getting better with Age

The Hubble Space Telescope started to be designed in 1975 and was supposed to be launched in the early 1980s. Due to the loss of the Challenger Shuttle it was launched in 1990 only. Its technology therefore is nearly 25 years old now. During these 25 years technological progress was impressive and it continues probably to be so. Any non-serviceable spacecraft would already be outdated at the beginning of its lifetime in orbit. The ingenious concept to update the scientific instruments as well as elements like onboard computers, tape recorders, solar panels, etc. kept HST abreast of technological progress. In other words rather than becoming obsolete HST gets better with increasing age. **Figure 26** gives an impression of the importance of the concept of serviceability for the ambitious increase in sensitivity.

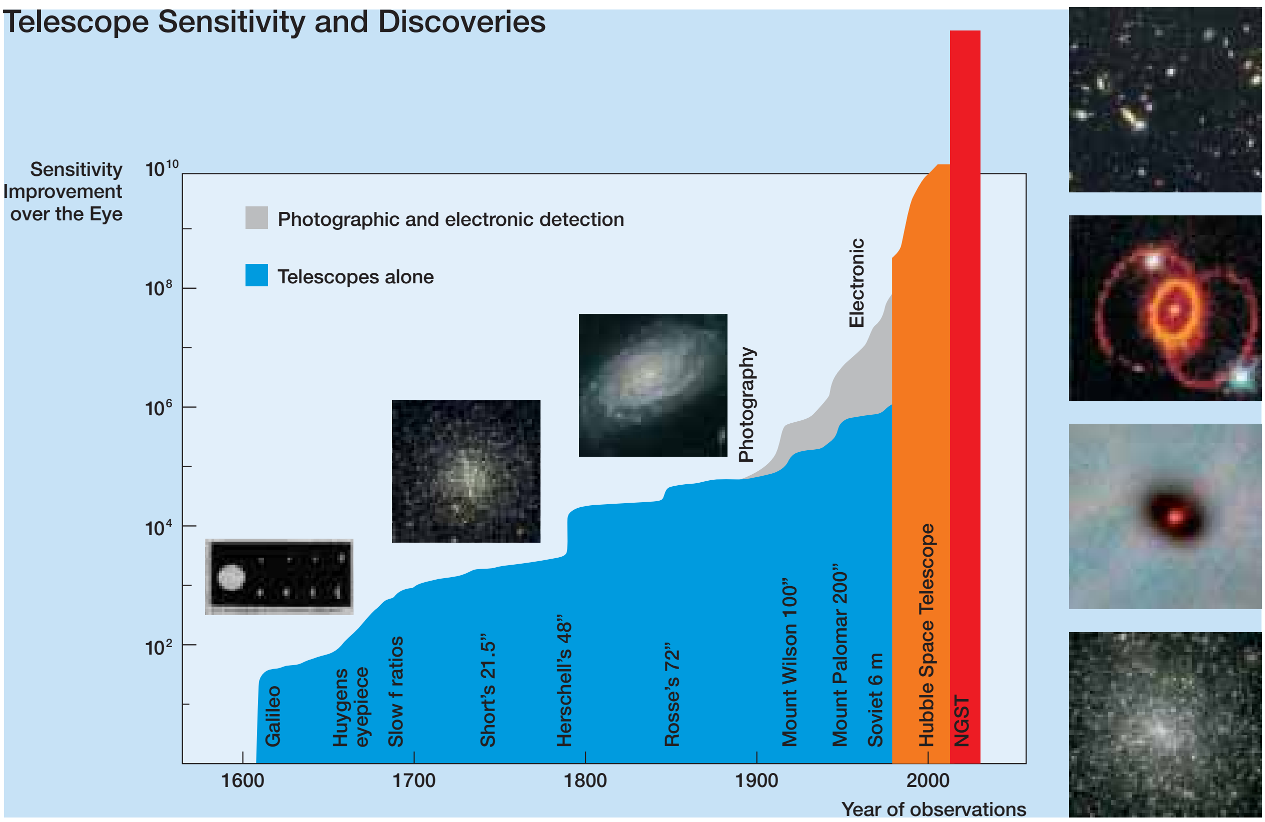


Figure 26 shows the evolution of observation sensitivity as compared to the naked human eye since Galileo Galilei's first astronomical observations in 1609. The striking result here is that since Galilei's initial advance, no other advance in technology has been proportionately as great as when HST became operational. It advances the state of the art by nearly two orders of magnitude.

It is important to note that the HST line in this figure is sloped upward. Each time the spacecraft is upgraded with new instruments its observing capabilities are increased, making HST a continuously improving facility. The right hand bar shows the expected gains to come from the James Webb Telescope (NGST), the successor of Hubble due to launch around 2011. (Credit: STScI)

After the Hubble Space Telescope

In sharp contrast to the spectacular images HST gathered from dying stars, its own end will be far from spectacular. Unlike most spacecraft Hubble has no onboard propulsion system, relying instead on gyroscopes and flywheels to maintain stability. Thus, at the end of its lifetime, special measures will have to be taken to prevent it from entering the Earth atmosphere in an uncontrolled way. Initially it was foreseen to be retrieved through a dedicated Shuttle mission, but, in the wake of the Columbia disaster, NASA is now forced to concentrate the Shuttle missions on the construction of the International Space Station. Therefore, plans have been presented lately to de-orbit the HST by means of a dedicated autonomous space tug. This spacecraft is intended to launch on a Delta 2 rocket, to grapple the HST and to guide it safely into the Earth's atmosphere over an unpopulated region. That will be the end of one of mankind's greatest achievements…

Several years ago NASA initiated a competition for the design of the successor instrument, which at that time was called the Next Generation Space Telescope. In 2003 the design by TRW Inc was chosen **(Figure 27)** and the mission was baptised James Webb Space Telescope in honour of NASA's second administrator. Unlike the HST, this instrument will not be serviceable, as it will be placed at the Lagrange L2 point that is 1.5 million km from the Earth leeward to the Sun. While this position is attractive, because it is in the Earth's shadow, hence less disturbed by solar radiation, it lacks servicing capabilities at least for the time being, since it is around 5 times farther from the Earth than the current Space Shuttle's reach.

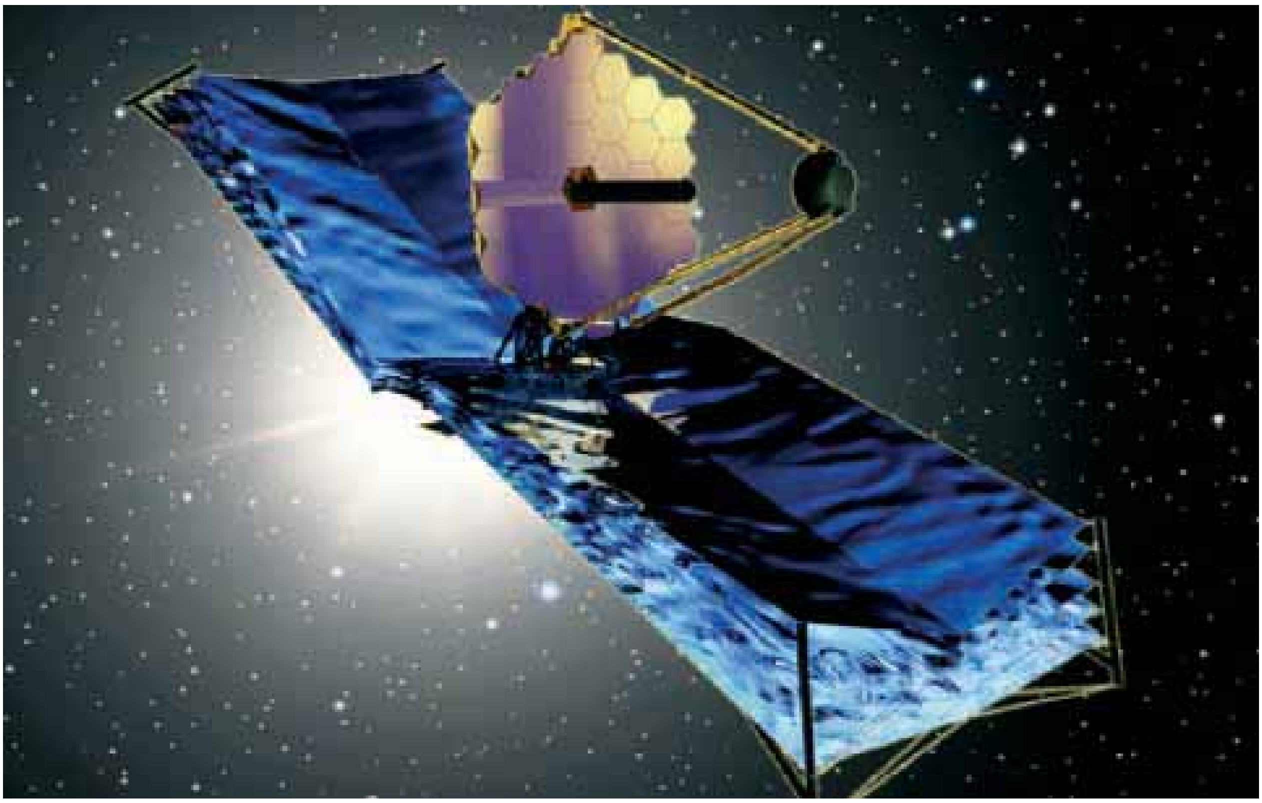


Figure 27: Artists view of the James Webb Telescope, the successor mission to the Hubble Space Telescope (Credit: TRW).

Conclusion

The unique advance permitted by HST amply sustains the European and American investment in this programme. The Hubble Space Telescope has truly revolutionised our knowledge of:

- the creation of galaxies,
- the distance scale of the universe,
- giant black holes in galaxies,
- the intergalactic medium,
- interstellar medium chemistry,
- extra solar planets,

and it is hoped that in the future years of its planned operation, the HST will bring us additional scientific surprises. It is an old wisdom that leading discoveries are often accidental: they follow the ability to observe, not necessarily the originally planned observation. Hubble has dramatically contributed to the modern revolution of astronomy permitted by space techniques and to our understanding of the universe.

Figure 28 shows another ring galaxy (AM 0644-741). The Hubble Space Telescope not only rendered invaluable services to scientists but made us also aware of the beauties of our world, much like that simple telescope, which stimulated Immanuel Kant some 250 years ago.

The author

Roger Maurice Bonnet received his diploma in physics and astronomy in 1968 from the University of Paris. He dedicated his doctorate to the imagery and spectroscopy of the Sun. He was involved in the development of new telescopes and cameras that were flown on balloons, rockets and satellites. In 1969 he founded the Laboratoire de Physique Stellaire et Planetaire of CNRS. There he was engaged in the development of the telescope for the Halley Multicolour Camera of the Giotto Mission under the responsibility of the Max Planck Institute at Lindau (Germany), which took the first high-resolution images of a cometary nucleus. R. M. Bonnet entered the field of international science policy in 1983, when the Council of the European Space Agency appointed him Director of the Science Programme. One of his most prominent achievements of that time was the formulation of the Horizon 2000 Plan. This programme consists of four large cornerstone missions and some smaller missions, amongst which is the European contribution to the Hubble Space Telescope. The transparent long term planning philosophy allowed the academic and the industrial communities to timely build up the required skills thereby placing ESA at the second most important space science agency in the world after NASA. Later, R. M. Bonnet adopted the same concept for the definition of the new Earth observation programme of ESA, known as the Living Planet Programme.

After his departure from ESA in April 2001 he served as Directeur General Adjoint Science at the French Centre National Etudes Spatiales (CNES). There he was asked by the French government to chair a National Commission to analyse and formulate the French space policy. In 2001, he was called by the Director General of ESA to advise him on how to formulate its Aurora programme, setting the goal of landing humans on Mars by 2030.

At the International Space Science Institute R. M. Bonnet succeeded Prof. J. Geiss as Executive Director in 2003. In the same year he was appointed President of the International Committee on Space Research (COSPAR). He is the author of over 150 scientific papers and textbooks and he holds the highest scientific and public service awards like for example the French Légion d'Honneur or the NASA Award for Public Service.

For R. M. Bonnet adventures like the Hubble Space Telescope are the convergence of many scientific and engineering talents in response to questions, which always have challenged the human mind. They realise missions thought impossible before, which open the brain of the young and less young children to the marvellous discovery of the universe.

International Space Science Institute

SPATIUM

Publikationsorgan des Vereins Pro ISSI

Nr. 13, November 2004

Woher kommen Kohlenstoff, Eisen und Uran?

Editorial

Zwei Dinge sind unendlich: Das Universum und die menschliche Dummheit; beim Universum bin ich mir noch nicht ganz sicher. Soweit Albert Einstein.

Unendlichkeit und Ewigkeit haben weder Anfang noch Ende und sprengen damit die heutigen Vorstellungen des Universums, dem ein zeitlicher Anfang und vielleicht auch ein apokalyptisches Ende zugeschrieben wird, und dessen Dimensionen vermutlich auch nicht unendlich sind.

Der sich seiner Hinfälligkeit bewusste Mensch hat seit je her versucht, der Vergänglichkeit zu entrinnen: wer auch immer etwas auf sich hielt, war bemüht, seiner Sterblichkeit zu entfliehen, sei es durch die berühmten, in Gold gearbeiteten Masken der ägyptischen Pharaonen bis hin zum modernen Menschen, der die Dienste der Medizin in Anspruch nimmt, um die in Falten gezeichnete Vergänglichkeit zu überwinden, für einen kurzen Augenblick wenigstens. Schon bei Heraklit findet sich die Erkenntnis, dass das hervorquellende Neue das Alte stets zu verdrängen hat und wir tun wohl auch heute gut daran, uns dies zu verinnerlichen, denn was heute ist, ist morgen schon nicht mehr.

Selbst das Gold, diese Chiffre der Unvergänglichkeit, ist zwar langlebig, hält aber auch nicht ewig und die radioaktiven Elemente führen uns den Zerfall nach den ihnen eigenen Gesetzen besonders drastisch vor Augen. Entstehung und Wandlung der Materie sind das Thema der vorliegenden Ausgabe des Spatium. Professor Rudolf von Steiger, Direktor am International Space Science Institute in Bern, hat am 6. April 2004 in einem faszinierenden Referat unsere Mitglieder in die Trilogie des *Werdens-Seins-Vergehens* der Materie eingeführt und wir freuen uns, heute darüber berichten zu können. Die vorliegende Nummer erscheint in deutscher Sprache: unsere Hommage an das Einstein-Jahr 2005 zur Feier des Annus Mirabilis 1905, in welchem Albert Einstein, damals Experte II. Klasse am Patentamt in Bern, unter anderem seine Abhandlung über die spezielle Relativitätstheorie veröffentlicht hat.

Hansjörg Schlaepfer
Zürich, Oktober 2004

Impressum

SPATIUM
Publikationsorgan des Vereins Pro ISSI
Erscheint ein- bis zweimal jährlich

INTERNATIONAL SPACE SCIENCE INSTITUTE

Verein Pro ISSI
Hallerstrasse 6, CH-3012 Bern
Tel. +41 (0)31 631 48 96
Fax +41 (0)31 631 48 97

Präsident
Prof. Heinrich Leutwyler, Universität Bern
Herausgeber
Dr. Hansjörg Schlaepfer, legenda schläpfer wort & bild, CH-8185 Winkel
Layout
Marcel Künzi, marketing · kommunikation, CH-8483 Kollbrunn
Druck
Schellenberg Druck AG
CH-8330 Pfäffikon

Die Supernova-Explosion, die den Krebs-Nebel bildete, konnte im Jahre 1054 erstmals beobachtet werden. Beinahe 1000 Jahre später entwickelt sich sein Zentrum immer noch mit Geschwindigkeiten von 5 Millionen Kilometern pro Stunde. Die orange gefärbten Stellen stammen von Wasserstoff, Stickstoff (rot), Schwefel (purpur) und Sauerstoff (grün). Quelle: William P. Blair (JHU) et al. Hubble Heritage Team (STScI/AURA), NASA.

Woher kommen Kohlenstoff, Eisen und Uran *)

Prof. Rudolf von Steiger, International Space Science Institute, Bern

Der Stein der Weisen

Der Alchemist James Price gab 1782 vor, mit Hilfe des Steins der Weisen, sowie einer Mixtur von weissem Pulver und Quecksilber Gold herstellen zu können; doch als ihn die Britische Royal Society auffordert, den Vorgang vor Zeugen zu wiederholen, begeht er vor ihren Augen Selbstmord…

Der Versuch, Gold aus anderen, weniger edlen Materialien herzustellen, war schon für die begnadeten Goldschmiede des alten Ägyptens ein stetes Ziel, denn Gold war knapp und in Gold gearbeitet wollten das Leben der grossen Pharaonen und ihre Werke der Nachwelt erhalten werden, wie die grossartige Maske des Pharao Tutenchamun (um 1330 v. Chr.) bis auf den heutigen Tag so eindrücklich beweist.

Bild 1: Maske des Pharao Tutenchamun um 1324 v. Chr.

Die Materie und ihre Wandlung blieben auch im klassischen Griechenland ein zentrales Thema. Anaximander (ca. 611–546 v. Chr.) postulierte die Existenz von Atomen als die Basis der Materie. Aristoteles (384–322 v. Chr.) dachte sich alle Dinge aus den vier Grundelementen Erde, Luft, Feuer und Wasser zusammengesetzt.

Im Jahre 288 v. Chr. gründete Ptolemäus die Bibliothek von Alexandria in Ägypten, den Treffpunkt der Weisen seiner Zeit. Diesem Schmelztiegel hellenistischer, ägyptischer und asiatischer Kultur verdanken wir nicht nur die ersten grossen wissenschaftlichen Errungenschaften des Abendlandes, sondern auch die Überlieferung antiken Wissens. Hier entdeckten die späteren Araber die griechischen Papyrusrollen mit Anleitungen zur Herstellung von Gold. Sie setzten die Suche der alten Ägypter nach dem glänzenden Metall fort und gaben der Wissenschaft einen Namen: «al kymia», das Land der alten Ägypter.

Auch im mittelalterlichen Europa zog der Stein der Weisen, jenes Symbol der Vollendung, dem die Kraft innewohnt, unedle Materie in reines Gold zu verwandeln, viele grosse und kleine Geister in seinen Bann. Mit ihrem rationalen Denken und Beobachten haben sie die Grundlagen für die heutige Chemie gelegt. Obwohl ihr Bemühen letztlich fruchtlos blieb, verdanken wir ihnen doch die inzwischen bestätigte Hypothese, dass die Materie aus einigen wenigen Grundbausteinen aufgebaut ist und somit wandelbar ist.

Tatsächlich:

Gold entsteht aus anderen Elementen, doch die dazu erforderlichen Energien standen den Alchemisten auch nicht annähernd zur Verfügung. Die Erkenntnisse der Kernphysik des 20. Jahrhunderts haben zwar Einblicke in die Welt des Mikrokosmos vermittelt, aber auch gezeigt, dass wir die Umwandlung unedler Elemente in das vollkommene Gold weiterhin den Sternen überlassen müssen…

*) Referat vom 6. April 2004 für die Mitglieder des Vereins Pro ISSI

Die Systematik der Elemente

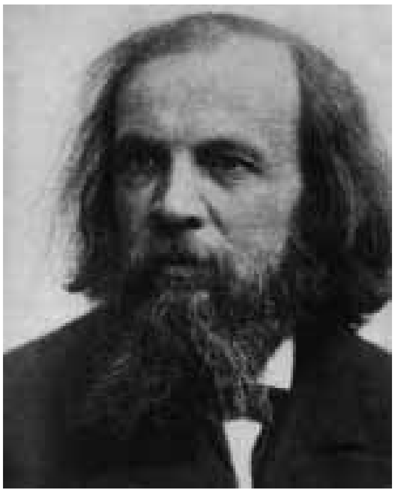
Den Übergang von mittelalterlicher Alchemie zur nachvollziehbaren Wissenschaft vollzog der gelernte Apotheker Johann Wolfgang Döbereiner (1780–1849). Zwar besass er kein akademisches Diplom; seine Genialität überzeugte aber Johann Wolfgang von Goethe so sehr, dass er ihm eine ausserordentliche Professur für Chemie, Pharmazie und Technologie an der Universität Jena vermittelte. Döbereiner ist der Vordenker des Periodensystems der Elemente, das der russische Chemiker Dmitri Mendeleev (1834–1907) später vollständig ausformulierte.

Das Periodensystem

Das Periodensystem ist eine systematische Anordnung aller chemischen Elemente nach Massgabe ihrer chemischen Eigenschaften. Es ermöglicht nicht nur die Aufzählung und Systematisierung aller bekannter Elemente; vielmehr konnten auch damals noch nicht entdeckte Elemente vorausgesagt werden.

Gruppe / Periode	1	2		3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1	1 H																		2 He
2	3 Li	4 Be												5 B	6 C	7 N	8 O	9 F	10 Ne
3	11 Na	12 Mg												13 Al	14 Si	15 P	16 S	17 Cl	18 Ar
4	19 K	20 Ca		21 Sc	22 Ti	23 V	24 Cr	25 Mn	26 Fe	27 Co	28 Ni	29 Cu	30 Zn	31 Ga	32 Ge	33 As	34 Se	35 Br	36 Kr
5	37 Rb	38 Sr		39 Y	40 Zr	41 Nb	42 Mo	43 Tc	44 Ru	45 Rh	46 Pd	47 Ag	48 Cd	49 In	50 Sn	51 Sb	52 Te	53 I	54 Xe
6	55 Cs	56 Ba	*	71 Lu	72 Hf	73 Ta	74 W	75 Re	76 Os	77 Ir	78 Pt	79 Au	80 Hg	81 Tl	82 Pb	83 Bi	84 Po	85 At	86 Rn
7	87 Fr	88 Ra	**	103 Lr	104 Rf	105 Db	106 Sg	107 Bh	108 Hs	109 Mt	110 Ds	111 Uuu	112 Uub	113 UUt	114 Uuq	115 Uup	116 Uuh	117 Uus	118 Uuo
*Lanthanoide			*	57 La	58 Ce	59 Pr	60 Nd	61 Pm	62 Sm	63 Eu	64 Gd	65 Tb	66 Dy	67 Ho	68 Er	69 Tm	70 Yb		
**Actinoide			**	89 Ac	90 Th	91 Pa	92 U	93 Np	94 Pu	95 Am	96 Cm	97 Bk	98 Cf	99 Es	100 Fm	101 Md	102 No		

Die Figur 2 stellt das Periodensystem der chemischen Elemente nach Dmitri Mendeleev dar. Die Elemente sind in der Reihenfolge zunehmenden Atomgewichts, links oben beginnend, zeilenweise aufgetragen. Die Zahlen oberhalb des Elementsymbols bezeichnen die Anzahl Protonen im Kern. Der Name des Periodischen Systems leitet sich ab von der auffallenden Periodizität, mit welcher die chemischen Eigenschaften der Elemente wiederkehren. Diese Periodizität basiert darauf, dass die Elektronen den Kern in Schalen umkreisen, und dass im Wesentlichen nur die Elektronen der äussersten Schale, die Valenzelektronen, für das Eingehen von Verbindungen verfügbar sind. Das Periodensystem ordnet die Elemente entsprechend dem Auffüllungsgrad der äussersten Schale an. So finden sich in der linken äusseren, blau markierten Spalte die Alkalimetalle mit einem einzigen Valenzelektron, welche Basen bilden, sowie rechts davon, rot markiert, die Erdalkalimetalle. In der hell-orangen Kolonne rechts aussen befinden sich die Edelgase. Da ihre äussere Elektronenschale voll ist und somit für chemische Verbindungen nicht zur Verfügung steht, sind sie alle reaktionsträge. Links davon befindet sich die Kolonne der säurebildenden Elemente von Fluor bis Jod. Gelb markierte Felder enthalten Nichtmetalle (elektrische Isolatoren), hellblaue Felder Metalle (elektrisch leitende Elemente) und hellrote Felder Halbmetalle (Halbleiter: Silizium, Germanium, etc.), welche als reine Elemente zwar nichtleitend sind, durch Verunreinigungen aber leitend werden.



Das Bild 3 zeigt Dmitri Mendeleev, den Entdecker des Periodensystems. Zu seinen Ehren ist das Element 101 Mendelevium benannt.

Das Atom

Wenn auch das periodische System schon ein immenser Schritt zum Verständnis der Materie war, so blieb es doch zunächst eine heuristische Darstellung, die sich einer theoretischen Begründung entzog. Erst die bahnbrechenden Erkenntnisse von Ernest Rutherford (1871–1937) ermöglichten die heutigen Vorstellungen über den inneren Aufbau der Materie. Auf Grund seines berühmten Streuexperiments von Alphateilchen (d.h. Heliumkernen) an Gold(!)folien postulierte er Atomkerne von 10^{-14} m Durchmesser, um die im Abstand von 10^{-10} m Elektronen kreisen. Materie besteht also aus Leere: einen Tennisball als Atomkern würden die Elektronen auf Bahnen von 800 m Radius im sonst leeren Raum umkreisen! Der Atomkern seinerseits ist aus Protonen und Neutronen aufgebaut.

Die Masse der Elementarteilchen

Die drei Bestandteile eines Atoms haben folgende Massen:

	Proton	Neutron	Elektron
Masse (kg)	$1{,}67252 \times 10^{-27}$	$1{,}67482 \times 10^{-27}$	$9{,}10908 \times 10^{-31}$
Atomgewicht µ	1,0072763	1,0086654	$5{,}48597 \times 10^{-4}$

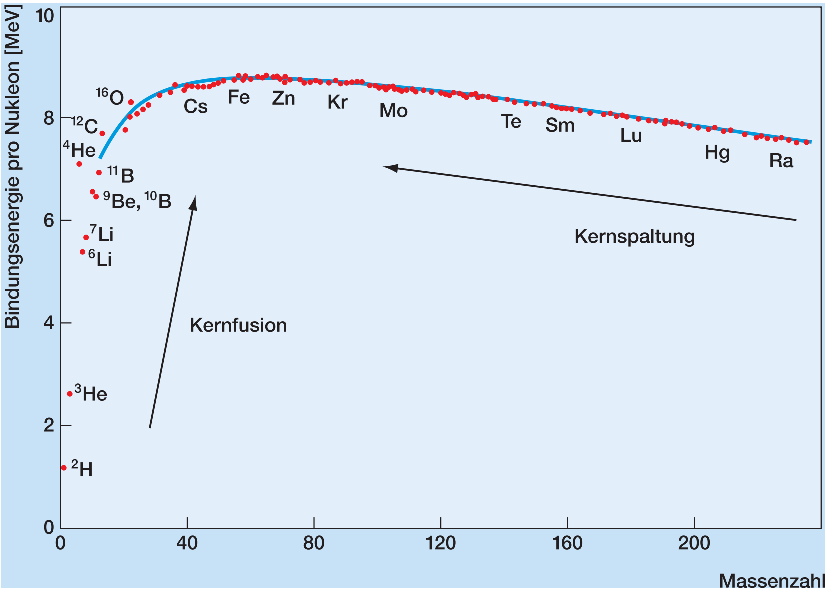
Für den täglichen Gebrauch sind diese Zahlen unhandlich. Man hat daher 1962 den Massenstandard $^1/_{12}$ des gewöhnlichen Kohlenstoffs (genauer des Kohlenstoffisotops ^{12}C einschliesslich seiner Elektronenhülle) mit einem Wert von $1{,}6605402 \times 10^{-27}$ kg eingeführt. Dieser Massenstandard erhält im Deutschen die Bezeichnung µ, im Englischen den Begriff atomic mass unit, kurz amu.

Demnach bestimmen die Protonen und Neutronen im Kern die Masse eines Atoms, wogegen die Masse der Elektronen vernachlässigbar ist.

Der Massendefekt

Wenn wir nun in Gedanken mittels einer Anzahl von Protonen und Neutronen einen Kern konstruieren, so zeigt sich, dass dessen Masse etwas geringer ist als die Summe der Massen seiner Bestandteile. Dieser Verlust, der Massendefekt, entspricht gemäss der Einstein'schen Äquivalenz von Energie und Masse der Bindungsenergie, mit welcher die Nukleonen (d.h. Protonen und Neutronen) im Atomkern gebunden sind. Bei der Zusammenführung eines Protons und eines Neutrons wird diese Bindungsenergie in Form eines γ-Quants frei. Umgekehrt muss die einem γ-Quantum entsprechende Energie zugeführt werden, um die beiden Nukleonen wieder voneinander zu trennen.

Wenn wir nun weitere Elementarteilchen in den Kern einbauen, stellen wir fest, dass die Bindungsenergie pro Elementarteilchen zunimmt bis zum Eisen und anschliessend wieder kleiner wird. Diesen Zusammenhang beschreibt die Massenformel des Deutschen Physikers Carl Friedrich von Weizsäcker (geboren 1912).



Die Figur 4 zeigt den Verlauf der Bindungsenergie pro Nukleon (Protonen und Neutronen) der verschiedenen stabilen Atomkerne in Abhängigkeit von der Massenzahl, d.h. der Anzahl Nukleonen im Kern (Weizsäckersche Masseformel). Die Bindungsenergie pro Nukleon hat ein Maximum beim Eisen, während sie in Richtung zu leichteren und schwereren Elementen abfällt.

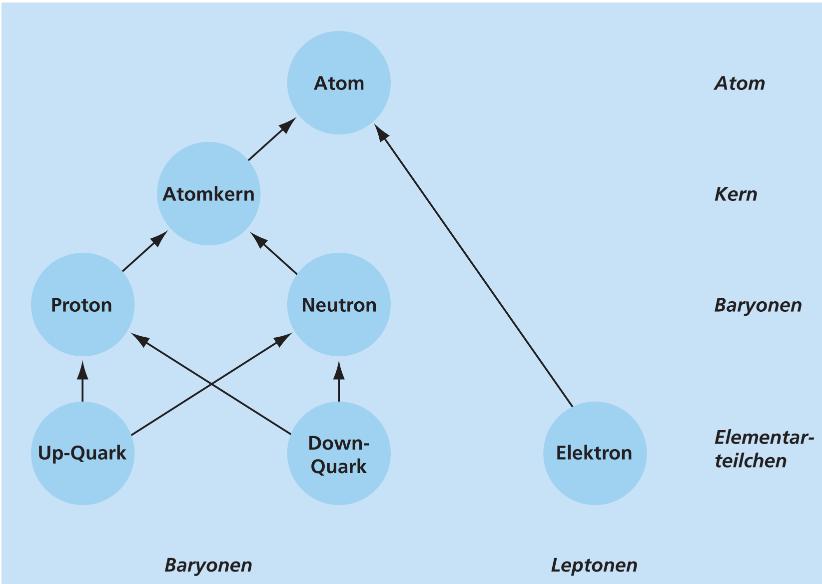
Ein einziges Gramm Eisen enthält die unvorstellbar hohe Bindungsenergie von 846 Giga-Joule, was etwa dem Heizwert von 20 Tonnen Heizöl entspricht. Die Energiegewinnung aus nuklearen Prozessen ist demnach um viele Grössenordnungen ergiebiger als in Verbrennungsprozessen. Aus der Weizsäckerschen Masseformel lässt sich ableiten, dass Energie entweder aus der Fusion leichter Atomkerne zu gewinnen ist oder aber durch Fission, bei welcher schwere Atomkerne in leichtere Elemente gespalten werden. Die in Sternen stattfindenden Kernfusionsprozesse liefern die Energie, um die schwereren Elemente als Eisen zu erzeugen.

Die Ladung der Elementarteilchen

Neben der Masse ist die elektrische Ladung eine zweite wesentliche Eigenschaft der subatomaren Teilchen. Auf jede elektrische Ladung wird in einem elektrischen Feld (das z.B. von einer anderen Ladung herrühren kann) eine Kraft ausgeübt, analog der Kraft auf eine Masse im Gravitationsfeld. Die Einheit der Ladung heisst Coulomb (C), benannt nach dem französischen Physiker Charles Augustin de Coulomb (1736–1806). Die drei Elementarteilchen besitzen die aus der untenstehenden Tabelle ersichtlichen elektrischen Ladungen.

Das Proton ist demnach positiv, das Elektron negativ geladen, und das Neutron ist ungeladen (neutral). Auch hier erweisen sich die sehr kleinen Werte als unpraktisch, zudem sind sie betragsmässig genau gleich. In Analogie zum Massenstandard hat man daher die Elementarladung e = 1.602×10^{-19} C definiert, womit die übersichtliche Situation gemäss der zweiten Zeile der untenstehenden Tabelle entsteht. Die Ursache dieser Symmetrie liegt im Bau der Teilchen begründet, denn auch die Neutronen und Protonen sind nicht elementar, sondern bestehen aus zwei Arten von Quarks, den Up-Quarks und den Down-Quarks, siehe **Figur 5**. Das Proton besteht aus zwei Up- und einem Down-Quark, beim Neutron ist es genau umgekehrt. Die Quarks tragen eine elektrische Ladung, wobei die Ladung des Up-Quarks positiv ist und zwei Drittel einer Elementarladung entspricht, während die Ladung des Down-Quarks einem Drittel der negativen Elementarladung entspricht.

	Proton	***Neutron***	***Elektron***
Elektrische Ladung (C)	$+1{,}602 \times 10^{-19}$	0	$-1{,}602 \times 10^{-19}$
Elektrische Ladung (e)	+ 1	0	– 1



Die Figur 5 zeigt stark vereinfacht den hierarchischen Aufbau der klassischen (baryonischen) Materie.

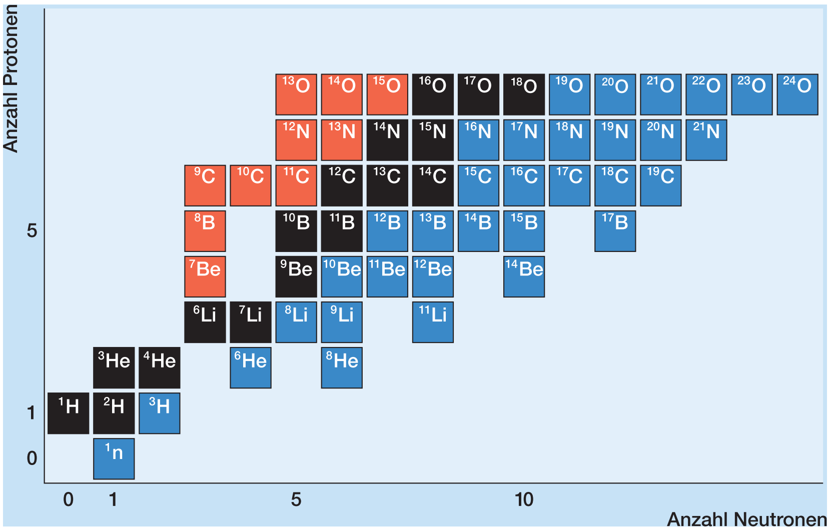
Die Isotope

Das chemische Verhalten eines Atoms wird durch die Zahl der Elektronen bestimmt. Beim neutralen Atom entspricht sie auch gerade der Zahl der Protonen im Atomkern. Über die Zahl der Neutronen im Kern ist damit aber noch nichts ausgesagt. Im einfachsten Fall enthält ein Atomkern gar kein Neutron, sondern nur ein einziges Proton, wie dies beim häufigsten Atom, dem Wasserstoff zutrifft. Es gibt aber auch Wasserstoffkerne, die ein Neutron enthalten. Dieser Wasserstoff wird als Deuterium bezeichnet, eines der beiden stabilen Isotope des Wasserstoffs. Das Tritium, ein zweites Isotop des Wasserstoffs, besitzt zwei Neutronen im Kern. Alle diese Atome sind chemisch nicht unterscheidbar, da sie sich in Verbindungen genau gleich verhalten. Physikalisch sind sie aber mit Hilfe von Massenspektrometern zu unterscheiden auf Grund ihrer unterschiedlichen Massen. Der Begriff Isotop stammt aus dem Griechischen und bezeichnet den «gleichen Ort» im Periodensystem: auf Grund ihrer gleichen Protonenzahl sind alle genannten Isotope des Wasserstoffs an der gleichen Stelle im Periodensystem zu finden.

Um ein Isotop zu beschreiben, verwendet man die Abkürzung: ^{A}X, wobei A Gesamtzahl von Protonen und Neutronen und X das chemische Zeichen des fraglichen Elements ist. Das Reinelement Wasserstoff hat daher das Zeichen ^{1}H, da es nur ein Nukleon (Proton) im Kern besitzt. Dagegen kennzeichnet ^{2}H das Deuterium, den «schweren Wasserstoff» mit je einem Proton und einem Neutron im Kern und ^{3}H das (instabile) Tritium mit einem Proton und zwei Neutronen. Da es von jedem der 112 Elemente zahlreiche Isotope gibt, existieren insgesamt über 1000 Nuklide. Davon sind 287 stabil, alle anderen zerfallen spontan, d. h. sie sind radioaktiv. Man nennt sie deshalb Radionuklide. Die Zeitperiode, innerhalb derer die Hälfte der vorhandenen Radionuklide eines Isotops zerfallen, wird Halbwertszeit genannt. Die Halbwertszeiten können zwischen Bruchteilen von Sekunden bis Milliarden von Jahren liegen. Die radioaktiven Isotope spielen daher zur Altersbestimmung in den verschiedensten Zweigen der Naturwissenschaften eine bedeutende Rolle, siehe zum Beispiel Spatium Nr. 8: Sun and Climate.

Die Nuklidkarte

Währenddem das Periodensystem die Atome nach ihrer Protonenzahl ordnet und damit alle Isotope eines Elements zusammenfasst, stellt die Nuklidkarte eine feinere Gliederung der einzelnen Elemente dar, wo auch die Isotope einzeln aufgeführt sind, siehe **Figuren 6 bis 8.**



Die Figur 6 zeigt den linken unteren Ausschnitt der Nuklidkarte. Jeder Zeile nach oben entspricht ein zusätzliches Proton im Kern. Jeder Spalte nach rechts entspricht ein zusätzliches Neutron im Kern. Die Nuklidkarte beginnt unten links mit dem Neutron. Auf der Zeile P=1 folgt in der Kolonne N=0 das Wasserstoffatom. Rechts davon finden sich seine Isotope Deuterium und Tritium. Die ersten beiden Isotope Wasserstoff und Deuterium sind stabil: ihre Felder sind schwarz markiert. Tritium hingegen ist instabil, es ist daher wie alle anderen instabilen Elemente farbig markiert, wobei die Farbe Hinweis auf den Zerfallsprozess (siehe weiter unten) gibt. Auf der nächsten Zeile finden sich alle Isotope mit zwei Protonen im Kern, also die Isotope des Heliums.

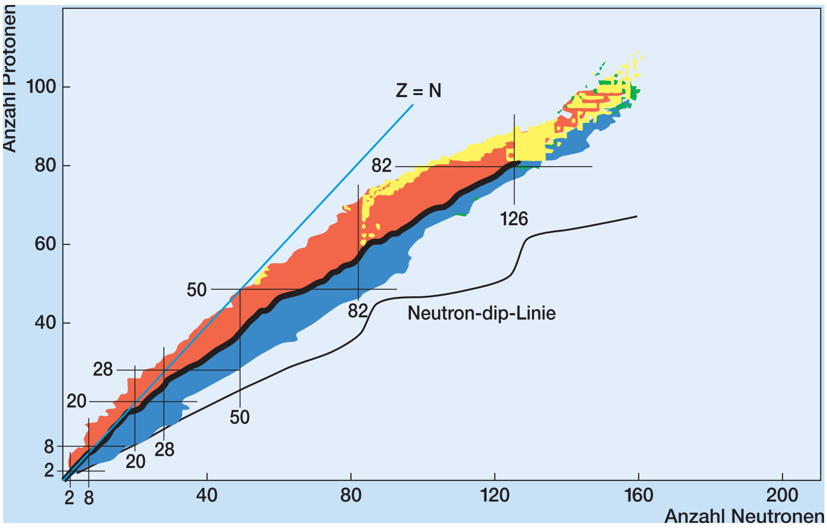
Zerfallsprozesse

Die instabilen Isotope zerfallen je nach ihrer Lage auf der Nuklidkarte mit verschiedenen Zerfallsprozessen. Beim β^--Zerfall verlässt ein schnelles Elektron den Kern. Dieses Elektron war vorher nicht im Kern vorhanden, vielmehr wird es beim Zerfall erzeugt, indem sich ein Neutron des Kerns in ein Proton umwandelt. Zusammen mit dem Elektron verlässt zudem ein Anti-Neutrino den Kern. Beim β^--Zerfall entsteht daher ein neues Element, das sich in der Nuklidkarte eine Zeile höher (d.h. ein Proton mehr) und eine Spalte weiter links (d.h. ein Neutron weniger) befindet. Nach diesem Prozess zerfallen alle in den Nuklidkarten blau markierten Isotope.

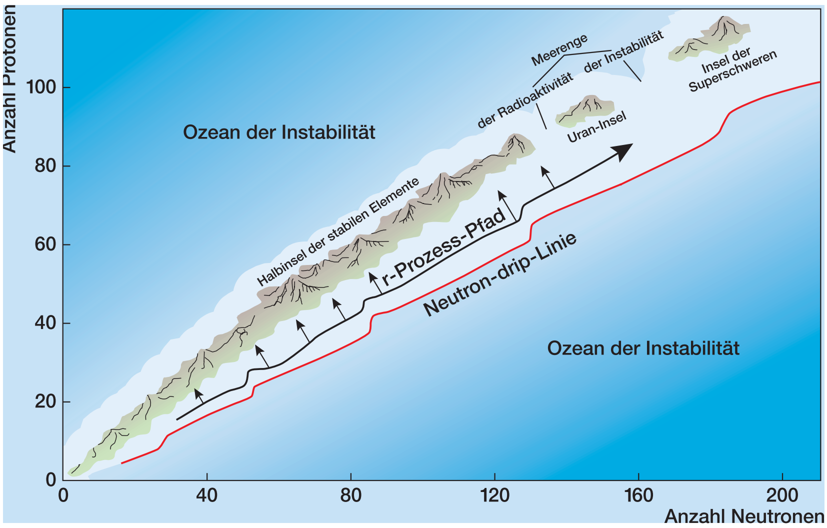
Beim β^+-Zerfall dagegen zerfällt im Kern ein Proton in ein Neutron. Dabei wird ein Positron, das Antiteilchen des Elektrons, und ein Neutrino erzeugt. Das entstandene Positron hat die gleichen Eigenschaften wie ein Elektron der β^--Strahlung, aber eine positive Ladung. Daher wird dieser Prozess als β^+-Zerfall genannt. Es entsteht ein neues Element, das sich in der Nuklidkarte eine Zeile tiefer und eine Spalte weiter rechts befindet. Nach diesem Zerfallsprozess zerfallen alle in den Nuklidkarten rot markierten Isotope.

Ein weiterer Zerfallsprozess ist der α-Zerfall, wobei ein Heliumkern bestehend aus je zwei Neutronen und Protonen den Kern verlässt. Dabei verliert der Atomkern vier Einheiten Masse und zwei Einheiten Ladung. Beim α-Zerfall entsteht ein neues Element, das sich in der Nuklidkarte zwei Zeilen tiefer und zwei Spalten weiter links befindet. Nach diesem Zerfallsprozess zerfallen alle schweren in der Nuklidkarte gelb markierten Isotope, siehe **Figur 7.**

Damit sind nun die Grundlagen für das Verständnis der Nukleosynthese geschaffen, der Lehre von der Entstehung der Elemente.

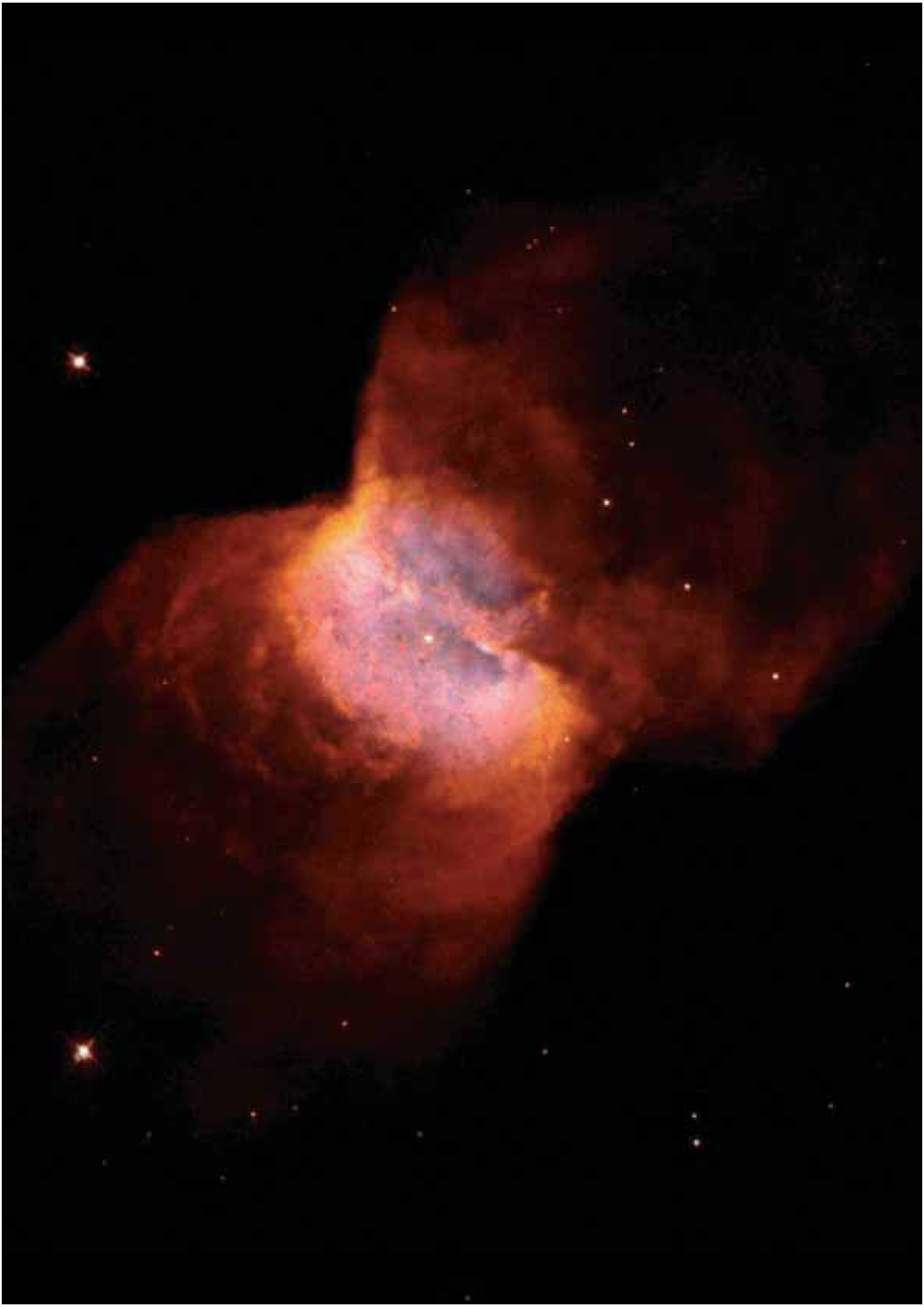


Die Figur 7 gibt einen Überblick über die gesamte Nuklidkarte. Deutlich ist die Abflachung der Orte der stabilen (schwarzen) Elemente zu erkennen: Bei niedrigen Protonenzahlen liegen sie dort, wo sich die Anzahl von Protonen und Neutronen etwa die Waage halten (Z = N). Dagegen sind bei Elementen mit grossen Protonenzahlen wesentlich mehr Neutronen erforderlich, um stabile Kerne zu bilden. Die Isotope der blauen Felder zerfallen mit dem β^--Zerfall, die roten mit dem β^+-Zerfall, wogegen die gelb markierten Elemente dem α-Zerfall unterliegen.



Die Figur 8 zeigt eine etwas anschaulichere Deutung der Nuklidkarte. In der Halbinsel der stabilen Elemente sind die Bindungsenergien der einzelnen Elemente gemäss der Weizsäckerschen Masseformel nach oben (als Höhe der Gebirge) aufgetragen. Die höchste Erhebung befindet sich demnach beim Eisen mit seiner grössten Bindungsenergie pro Nukleon. Die als Neutron-drip bezeichnete Linie stellt das Ufer zum Ozean der Instabilität dar. In diesem Ozean sind alle nicht existierenden Isotope versammelt, deren Bindungsenergie negativ ist und deren Erhebung daher unter dem Meeresspiegel verschwindet. Das Ufer ist definiert durch die Kombinationen der Protonen- und Neutronenzahlen, welche für die Bindungsenergie in der Masseformel den Wert Null ergeben (Meereshöhe). Am oberen Ende der Halbinsel folgt die Meerenge der Radioaktivität, ein Bereich, wo es überhaupt keine stabilen Elemente gibt. Im Bereich noch grösserer Protonenzahlen liegt die Uran-Insel. Dann folgt die Meerenge der Instabilität und die hypothetische Insel der superschweren Elemente, deren Existenz bisher aber noch nicht nachgewiesen werden konnte. (Auf den r-Prozess-Pfad wird später noch eingegangen).

Die Figur 9 auf der nächsten Seite zeigt den Nebel NGC 2346. Er besteht aus Gas und Staub und ist wesentlich grösser als unser Sonnensystem. Im Zentrum befinden sich zwei Sterne, die sich innerhalb von nur 16 Tagen umkreisen. Der grössere Stern ist ein Roter Riese, in welchem je zwei ^{4}He-Kerne zum instabilen ^{8}Be verschmelzen, das trotz seiner kurzen Halbwertszeit von 10^{-16} Sekunden sich gelegentlich mit einem weiteren ^{4}He Kern zu ^{12}C verbindet. Unsere Sonne wird in einigen Milliarden Jahren ebenfalls zu einem roten Riesen werden und dabei die inneren Planeten, einschliesslich der Erde, in sich aufnehmen. Quelle: Massimo Stiavelli (STScI), Inge Heyer (STScI) et al., and the Hubble Heritage Team (AURA/ STScI/ NASA).



Die Entstehung der Elemente

Der Urknall *

Zu Beginn des 20. Jahrhunderts stellte der amerikanische Astronom Edwin Hubble (1889–1953) fest, dass sich die fernen Nebel von uns wegbewegen und zwar umso schneller je weiter sie bereits von uns weg sind. Die Grundlage für die spätere Urknalltheorie war gefunden, denn wenn sich diese Galaxien jetzt von uns entfernen, müssen sie zu einem früheren Zeitpunkt nahe beisammen gewesen sein. Aus dem Verhältnis ihrer Geschwindigkeit und ihrer Entfernung lässt sich der Zeitpunkt des Anfangs des Universums ermitteln. Die heute gültige Schätzung für sein Alter ist 14±2 Milliarden Jahre.

Die Hypothese von Edwin Hubble widersprach zu ihrer Zeit allen gängigen Vorstellungen von einem stabilen Universum. Selbst Albert Einstein (1879–1955) sah sich gezwungen, in seinem kosmologischen Modell eine Korrekturgrösse einzuführen, obwohl es richtigerweise die Möglichkeit der Expansion voraussagte. Erst nachdem er von den Erkenntnissen von Edwin Hubble gehört hatte, wurde ihm dieser Fehler bewusst und er bezeichnete ihn als die grösste Eselei seines Lebens.

Die Theorie des Urknalls postuliert einen Anfang von Raum und Zeit und ein Universum von zunächst unvorstellbar kleinem Volumen und unvorstellbar hoher Temperatur und Dichte, wo die uns bekannten Naturgesetze versagen. Doch schon Bruchteile einer Sekunde später, bei der Zeit von 10^{-42} Sekunden nach dem Urknall, können diese Gesetze das Verhalten des Universums beschreiben. Die nun folgende Geschichte des Universums ist eine Kette a priori völlig unwahrscheinlicher Ereignisse, von denen jedes einzelne in der Lage gewesen wäre, die weitere Entwicklung zu einem vorzeitigen Ende zu bringen. Da wir nun aber einmal da sind, erscheinen a posteriori diese Ereignisse nicht mehr als Zufälle, sondern als Notwendigkeiten, wie das Anthropische Prinzip feststellt.

Der erste Schritt: H ➔ He

Bereits 1 Millisekunde nach dem Big Bang wäre es zwar für die Existenz von stabilen Heliumkernen schon kühl genug gewesen, weil diese sehr stark gebunden sind. Die Dichte der Materie war aber schon so gering, dass die direkte (Vierkörper-)Reaktion zweier Protonen und zweier Neutronen zu ^{4}He nicht mehr möglich war. Stattdessen musste das Helium in Ketten von Zweikörper-Reaktionen gebildet werden. Der erste Schritt dazu ist die Verbindung eines Protons und eines Neutrons zu Deuterium, was erst bei Temperaturen unter 30 Milliarden K möglich ist, weil es nur schwach gebunden ist **(siehe Figur 4)**. Das Deuterium seinerseits wird anschliessend sehr schnell zu Helium-4 weiter verbrannt. Weil dieser Kern sehr gut gebunden ist, enden dort die meisten Reaktionen, es entsteht also hauptsächlich ^{4}He und die überzähligen Protonen verbleiben als Wasserstoffkerne. Kurz darauf kommt aber der Fusionsprozess zum Stillstand. Eine Reaktion von ^{4}He mit einem Nukleon würde zu einem Kern mit Massenzahl 5 führen, eine Reaktion von zwei ^{4}He-Kernen zu einem solchen mit Massezahl 8. Es gibt jedoch keine genügend stabilen Kerne mit diesen Massenzahlen **(siehe Figur 11)**. Zu jenem Zeitpunkt bestand das Universum zu drei Vierteln aus Wasserstoffkernen, zu einem Viertel aus Helium-4-Kernen, und Spuren von Deuterium, Helium-3 und Lithium-7.

Das dunkle Zeitalter

Etwa 350 000 Jahre nach dem Urknall betrug die Temperatur im

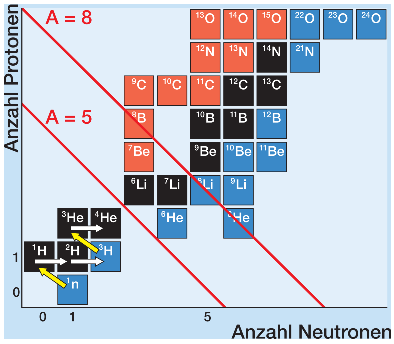
Die Figur 10 auf der vorangehenden Seite zeigt den Katzenaugen-Nebel NGC 6543. Es ist das bisher letzte mit dem Hubble Space Telescope (HST) aufgenommene Bild dieses Nebels, der seit 1994 immer wieder in seiner Entwicklung beobachtet werden konnte. Dabei handelt es sich um einen der komplexesten, je untersuchten Nebel: er enthält derzeit mindestens elf konzentrische Strukturen, die sich wie Blasen in den Raum hinaus entwickeln. Nach heutiger Interpretation folgen sich diese Blasen in Abständen von 1500 Jahren. Jede dieser Schalen enthalten Staub, dessen Masse etwa derjenigen aller Planeten des Sonnensystems entspricht. Quelle: R.Corradi (Isaac Newton Group of Telescopes, Spanien) and Z.Tsvetanov (NASA).

*) Siehe auch Spatium Nr.1 von J.Geiss, Nr.3 von G.A.Tammann, und Nr.7 von K.Pretzl

Universum noch 10 000 K. Die Atomkerne waren nun in der Lage, Elektronen zu binden und somit elektrisch neutrale Atome zu bilden: das Universum wurde für elektromagnetische Strahlung transparent. Durch die Expansion kühlte es sich immer mehr ab und wurde damit auch immer dunkler, das dunkle Zeitalter begann. Der abgekühlte Überrest dieser 10 000 K heissen Strahlung ist heute noch als kosmische Hintergrundstrahlung jenseits der fernsten Galaxien erkennbar.

Die ersten Sterne

Ungefähr 700 Millionen Jahre nach dem Urknall war die Finsternis zu Ende. Winzige Fluktuationen der Hintergrundstrahlung erzeugten kleine lokale Dichteunterschiede in der Verteilung der Materie. Diese führten zu lokal erhöhten Gravitationsfeldern, die ihrerseits die Dichteunterschiede noch weiter verstärkten. In den Zentren erhöhter Dichte sammelte sich immer mehr Materie aus der Umgebung an und im Verlauf von Millionen Jahren hatten sich örtlich Wolken der hundertfachen Masse der heutigen Sonne zusammengeballt. Die Kontraktion der Wolke führte zu hohem Druck und hoher Temperatur in ihrem Innern, wodurch schliesslich erstmals seit dem Big Bang wieder Kernfusionsprozesse in Gang kommen konnten und dadurch Energie freisetzten: Die ersten leuchtenden Sterne waren entstanden. Wie im Urknall verbrannte zuerst Wasserstoff zu Helium, aber im Unterschied zu damals ist in Sternen die Dichte



Die Figur 11 zeigt die möglichen Fusionsreaktionen während der ersten Sekunden des Universums. Zu Beginn gibt es nur einzelne Protonen und Neutronen zu etwa gleichen Teilen. Mit der Zeit beginnen die Protonen zu überwiegen, denn ein Neutron kann zu einem Proton zerfallen und somit einen Wasserstoffkern bilden. Aus dem Wasserstoffkern ^{1}H wird durch Einlagerung eines Neutrons das Deuterium ^{2}H. Ein weiteres Neutron führt dann zum Tritium ^{3}H, das sich im β^--Zerfall zum ^{3}He wandelt. Nach Einlagerung eines Neutrons entsteht ^{4}He. Hier endet die Nukleosynthese vorläufig, da infolge der Masselücke bei A = 5 die Einlagerung weiterer Neutronen nicht zu stabilen Elementen führen kann.

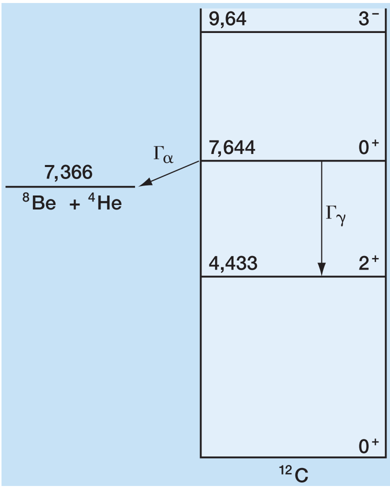
viel höher, so dass mit weiteren Reaktionen die Massenlücken in der Nuklidkarte übersprungen und schwerere Elemente gebildet werden konnten.

Der zweite Schritt: He ➔ C

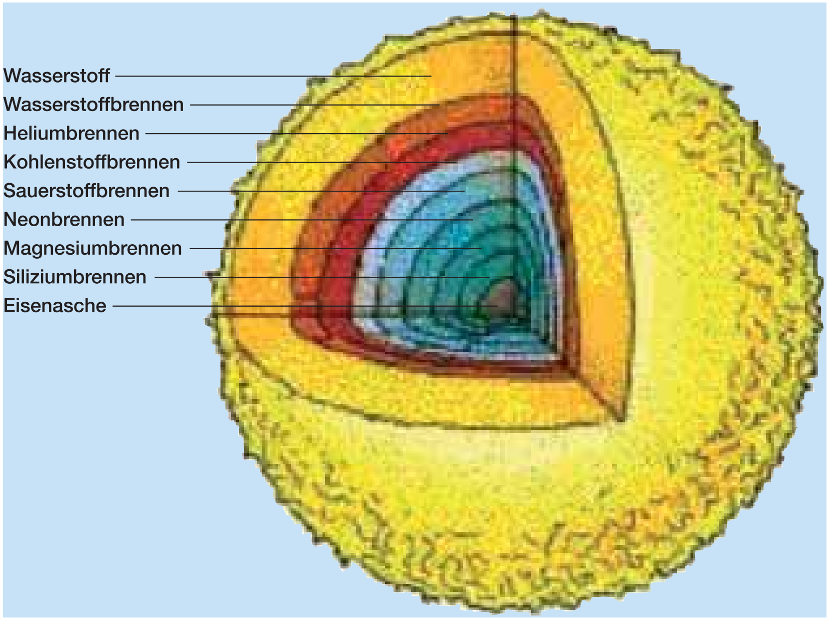
Diese ersten Sterne erzeugten zunächst durch Wasserstoffbrennen Helium, das bereits vom Urknall her im Universum vorhanden war, sie trugen also noch nichts Neues bei. Die grössten unter ihnen hatten nach etwa 50 Millionen Jahren im Innern ihren Vorrat an Wasserstoff verbrannt und wurden zu Roten Riesen, Sterne gewaltiger Größe, aber geringer Dichte. In ihrem Zentrum sind die Bedingungen geeignet, dass zwei ^{4}He-Kerne zunächst zum instabilen ^{8}Be verschmelzen und dieses sehr kurzlebige Isotop (Halbwertszeit von 10^{-16} Sekunden) durch Einbau eines dritten ^{4}He-Kerns den stabilen Kohlenstoff ^{12}C bilden kann. Dafür ist aber eine wichtige Bedingung erforderlich: Die Resonanz des ^{12}C-Kerns mit dem Paar ^{4}He und ^{8}Be.

Die Resonanz von ^{12}C

Fred Hoyle (1915–2001) postulierte auf Grund des anthropischen Prinzips den angeregten Zustand von ^{12}C lange bevor er experimentell nachgewiesen werden konnte. In einem angeregten Zustand bewegen sich die Nukleonen im Kern mit einer kinetischen Energie, die nur wenige diskrete Werte über dem Grundzustand annehmen kann. Das Paar ^{8}Be und ^{4}He besitzt nun zufälligerweise eine gemeinsame Energie, die ziemlich genau einem der angeregten Zustände des ^{12}C-Kerns entspricht **(siehe Figur 12)**. Mit einer Wahrscheinlichkeit von etwa 1:1000 gelingt es diesem Paar aufgrund seiner Resonanz mit dem ^{12}C-Kern, ein stabiles Kohlenstoffatom ^{12}C zu bilden. Bei den restlichen zerfällt das ^{8}Be wieder in Heliumkerne. Da die weitere Kernsynthese, von ^{12}C + ^{4}He zu ^{16}O und weiter, nicht resonant verläuft, bleibt genügend Kohlenstoff übrig. Im anderen Falle wäre der für den Aufbau des Lebens so wichtige Kohlenstoff im Universum praktisch nicht vorhanden…



Figur 12 zeigt die Bindungsenergie des Paares ^{8}Be und ^{4}He (links) und die Energiezustände des Kohlenstoffkerns ^{12}C (rechts). Die Bindungsenergie von ^{8}Be und ^{4}He zusammen mit ihrer kinetischen Energie kann den angeregten Zustand von ^{12}C auf dem Niveau 7.644 MeV erreichen, sodass die resonante Reaktion von ^{8}Be und ^{4}He zu ^{12}C möglich wird.



Die Figur 13 zeigt einen schematischen Schnitt durch einen aktiven Riesenstern, mit den verschiedenen Zonen, wo die einzelnen Fusionsprozesse stattfinden. Der Abfall, aus dem keine weitere Energie bezogen werden kann, ist die Eisenasche, die sich auf Grund ihrer hohen Dichte im Zentrum des Sterns anhäuft.

Der dritte Schritt: C ➔ Fe

Viele der ersten Sterne waren wesentlich massiver als die heutige Sonne. In ihrem Innern geht während des Stadiums als Roter Riese der Heliumbrennstoff langsam zur Neige. Der Stern kontrahiert abermals, bis die Temperatur im Zentrum soweit angestiegen ist, dass die nächste Brennstufe zündet, das Kohlenstoffbrennen. (Weniger massive Sterne wie unsere Sonne enden dagegen nach dem Heliumbrennen als Weisse Zwerge, weil die nötige Temperatur für das Kohlenstoffbrennen nicht erreicht werden kann.) So zünden im Zentrum in immer rascherer Folge weitere Brennstufen, während die anderen in Schalen weitergehen – der Stern erhält eine Zwiebelschalenstruktur **(siehe Figur 13)**. Auf diese Weise werden alle weiteren Elemente bis hin zum Eisen erzeugt. Wenn die Kette dieser Fusionsprozesse beim Eisen angelangt ist, gibt es keine Reaktionen mehr, aus denen sich Energie gewinnen lässt, um den Stern zu stabilisieren. Er kollabiert schliesslich infolge seiner eigenen Gravitation und endet als Supernova vom Typ II. Er enthält bereits alle Elemente von Wasserstoff bis Eisen, eine beachtliche Vielfalt von

Beim Kohlenstoffbrennen verschmelzen zwei Kohlenstoffkerne zu einem angeregten Magnesiumkern, aus dem verschiedene Tochterprodukte entstehen können

$^{12}C + ^{12}C \rightarrow ^{24}Mg^{*}$ $\quad$ ^{23}Mg, ^{20}Ne, ^{23}Na

Beim Neonbrennen wird ein Neonkern durch ein energiereiches Photon in einen Sauerstoff- und einen Heliumkern gespalten. Letzterer kann mit einem zweiten Neonkern zu Magnesium reagieren, also netto

$^{20}Ne + ^{20}Ne \rightarrow ^{16}O + ^{24}Mg$

Beim Sauerstoffbrennen verschmelzen zwei Sauerstoffkerne zu einem angeregten Schwefelkern, aus dem wiederum verschiedene Tochterprodukte entstehen können

$^{16}O + ^{16}O \rightarrow ^{32}S^{*}$ $\quad$ ^{31}S, ^{31}P, ^{30}P, ^{28}Si

Schliesslich setzt, ausgehend von der Gruppe um ^{28}Si, ein Strom von Gleichgewichtsreaktionen ein bis hin zur Gruppe um Fe.

26 verschiedenen Atomen, aber alle schwereren Elemente als Eisen fehlen noch.

Neutronen

Aus der Weizsäckerschen Formel geht hervor, dass die Erzeugung von Elementen, die schwerer sind als Eisen, bedeutende Energiemengen erfordert. Dazu stehen verschiedene kosmische Prozesse zur Verfügung, wobei aber in allen Fällen die Neutronen die Hauptrolle spielen.

Der vierte Schritt: Fe ➔ U

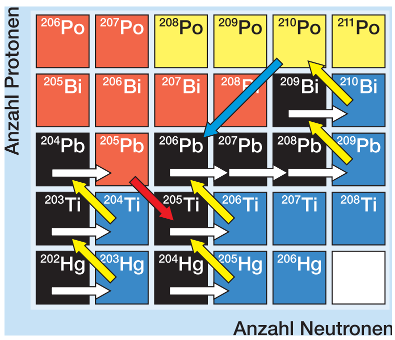
Supernovae sind nicht nur das Ende einer Kette von Prozessen, die zum Eisen führen, sondern auch der Anfang weiterer Reaktionen, welche die Synthese von Atomkernen höherer Ordnungszahlen ermöglichen. In ihren abgestossenen Hüllen entstehen viele Neutronen im explosiven Sauerstoff- und Siliziumbrennen. Da Neutronen keine elektrische Ladung besitzen, können sie ungehindert in die vorhandenen Atomkerne eindringen (wobei diese ein Gammaquantum abgeben). Dabei erhöhen sich die Massezahl und die Neutronenzahl je um den Wert 1. Wenn dabei ein instabiler Kern entsteht, zerfällt dieser, indem ein Neutron durch den β^--Zerfall in ein Proton umgewandelt und somit ein neues Element gebildet wird. Freie Neutronen kommen nicht nur in Supernovae vor, sondern auch in Roten Riesen während des Heliumbrennens, wo sie in Nebenreaktionen entstehen. Die schweren Elemente können somit auch dort durch Neutronenaddition gebildet werden, sobald diese etwas Eisen enthalten. Entsprechend der unterschiedlichen Zeitskalen werden zwei Neutroneneinfangprozesse unterschieden:

- Der s-Prozess (slow) addiert einzelne Neutronen in Zeiträumen von einigen 1000 Jahren.

- Der r-Prozess (rapid) addiert eine grosse Zahl von Neutronen innerhalb Bruchteilen von Sekunden.

Der s-Prozess

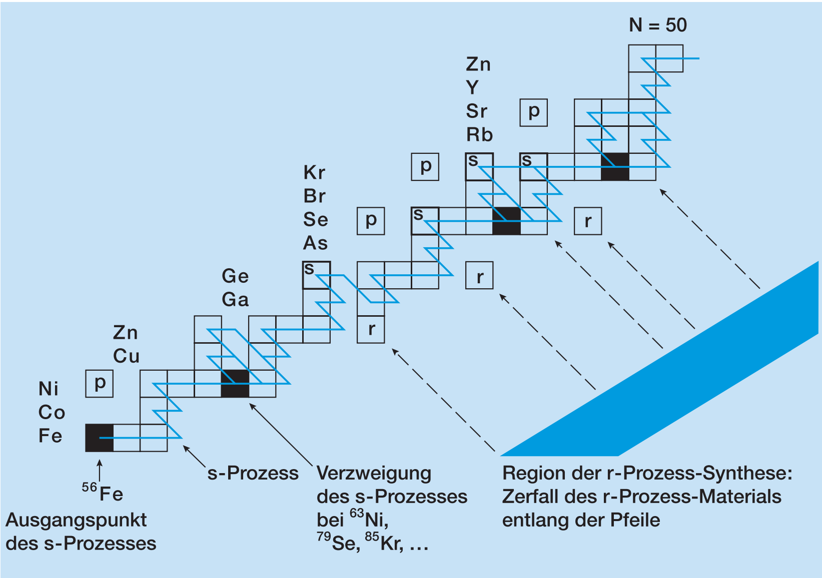
Beim s-Prozess werden über grosse Zeiträume Neutronen im Kern aufgenommen, so langsam, dass ein entstehender instabiler Kern zerfällt, bevor ein weiteres Neutron addiert wird. Der Prozess beginnt bei der Eisengruppe: Durch Einbau eines Neutrons z.B. in ^{56}Fe entsteht zunächst das Isotop ^{57}Fe, das nächste Neutron führt auf ^{58}Fe und das drittte zum instabilen Isotop ^{59}Fe. Dieses zerfällt innert 44 Tagen zum stabilen Kobalt (^{59}Co). Von dort führt der Prozess immer weiter entlang der stabilen Elemente in Richtung der schweren Atome. Allerdings können auf diese Weise nicht alle schweren Elemente gebildet werden; vielmehr endet der Prozess beim letzten wirklich stabilen Isotop auf der Nuklidkarte, dem Wismuth ^{209}Bi, **(Figur 14)**. Der s-Prozess konnte erstmals 1952 nachgewiesen werden: das in Roten Riesen beobachtete radioaktive Technetium kann aufgrund seiner Halbwertszeit von wenigen Millionen Jahren erst kurz zuvor in diesem s-Prozess entstanden sein.



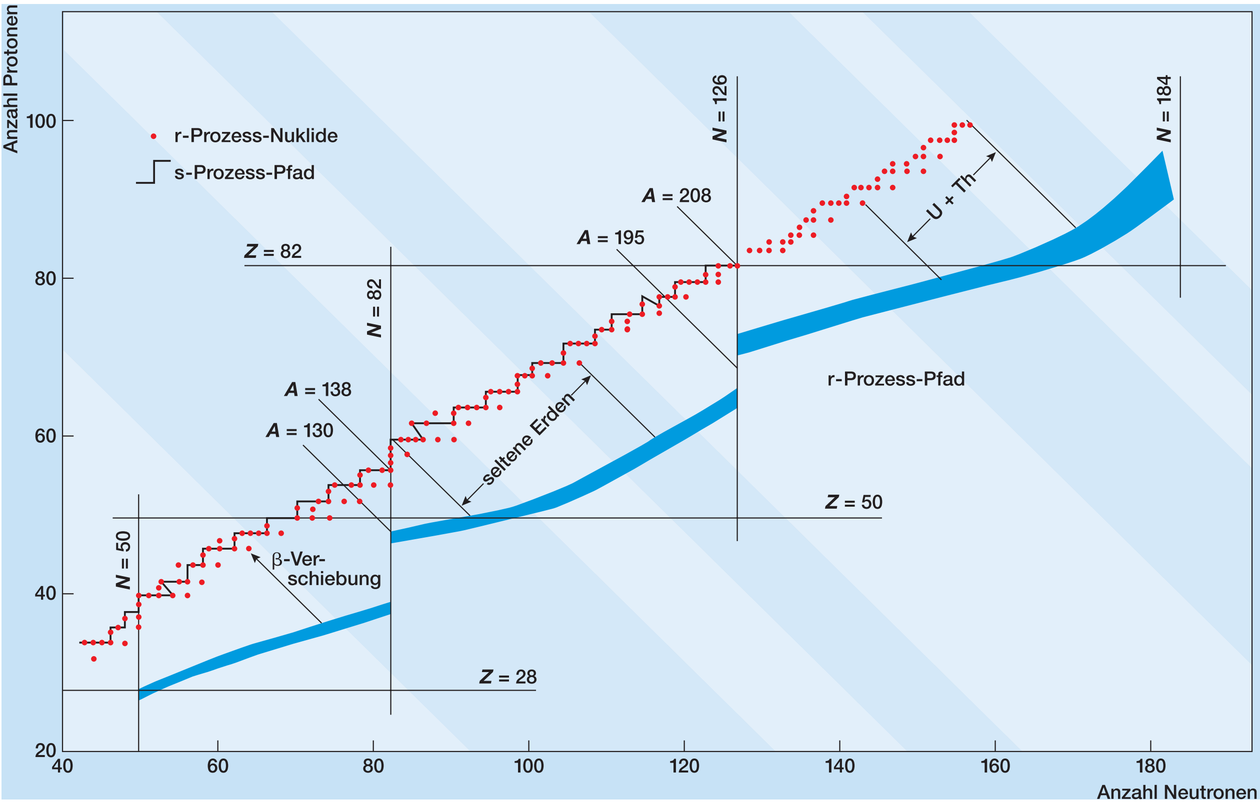
Die Figur 14 zeigt die letzten Schritte des s-Prozesses bei ^{209}Bi. Nach Integration eines weiteren Neutrons im Kern des ^{209}Bi wird dieses zum instabilen ^{210}Bi, welches mit β^- zu ^{210}Po zerfällt. Dieses ist ebenfalls instabil und zerfällt unter Abgabe zweier Neutronen und zweier Protonen (α-Zerfall) zu Blei (^{206}Pb), das nach Aufnahme weiterer Neutronen und β^--Zerfall wieder zu ^{209}Bi wird. Dieser geschlossene Kreislauf beendet den s-Prozess in Richtung der schwereren Elemente.

Der r-Prozess

Für die Erklärung der Existenz der schwereren Elemente als ^{209}Bi, wie Uran und Thorium, ist eine weitere Reaktion erforderlich, denn der s-Prozess ist viel zu langsam, um die Lücke zwischen Wismuth und Thorium zu überbrücken. Als r-Prozess bezeichnet man die sehr rasche Integration von Neutronen im Kern, schneller als die dabei entstehenden instabilen Isotope zerfallen können, siehe **Figur 16**. Erst der r-Prozess ermöglicht die Überquerung der Meerenge der Radioaktivität in **Figur 8**. Er setzt hohe Neutronendichten und hohe Temperaturen voraus, damit in Sekundenbruchteilen sehr viele



Die Figur 15 vergleicht den r- und den s-Prozess. Der s-Prozess verläuft entlang der Reihe der stabilen Elemente in Richtung schwererer Elemente. Wenn ein instabiles Isotop entsteht, zerfällt es im β^--Zerfall zu einem stabilen Isotop, das seinerseits wieder Ausgangspunkt für einen weiteren Schritt auf dem s-Prozessweg ist. Der s-Prozess besitzt Gabelungen zum Beispiel bei ^{63}Ni, ^{79}Se, ^{85}Kr, weil dort die Zerfallszeit etwa gleich ist wie die Zeit bis zur Addition des nächsten Neutrons. Manche stabile Isotope werden vom s-Prozessweg jedoch nicht erreicht (markiert mit r) und können nur durch den r-Prozess erzeugt werden. Der r-Prozesspfad führt entlang den radioaktiven Isotopen im blauen Bereich, aus denen dann durch β^--Zerfall die stabilen, neutronenreichen Isotope entstehen.



Die Figur 16 zeigt schematisch den r-Prozesspfad in der Nuklidkarte. Dieser liegt rechts des s-Prozesspfades bei den neutronenreichen, instabilen Isotopen der schweren Elemente im Bereich der blauen Felder. Von dort gelangen sie durch β^--Zerfall zu den stabilen, neutronenreichen Isotopen.

Figur 17 Vor etwa 8000 Jahren explodierte in unserer Milchstrasse der Stern IC 443. In seinen Hüllen herrschten in der Folge von explosivem Sauerstoff- und Siliziumbrennen hohe Neutronendichten, woraus im r-Prozess schwere Elemente entstanden, die nun in den Weltraum verteilt werden. (Quelle: Jean-Charles Cuillandre [CFHT], Hawaiian Starlight, CFHT)

Neutronenanlagerungen stattfinden können. Diese Bedingungen sind in abgestossenen Hüllen von Supernovae gegeben, wo die Neutronen von explosivem Sauerstoff- und Siliziumbrennen stammen. Nach Abschluss der intensiven Neutronenbestrahlung zerfallen die entstandenen, sehr neutronenreichen Kerne sukzessive durch β^--Zerfall, bis stabile Kerne erreicht werden. Auf diese Weise entstehen die stabilen, neutronenreichen Isotope von Eisen bis Blei rechts des s-Prozessweges in der **Figur 15** und eben die langlebigen (fast stabilen) Isotope von Uran und Thorium.

Der p-Prozess

Nebst den beiden genannten Prozessen, aber wesentlich seltener, kommt der p-Prozess vor, wobei p für Proton steht. Bei extrem hohen Temperaturen (über einer Milliarde K) können in Supernovae Protonen mit Atomkernen reagieren und somit protonenreiche Isotope der schweren Elemente mit Massenzahlen über 56 (Eisen) erzeugen, siehe die mit p bezeichneten Isotope in der **Figur 15.** Die bereits hohe Anzahl von Protonen dieser Kerne bewirkt jedoch eine starke elektrische Abstossung eines zusätzlichen Protons. Deshalb müssen diese sehr hohe Geschwindigkeiten (hohe Temperaturen) besitzen, um in den Kern eindringen zu können. Die beobachteten geringen Häufigkeiten solcher protonenreichen Isotope bestätigen die Vermutung, dass der p-Prozess sehr selten auftritt.

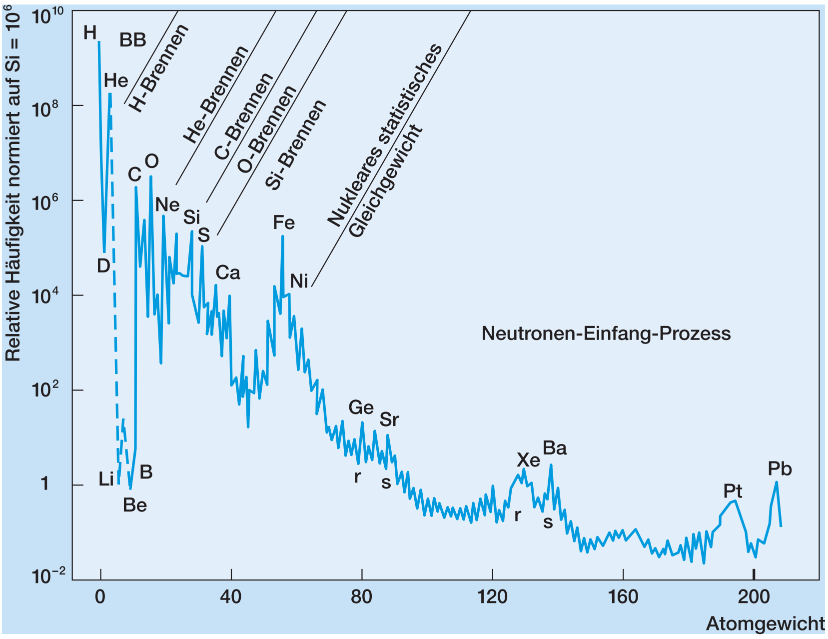
Die Häufigkeiten im Sonnensystem

Die Häufigkeit der Elemente im Sonnensystem lässt sich unter anderem durch die Analyse des Sonnenwindes ermitteln. Dieser besteht hauptsächlich aus Protonen und Elektronen, aber auch Kerne von Helium und schwererer Elemente kommen vor. Er widerspiegelt in erster Näherung und mit wenigen Ausnahmen die Zusammensetzung der Sonne, da diese rund 99,9% der Masse im Sonnensystem besitzt. Die heute feststellbaren Häufigkeiten sind ein angenähertes Abbild des damaligen Zustandes im lokalen interstellaren Raum, wo die Sonne vor 4,6 Milliarden Jahren entstand. Vor ihr waren schon Generationen von Sternen zu Supernovae geworden und hatten die Kette der Nukleosynthese durchgearbeitet. Von diesen früheren Sternen sind die schwereren Elemente als Eisen die wichtigsten Zeugen. Ihre relative Seltenheit zeigt jedoch, dass der Prozess der Nukleosynthese oder die galaktische chemische Evolution erst am Anfang ist: es stehen noch viele leichte Elemente zur Verfügung, welche die Energie liefern, um in den kommenden Milliarden von Jahren weitere schwere Elemente zu bilden.

In der untenstehenden Tabelle sind die 25 häufigsten Isotope aufgelistet, aus denen unser Sonnensystem und damit wir selber bestehen, sowie die Prozesse, die zu deren Bildung

Rang	A	Element	Anteil	Ursprung								
				Big Bang	H-Brennen	He-Brennen	C-Brennen	O-Brennen	Si-Brennen	Ne-Brennen	s-Prozess	NSE
1	1	H	7×10^5	■								
2	4	He	3×10^5	■	■							
3	16	O	9592			■						
4	12	C	3032			■						
5	20	Ne	1548				■					
6	56	Fe	1169									■
7	14	N	1105		■							
8	28	Si	653					■				
9	24	Mg	513				■					
10	32	S	396					■				
11	22	Ne	207			■						
12	26	Mg	79								■	
13	36	Ar	77					■	■			
14	54	Fe	72						■			■
15	25	Mg	69				■					
16	40	Ca	60						■			
17	27	Al	58				■					
18	58	Ni	49						■			■
19	13	C	37		■							
20	3	He	35	■	■							
21	29	Si	34				■			■		
22	23	Na	33				■					
23	57	Fe	28								■	■
24	30	Si	23								■	
25	2	H	2	■								

Die Tabelle 1 zeigt die 25 häufigsten Isotope im Sonnensystem und ihre Entstehungsprozesse. Ihre Häufigkeit ist als relativer Massenanteil angegeben. Im nuklearen, statistischen Gleichgewicht (NSE) entstehen die verschiedenen Eisen-Isotope sowie ^{58}Ni.



Die Figur 18 zeigt die Häufigkeit der Elemente im Sonnensystem. Die verschiedenen Kernfusionsprozesse (H-Brennen bis Si-Brennen) produzieren die Elemente bis zum Eisen, einschliesslich den Kohlenstoff. Diese sind relativ häufig. Die schwereren Elemente sind in Supernovae im r-Prozess (Beispiele Ge, Xe, Pt) und in Roten Riesen im s-Prozess (Beispiele Sr, Ba, Pb) entstanden.

geführt haben. Es ist erstaunlich, welche Komplexität vonnöten ist, um diese Materie zu erzeugen, sie umfasst alle hier besprochenen Prozesse. Mit Ausnahme des Wasserstoffs (Rang 1) und des Deuteriums (Rang 25) ist das gesamte Material, aus dem wir bestehen, durch Sterne und Supernovae gekocht worden. Wir sind also im buchstäblichen Sinn Kinder der Sterne.

Woher stammen Kohlenstoff, Eisen und Uran?

Die Frage kann nun beantwortet werden:

- Der Kohlenstoff stammt **aus Roten Riesen,** wo durch die Verschmelzung von zwei ^{4}He Atomen zunächst das instabile ^{8}Be erzeugt wird, das sich mit einem weiteren ^{4}He-Kern auf Grund der Resonanzreaktion zum ^{12}C-Kern weiterentwickeln kann.

- Das Eisen entsteht als Endprodukt der Kernfusionsprozesse **in massereichen Sternen** vor ihrer Explosion zu Supernovae.

- Das Uran schliesslich stammt aus dem r-Prozess in den **Hüllen von Supernovae** (oder auch aus verschmelzenden Neutronensternen).

Ausblick

Die Theorie der Nukleosynthese verbindet die Kernphysik mit ihren kleinsten Dimensionen und die Astrophysik mit ihren grössten Dimensionen des Universums. Wir können heute im Labor Produkte untersuchen, die in den Explosionen von Supernovae vor Milliarden von Jahren irgendwo im Universum entstanden sind. Diese Beobachtungen stützen die These von der Universalität der Naturgesetze, wonach diese unabhängig von Ort und Zeit überall gegolten haben und auch in aller Zukunft weiter gelten werden. Auch wenn die Wissenschaft bereits viele Erkenntnisse über die Entstehung der Elemente gebracht hat, so bleiben doch für künftige Generationen von Naturwissenschaftern zahllose faszinierende Fragestellungen, die es Wert sind, weiter erforscht zu werden.

Der Autor

Rudolf von Steiger stammt aus einer alten Stadtberner Familie. Er besuchte das Gymnasium in Bern und schloss mit der Maturität vom Typ B ab. Anschliessend studierte er Physik, Mathematik und Astronomie an der Universität Bern. Seine Diplomarbeit in theoretischer Physik bei Prof. P. Hajicek (1984) befasste sich mit Fragen der Allgemeinen Relativitätstheorie. Seine Doktorarbeit bei Prof. Johannes Geiss widmete er Modellen zur Fraktionierung der Häufigkeiten von Elementen und Isotopen in der solaren Chromosphäre (1988). Anschliessend arbeitete R. von Steiger an der Universität Bern als Forschungsassistent. Ein Stipendium des Nationalfonds ermöglichte ihm 1990 ein Studienjahr an Institut für Physik und Astronomie der Universität Maryland, College Park (USA). Nach Bern zurückgekehrt war er wieder als Forschungsassistent bei J. Geiss tätig, wo vor allem die Auswertung und die Interpretation der Daten des Solar Wind Ion Composition Spectrometer (SWICS) auf der Weltraumsonde Ulysses im Vordergrund seiner Interessen standen. In dieser Zeit erhielt Rudolf von Steiger auch erste Lehraufträge, insbesondere im Bereich der experimentellen Kosmologie und der Physik. 1995 habilitierte er sich an der philosophisch-naturwissenschaftlichen Fakultät der Universität Bern mit einer Arbeit über die Zusammensetzung des Sonnenwindes.

Als 1995 das International Space Science Institute geschaffen wurde, übernahm Rudolf von Steiger unter der Leitung von J. Geiss die Funktion des Senior Scientists. Neben eigenen wissenschaftlichen Arbeiten gestaltete er aktiv das Wissenschaftsprogramm des ISSI und sorgte für die Herausgabe zahlreicher Bücher, die über den Stand der Forschung in ausgewählten Gebieten der Weltraumforschung berichten, eine Reihe, die inzwischen grosse internationale Beachtung gefunden hat. Ein weiteres Sabbatical führte ihn 1998 als Visiting Research Scientist, ans Institut für atmosphärische, ozeanische und Weltraumwissenschaften an der Universität von Michigan in Ann Arbor. Seit 1999 ist er Direktor am International Space Science Institute (ISSI) in Bern und Extraordinarius an der Universität Bern. Verschiedene Lehraufträge führten ihn unter anderem zurück an das Gymnasium Kirchenfeld, wo er selber die Maturität erlangt hatte und an die Universität Freiburg, wo er als Gastprofessor über den Aufbau der Materie berichtete.

Die stark im Humanismus verankerte Ausbildung öffnete Rudolf von Steiger den Blick nicht nur für die Naturwissenschaften, sondern auch für Sprachen, von denen er neben seiner Muttersprache Deutsch auch Latein, Englisch, Französisch, Italienisch und Japanisch beherrscht. Ein weiteres Interessensgebiet von Rudolf von Steiger ist die Musik. Er setzt die mit technischen Empfängern aufgenommenen Signale der Sonne oder die Sequenzen des menschlichen Genoms als elektronische Klänge um und interpretiert sie gemeinsam mit Musikern. Rudolf von Steiger ist verheiratet und lebt heute in Bugdorf (BE).

INTERNATIONAL SPACE SCIENCE INSTITUTE

SPATIUM

Publikationsorgan des Vereins Pro ISSI

Nr. 14, Juni 2005

Grundlagen der Physik im extraterrestrischen Test

Editorial

Eine kurze Begegnung im Oktober 1798 änderte das Leben des jungen schottischen Botanikers und Militärarztes Robert Brown grundsätzlich: Er lernte Sir Joseph Banks, den grossen Naturforscher kennen, der ihn gleich für seine Forschungsreise mit dem Segelschiff *Investigator* anheuerte. Die Reise führte ihn nach Australien und Tasmanien, und als die *Investigator* vier Jahre später wieder in London anlegte, hatte sie Tausende von Pflanzen an Bord, von denen die meisten noch unbekannt waren.

Fünf Jahre benötigte Robert Brown, um das umfangreiche Material zu beschreiben und zu klassifizieren. Dazu nutzte er das Mikroskop, um möglichst genaue Hinweise für die systematische Einordnung der Pflanzen zu finden. Eine überraschende Beobachtung gelingt ihm 1827: Er stellt fest, dass der Pollen der *Clarkia pulchella* aus kleinen Partikeln besteht, die sich offensichtlich in unaufhörlicher Bewegung befinden. Der Biologe denkt zunächst an kleinste Lebewesen, die sich mit eigener Kraft fortbewegen; daher untersucht er im Weiteren auch Pollenkörner, die während Monaten in Alkohol konserviert wurden, nur um wieder dieselbe dauernde Bewegung festzustellen. Selbst zu Staub zerriebener Fels zeigt das gleiche Verhalten. Brown, der systematische Forscher, zieht daraus den Schluss, dass es sich hier um ein physikalisches und nicht um ein biologisches Phänomen handelt.

Achtzig Jahre später stimmte der junge Albert Einstein dieser Interpretation zu. Er deutete die Brown'sche Bewegung als die Folge von Stössen der sich in thermischer Bewegung befindlichen unsichtbaren Moleküle der Flüssigkeit auf die kleinen sichtbaren Partikel. Diese Erkenntnis publizierte er 1905 zusammen mit drei weiteren Arbeiten in den Annalen der Physik, womit er ein neues Kapitel der Physik aufschlug.

Seit Einstein bemühen sich die Wissenschafter, in immer komplexeren Experimenten die Aussagen des Einstein'schen Gedankengebäudes zu überprüfen oder – noch besser – zu widerlegen. Denn bis heute ist es nicht gelungen, seine Relativitätstheorie mit der Quantenmechanik in Einklang zu bringen. Daher suchen die Physiker selbst nach kleinsten Unstimmigkeiten zwischen der Relativitätstheorie und dem Experiment, um allfällige Hinweise darauf zu erhalten, wo anzusetzen ist, um vielleicht doch noch die *Grand Unified Theory* zu finden, von der die Physiker seit Einstein träumen. Diesem Thema ist die vorliegende Ausgabe des Spatiums gewidmet.

Wir danken Herrn Professor Martin C.E. Huber, Präsident der Europäischen Physikalischen Gesellschaft, für die freundliche Erlaubnis, sein Referat vom 21. Oktober 2004 für Pro ISSI in erweiterter Form wiedergeben zu dürfen.

Hansjörg Schlaepfer
Zürich, Juni 2005

Impressum

SPATIUM
Publikationsorgan des Vereins Pro ISSI
Erscheint zweimal jährlich

INTERNATIONAL SPACE SCIENCE INSTITUTE

Verein Pro ISSI
Hallerstrasse 6, CH-3012 Bern
Tel. +41 (0)31 631 48 96
www.issi.unibe.ch/pro-issi.html
(hier sind alle bisherigen Nummern des Spatiums als pdf verfügbar)

Präsident
Prof. Heinrich Leutwyler, Universität Bern
Herausgeber
Dr. Hansjörg Schlaepfer, legenda schläpfer wort & bild, CH-8185 Winkel
Layout
Marcel Künzi, marketing · kommunikation, CH-8483 Kollbrunn
Druck
Schellenberg Druck AG
CH-8330 Pfäffikon

Titelbild
Dieses Bild zeigt die Wirkung hoher Massenkonzentrationen auf die Ausbreitung des Lichts: Im Vordergrund ist der Galaxienhaufen Abell 1689 als gelbe Objekte zu sehen. Die Masse seiner Galaxien ist so gross, dass er das Licht der dahinter liegenden Galaxie ablenkt und verzerrt, sodass sie mehrfach als blaue Kreissegmente sichtbar wird. (Quelle: NASA/ESA, STScI)

Grundlagen der Physik im extraterrestrischen Test[1]

Einleitung

Der österreichische Wissenschaftsphilosoph Karl Popper[2] schreibt 1972 in seiner «Objektiven Erkenntnis»[3]:

«… die neue Theorie kann wie jede unwiderlegte Theorie auch falsch sein. Der Theoretiker wird also alles versuchen, unter den nicht widerlegten Theorien falsche herauszufinden, ‹dingfest zu machen›. …»

«Er wird … versuchen, strenge Prüfungen sowie zwischen Theorien entscheidende Prüfungssituationen zu entwerfen.» …

«Ziel des Experimentes muss es … sein, eine Theorie zu falsifizieren, um auf diesem Weg allenfalls auf eine wahre Theorie zu stossen …»

Albert Einstein[4], der Schöpfer der Relativitätstheorie, war sich der epochalen Bedeutung seines Werkes wohl bewusst. Er selbst hat Experimente vorgeschlagen, um die Gültigkeit seiner Theorien zu überprüfen, ihre Aussagen allenfalls zu falsifizieren. Es ist besonders seit den Sechzigerjahren eines der zentralen Anliegen der Physik, die Relativitätstheorie immer «strengeren Prüfungen» zu unterwerfen und damit in den innersten Kern unseres physikalischen Weltbildes vorzustossen. Einerseits gibt das Universum selbst mit seiner unvorstellbaren Grösse und seinen riesigen Massen den idealen Rahmen für solche Tests. Andererseits eröffnet die Weltraumtechnik die Möglichkeiten, Experimente unter praktisch störungsfreien Bedingungen durchzuführen und zu Messgenauigkeiten vorzudringen, die weit über denjenigen irdischer Versuchsvorrichtungen liegen. Der Zugang zu solch exquisiten Experimentiertechniken ermöglicht es, sowohl die Richtigkeit der Theorien als auch die ihnen zu Grunde liegenden Annahmen mit stets grösserer Genauigkeit zu überprüfen. Das vorliegende Spatium ist dem Menschen Albert Einstein und den Bemühungen der Physiker gewidmet, seine Theorien zu falsifizieren.

« Eine Theorie sollte so einfach wie möglich sein, aber nicht einfacher. »

Spezielle und Allgemeine Relativitätstheorie

Die Spezielle Relativitätstheorie (1905) beschreibt die physikalischen Gesetze im Spezialfall unbeschleunigter Referenzsysteme, der sogenannten Inertialsysteme.

Die Allgemeine Relativitätstheorie (1916) beschreibt die physikalischen Gesetze im allgemeinen Fall, also auch für beschleunigte Referenzsysteme oder solche, die sich im Einflussbereich eines Gravitationsfeldes befinden.

Der entscheidende Gedanke der Relativitätstheorie ist der, dass alle physikalischen Gesetze in beschleunigten Systemen dieselben sind, unabhängig davon, ob die Beschleunigung von der Schwerkraft oder von einer Änderung der Geschwindigkeit stammt.

Bild 1: Der junge Albert Einstein im *annus mirabilis* 1905

[1] Nach dem Pro ISSI Referat vom 21. Oktober 2004
[2] Sir Karl R. Popper, 1902 Wien bis 1994 London; Philosoph und Wissenschaftstheoretiker
[3] K. Popper: Objektive Erkenntnis, campe paperback, 1993
[4] Albert Einstein, 1879 Ulm bis 1955 Princeton; Physiker, Nobelpreis für Physik 1922

Die Zeit vor der Relativitätstheorie

Am 14. März 1879 kommt in Ulm ein Knabe zur Welt, der die Naturwissenschaften in ihren Fundamenten erschüttern wird: Albert Einstein. Mit vier Jahren erhält er von seinem Onkel einen Kompass, der mit seiner stets nach Norden zeigenden Nadel den jungen Albert fasziniert. Vielleicht ahnt er dabei ein erstes Mal, dass etwas hinter den Dingen unsichtbar verborgen sein muss, das die Welt beherrscht. Von einem Freund der Familie erhält er im Laufe der Jahre eine Vielzahl populärwissenschaftlicher Bücher, die der heranwachsende Junge mit grossem Interesse verschlingt. In der Schule in München erbringt er zwar ausgezeichnete Leistungen in Mathematik und auch in den klassischen Sprachen, doch die wie «Feldwebel agierenden Lehrer» werden ihm zunehmend zum grossen Ärgernis. Ein eifriges Selbststudium hilft, den ihm völlig ungenügend erscheinenden Unterricht in Mathematik zu ertragen. Die Mittelschule bricht er unvermittelt ab, reist zu seinen Eltern, die sich mittlerweile in Mailand niedergelassen haben, und beginnt sich dort im Selbststudium für die Aufnahmeprüfung am Polytechnikum in Zürich vorzubereiten.

Diese findet im Oktober 1895 statt, endet für den erst 16-Jährigen aber mit einem Fiasko: Albert hatte die Sprachen vernachlässigt und sich zu sehr auf die ihn so faszinierenden naturwissenschaftlichen Fächer konzentriert. Für einen zweiten Anlauf schreibt er sich in der Abschlussklasse der Kantonsschule Aarau ein. Ihr liberales Klima sagt ihm sehr zu. Die herzliche Aufnahme durch die Familie des Kantonsschullehrers Winteler, vor allem aber die Tochter Marie, Einsteins erste grosse Liebe, haben es ihm angetan, und der junge Freigeist fühlt sich in Aarau überaus wohl. Das Ergebnis ist entsprechend: Eine hervorragende Maturität, wenn auch mit einer ungenügenden Note in der französischen Sprache…

« Mir geht es gut; ich bin ehrwürdiger eidgenössischer Tintenscheisser mit ordentlichem Gehalt: Daneben reite ich auf meinem alten mathematisch-physikalischen Steckenpferd und fege auf der Geige. »

Zu jener Zeit verfasst Albert Einstein seine erste wissenschaftliche Arbeit über den Äther, ein hypothetisches, das Vakuum ausfüllendes Medium, das damals zur Erklärung der Ausbreitung der elektromagnetischen Strahlung postuliert wurde. Im Oktober 1896 immatrikuliert er sich am Polytechnikum in der «Schule für Fachlehrer mathematischer und naturwissenschaftlicher Richtung». Die obligatorischen Lehrveranstaltungen lässt er mit mässiger Begeisterung über sich ergehen. Er fühlt sich eher zu der einen oder anderen fakultativen Vorlesung hingezogen, wie zum Beispiel zur Urgeschichte des Menschen oder zur Geologie der Gebirge. Nach vier Jahren schliesst Albert Einstein sein Studium mit dem Diplom ab, doch trotz intensiver Bemühungen gelingt es ihm nicht, eine Stelle als Assistent am Poly oder an einer anderen europäischen Hochschule zu finden. Daher muss er sich mehr schlecht als recht zunächst als Hilfslehrer am Technikum Winterthur, dann als Lehrer an einer Privatschule in Schaffhausen und als freier Privatlehrer in Bern durchschlagen. Dem gebürtigen Deutschen, der seine Staatsbürgerschaft nach der Übersiedlung in die Schweiz abgelegt hat, gewährt die stadtzürcherische Einbürgerungskommission das Schweizer Bürgerrecht und – mit Blick auf seine

Die Konstanz der Lichtgeschwindigkeit

Zentrales Element der Relativitätstheorie ist die absolute Konstanz der Lichtgeschwindigkeit von 299’792’458 m/s im Vakuum.
Sie bedingt, dass sich in einem bewegten Bezugssystem Raum und Zeit in Abhängigkeit seiner Geschwindigkeit so verändern, dass die Lichtgeschwindigkeit erhalten bleibt: Die Zeit dilatiert und die Raumdimensionen verkürzen sich; die Lichtgeschwindigkeit ist konstant, aber Raum und Zeit sind relativ.

bescheidenen finanziellen Verhältnisse – erst noch zu einem reduzierten Tarif. Dies und vertrauensvolle Beziehungen ermöglichen ihm, am 23. Juni 1902, eine Anstellung als Technischer Experte III. Klasse am Eidgenössischen Amt für Geistiges Eigentum in Bern zu finden. Nun sind auch die finanziellen Verhältnisse geregelt, sodass er zu Beginn des Jahres 1903 seine ehemalige Kommilitonin Mileva Maric heiraten kann. Das junge Paar fühlt sich in der Dachwohnung an der Tillierstrasse 18 in der «altertümlichen, urgemütlichen» Stadt Bern sehr wohl.

Es sind die Widersprüche der Physik seiner Zeit, die Albert Einstein zu den grossen Entdeckungen anregen: Während die klassische Mechanik von Isaac Newton[5] von einer zeitverzugslosen Wirkung der Gravitation auf Massen ausgeht, postuliert die neuere Feldtheorie von James Clerk Maxwell[6] die Ausbreitung elektromagnetischer Felder mit Lichtgeschwindigkeit. Auch der Begriff des Äthers führt zu Widersprüchen, denn das Licht einer bewegten Quelle müsste sich darin mit grösserer als Lichtgeschwindigkeit ausbreiten, was undenkbar ist, hatten doch Albert Michelson[7] und Edward Morley[8] erst vor wenigen Jahren in ihrem wegweisenden Experiment gezeigt, dass die Lichtgeschwindigkeit die gleiche ist, unabhängig davon, ob es sich in Richtung der Bahn der Erde um die Sonne – immerhin mit einer Geschwindigkeit von etwa 30 Kilometern pro Sekunde – oder quer dazu ausbreitet.

« Nur wer nicht sucht, irrt nicht. »

Oft diskutiert Einstein zu jener Zeit bis spät in die Nacht hinein oder auf Ausflügen am Sonntag mit seinen Freunden, die ihren Kreis scherzhaft *Akademie Olympia* nennen. Albert Einstein arbeitet an verschiedenen Themenkreisen gleichzeitig, und im Frühjahr 1905 publiziert er innerhalb weniger Monate fünf Arbeiten, welche der Physik die grosse Wende bringen: Die erste Arbeit gilt den energetischen Eigenschaften des Lichts. Hier führt er eine zum bisherigen Wellenbild alternative Erscheinungsform des Lichts ein: Energiepakete, die wir heute Photonen nennen. Dafür erhält er 1922 den Nobelpreis für Physik.

Die zweite Arbeit, seine Dissertation, betrifft die Bestimmung der Grösse von Molekülen. Die dritte damit verbundene Publikation erklärt die Brown’sche Bewegung von in ruhender Flüssigkeit suspendierten Körpern durch Stösse der unsichtbaren Moleküle, die sich gemäss den Gesetzen der Thermodynamik bewegen. Die vierte Arbeit erst behandelt die Elektrodynamik bewegter Körper: Die Grundlage der Speziellen Relativitätstheorie. Im Herbst veröffentlicht er noch einen fünften Artikel, in welchem er die wohl berühmteste Formel der Naturwissenschaften

$$E = m \cdot c^2$$

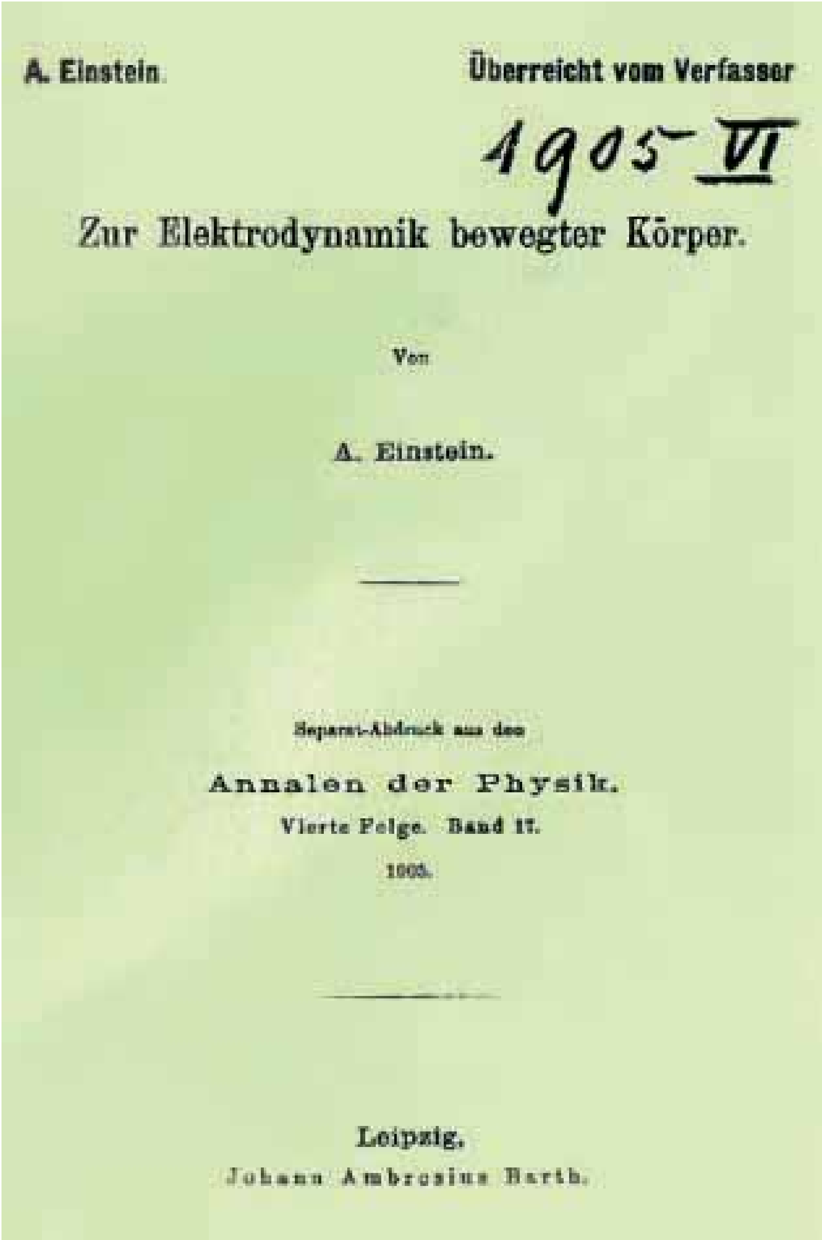
herleitet. Vier dieser Arbeiten erscheinen in den Annalen der Physik zwischen März und Dezember

[5] Sir Isaac Newton, 1643 Woolsthorpe bis 1727 London; Physiker
[6] James Clerk Maxwell, 1831 Edinburg bis 1879 Cambridge; Physiker
[7] Albert Michelson, 1852 Strelno (Posen) bis 1931 Pasadena; Physiker
[8] Edward Morley, 1838 Newark, New Jersey, bis 1923; Physiker

1905, die Arbeit über die Grösse der Moleküle reicht er als Doktorarbeit an der Universität Zürich ein. Diese ungeheure Leistung kennt in der gesamten Wissenschaftsgeschichte nichts Ebenbürtiges: Man spricht vom Jahr 1905 als dem *annus mirabilis,* dem Wunderjahr. So überragend die bisherigen Erkenntnisse von Albert Einstein auch sind, so zermürbend und rastlos ist der ihm noch bevorstehende, weitere elf Jahre dauernde Weg, der ihn erst auf die Höhe seines wissenschaftlichen Denkens bringen wird. Er versucht sich zunächst an der Universität Bern als Privatdozent für theoretische Physik zu habilitieren; erfolglos, denn er hat dem Gesuch zwar viele veröffentlichte Arbeiten, aber keine eigentliche Habilitationsschrift beigelegt. Erst nachdem Einstein der Forderung mit der unveröffentlichten, aber leider verschollenen Arbeit «Folgerungen aus dem Energieverteilungsgesetz der Strahlung schwarzer Körper, die Konstitution der Strahlung betreffend» nachkommt, sind die Formalitäten erfüllt, und Einstein beginnt Vorlesungen zu halten. Doch deren Erfolg hält sich in engen Grenzen: Nur wenige Hörer verirren sich in seine Lektionen.

Ein Jahr später folgt er einer Berufung der Universität Zürich als Extraordinarius. Ein weiteres halbes Jahr später übernimmt Einstein eine ordentliche Professur an der Universität Prag. Nach einem weiteren Jahr übersiedeln die Einsteins wieder nach Zürich und zwei Jahre später übernimmt Einstein eine Stelle in Berlin.

In dieser turbulenten Zeit arbeitet sich Einstein Schritt für Schritt voran: Zunächst postuliert er, dass die Physik in einer beschleunigten Umgebung dieselbe ist, gleichgültig ob die Beschleunigung durch Gravitation entsteht oder durch Änderung der Geschwindigkeit erzeugt wird. Man nennt dies das starke Äquivalenzprinzip (1907). Dieses beinhaltet auch die Äquivalenz von schwerer und träger


A. Einstein

Überreicht vom Verfasser

1905 VI

Zur Elektrodynamik bewegter Körper.

Von

A. Einstein.

Separat-Abdruck aus den

Annalen der Physik.

Vierte Folge. Band 17.

1905.

Leipzig.

Johann Ambrosius Barth.

Bild 2: Titelblatt der Publikation von Albert Einstein über die Elektrodynamik bewegter Körper. Diese Arbeit ist die Grundlage der Speziellen Relativitätstheorie.

Die Äquivalenz der Massen

Wenn sich ein Beobachter im freien Fall, zum Beispiel in einer frei fallenden Liftkabine, befindet und einen Apfel loslässt, so wird der Apfel bei der Hand bleiben. Wenn er sich für dieses Experiment an Bord eines um die Erde kreisenden Raumfahrzeuges begibt, so wird er das gleiche Resultat feststellen: In beiden Fällen fühlt er Schwerelosigkeit, in beiden Fällen wird er behaupten, keiner Gravitation ausgesetzt zu sein.

Wenn er sich dagegen auf dem Erdboden befindet, wird er feststellen, dass sein Apfel mit einer Beschleunigung von 9,8 m/s^2 zu Boden fällt. Entsprechend wird er an Bord eines mit 9,8 m/s^2 beschleunigten Raumfahrzeugs dasselbe feststellen.

Das Äquivalenzprinzip sagt aus, dass die – sich der Beschleunigung widersetzende – träge Masse und die – vom Gravitationsfeld angezogene – schwere Masse unabhängig von ihrer Zusammensetzung einander strikt proportional sind. Für Albert Einstein war dies eine Annahme, die damals erst mit einer Genauigkeit von 10^{-5} bestätigt war. Heute ist das Äquivalenzprinzip bis zu einer Genauigkeit von 10^{-13} überprüft.

Masse, das sogenannte schwache Äquivalenzprinzip. Dann erarbeitet er die nicht-euklidische Geometrie für gekrümmte Räume und stellt schliesslich die Feldgleichungen der Allgemeinen Relativitätstheorie auf (1916).

Es ist ein äusserst anstrengender Weg: Einstein verirrt sich einige Male, muss umkehren, um wieder von Neuem anzufangen. Hartnäckigkeit und überragende physikalische Intuition verhelfen ihm letztendlich doch zum krönenden Abschluss: Das wunderbare Werk der Allgemeinen Relativitätstheorie ist gelungen. Allerdings bezahlt er einen hohen Preis dafür: Das Verhältnis zu Mileva hat sich in den ruhelosen Jahren so sehr verschlechtert, dass sie mit den beiden Söhnen von Berlin zurück nach Zürich reist. Endgültig trennen sie sich im Sommer 1914, und Albert Einstein ist fortan allein, fern von seinen geliebten Söhnen Hans Albert und Eduard…

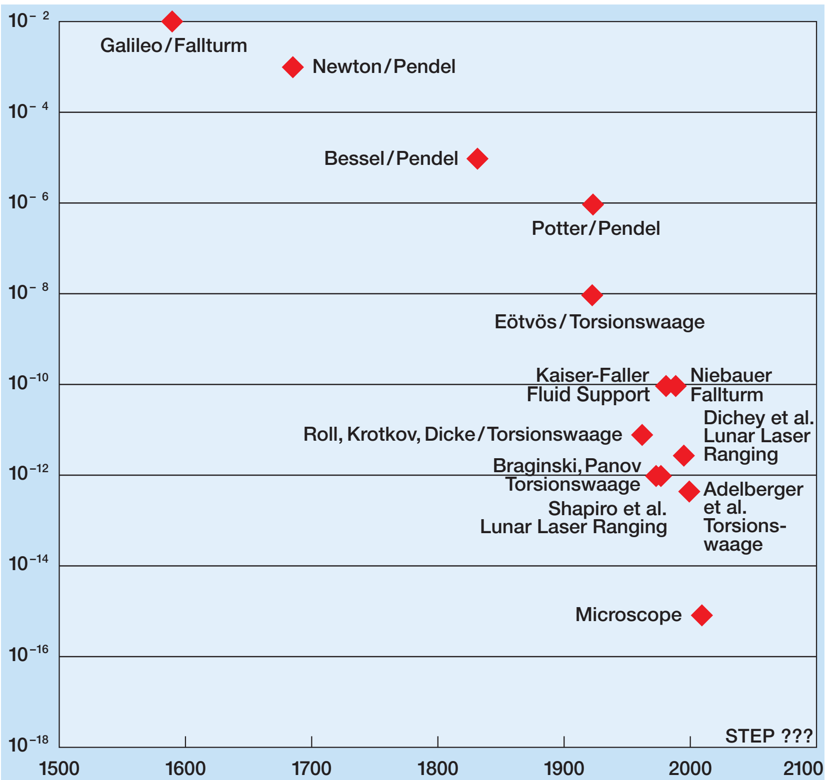


Bild 3: Die zeitliche Entwicklung der Messgenauigkeit zur Verifikation des Äquivalenzprinzips seit Galileo Galilei. Die heutige Genauigkeit liegt bei 10^{-13}. Das entspricht etwa der Wellenlänge von rotem Licht verglichen mit dem Durchmesser der Erde.

Die frühen Tests der Allgemeinen Relativitätstheorie

Die Kenntnis der Newton'schen Gesetze reicht in unserem Alltag längstens aus, um zurechtzukommen, denn hier sind die Geschwindigkeiten stets klein im Vergleich zur Lichtgeschwindigkeit, und das Gravitationsfeld der Erde ist überall auf ihrer Oberfläche annähernd konstant und verhältnismässig schwach. Daher spielen relativistische Effekte keine merkliche Rolle; zum Glück muss man sagen, denn die für das Verständnis der Newton'schen Gesetze erforderliche Mittelschul-Mathematik ist den meisten Menschen zugänglich, was bei der speziellen Relativitätstheorie gerade noch, bei der Allgemeinen Relativitätstheorie aber nicht mehr zutrifft.

Den Zeitgenossen Einsteins geht es gleich; seine Gedanken werden selbst in der Fachwelt zunächst nur mit äusserster Skepsis aufgenommen. Den ersten Durchbruch ermöglicht die Klärung eines seit Jahrzehnten offenen Problems, nämlich der Wanderung des Perihels des Merkurs, das heisst des Punktes der Merkurbahn, welcher der Sonne am nächsten liegt.

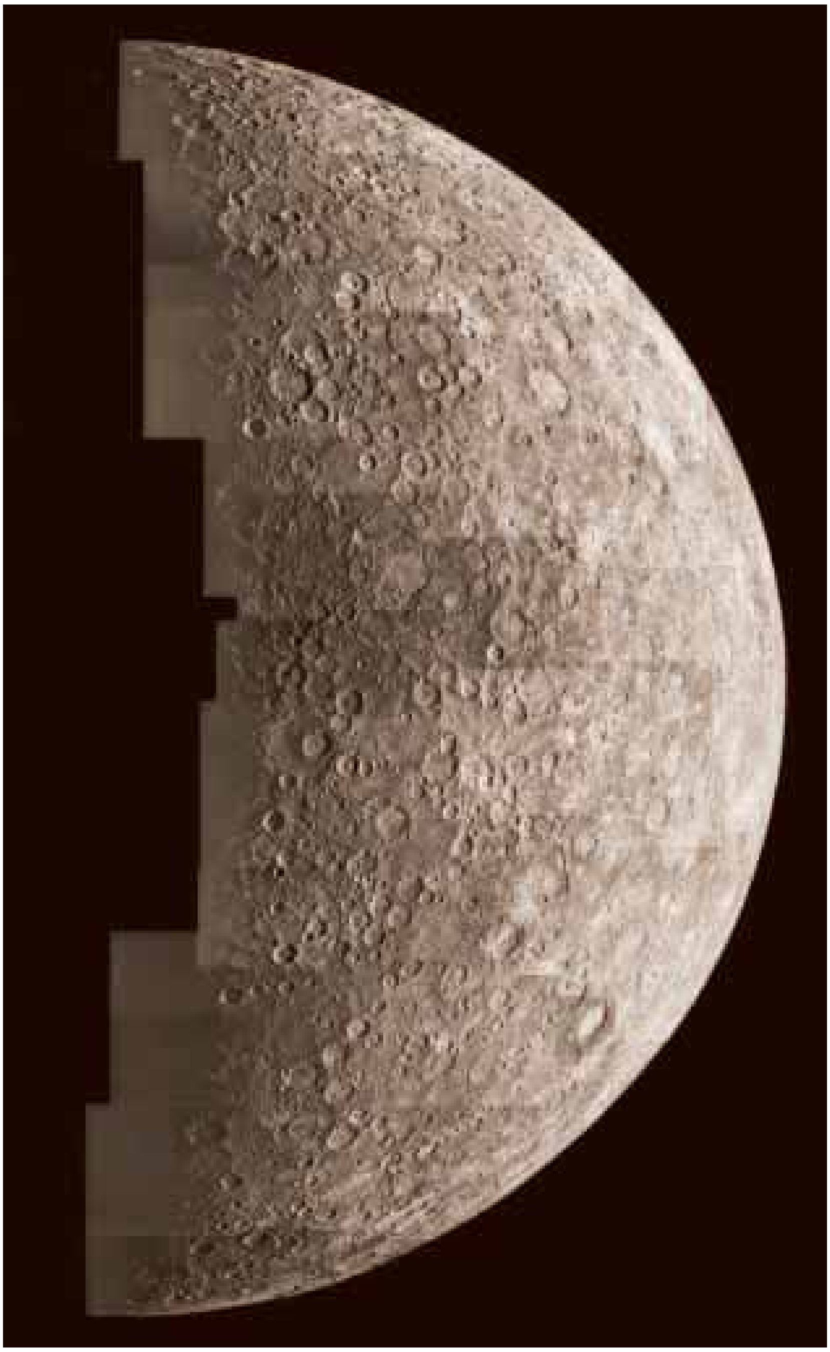


Bild 4: Merkur, der innerste Planet, aufgenommen 1974 von der amerikanischen Sonde Mariner 10. Er umkreist die Sonne auf einer elliptischen Bahn hoher Exzentrizität: Das Aphel, der sonnenfernste Punkt der Umlaufbahn, misst 69,82 Mio km, das Perihel (der sonnennächste Punkt) 46 Mio km. In diesem Bereich ist das Gravitationsfeld der Sonne so stark, dass sich das Perihel schneller um die Sonne dreht, als es die klassische Mechanik erwarten lässt. Dies war der erste Hinweis auf die Gültigkeit der Allgemeinen Relativitätstheorie.

Die Drehung des Perihels des Merkurs

Nach dem ersten Kepler'schen[9] Gesetz bewegen sich die Planeten auf Ellipsen um die Sonne. Das gilt, solange es sich um einen einzigen Planeten handelt, der um die Sonne kreist. Im Sonnensystem beeinflussen sich aber die Planeten gegenseitig, was zu einer laufenden Bahnänderung namentlich auch des Merkurs führt. Auch das kann die Newton'sche Physik noch erklären.

Bereits 1855 erkannte Urbain Le Verrier[10], dass die Bahn des Merkurs von diesen Erwartungen abweicht. Er versuchte, die Diskrepanz mit einem bisher nicht beobachteten Planeten oder einem unsichtbaren Asteroidengürtel zu erklären. De Sitter[11] stellte um 1880 auf Grund präziserer Messungen fest, dass das Perihelion um etwa 531 Bogensekunden pro Jahrhundert wandert, während die klassische Newton'sche Mechanik einen etwas kleineren Wert voraussagt. Die Allgemeine Relativitätstheorie erklärt die fehlenden 43 Bogensekunden mit der Dilatation der Zeit im Gravitationsfeld der Sonne, womit sie in Wissenschaftskreisen ein erstes Mal für Aufmerksamkeit sorgt. Damit wird die Erklärung der Drehung des Merkur-Perihels zum ersten grossen Erfolg der Relativitätstheorie.

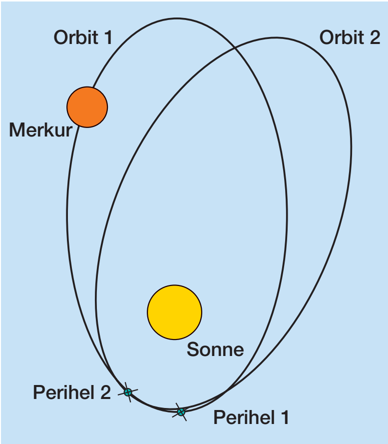


Bild 5: Die zeitliche Veränderung der Merkurbahn um die Sonne: Das Perihel des Merkurs wandert im Laufe der Zeit um die Sonne nicht nur auf Grund der Einflüsse der (hier nicht eingezeichneten) Nachbarplaneten, sondern auch infolge relativistischer Effekte im Gravitationsfeld der Sonne.

Die Ablenkung des Lichts an massereichen Sternen

Von Albert Einstein stammt dagegen die Idee, die Änderung der Ausbreitungsrichtung des Lichts im Bereich des Gravitationsfeldes der Sonne zu bestimmen, denn die klassische Mechanik sagt eine Ablenkung voraus, die halb so gross ist wie die von der Allgemeinen Relativitätstheorie vorhergesagte. Der Unterschied ist zwar im Bereich der Sonne klein, er ist zu jener Zeit aber schon messbar. Dazu ist eine totale Sonnenfinsternis erforderlich, bei welcher der Mond die Sonnenscheibe vollständig bedeckt, so dass die Sterne sichtbar werden. Zur Beobachtung der Sonnenfinsternis vom 21. August 1914 macht sich eine Expedition nach Russland auf. Sie endet mit der Gefangennahme der deutschen Expeditionsteilnehmer. Zwar können diese später gegen russische Kriegsgefangene ausgetauscht werden, doch die Sonnenfinsternis ist mittlerweile vorüber.

Eine weitere Finsternis ist für den 29. Mai 1919 in äquatorialen Breiten vorausgesagt. Sir Arthur Eddington[12], ein engagierter Verfechter der Relativitätstheorie, versucht, die sich zusehends verschlechternden Beziehungen Englands zu Deutschland nach der Besetzung des neutralen Belgiens durch Zusammenarbeit auf wissenschaftlicher Ebene zu verbessern. Er organisiert zwei Expeditionen, von denen die eine in

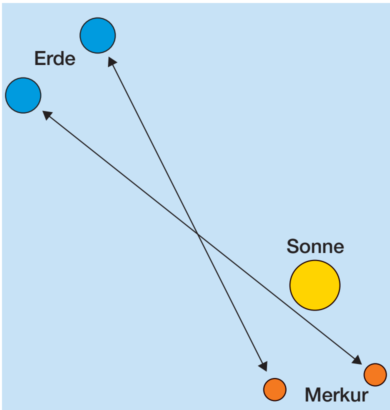


Bild 6: Zur Geometrie des Versuchs von I.I. Shapiro: Die Laufzeit eines Radarsignals von der Erde zum Merkur und zurück wird zunächst bei grossem Abstand der Achse Merkur/Erde von der Sonne gemessen und ein weiteres Mal, wenn die Achse nahe an der Sonne vorbeigeht. Im zweiten Fall wird das Signal im Gravitationsfeld der Sonne um insgesamt 0,2 ms gegenüber der von der klassischen Mechanik vorhergesagten Laufzeit verzögert.

[9] Johannes Kepler, 1571 Weil bis 1630 Regensburg; Mathematiker und Astronom
[10] Urbain Le Verrier, 1811 St. Lô bis 1877 Paris; Mathematiker und Astronom
[11] Willem de Sitter, 1872 Sneek bis 1934 Leiden; Astronom
[12] Sir Arthur Eddington, 1882 Kendal bis 1944 Cambridge; Physiker und Astronom

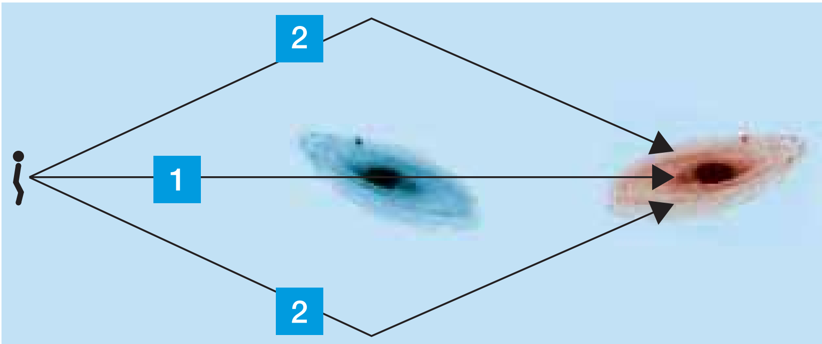


Bild 7: Zur Geometrie der Gravitationslinsen. In Abwesenheit der vorderen Galaxie sieht der Beobachter die Galaxie im Hintergrund in der Richtung 1. Unter dem relativistischen Einfluss der Gravitation der Galaxie im Vordergrund wird das Licht der Galaxie im Hintergrund umgelenkt, und der Beobachter sieht ihr Bild in den Richtungen 2, d. h. rund um die vordere Galaxie. Wären die beiden Galaxien praktisch punktförmig und lägen sie exakt auf derselben Linie, so entstünde ein kreisförmiges Bild der Hintergrundgalaxie. Die ausgedehnten Massenverteilungen von Galaxien führen zu einem verzerrten Bild, wobei das umgelenkte Licht vor allem in Form von Kreissegmenten erscheint. Das Titelbild der vorliegenden Nummer zeigt diesen Fall.

den Norden Brasiliens, die andere zur Insel Principe im Golf von Guinea ausgesandt wird. Trotz schwierigen Wetterbedingungen entstehen sieben brauchbare fotografische Platten. Sie werden ausgemessen und die Sternpositionen mit älteren Fotos des Nachthimmels, bei denen die Sonne weit weg von der Sichtlinie war, verglichen. Anfang September 1919 hält Eddington in Bournemouth einen Vortrag, in welchem er das vorläufige Ergebnis bekannt gibt: Die beobachteten Positionsverschiebungen der Sterne in Sonnennähe stimmen mit der Vorhersage der Allgemeinen Relativitätstheorie überein! Dieser augenfällige Beweis bringt Einstein die allgemeine Anerkennung seiner neuen Ideen, nicht nur in wissenschaftlichen Kreisen, sondern nun auch in der breiteren Öffentlichkeit…

Der Shapiro-Effekt [13]

Einen ähnlichen Versuch führt etwa 50 Jahre später I. I. Shapiro durch: Er sendet Radarsignale zu den Planeten Venus und Merkur und misst an den reflektierten Signalen die Verzögerung infolge der Dilatation der Zeit im Gravitationsfeld der Sonne. Die erwartete Verzögerung beträgt 0,2 Tausendstel Sekunden, was er mit einer Genauigkeit von etwa 10% bestätigen kann.

Auch die Natur selber führt relativistische Experimente durch, die von aufmerksamen Astronomen beobachtet werden können. Es handelt sich um die Gravitationslinsen: Hohe Dichten von Materie, die zwischen einer Galaxie im Hintergrund und dem Beobachter liegen, führen zu einer Ablenkung des Lichts der Galaxie, sodass diese mehrfach beobachtet werden kann, siehe **Bild 7**. Das berühmte Einsteinkreuz, wie es in der Abbildung 11 des Spatium Nr. 3 dargestellt ist, kommt auf diese Weise zustande.

Die Gravitations-Rotverschiebung

Eine weitere «strenge Prüfung» im Popper'schen Sinne betrifft die Rotverschiebung des Lichts, die auftritt, wenn sich die Strahlungsquelle in einem stärkeren Gravitationsfeld befindet als der Beobachter. Dort laufen die Uhren langsamer, und entsprechend sendet ein Atom seine Spektrallinien auch mit einer etwas langsameren Frequenz aus. Dies kann man mit modernen Spektrometern im Sonnenspektrum beobachten.

Das präziseste bisher durchgeführte Experiment zur Bestimmung der Rotverschiebung kommt überraschenderweise ohne das Gravitationsfeld eines Sterns aus, aber nützt die Potenzialdifferenz auf unterschiedlichen Höhen im Gravitationsfeld der Erde. Es stammt von Pound-Rebka-Snider (1960 bis 1965) und besteht darin, dass am Fuss des Laborturms der Harvard Universität angeregtes Eisen Photonen emittiert, die an der Spitze des Turms von einem auf dem Mössbauer-Effekt beruhenden Empfänger detektiert werden. Die Photonen bewegen sich entgegen dem Schwerefeld der Erde und verlieren dabei etwas Energie, was zu einer zwar äusserst geringen Rotverschiebung führt, die aber mit diesem empfindlichen Messverfahren doch zweifelsfrei nachgewiesen werden kann.

---

[13] Irwin I. Shapiro, 1929 New York; Physiker

Die extra-terrestrischen Tests

Seit etwa 50 Jahren steht die erforderliche Technik zur Verfügung, um Experimente im Weltraum, aber auch in grossen Teilchenbeschleunigern durchzuführen und die Relativitätstheorie wie auch die Quantenmechanik um viele Grössenordnungen genauer zu testen, als es mit den bisher besprochenen Versuchsanordnungen möglich ist. Und ohne dass wir es bemerkt hätten, hat sich die Thematik der relativistischen Zeit schon in unseren Alltag eingeschlichen, denn die Navigation mittels Satelliten ist eine alltägliche Anwendung der Relativitätstheorie. Diese Navigationssysteme erfordern Uhren höchster Genauigkeit. Damit alle Uhren im System dieselbe Zeit anzeigen, müssen ihre relativistischen Frequenzabweichungen ständig kompensiert werden. Doch kehren wir zunächst zum ersten Test im Weltraum, zur Gravity Probe A, zurück.

Die Gravity Probe A

Ende der 70er-Jahre startet die amerikanische Raumfahrtbehörde NASA auf einer Scout-Rakete das Experiment Gravity Probe A. Der Versuchsaufbau besteht aus zwei Wasserstoff-Masern (siehe Kasten) als Zeitnormale höchster Präzision, wovon der eine auf dem Erdboden stationiert ist, währenddem der zweite mit der Rakete auf eine Höhe von 10'000 Kilometern gebracht wird. Während des Fluges werden die Taktfrequenzen der beiden Uhren laufend miteinander verglichen. Die Geschwindigkeit der Rakete führt zu einer Verlangsamung der Uhr an Bord der Rakete gemäss der Speziellen Relativitätstheorie. Andererseits bewirkt das mit zunehmender Höhe schwächer werdende Gravitationsfeld der Erde eine stets raschere Gangart der Borduhr entsprechend der Allgemeinen Relativitätstheorie. Die Messungen bestätigen diese mit einer Genauigkeit von einem Teil in 100'000.

Maser und Laser

Der Maser (Microwave Amplification by Stimulated Emission of Radiation) und der Laser (Light Amplification by Stimulated Emission of Radiation) sind Verfahren zur Erzeugung kohärenter elektromagnetischer Strahlung.

Im Bohr'schen Atommodell umkreisen die Elektronen den Kern, und zwar je nach ihrem Energiezustand in unterschiedlichen Bahnen. Energiezufuhr durch ein Photon bewirkt, dass die Elektronen in energiereichere Bahnen übergehen (angeregte Zustände), aus denen sie normalerweise spontan wieder in den Grundzustand fallen unter Abgabe eines Photons.

Einstein hat 1915 das Konzept der stimulierten Emission eingeführt: Hier wird ein angeregtes Atom durch ein Photon zu einem Übergang in einen tieferen Energiezustand stimuliert, was zu einer Verstärkung führt: Aus einem Photon werden zwei. Dieser Vorgang wird in Lasern mehrfach wiederholt, wobei die entstehenden Lichtwellen im Takt einer wohldefinierten Frequenz sind. Damit lassen sich Uhren höchster Präzision herstellen.

Satellitennavigation

Die 24 Satelliten des zukünftigen europäischen Galileo-Systems umkreisen in 12 Stunden die Erde einmal. Die Uhren an Bord der Satelliten verlieren auf Grund ihrer Geschwindigkeit 7 µs pro Tag im Vergleich zu den Referenzuhren am Boden (Spezielle Relativitätstheorie). Die in der Umlaufbahn der Satelliten geringere Schwerkraft der Erde führt andererseits zu einem Gewinn von 45 µs pro Tag (Allgemeine Relativitätstheorie). Somit ergibt sich eine Differenz von 38 µs pro Tag, was bei der Lichtgeschwindigkeit von ca. 300'000 km/s zu Fehlmessungen von bis zu 11'400 m führen würde, wenn die relativistischen Effekte nicht kompensiert würden.

Das ACES (Atomic Clock Ensemble in Space) Experiment

Von allen physikalischen Grössen kann die Zeit mit Abstand am genauesten gemessen werden. Diesen Vorsprung weiter voranzutreiben, ist eines der Ziele des ACES-Experiments, das für die Installation als externe Nutzlast des Europäischen Columbus Labors an Bord der Internationalen Raumstation ISS vorgesehen ist. Es besteht aus zwei Atomuhren mit unterschiedlichen Atomen: Das PHARAO (Projet d'horloge atomique par refroidissement d'atomes en orbite) des französischen Centre National d'Etudes Spatiales (CNES) basiert auf Caesium-Atomen, die mittels Laserstrahlung auf eine Temperatur von wenigen Millionstel Grad über dem absoluten Nullpunkt gekühlt werden. Diese Temperatur und das praktische Fehlen von Beschleunigungen an Bord der ISS ermöglichen eine Stabilität besser als 3×10^{-16} über einen Tag. Die zweite Uhr wird vom Observatoire Cantonal de Neuchâtel in Zusammenarbeit mit Contraves Space entwickelt. Es handelt sich dabei um den Space Hydrogen Maser (SHM), der auf Wasserstoffatomen basiert. Während PHARAO eine hervorragende Langzeitstabilität besitzt, ist ihm der Wasserstoffmaser über Zeiträume von weniger als einer Stunde überlegen. Er dient daher als Referenz für den Caesium Maser. Zusammen bilden PHARAO und SHM eine Uhr, die genauer ist als alle bisherigen Uhren: Sie hat eine Abweichung von weniger als einer Sekunde in 100 Millionen Jahren!

Derartige Genauigkeiten sind erforderlich, wenn Verkehrsflugzeuge künftig mit satellitenbasierten Navigationssystemen sicher landen sollen.

Die Signale der beiden Uhren gelangen zunächst zu einer gemeinsamen Datenverarbeitung an Bord der Raumstation und anschliessend zu Forschungslaboratorien in aller Welt. Das Experiment verfolgt damit auch das Ziel einer genauen zeitlichen Synchronisation von globalen Datennetzwerken.

Gemäss der Allgemeinen Relativitätstheorie ist zu erwarten, dass die beiden Uhren schneller gehen werden als auf der Erde, da auf der Höhe der Raumstation das Gravitationspotenzial der Erde geringer ist; andrerseits folgt aus der Speziellen Relativitätstheorie, dass die Uhren wegen ihrer Geschwindigkeit gegenüber fest auf der Erdoberfläche montierten Uhren etwas langsamer laufen. Beide Effekte lassen sich aber aufgrund der bekannten Position und Geschwindigkeit der Uhren in der Umlaufbahn berechnen. Mit dieser Versuchsanordnung wird auch das Äquivalenzprinzip überprüft, das besagt, dass sich die beiden Uhren am Boden und im Orbit der Raumstation gleich verhalten, obwohl sie für die Zeitmessung verschiedene Atome benützen. Eine allfällige Drift der Taktfrequenzen der beiden Uhren im Orbit, das heisst in einem anderen Gravitationspotenzial, käme daher einer Verletzung des Äquivalenzprinzips gleich.

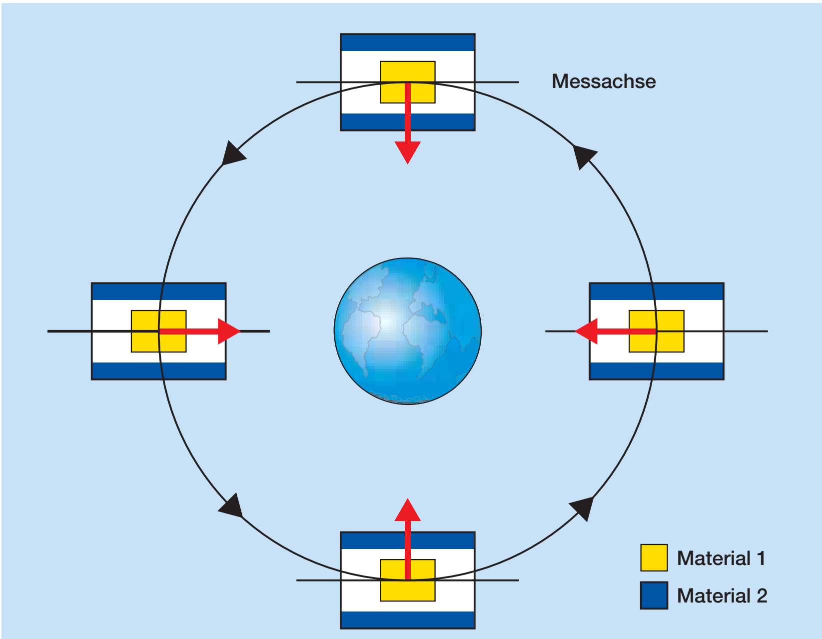


Bild 8: Das Messprinzip des Microscope-Satelliten: Zwei Testmassen aus unterschiedlichen Materialien (gelb und blau eingetragen) sind konzentrisch angeordnet. Sie können sich gegenseitig entlang der gemeinsamen Achse verschieben. Eine allfällige Verletzung des Äquivalenzprinzips würde zu einer Auslenkung der inneren gegenüber der äusseren Testmasse im Rhythmus des Erdumlaufs führen.

Die Microscope-Mission

Das Äquivalenzprinzip und damit auch die Äquivalenz von träger Masse und schwerer Masse ist der zentrale Kern, sozusagen das Dogma der Allgemeinen Relativitätstheorie. Für die Verifikation des Äquivalenzprinzips hatten Wissenschafter schon 1989 die Mission «Satellite Test for the Equivalence Principle» (STEP) vorgeschlagen. Die Mission ist jedoch bis heute nicht realisiert worden, dagegen hat die französische Raumfahrtbehörde CNES die Initiative ergriffen und das Projekt Microscope (MICROSatellite à Traînée Compensée pour l'Observation du Principe d'Equivalence) in ihr Programm aufgenommen. Dieses Experiment wird auch von der ESA unterstützt; es befindet sich derzeit in Entwicklung und soll im März 2008 in eine Erdumlaufbahn gebracht werden.

Das Messverfahren

Das Konzept der Mission gleicht dem Versuch von Galileo Galilei, der um 1590 zwei Körper unterschiedlicher Zusammensetzung gleichzeitig vom Schiefen Turm in Pisa fallen liess und feststellte, dass sie exakt zur gleichen Zeit auf dem Boden auftreffen. Auch beim Microscope-Experiment handelt es sich um den freien Fall zweier unterschiedlicher Testmassen aus reinem Platin, resp. aus einer Platin-Titan-Legierung. Sie befindet sich in einer Erdumlaufbahn und unterliegen dem Gravitationsfeld der Erde: Ihre schweren Massen erfahren demnach die gleiche Anziehung. Beide bewegen sich mit einer Geschwindigkeit von etwa 20'000 Kilometern pro Stunde um die Erde, und ihre trägen Massen widersetzen sich in gleicher Weise der durch das Gravitationsfeld erzwungenen Kreisbahn. Wenn nun die träge Masse ungleich der schweren Masse wäre, dann würden mit der Zeit die Bahnen der beiden Testmassen auseinanderlaufen. Um den Versuch zu vereinfachen, sind die Testmassen nur entlang einer Richtung frei beweglich. Präzise kapazitive Sensoren messen laufend ihre Lage. Sobald sie kleinste Unterschiede erkennen, werden Rückführkräfte ausgelöst, sodass die Testmassen in ihrer ursprünglichen Lage bleiben. Sollten periodische Rückführkräfte im Rhythmus der Umlaufbahn auftreten, so liesse sich daraus eine Verletzung des Äquivalenzprinzips bestimmen. Dazu sind die Testmassen vor äusseren Einflüssen, wie zum Beispiel den restlichen Atomen der Atmosphäre und der Strahlung der Sonne, zu schützen. Dies wird erreicht, indem der Satellit mittels feiner Triebwerke so gesteuert wird, dass die Testmassen stets in einer rein der Gravitation unterliegenden Umlaufbahn sind.

Der Satellit

Der Satellit basiert auf der Myriade-Plattform des CNES und findet mit seinen nur 120 Kilogramm Gewicht an Bord einer Zusatzstruktur der Ariane 5 Platz.

Bild 9 zeigt den Microscope-Satelliten in einer Darstellung von D. Ducros.

Gravitationswellen

Nach den Vorstellungen der klassischen Physik wirkt die Gravitationskraft über grössere Entfernungen zwar mit abnehmender Stärke, aber ohne Zeitverzug. Nach relativistischer Anschauung dagegen ist Gravitation ein sich mit Lichtgeschwindigkeit ausbreitendes Feld, was grundsätzlich die Existenz von Wellen beinhaltet.

Als in Bern im Jahre 1955 unter dem Vorsitz von Wolfgang Pauli[14] eine Konferenz zum 50. Jubiläum von Einsteins *annus mirabilis* stattfindet, gehen die Meinungen darüber, wie real Gravitationswellen seien, noch weit auseinander. Hermann Bondi schreibt in seinen Erinnerungen[15], dass er damals von Markus Fierz[16] mit den Worten «the problem of gravitational waves is ready for solution, and you are the person to solve it» aufgefordert worden sei, sich mit diesem Problem zu beschäftigen. Und tatsächlich haben Bondi und seine Mitarbeiter in den folgenden Jahren die Theorie der Gravitationswellen erarbeitet. Ihre Bemühungen finden 1962 ihren krönenden Abschluss.

Bis heute ist es nicht gelungen, Gravitationswellen direkt nachzuweisen. Aber zwei astronomische Entdeckungen der 60er- und 70er-Jahre haben zu einem indirekten Nachweis der Existenz von Gravitationswellen geführt. Beiden wurden Nobelpreise zuerkannt. Antony Hewish[17] und seine Studentin Jocelyn Bell finden 1967 ein neuartiges Himmelsobjekt, einen Pulsar, d.h. eine Radioquelle, die äusserst regelmässig kurze Pulse aussendet. In einer intensiven Suche nach weiteren Pulsaren entdecken Russell Hulse und Joseph Taylor 1974 den Doppelpulsar PSR 1913+16, in welchem ein Pulsar, zusammen mit einem etwa gleich schweren Begleiter, ein enges Doppelsternsystem bildet.

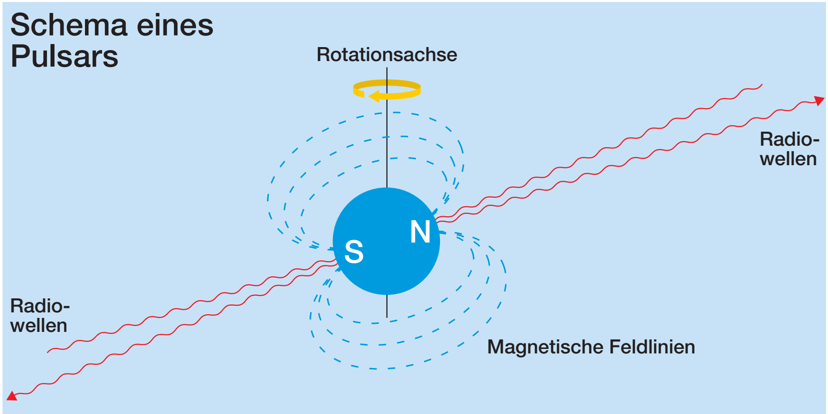


Bild 10: Ein Pulsar rotiert und sendet Radiowellen in zwei Strahlenbündeln aus, die, ähnlich wie beim Lichtbündel eines Leuchtturms, an einem weit entfernten Beobachter vorbeiflitzen und von diesem als Radio- und Lichtpulse registriert bzw. empfunden werden. Pulsare sind äusserst kompakte Objekte, sogenannte Neutronensterne. Diese sind etwa gleich schwer wie die Sonne (genauer gesagt: Sie haben eine mit der Sonne vergleichbare Masse); aber dieselbe Masse, die in der Sonnenkugel einen Durchmesser von 1,2 Millionen Kilometern einnimmt, ist in Neutronensternen enorm verdichtet – sie ist in einen Durchmesser von lediglich etwa 10 km gepackt.

Dass es sich um ein Doppelsternsystem handelt, bei dem die zwei Objekte umeinander kreisen, schliessen Hulse und Taylor aus der regelmässigen Variation der Pulsfolge: Während sich der Pulsar von uns wegbewegt, verlangsamt sich scheinbar die Folge seiner Pulse, und wenn er wieder auf uns zukommt, treffen die Pulse wieder schneller hintereinander ein. Diese Modulation der Pulsfrequenz beruht auf dem Dopplereffekt: Wenn der Pulsar von uns wegfliegt, vergrössert sich der Zeitabstand zwischen zwei Pulsen, weil der spätere Radiopuls einen etwas längeren Weg zurücklegen muss als sein Vorgänger, bis er bei uns eintrifft. Das Umgekehrte ist der Fall, wenn der Pulsar auf uns zukommt.

---

14 Wolfgang Pauli, 1900 Wien bis 1958 Zürich; Physiker, Nobelpreis für Physik 1945.
15 Sir Hermann Bondi, 1919 Wien; Mathematiker und Kosmologe: *Science, Churchill and Me,* Pergamon Press, 1990
16 Markus Fierz, 1912; Physiker
17 Antony Hewish, 1924 Fowey, Cornwall; Physiker, Nobelpreis für Physik 1974

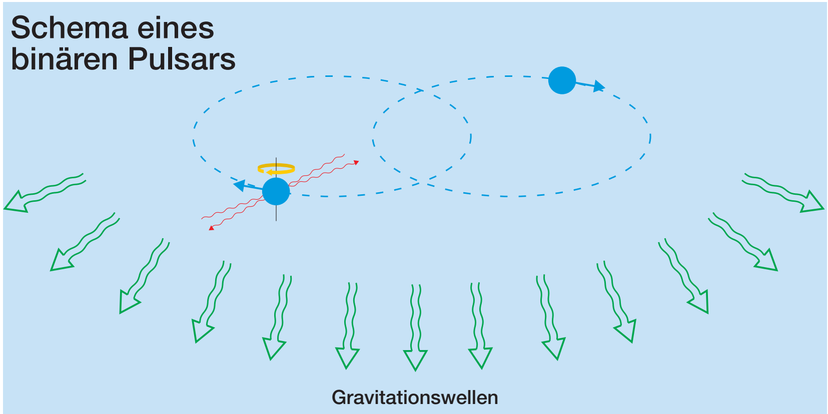


Bild 11 zeigt den von Hulse und Taylor entdeckten Doppelpulsar PSR 1913+16 in einer Entfernung von 21'000 Lichtjahren. Er besteht aus zwei etwa gleich grossen Neutronensternen, die sich innerhalb von nur acht Stunden umkreisen. Einer der beiden Neutronensterne wird als Pulsar beobachtet; er rotiert etwa 17 Mal pro Sekunde um seine eigene Achse. Der Abstand der beiden Neutronensterne im Doppelpulsar ist klein: lediglich einige Male die Distanz von der Erde zum Mond.

Taylor und seine Mitarbeiter beobachten den Doppelpulsar über mehrere Jahre und finden, dass sich die Pulsfolge langsam verändert. Sie können eindeutig nachweisen, dass das System stetig Energie verliert und sich dabei entsprechend den Voraussagen der Allgemeinen Relativitätstheorie verändert. Die verlorene Energie entspricht genau dem Wert, den das System nach der Theorie von Bondi in Form von Gravitationswellen abstrahlt. Damit liegt ein erster indirekter Beweis für die Existenz von Gravitationswellen vor.

Gravitationswellen entstehen, wenn sich Massen relativ zueinander bewegen. Dabei verändern sie in ihrer Umgebung das Raumzeit-Kontinuum. Diese Störungen des Raumzeit-Kontinuums breiten sich als Gravitationswellen aus, ähnlich wie wenn von einem Stein, der ins Wasser geworfen wurde, an der Oberfläche Wellen erzeugt werden. Allerdings breiten sich Gravitationswellen mit Lichtgeschwindigkeit aus. Aber selbst die Analogie mit elektromagnetischen Wellen ist nicht vollständig. Der tiefere Grund dafür liegt in der Symmetrie der Gravitation. Diese ist eine von der Elektrizität grundsätzlich verschiedene Erscheinung. Dort ziehen sich positive und negative Ladungen an, während sich gleichpolige abstossen. In der Gravitation dagegen gibt es nur eine Art «Ladung», die Masse, und nur Anziehung, keine Abstossung. Dies bedeutet auch, dass Gravitationswellen nur entstehen, wenn sich Massen asymmetrisch gegeneinander bewegen. Beispielsweise erzeugt ein Stern, der so pulsiert, dass sich lediglich sein Radius periodisch verändert, keine Gravitationswellen.

Faszinierend am Doppelpulsar von Hulse und Taylor ist, dass man hier zum ersten Mal ein Objekt beobachten kann, das starker Gravitation unterworfen ist. Die Effekte der Allgemeinen Relativitätstheorie führen hier zu erheblichen Abweichungen von der Newton'schen Mechanik. So beträgt im Doppelpulsar die Drehung des Periastrons 4,2° pro Jahr – rund 35'000 mal schneller als die dazu analoge Drehung des Merkurperihels von 43 Bogensekunden pro Jahrhundert!

Gravitationswellen sind nicht nur deshalb interessant, weil sie eine der wenigen bis heute nicht direkt nachgewiesenen Voraussagen der Allgemeinen Relativitätstheorie sind, sondern auch, weil sie uns Einblicke in Objekte geben, in denen starke Gravitation herrscht. In der Tat eröffnen Gravitationswellen ein völlig neues Fenster zum Universum, da sie ein von elektromagnetischen Wellen grundsätzlich verschiedenes Phänomen sind.

Gravitationswellen enthalten zwar enorme Energien, sie verzerren aber das Raumzeit-Kontinuum nur geringfügig, das heisst, sie haben nur eine geringe Wechselwirkung mit der Materie. Typisch sind Verzerrungen des Raums um einen Teil in 10^{22}, d.h. eine kilometerlange Strecke, wie sie in den derzeit sich im Bau befindlichen erdgebundenen Gravitationswellen-Detektoren üblich ist, ändert sich lediglich um einen Zehntausendstel eines Kerndurchmessers (10^{-19} m). Dies macht es so schwierig, Gravitationswellen zu detektieren. Die

Bild 12: Der Spiralnebel NGC 613 enthält, wie unsere Milchstrasse, ein Schwarzes Loch in ihrem Zentrum. Diese seltsamen Objekte sind eine der erstaunlichsten Voraussagen der Allgemeinen Relativitätstheorie. Sie können entstehen, wenn ein massiver Stern unter seiner eigenen Gravitation in sich zusammenfällt. Dann krümmt er den Raum in seiner Umgebung derart, dass selbst das Licht nicht mehr entweichen kann. Schwarze Löcher können dann auch weitere Sterne auffressen und zu Monstern mit Millionen von Sonnenmassen anwachsen. (Bild: ESO)

geringe Wechselwirkung zwischen Gravitationsstrahlung und Raum und Materie macht sie andererseits auch besonders attraktiv: Gravitationswellen werden praktisch nicht absorbiert oder gestreut, wenn sie das Weltall durchqueren.

Daher ist anzunehmen, dass Gravitationswellen tiefere Einblicke ins Universum erlauben als elektromagnetische Wellen. Beispielsweise zeigt die kosmische Hintergrundstrahlung das frühe Universum im Alter von 300'000 Jahren, als es sich auf etwa 3000 K abgekühlt hatte, so dass sich neutrale Atome bilden konnten und der Kosmos für elektromagnetische Strahlung durchlässig wurde. Noch früher war es von Plasma, das heisst elektrisch geladenen Teilchen, ausgefüllt und damit für elektromagnetische Strahlung undurchlässig. Im Gegensatz dazu ist dieses frühere Universum für Gravitationswellen durchsichtig. Die direkte Beobachtung von Gravitationswellen öffnet daher in Zukunft ein Fenster näher zum Urknall. Auch die unmittelbare Umgebung von schwarzen Löchern wird mittels Gravitationswellen der Beobachtung zugänglich sein, bei denen uns heisse, ionisierte Materie den Einblick mit Hilfe elektromagnetischer Strahlung verwehrt.

Die LISA-Mission

Die Weltraummission LISA (Laser Interferometer Space Antenna) ist ein gemeinsames Projekt von ESA und NASA, ein Observatorium zur Beobachtung von Gravitationswellen mit niedrigen Frequenzen oder, was dasselbe ist, Gravitationswellen mit grossen Wellenlängen. LISA nützt mit seinen 5 Millionen Kilometer langen Armen die Weite des Weltraums und ermöglicht damit die Beobachtung auch langwelliger Gravitationswellen. Fern von den lokalen Schwankungen der Gravitation auf der Erde infolge seismischer Aktivitäten, denen alle erdgebundenen Observatorien unterworfen sind, kann LISA Gravitationswellen im Bereich von 10^{-4} Hz bis 10^{-1} Hz feststellen.

Die Beobachtung niederfrequenter Gravitationswellen ermöglicht das Studium von astrophysikalischen Objekten, die über lange Zeit – von Monaten bis Äonen – detektierbare Gravitationswellen ausstrahlen. Für Beobachtung vom Weltraum aus ist der Himmel übersät von Objekten,

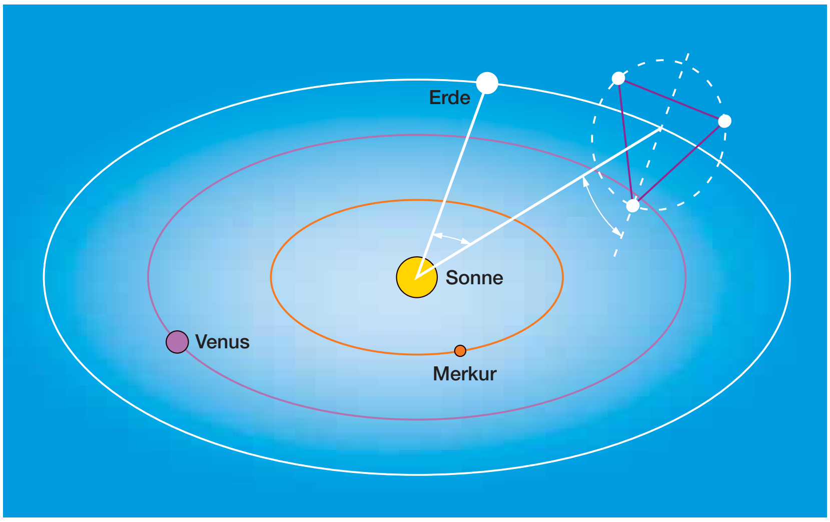


Bild 13 zeigt die Konfiguration der LISA-Mission. Sie besteht aus drei identischen Satelliten, die ein gleichseitiges Dreieck mit 5 Millionen Kilometern Seitenlänge bilden, das sich im Abstand von 50 Millionen Kilometer von der Erde mit ihr um die Sonne bewegt. Die Ebene des von den Satelliten aufgespannten Dreiecks ist gegenüber der Ebene der Erdumlaufbahn um 60° geneigt.

von denen wir aufgrund anderweitiger, klassischer astronomischer Beobachtungen wissen, dass sie niederfrequente Gravitationswellen aussenden. Dagegen sind die Quellen, deren höherfrequente Gravitationswellen vom Erdboden aus detektiert werden können, zum vornherein unbekannt und sie leuchten nur kurz auf. Damit ist die exakte Position dieser Quellen schwieriger zu bestimmen.

Das Messverfahren

LISA misst die Veränderung der räumlichen Dimensionen durch Gravitationswellen. Ihr Messprinzip beruht auf der laufenden, exakten Vermessung der gegenseitigen Lage der Testmassen in drei je 5 Millionen Kilometer entfernten Satelliten, siehe **Bild 13**. Es handelt sich dabei nicht um eine statische, sondern um eine dynamische Vermessung, bei der – anstatt Abweichungen von einer mittleren Distanz – Abweichungen von einer mittleren Frequenz gemessen werden. Frequenzen kann man äusserst genau messen.

Jeder Satellit besitzt zwei Teleskope, die auf die beiden andern Satelliten ausgerichtet sind, siehe **Bild 15**. Jedes Teleskop enthält je einen Laser-Sender und -Empfänger. Jedes Teleskop enthält zudem eine aus einer Gold-Platin-Legierung gefertigte würfelförmige Testmasse von etwa 4 cm Kantenlänge. Das Lichtsignal eines Lasersenders wird über das Teleskop auf einen der gegenüberliegenden

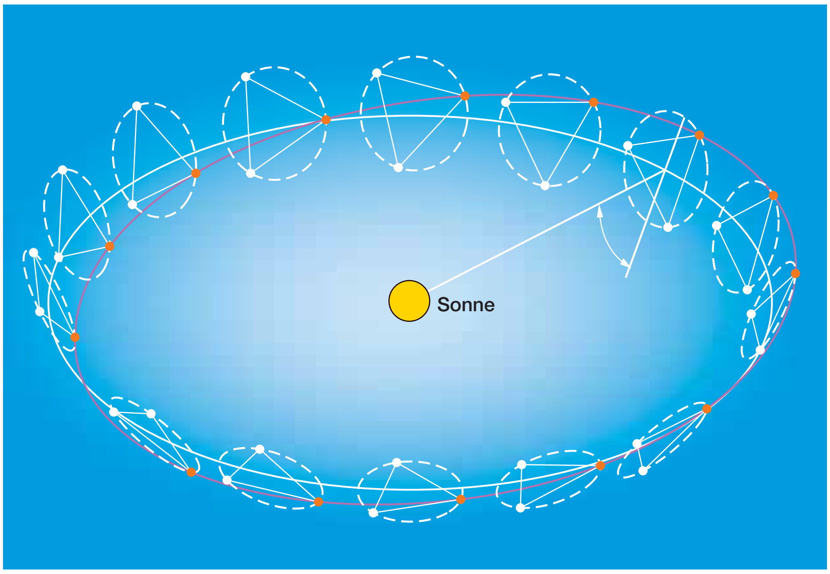


Bild 14 zeigt die Lage der drei LISA-Satelliten im Laufe eines Jahres. Sie befinden sich in natürlichen, heliozentrischen Umlaufbahnen, sind aber gegenüber der Ekliptik leicht geneigt, sodass sich ein bestimmter Satellit ein halbes Jahr über und ein halbes Jahr unter der Ekliptik befindet. Das von den Satelliten gebildete Dreieck dreht sich dabei im Laufe eines Jahres einmal um seinen Mittelpunkt. Da die Umlaufbahnen der Satelliten, wie diejenige der Erde, leicht exzentrisch sind, verändert sich auch der Abstand der Satelliten, und damit die Armlänge der Interferometer. Dies erleichtert die Messung tieffrequenter Gravitationswellen wesentlich.

Satelliten gerichtet. Dort befindet sich ebenfalls ein Teleskop, welches die eintreffende Strahlung sammelt und, nach Reflexion an der Testmasse, auf einen Lasertransponder abbildet. Dieser verstärkt das empfangene Signal und sendet es über das Teleskop zum ersten Satelliten zurück. Um die Veränderung des Raumes durch Gravitationswellen zu messen, wird nicht die gegenseitige Entfernung der Satelliten ermittelt, sondern deren zeitliche Veränderung. Jeder der drei LISA-Satelliten befindet sich in einer natürlichen, erdähnlichen Umlaufbahn um die Sonne. Da die Erdbahn nicht exakt kreisförmig, sondern leicht exzentrisch ist, ändert sich im Laufe der Zeit der Abstand der Erde und der Satelliten zur Sonne und, in geringerem Mass allerdings, auch der gegenseitige Abstand der Satelliten. Dies hat zur Folge, dass zwischen Sendesignal und Empfangssignal eine zeitlich rasch veränderliche Interferenz eintritt, die eine Grundfrequenz definiert. Verzerrt nun eine Gravitationswelle den Raum, so erzeugt sie kleine periodische Abweichungen von dieser Grundfrequenz, das heisst, sie moduliert die Grundfrequenz. Solche Modulationen sind leicht von der Grundfrequenz zu trennen: Damit kann sowohl die Amplitude wie auch die Frequenz der Gravitationswelle sehr genau gemessen werden.

Bild 15 zeigt einen der drei LISA-Satelliten. Diese bestehen aus einem gleichseitigen Y, in dessen beiden Armen je ein Teleskop mit seinen Testmassen und je ein Laser-Sender- und -Empfänger untergebracht sind. Im unteren Teil des Y befinden sich die Bordelektronik, die Stromversorgung und die Telemetrieelektronik.

Die Satelliten

Ähnlich wie bei der Microscope-Mission müssen auch bei LISA Vorkehrungen getroffen werden, um die Testmassen gegen äussere Störungen abzuschirmen, sodass letztlich nur noch gravitationelle Kräfte auf sie einwirken. Dazu werden kleine Triebwerke eingesetzt, welche eine aktive Korrektur der Flugbahnen der Satelliten so ermöglichen, dass die auf sie einwirkenden nicht-gravitationellen Effekte, wie zum Beispiel der Lichtdruck der Sonnenstrahlung, kompensiert werden. Diese Triebwerke müssen Schübe im Bereich von einigen μN (ein Zehnmillionstel eines Gramms) erzeugen. Um diese Triebwerke und das dazu gehörende Navigationssystem im Weltraum selbst zu testen, sieht das Programm eine Vorläufermission (LISA Pathfinder) vor. Erst anschliessend wird LISA entwickelt und voraussichtlich im Jahre 2013 von einer amerikanischen Trägerrakete in ihre Umlaufbahnen gebracht werden.

Die Gravity Probe B

In der Newton'schen Physik sind Raum und Zeit unabhängige Grössen. In relativistischer Sicht sind dagegen Raum und Zeit untrennbar miteinander verknüpft. Nach der Allgemeinen Relativitätstheorie krümmt die Masse die Raumzeit. In einem zweidimensionalen Modell führt eine Masse, ein Stern beispielsweise, zu einer Delle in der Raumzeit, die um so tiefer ist, je höher die Konzentration der Masse ist, **siehe**

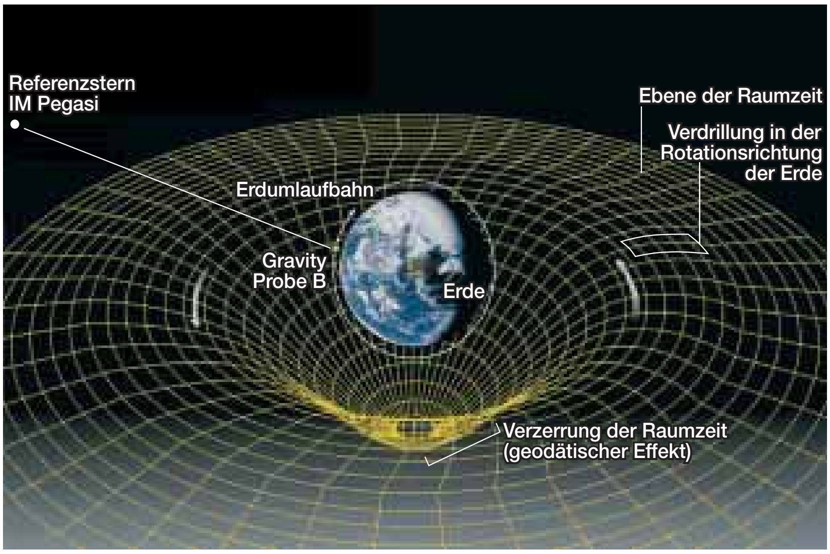


Bild 16 veranschaulicht die Krümmung der (hier zweidimensional angenommenen) Raumzeit durch die Erde. In der relativistischen Vorstellung von Lense-Thirring krümmt die Masse der Erde einerseits die Raumzeit in ihrem Umfeld und führt zum geodätischen Effekt, andererseits verursacht ihre Rotation eine Verdrillung der Raumzeit (frame dragging). Die quantitative Bestimmung dieser beiden Effekte ist das Ziel der Gravity Probe B Mission.

Bild 15. Nun haben Lense[18] und Thirring[19] wenige Jahre nach der Publikation der Allgemeinen Relativitätstheorie gezeigt, dass eine Masse nicht nur zu einer Krümmung der Raumzeit führt, sondern diese auch mit sich zieht (verdrillt), wenn sie rotiert. Dieser Effekt wird als Lense-Thirring-Effekt bezeichnet.

Der Lense-Thirring-Effekt zählt, wie die Gravitationswellen, zu den bisher experimentell noch nicht verifizierten Voraussagen der Relativitätstheorie, was ihn natürlich besonders attraktiv macht. Die Gravity Probe B der NASA wurde 2004 in eine Erdumlaufbahn gebracht, um diesen subtilen Effekt experimentell nachzuweisen.

Das Messprinzip

Die Aufgabe besteht in der exakten Bestimmung der lokalen Ausrichtung der Raumzeit auf der Umlaufbahn des Satelliten um die Erde. Als Sensoren eignen sich Kreisel, denn ein rotierender Kreisel hat die Eigenschaft, dass er ohne äussere Einflüsse seine Drehrichtung nicht ändert, solange er rotiert.

Die Gravity Probe B besteht daher aus vier hochpräzisen Kreiseln, deren exakte Ausrichtung im Universum laufend zu bestimmen ist. Dazu besitzt der Satellit ein Teleskop, das auf einen weit entfernten Stern als Referenz ausgerichtet ist. Wenn der Satellit im Orbit ist, werden die Kreisel in Rotation gebracht und ihre Drehrichtung mit der Richtung zum Referenzstern verglichen. Im Verlauf der etwa ein Jahr dauernden Messung wird die Drehrichtung der Kreisel

Der Lense-Thirring-Effekt

Schon kurz nach der Publikation der Allgemeinen Relativitätstheorie durch Albert Einstein fanden die beiden österreichischen Physiker Lense und Thirring einen relativistischen Effekt, der im Englischen als «frame dragging effect» bezeichnet wird. Auf Deutsch heisst er Lense-Thirring-Effekt.

Dieser bezeichnet das Phänomen, dass eine rotierende Masse die Raumzeit in ihrer Umgebung mitrotieren lässt. Man kann sich dies so vorstellen, dass eine sich drehende Kugel in zähflüssigen Honig geworfen wird. Dabei nimmt die Kugel bei ihrer Drehung den sie umgebenden Honig teilweise mit, und zwar je stärker, je näher zur Kugel sich der Honig befindet.

[18] Joseph Lense, 1890 Wien bis 1985 München; Physiker
[19] Hans Thirring, 1888 Wien bis 1976 Wien; Physiker

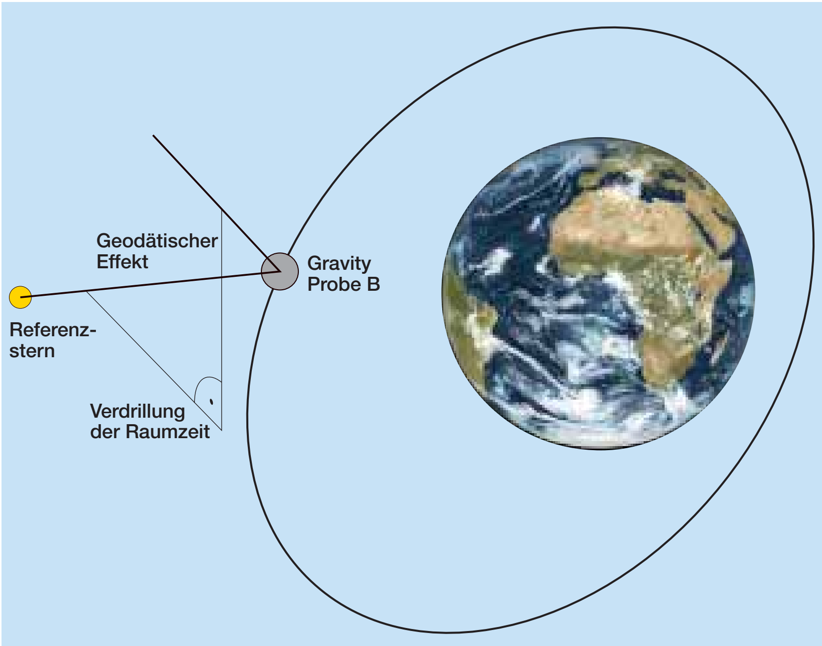


Bild 17: Das Konzept der Gravity Probe B. Der Satellit umkreist die Erde. Er besitzt Kreisel, welche die Eigenschaft haben, laufend der Richtung der lokalen Raumzeit zu folgen. Der Stern IM Pegasi dient als räumliche Referenz für die Messung. Gemäss der Allgemeinen Relativitätstheorie ist eine Drift der Kreisel bezogen auf die Achse zum Referenzstern von 6,6 Bogensekunden pro Jahr in Richtung des Nordpols zu erwarten (geodätischer Effekt). Der Lense-Thirring-Effekt seinerseits verursacht eine zusätzliche Drift in azimutaler Richtung von 0,042 Bogensekunden pro Jahr.

laufend beobachtet: Wenn die lokale Raumzeit durch die Gravitation der Erde verformt ist und wenn jene infolge der Erdrotation auf Grund des Lense-Thirring-Effektes verdrillt ist, dann muss sich im Verlauf der Zeit eine Differenz zur anfänglichen Richtung der Kreisel bezogen auf den Referenzstern ergeben. Aus dieser Differenz lassen sich die lokale Krümmung sowie das Ausmass der Verdrillung der Raumzeit ermitteln. Die zu messenden Differenzen sind allerdings äusserst klein: Gemäss der Allgemeinen Relativitätstheorie werden sich die Kreisel infolge der durch die Masse der Erde gekrümmten Raumzeit um 6,6 Bogensekunden[20] pro Jahr und auf Grund des Lense-Thirring-Effektes um 42 Tausendstel einer Bogensekunde pro Jahr ändern. Dies entspricht einer vollen Umdrehung in 30 Millionen Jahren! Das Experiment ist so konzipiert, dass die erwarteten Verschiebungen rechtwinklig zueinander stehen, sodass sie experimentell getrennt bestimmt werden können.

Der Satellit

Der Satellit Gravity Probe B umkreist derzeit die Erde auf einer Höhe von 600 km. Er besitzt aus Gründen der Redundanz vier hochpräzise Kreisel. Diese sind – wie bei den bisher besprochenen Satelliten – vor externen Einflüssen geschützt. Bei den Kreiseln muss zudem sichergestellt sein, dass die Masseverteilung so homogen ist, dass das Gravitationsfeld der Erde nicht zu einer Präzession führt. Die Kreisel sind daher nahezu perfekte Kugeln von höchster Homogenität.

Weil sich der Stern, auf den das Teleskop der Gravity Probe B gerichtet ist, am Himmel leicht bewegt, muss auch diese Bewegung gegenüber weit entfernten Objekten geringer räumlicher Ausdehnung (sogenannten Quasaren[21]) auf Tausendstel einer Bogensekunde genau vermessen werden. Erst nachdem sowohl die Messdaten des Satelliten als auch die Beobachtungen des Referenzsterns unabhängig voneinander ausgewertet sind, können die Resultate miteinander verglichen und die wissenschaftlichen Erkenntnisse bekannt gegeben werden, was etwa im Spätsommer 2006 der Fall sein dürfte.

Bild 18: Die gewaltigen Spiralarme der Galaxie M101 haben einen Durchmesser von 170'000 Lichtjahren. Nach derzeitigem Wissen hat eine nahegelegene Galaxie zu Wellen von kondensiertem Gas geführt, welche um das Zentrum der M101 kreisen. Dabei wird das vorhandene Gas komprimiert, was zur Entstehung neuer Sterne führt.

[20] Eine Bogensekunde entspricht etwa dem Winkel, den ein menschliches Haar in 20 Metern Entfernung einnimmt.
[21] Quasare sind sehr leuchtstarke und meist weit entfernte aktive Galaxien.

Die Zeit nach der Allgemeinen Relativitätstheorie

Wenn Albert Einstein mit der Veröffentlichung seiner epochalen Erkenntnisse zwar in wissenschaftlichen Kreisen zu Berühmtheit gelangt ist, so verschafft ihm doch erst die Bestätigung der vorausgesagten Lichtablenkung bei der Sonnenfinsternis von 1919 den Durchbruch in der breiten Öffentlichkeit. Die Beobachtungen von Sir Arthur Eddington von 1919 bestätigen die Voraussagen der Allgemeinen Relativitätstheorie, und die Nachricht geht wie ein Lauffeuer über die ganze Welt. Die Berliner Illustrierte Zeitung beispielsweise kommentiert dieses Ereignis auf ihrer Titelseite mit «Eine neue Grösse der Weltgeschichte: Albert Einstein, dessen Forschungen eine völlige Umwälzung unserer Naturbetrachtung bedeuten und den Erkenntnissen eines Kopernikus, Kepler und Newton gleichwertig sind». Damit ist der Höhepunkt seiner Anerkennung in Deutschland erreicht, doch als Jude und überzeugter Pazifist warnt er vor dem aufkommenden Nationalsozialismus und wird politisch zunehmend isoliert. Bei Einsteins Rückkehr von einer Reise in die USA kommt 1933 Hitler an die Macht. Einstein kehrt nicht mehr nach Berlin zurück, gibt aus Protest seinen deutschen Pass ab und verreist unverzüglich wieder nach Princeton, USA, wo er die zweite Hälfte seines Lebens verbringt.

« Meine Religiosität besteht in der demütigen Bewunderung des unendlich überlegenen Geistes, der sich in dem wenigen offenbart, was wir mit unserer schwachen und hinfälligen Vernunft von der Wirklichkeit zu erkennen vermögen. »

Auch im amerikanischen Exil engagiert sich Albert Einstein politisch. Er, der Pazifist, rät Präsident Roosevelt zur Entwicklung der Atombombe, um darin den Deutschen zuvorzukommen, ein Schritt, den er später aufs Tiefste bereut. (Ob Roosevelt den Brief von Einstein allerdings jemals gesehen hat, wird heute bezweifelt.)

In wissenschaftlicher Hinsicht ist 1916 mit der Publikation der Allgemeinen Relativitätstheorie der Zenith erreicht. Zwar hat Einstein selber die Grundlagen der Quantenmechanik geschaffen, die nun von einer neuen Generation von Physikern weiterentwickelt wird, doch er kann sich mit der von ihr postulierten Zufälligkeit nicht mehr anfreunden: «Gott würfelt nicht», sagt er. Es kann nicht sein, wie der junge Heisenberg[22] postuliert, dass sich Position und Geschwindigkeit eines Elementarteilchens nicht exakt bestimmen lassen, sondern nur deren Wahrscheinlichkeiten. In dieser Haltung erweist sich Albert Einstein als der letzte der grossen klassischen Physiker.

Einen wesentlichen Teil seiner zweiten Lebenshälfte widmet Albert Einstein dem Versuch, die Relativitätstheorie und die Quantentheorie zu einer einheitlichen Theorie zusammenzufassen. Seine Bemühungen

[22] Werner Karl Heisenberg, 1901 Würzburg bis 1976 München; Physiker, Nobelpreis für Physik 1932

bleiben vergeblich, die Zeit ist noch nicht reif dafür. Sie ist es auch heute noch nicht: Noch immer suchen die Physiker intensiv nach der *Grand Unified Theory*. Dennoch hat Einstein fundamentale Beiträge zum Verständnis des Universums geleistet.

In seinem mathematischen Modell hat Albert Einstein die berühmt gewordene kosmologische Konstante Λ eingeführt, um das dem damaligen Paradigma entsprechende stationäre Universum zu beschreiben. Doch im Jahre 1928 entdeckt Edwin Hubble[23], dass sich die Galaxien von uns entfernen, und zwar umso schneller, je weiter sie von uns weg sind: das Universum befindet sich also doch in einem dynamischen Zustand. Unter dem Eindruck dieser überraschenden Neuentdeckung glaubt Einstein, die Einführung der kosmologischen Konstanten sei die «grösste Eselei seines Lebens». Neueste Messungen, unter anderem mit dem Hubble Space Telescope, legen jedoch den Schluss nahe, dass sich das Universum sogar in einer beschleunigten Expansion befindet, was mit dem kosmologischen Modell nur beschrieben werden kann, wenn man die verschmähte kosmologische Konstante wieder einführt…

« Das Unverständlichste am Universum ist im Grunde, dass wir es verstehen. »

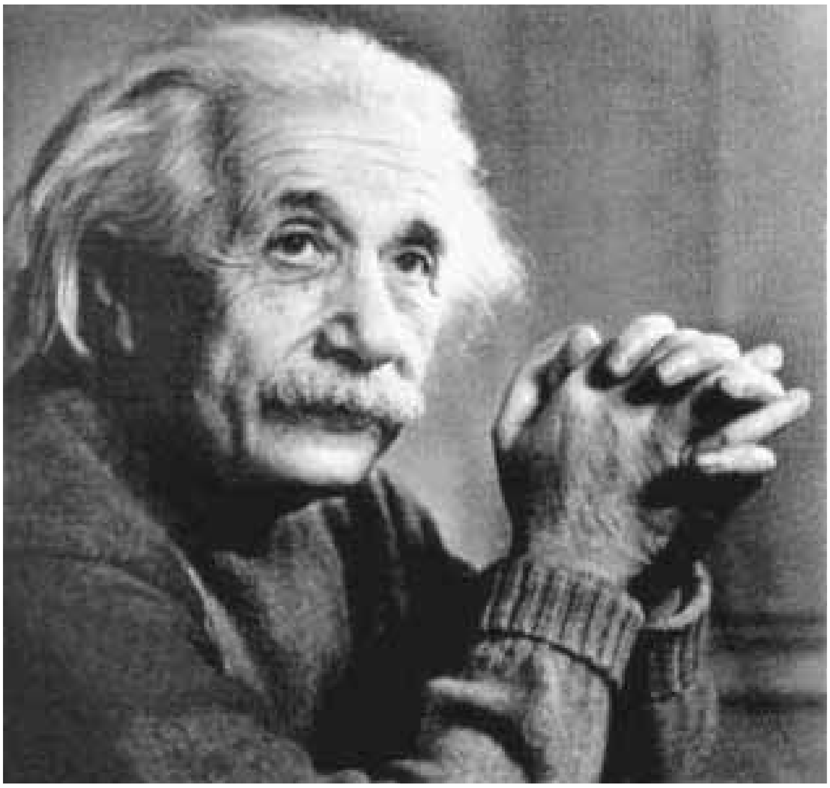


Bild 19: Albert Einstein (1954)

Am 13. April 1955 überfallen den 76-Jährigen heftige Schmerzen. Aber er lehnt die von den Ärzten dringend empfohlene Operation ab. Am 18. April gegen 1 Uhr nachts schliesst Albert Einstein seine Augen für immer. Eine Krankenschwester ist bei ihm. Sie hört seine letzten Worte, doch sie kann kein Deutsch. So verhallen seine Worte unverstanden. «The last words of the intellectual giant were lost in the world», berichtet die New York Times am Tag darauf.

23 Edwin Powell Hubble, 1889 Marshfield bis 1953 San Marino; Jurist und Astronom

Der Autor

U. Markus, Zürich

Martin C. E. Huber ist Bürger der Stadt Basel, wo er die Primarschule und das Humanistische Gymnasium besuchte. Er studierte zunächst an der Eidgenössischen Technischen Hochschule (ETH) in Zürich und diplomierte mit einer Arbeit in experimenteller Festkörperphysik; anschliessend kehrte er nach Basel zurück und schloss sein Universitätsstudium 1963 bei Prof. Miescher mit einer Dissertation in Molekülspektroskopie ab. Nach einem weiteren Jahr in Basel, in dem er den wohl ersten Moleküllaser der Schweiz baute, erhielt er ein Stipendium der European Space Research Organisation (ESRO) für einen Forschungsaufenthalt am Harvard-Smithsonian Center for Astrophysics in Cambridge (Massachusetts, USA). Vorerst arbeitete er dort in Laboratoriumsastrophysik, d. h. an der Messung von astrophysikalisch relevanten Eigenschaften von Atomen. Er besorgte auch die radiometrische Kalibrierung der Harvard-Instrumente auf dem Orbiting Solar Observatory, OSO-6, und dem Apollo Telescope Mount auf der damaligen Weltraumstation Skylab.

Gegen Mitte der 70er-Jahre kehrte M. Huber an die ETH Zürich zurück, wo er weiterhin spektroskopische Arbeiten durchführte und effiziente optische Systeme für Sonnenbeobachtungen im Weltraum entwickelte. Dort erwarb er auch die *venia legendi* und erhielt später den Titel eines Professors.

Als Vorsitzender der Solar System Working Group der Europäischen Weltraumorganisation ESA beteiligte er sich in den 80er-Jahren an der Vorbereitung des langfristigen ESA-Wissenschaftsprogramms Horizont 2000. Bei der gemeinsamen ESA / NASA-Mission Solar and Heliospheric Observatory SOHO ist er Co-Investigator bei drei Experimenten.

Von 1987 bis 2001 arbeitete Martin Huber in der ESA als Leiter des Wissenschaftsdepartements und als Wissenschaftsberater. Danach war er gelegentlich als Gast am Smithsonian Astrophysical Observatory in Cambridge, Mass., und am ISSI tätig. Seit 2003 ist er regelmässiger Gast des Laboratoriums für Astrophysik am Paul Scherrer Institut in Villigen. Martin Huber ist Mitherausgeber des 2001 erschienenen umfassenden Werkes The Century of Space Science und Redaktor der Zeitschrift The Astronomy and Astrophysics Review.

Seine Wahl zum Präsidenten der Europäischen Physikalischen Gesellschaft EPS, und insbesondere das von der EPS initiierte und von der UNESCO wie auch den Vereinten Nationen unterstützte Weltjahr der Physik, gibt Martin Huber zahlreiche Gelegenheiten, der Öffentlichkeit nicht nur die technologische, sondern auch die kulturelle Bedeutung der Physik näher zu bringen. Es ist sein Wunsch, dass vor allem die jüngere Generation die Bedeutung der Naturwissenschaften wieder vermehrt anerkennt; kurz gesagt: er hofft, dass ein Bekenntnis vollkommener Ignoranz in physikalischen Dingen in naher Zukunft absolut *uncool* sein wird.

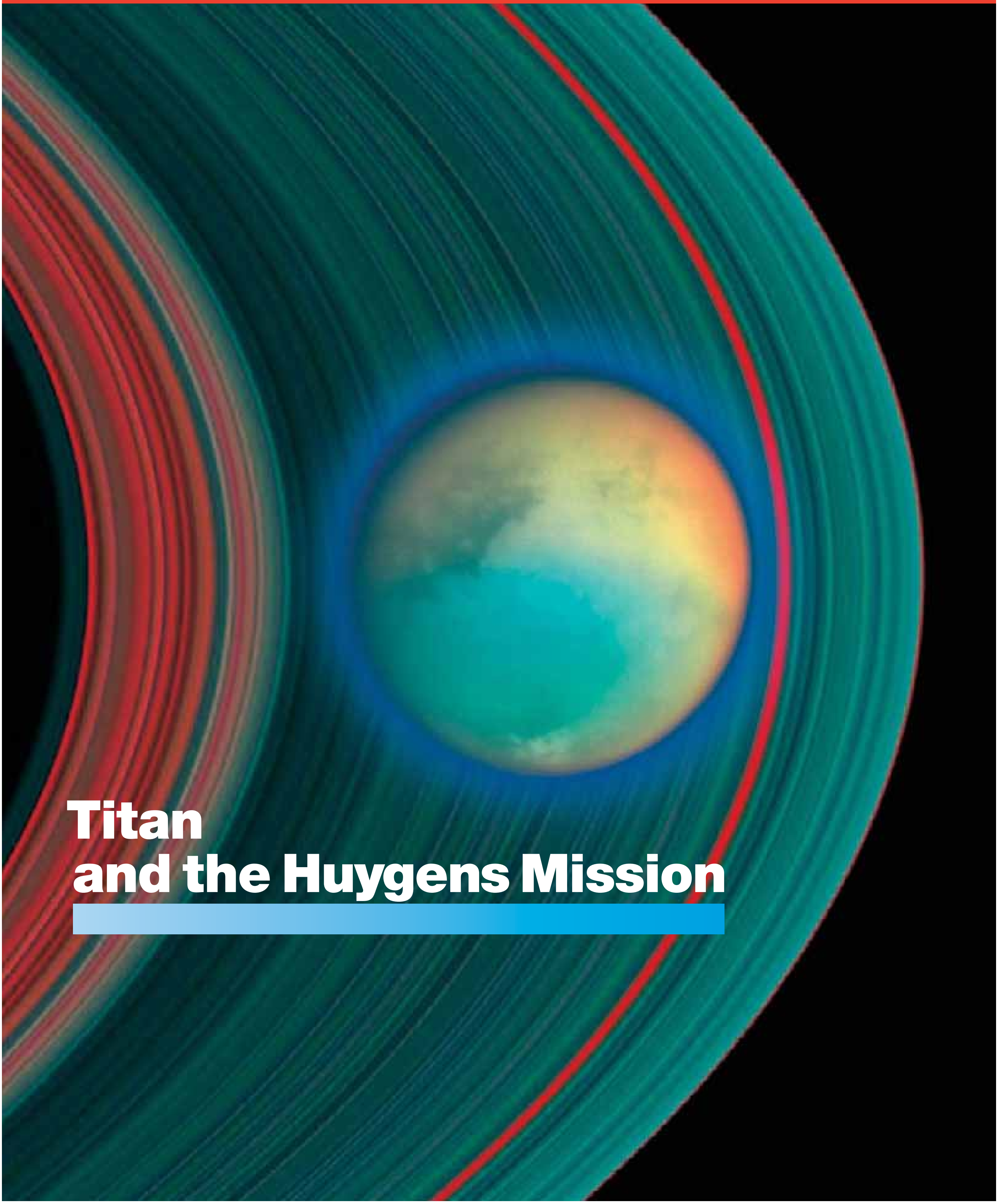
Als Amateur spielt Martin Huber in einem Klaviertrio – in dem nunmehr seit über 40 Jahren bestehenden Trio de Satigny – den Violoncellopart.

INTERNATIONAL SPACE SCIENCE INSTITUTE

SPATIUM

Published by the Association Pro ISSI

No 15, November 2005



Titan and the Huygens Mission

Editorial

Impressum

INTERNATIONAL SPACE SCIENCE INSTITUTE

Titan and the Huygens Mission[1]

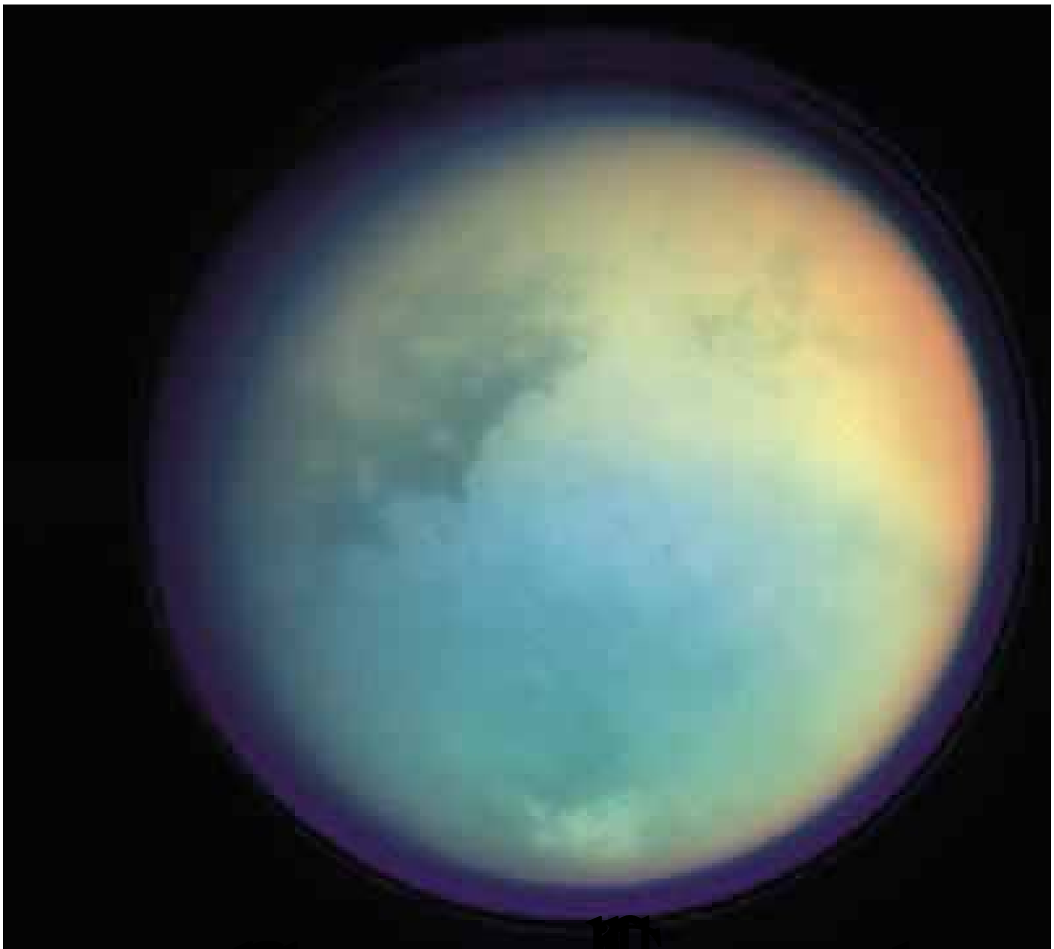
Introduction

[illegible]

Figure 1: A [illegible]lf-made telescope enabled the [illegible]h scientist Christiaan Huygens [illegible]

Titan, the Mysterious Moon

(figure



close-up view of Titan

	Earth	Titan
Atmospheric composition	78% N_2, 21% O_2, <1% H_2O	98% N_2, 2%CH_4
Surface pressure	1 bar	1.5 bar
Atmospheric thickness (top of stratosphere)	50 km	300 km
Clouds/rain	H_2O	CH_4
Radius	6,371 km	2,575 km
Axial tilt	23.5°	26.7°
Distance from Sun	1 AU	9.5 AU
Solar irradiance	1,368 W/m^2	15 W/m^2

The Objectives for the Huygens Mission

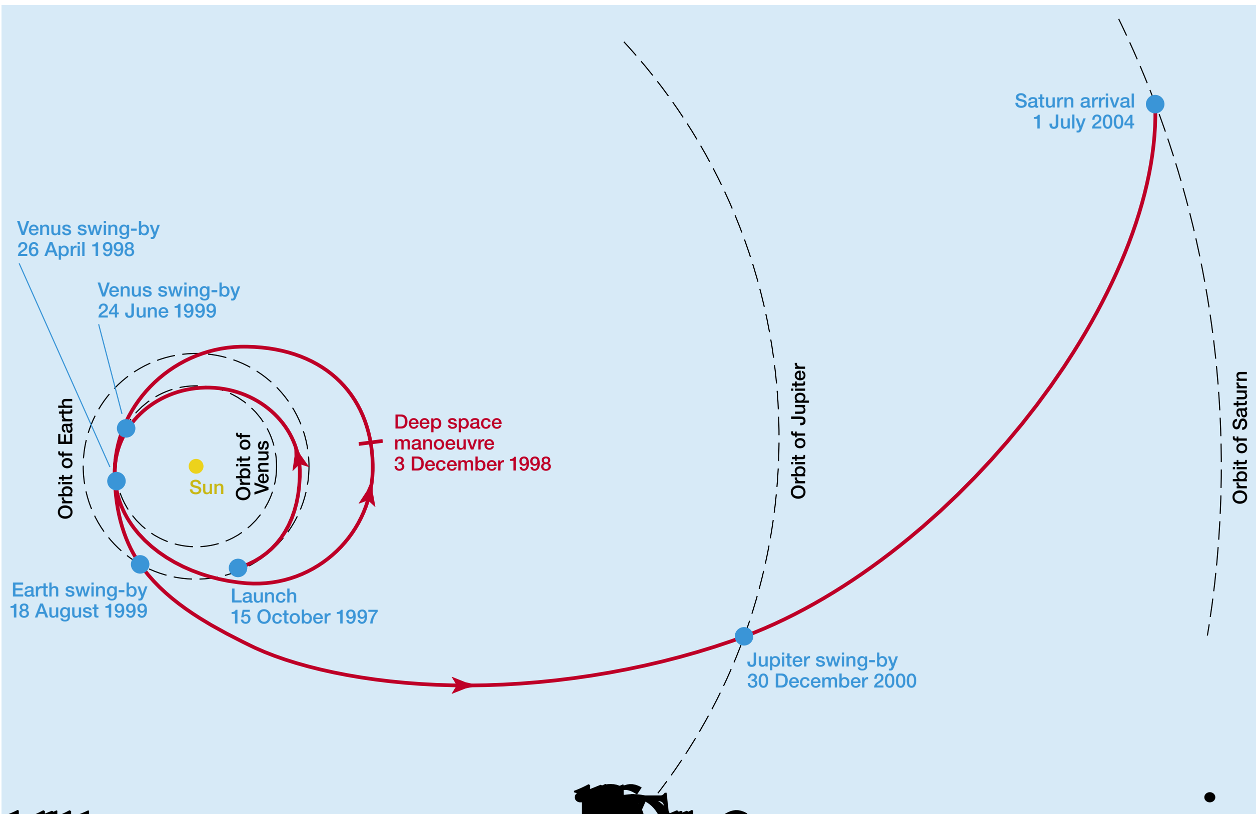
The Surface

The Interior

The Atmosphere

The Cassini-Huygens Mission

The Challenges



The complex trajectory of Cassini-Huygens

The Technical Solutions

[illegible] Huygens Spacecraft

4 The Huygens probe

(figure 5).

Installation … the Multi-Layer Insulation (MLI) blank…

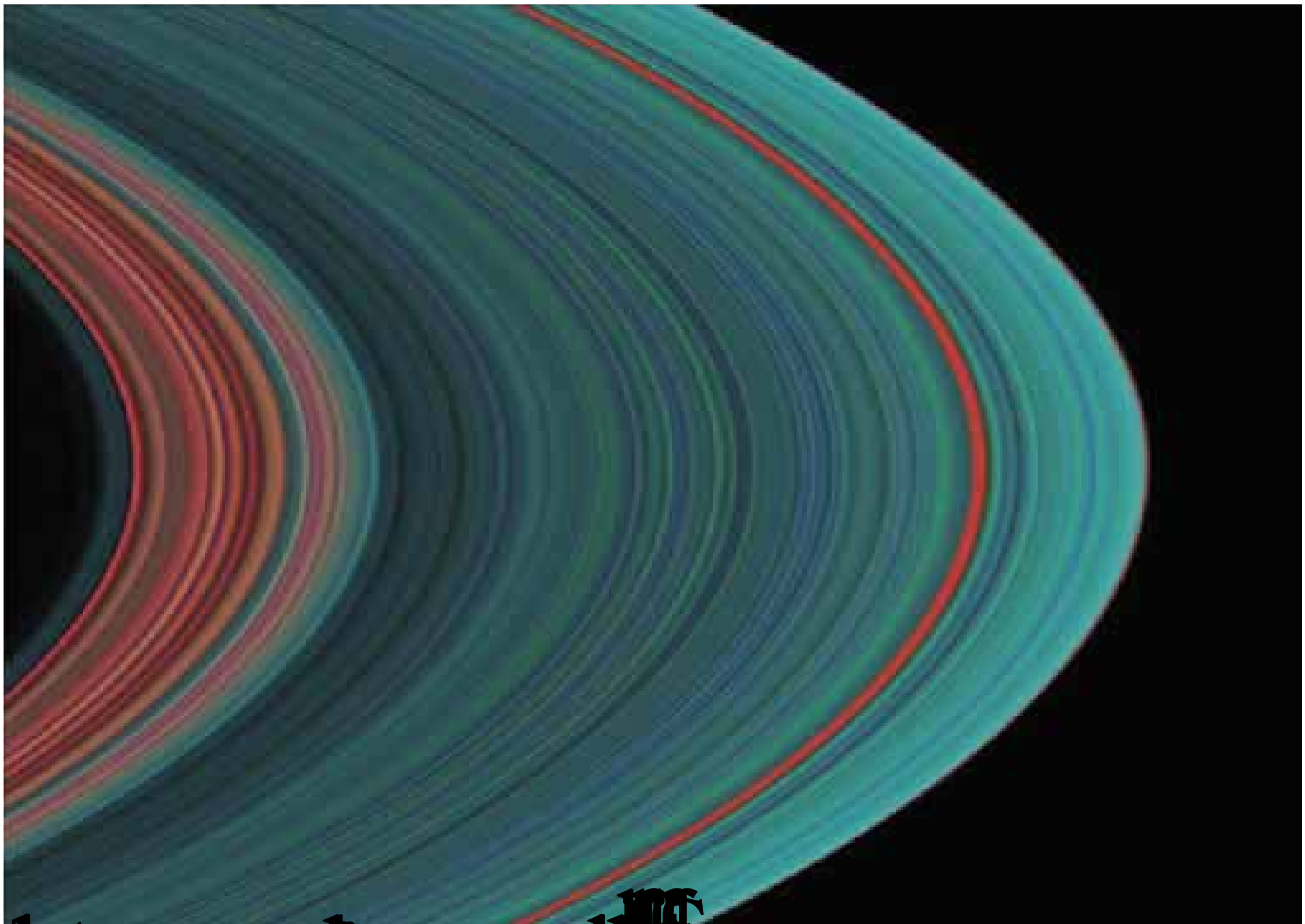
A demanding mission like Huygens-Cassini

see figure 7.

see fig-

The trajectory

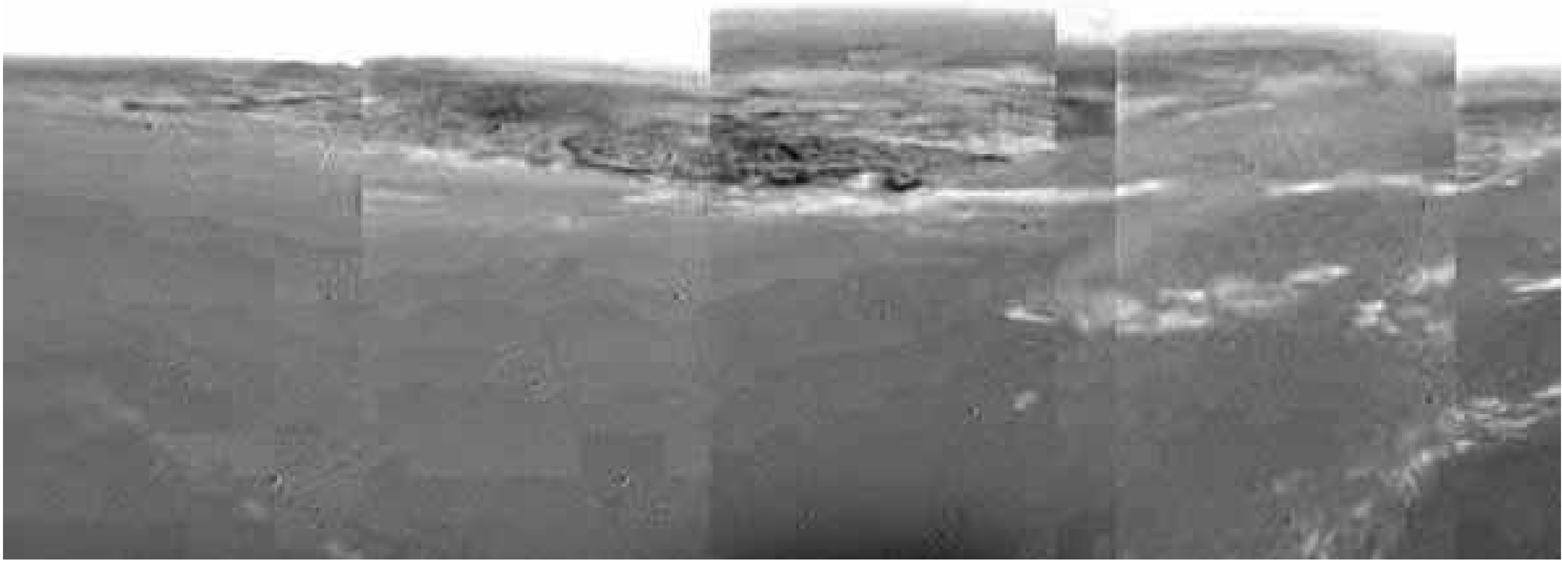
figure 6.



[illegible]: *False colours help to* [illegible]*ght subtle differen*[illegible]

…rtist's concept of the Cassini-Huygens orbiter

(figure 9).



F[illegible]: The Huyge[illegible] [illegible]obe during t[illegible]ase of entry.

Figure 10 shows the vi[illegible] [illegible]f [illegible]uyges f[illegible] height of 30 km.

Titan: the Results so Far

The interior

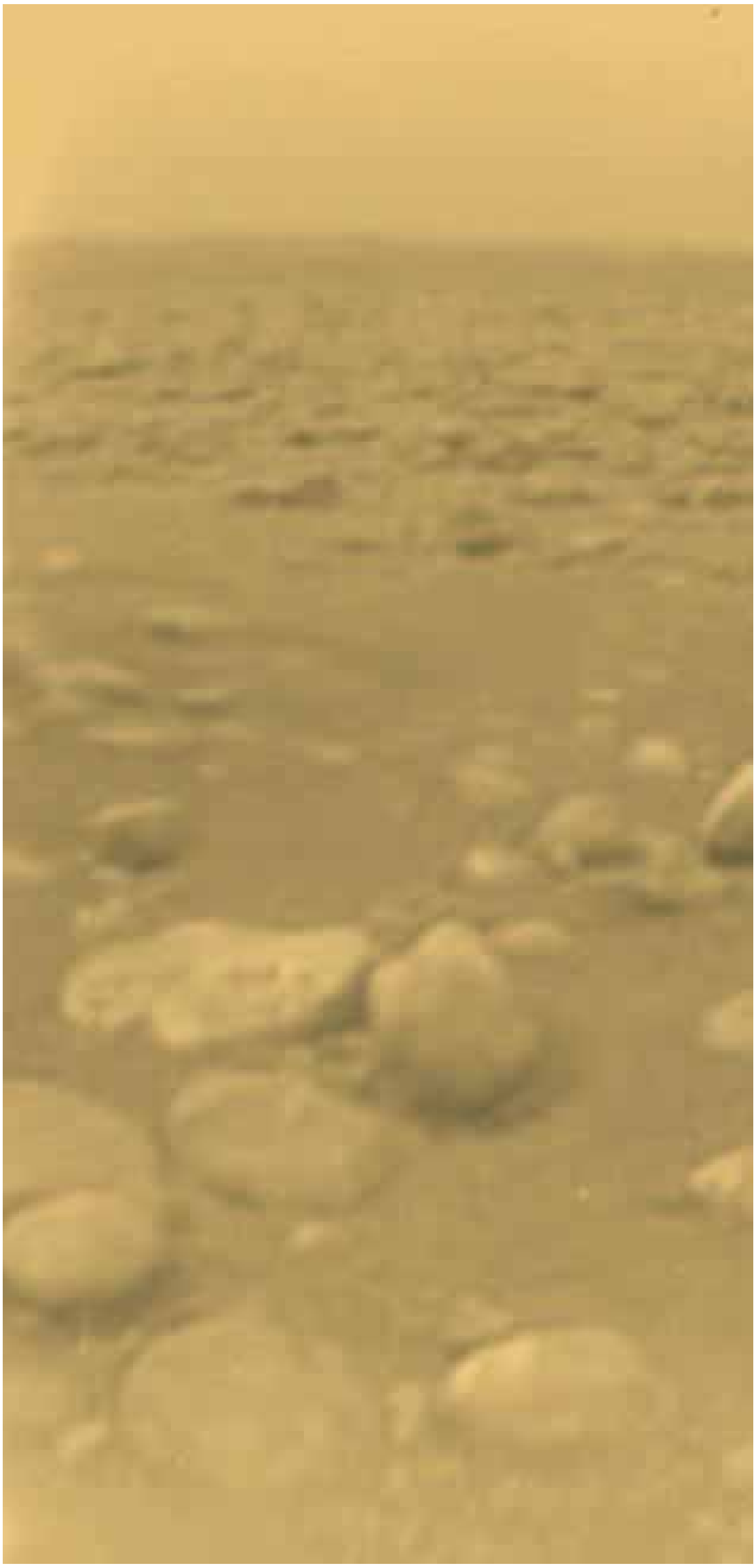


Figure 1: The first s… recorded by t… …gen's probe

The Surface

Impact Craters

Figure 12: The dark patch

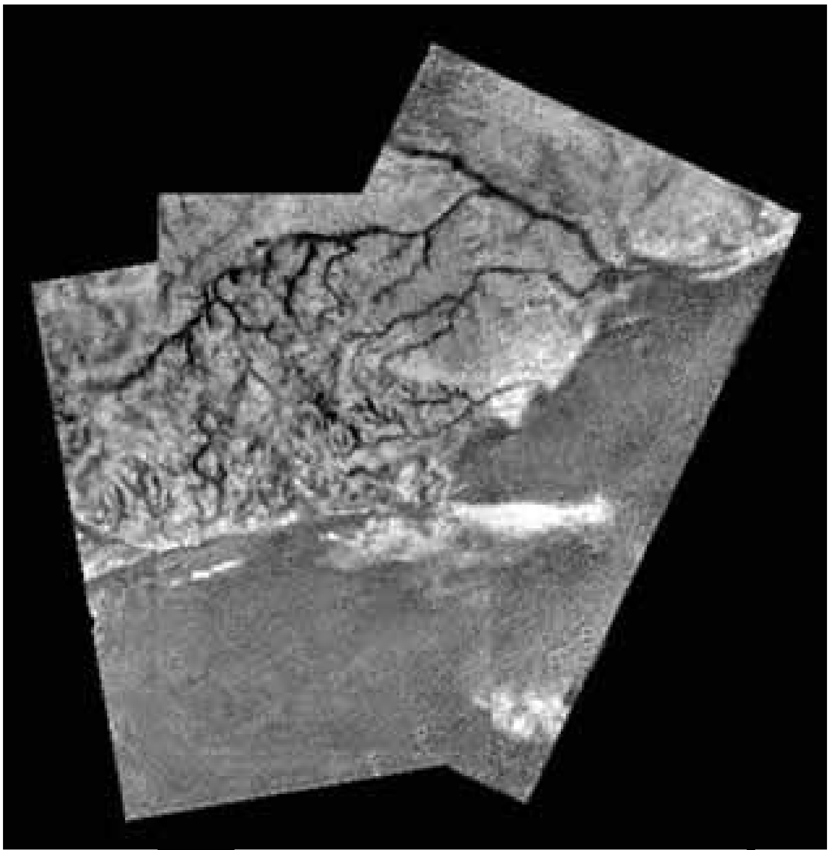
Liquid Surfaces and Volcanoes

see figure 12.

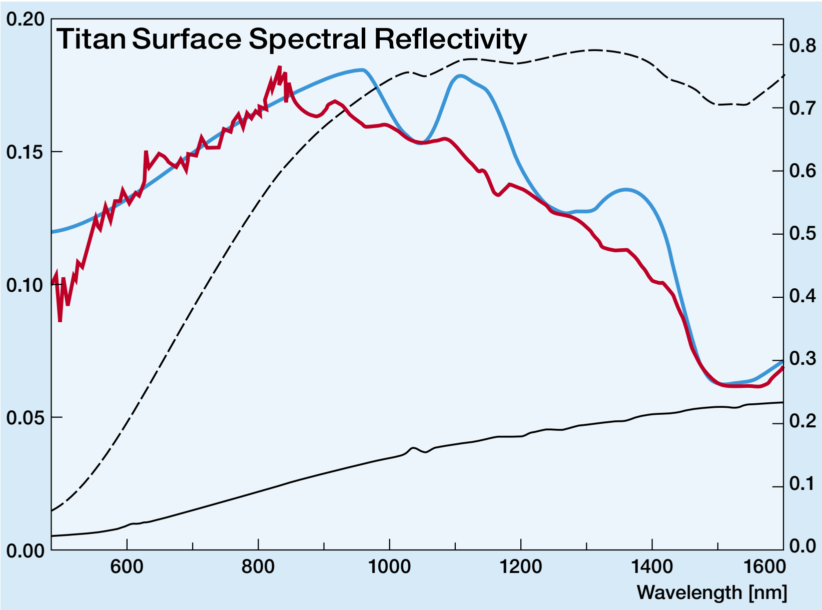
Dunes, Tectonic Faults and Fluvial Systems

see figure 13.

...ectivity



F... 13: Met... ...uds and rain, evaporating lakes, flowin... ...ers, a... ...ater i... ...anoes



... shows the surface reflectivity ...

The Atmosphere

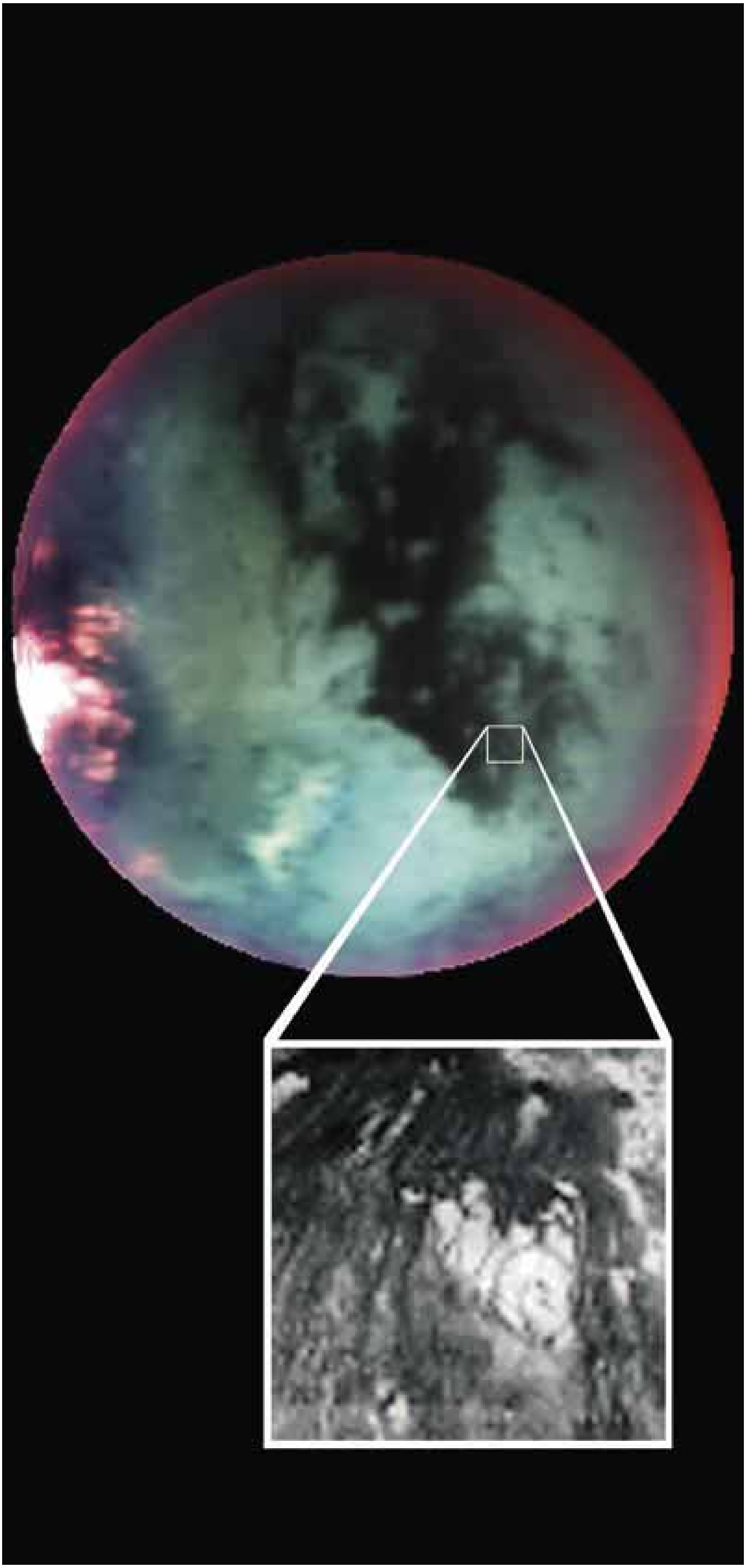
... Composition

see figure 15.

Clouds and Atmospheric Dynamics

[illegible]: This image taken by the Cassini orbiter [illegible]

[illegible]: The domed feature [illegible]



Outlook

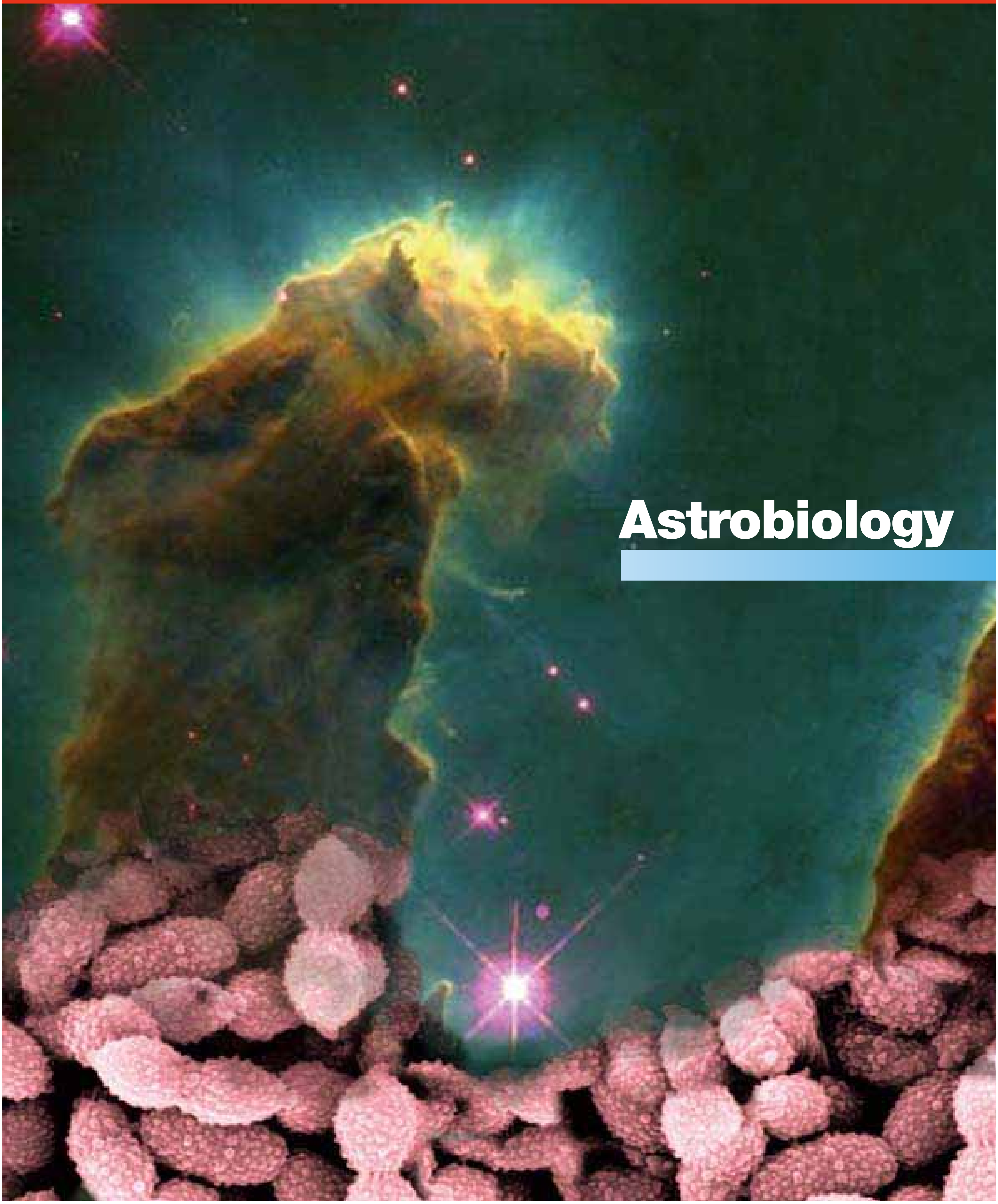
The Author

INTERNATIONAL
SPACE
SCIENCE
INSTITUTE

SPATIUM

Published by the Association Pro ISSI

No 16, July 2006



Astrobiology

Editorial

Are we the proud subjects of our will or the wretched victims of our drives? Charles Darwin, the British clergyman and naturalist, found the origin of the species – including men – in the continued natural selection process ruling the randomly mutating gene pools of living systems. In other words, the variety of species – including men – is seen as the result of a simple albeit very efficient trial and error process adapting the species to different and ever changing environmental conditions. Are we nothing else than the output of a soulless evolution machinery?

Sigmund Freud, the Austrian psychiatrist, identified in us an ego struggling for a niche between the super-ego coined by the authorities we experienced in our youth and the id, the instinctual heritage from our distant ancestors. This is good news, as the ego is, at least in principle, able to anticipate the future and to act responsibly in the here and now, while the evolution process is strictly governed by the conditions prevailing in the past. If we take, however, the number of species living on this planet as the measure of success, then we find no much remains of the proud responsible ego: the rate of extinction caused by human activities is estimated to be similar that of the major mass extinctions in the Earth's long history.

Sometimes we have to go far to come close. Astrobiology tells us what an isolated and outstanding value life represents in the universe: against all efforts so far, no traces of life have been found beyond the Earth. Oliver Botta, the young scientist at ISSI, presented the fascinating quest for life in the universe in the frame of the Pro ISSI lecture on 25 October 2005. We are indebted for his kind permission to submit our readers herewith a revised version of his talk. May the present issue of Spatium on astrobiology enhance our awareness of the unique value of life here on Earth!

Hansjörg Schlaepfer
Zürich, June 2005

Impressum

SPATIUM
Published by the
Association Pro ISSI
twice a year

INTERNATIONAL SPACE SCIENCE INSTITUTE

Association Pro ISSI
Hallerstrasse 6, CH-3012 Bern
Phone +41 (0)31 631 48 96
see
www.issi.unibe.ch/pro-issi.html
for the whole Spatium series

President
Prof. Heinrich Leutwyler,
University of Bern
Publisher
Dr. Hansjörg Schlaepfer,
legenda schläpfer wort & bild,
CH-8173 Neerach
Layout
Marcel Künzi, marketing · kommunikation, CH-8483 Kollbrunn
Printing
Schellenberg Druck AG
CH-8330 Pfäffikon

Front Cover: An image of the Eagle nebula (Credit: NASA) is combined in this artist's impression with a colony of bacteria.

Astrobiology[1]

Dr. Oliver Botta, NASA Goddard Space Flight Center and International Space Science Institute

Introduction

Natural science has made overwhelming progress in the past few centuries: Galilei, Newton, Einstein are among the names standing for quantum leaps in our knowledge of the physical world. Yet, we are far from understanding one of the most common phenomena here on Earth: life. Charles Darwin's[2] concept of evolution based on random mutation of genetic properties and selection by the environment was a major leap forward **(Fig. 1)**. Watson[3] and Crick[4] found the double helix structure of the deoxyribonucleic acid (DNA) helping to understand these processes on the molecular level. But despite all this progress, the most fundamental questions remain unanswered: What is life? How did life emerge on Earth? What are the conditions for life to emerge? How did the environment influence the evolution of life on Earth and how did life in turn influence the environment? These and many other questions are open: the stage is set for astrobiology. It requires an interdisciplinary approach with contributions from all branches of science to unravel the answers to these questions. Astrobiology (sometimes also called xenobiology or exobiology) is the attempt to bring together scientists to combine their knowledge with a twofold goal: to understand the origin of life on Earth and to search for life outside of the Earth.

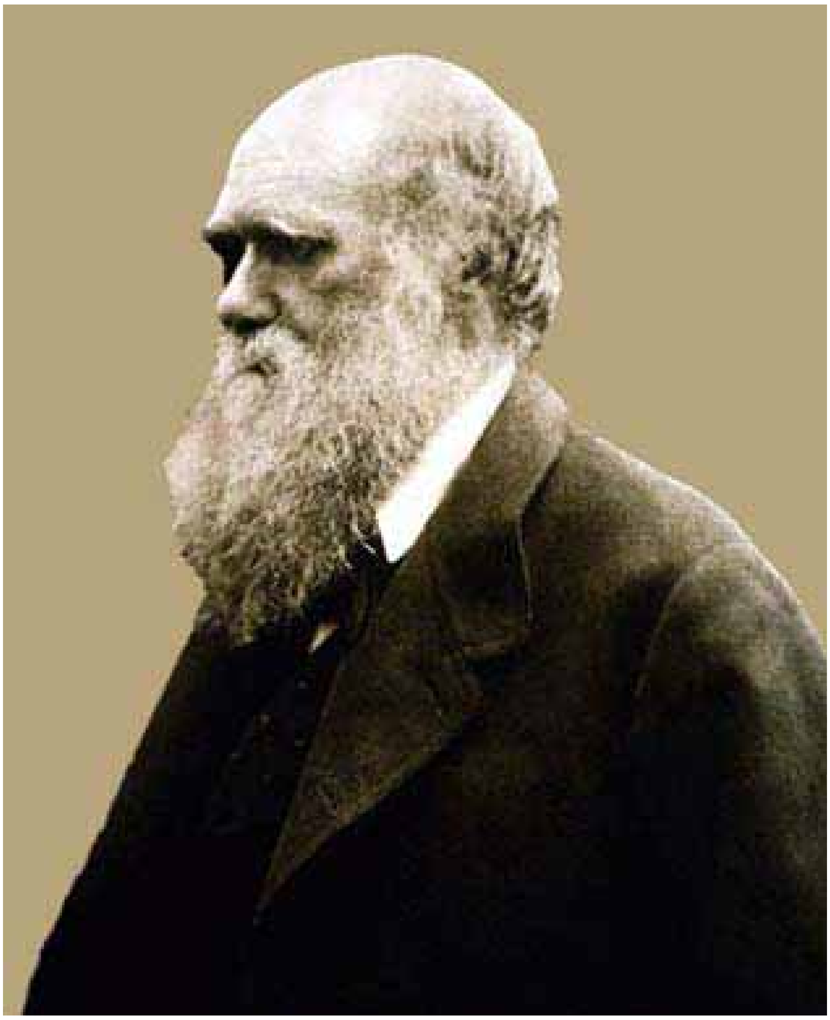


Fig. 1: Charles Darwin (1868)

---

[1] This article is based on a Pro ISSI lecture by Dr. Oliver Botta on 25 October 2005. In addition, we owe Dr. Frank Rutschmann, University of Zurich, significant contributions regarding biology.
[2] Charles Darwin, 1809 Shrewsbury–1882 London, clergyman and natural scientist
[3] James Dewey Watson, 1928 Chicago, biochemist, Nobel laureate 1962
[4] Francis Harry Compton Crick, 1916 Northampton–2004 San Diego, physicist, Nobel laureate 1962

What is Life?

Life is everywhere on Earth: heat-loving micro-organisms (called hyperthermophiles) populate hot springs with temperatures of more than 100 °C, alkaliphiles (alkalic-environment-preferring organisms) are found in soda lakes, piezophiles (pressure-resistant organisms) have been reported from the Mariana Trench, the deepest sea floor on Earth, some 10'000 metres below sea level. Life is everywhere, yet the most fundamental problem is our current inability to provide an all-encompassing definition, which we could use to direct the search for life elsewhere. In this situation, NASA adopted for its current roadmap for astrobiology the formulation of life as *a self-sustained chemical system capable of undergoing Darwinian evolution.* In the following chapters, we will adhere to this definition.

There are two basically different approaches to the understanding of the origin of life on Earth. The "top-down" approach looks at the current organisms, in particular in extreme environments, and also at the fossil record, to understand the molecular and phenological history of life and to attempt to extrapolate this knowledge towards the earliest forms of life on this planet. The "bottom-up" approach tries to understand how the planetary system, including the Earth, has formed and what the conditions on the surface were in the first billion years of its history. One also tries to understand which molecular building blocks of life could have been present on the early Earth, and how and where they were formed. From that knowledge one attempts to extrapolate towards the earliest traces of life which can be observed today.

The Building Blocks of Life

From atoms and molecules…

On the level of atoms, life is nothing else than a complex arrangement of a very limited set of chemical elements. This set comprises the biogenic elements carbon C, hydrogen H, nitrogen N, oxygen O, phosphorus P and sulphur S. Of every 200 atoms in our body, for example, 126 are hydrogen, 51 oxygen, 19 carbon, three are nitrogen and the remaining atom is one of all other elements. We have inherited all the hydrogen atoms, 63% of the material constituting our body, directly from the Big Bang fourteen billion years ago. The other elements up to iron have been synthesized by the main sequence stars burning hydrogen to helium and helium to carbon, etc., while the heaviest elements were created and expelled into space during supernova explosions[5]. As the first supernovae probably occurred just 500 million years after the Big Bang, life has had, at least in principle, plenty of time and space to evolve since then.

Amongst the six biogenic elements, carbon plays a central role thanks to its outstanding ability to form complex macromolecules containing thousands of atoms, which are stable in the temperature range where water is a liquid. On the other hand, the energy content of the carbon-carbon bond is low enough allowing water to dissolve and rearrange the carbon compounds.

Two other biogenic elements, namely hydrogen and oxygen, are important when combined to the water molecule H_2O. Water is a key ingredient for life as we know it as a solvent allowing molecules to be transported and chemical reactions to take place. Water has some unique properties including the fact that its solid form (ice) is lighter than its liquid state. In addition, water is very common in the universe: water is produced in the cold molecular clouds of the interstellar medium, where the oxygen atoms freeze on the surface of silica dust grains and eventually capture two hydrogen atoms to form the water molecules. Comets consist of a mixture of dust grains embedded in water ice. As they are the most pristine objects in the solar system, they testify that water has been present in remarkable quantities in the early solar system already.

In addition to water, a variety of other solvents, such as formamide ($CHONH_2$) or ammonia (NH_3), have similar solubility properties under certain conditions, but as

[5] See Spatium 13: Woher kommen Kohlenstoff, Eisen und Uran, Ruedi von Steiger, November 2004

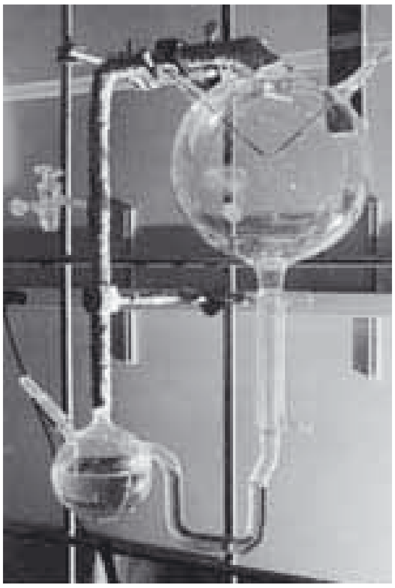


Fig. 2: The original laboratory equipment used by Stanley Miller (1953).

they are less abundant in the universe than water, they did not play a role for the emergence of life on Earth.

In search for experimental evidence that the elements and simple compounds present on the early Earth can form complex organic molecules, Miller and Urey (1953) conceived an experiment **(Fig. 2)** which produced a wealth of complex organic molecules from water, methane, ammonia and hydrogen. In his apparatus, Miller boiled water and circulated the vapour through a glass vial, past an electric spark, then through a cooling jacket that condensed the vapour and directed it back into the boiling flask. This process yielded for the most part black, tarry organic residue, which was difficult to characterize back in these days. Most surprisingly, however, Miller was able to identify in this mixture the presence of amino acids, which are the very basic building blocks of proteins. This was the first experimental evidence that biologically relevant molecules could be synthesized from simple gases under conditions as they were thought to have occurred on the early Earth.

Later, it was found that the early atmosphere of the Earth was probably composed of other gases, from which under the same conditions significantly lower amounts of amino acids and other organic molecules are formed. This finding led to an increased interest among astrobiologists in the possibility that extraterrestrial organic matter could have been delivered to the early Earth by impacts of comets, asteroids, meteorites and interplanetary dust particles to contribute the first prebiotic building blocks of life.

…over macromolecules…

The next major step consists of combining many relatively simple molecules, such as for example amino acids, to complex macromolecules. In modern biology, the formation of macromolecules, which involves many individual steps, requires a sophisticated biochemical machinery. The synthesis of these molecules under prebiotic conditions has turned out to be one of the biggest challenges in the understanding of the origin of life. Biological macromolecules in all forms of life on Earth come in the following classes:

- The *carbohydrates* are macromolecules based on a carbon structure with many hydroxyl groups (–OH) attached. Large carbohydrate structures are used for energy storage (for example sugar or starch) as well as providing structural support in living systems (for example cellulose in the cell wall of plants).

- The *proteins* are the most diverse and functionally versatile biological macromolecules. The proteins consist of a chain of amino acids. Every amino acid is connected to one of 20 different residues, small compounds such as for example CH_3 or CH_2OH. Most proteins of living systems on Earth are composed of the same 20 different amino acids. Proteins are either important structural molecules of cells or possess catalytic properties. In the latter case they are called enzymes.

- The *nucleic acids* are the largest macromolecules existing in living organisms. They consist of individual nucleotides linked together to form long linear polymers. A nucleotide is a chemical compound that consists of a heterocyclic base, a sugar, and one or more phosphate groups. The nucleic acids come in two flavours, the deoxyribonucleic acid (DNA, **see Figure 3)** and the ribonucleic acid (RNA) discussed below.

Both the DNA and the RNA consist of two complementary strands that coil about each other like a spiral staircase to form a double

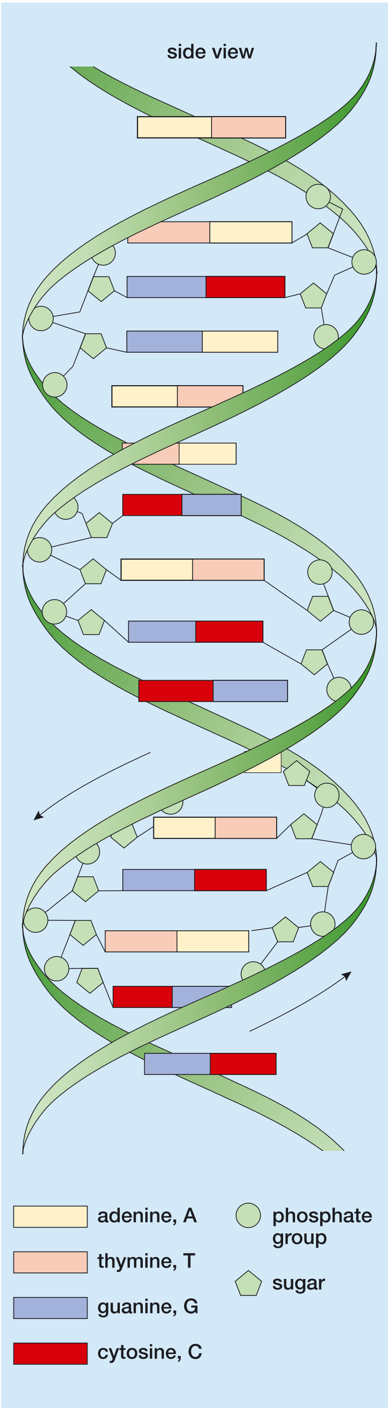


Fig. 3: The structure of the DNA double helix. (Credit: I. Gilmour, M. A. Sephton, Editors: An Introduction to Astrobiology, Cambridge University Press, 2003)

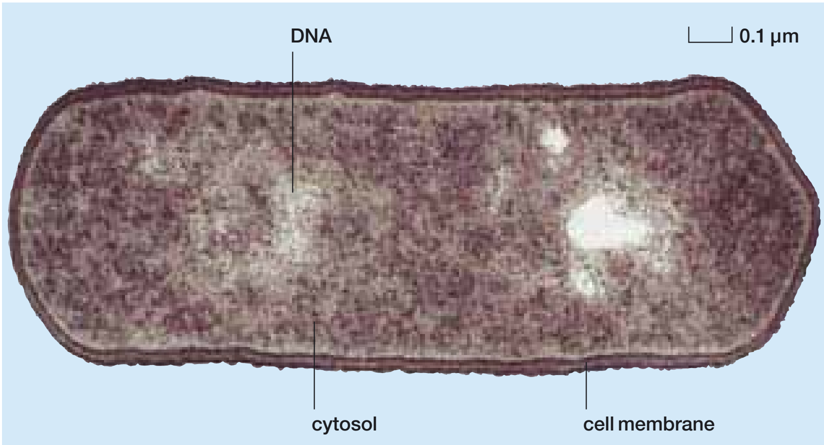


Fig. 4: The basic building blocks of a simple cell. (Credit: I. Gilmour, M. A. Sephton, Editors: An Introduction to Astrobiology, Cambridge University Press, 2003)

helix. Hydrogen bonds between the nucleotides form the steps of the staircase **(Fig. 3)**. They consist always of pairs: adenine in one strand is always linked to thymine in the other, and guanine in one strand is always paired with cytosine in the other. The bases are attached to a backbone made out of sugar molecules, which in turn are connected together along the exterior of the helix by phosphate groups.

Intriguingly, the DNA of all known forms of living systems is based on only these four bases constituting the universal genetic code. The code is expressed in the form of the specific sequence of the base molecules. The genetic code directs the production of proteins needed for the structure and the functions of the cell and for its own replication machinery. It represents the entire construction plan of a living organism.

…to self-replication…

Replication is one of the core principles of life. It allows overcoming the limited lifetime of all material entities. In addition, replication is central for evolution by providing a set of similar copies of an original on which the process of selection can act.

Under the effect of a large number of enzymes, the DNA can replicate in a very complicated process. In the vast majority of cases, the replication process yields an exact copy of the original DNA, so the replica is identical to the original. Sometimes, however, the sequence of the nucleotides of the copy is different from the original. It is this accidental change which is the underlying mechanism for the Darwinian evolution. A sequence of such random mutations may become prevalent in the gene pool of a population, if it is favoured by the natural selection process.

In contrast to the DNA, some RNA molecules were found to possess catalytic properties enabling to replicate themselves (self-replication). In a complicated laboratory procedure called *in-vitro selection,* some RNA molecules were found that showed the unique capacity to catalyze their own replication. It is thought that a selection process favoured RNA thanks to its replication potential. This led to the "RNA World Hypthesis", stating that on the early Earth the first replicating systems were RNA molecules. Based on its replication and evolution capacity, the RNA is seen as the most basic living system as per the definition above. That system did not contain a cell, but solely existed in suitable media providing the necessary supply of organic matter. Eventually, according to the RNA World Hypthesis, the genetic storage functionality of RNA was taken over by the more stable DNA, and the catalytic activity by the proteins, which have a wider range of functionalities. However, there still remain (major) problems even in this relatively straightforward scenario, for example how the building blocks of RNA, the ribonucleotides, were synthesized under prebiotic conditions. Another unsolved question is how the very reactive ribonucleotides were protected from simply being hydrolyzed in water thereby losing their reactivity.

… to living systems …

While the DNA in modern organisms stores its genetic information, it cannot self-replicate without being provided with monomeric building blocks to construct a second copy. These building blocks have to be synthesized by the organism or supplied by the surrounding medium. It is the cell, which provides the DNA the required infrastructure for its replication. The next giant step forward in the history of life was the encapsulation of bio-molecules inside compartments, where they are protected and concentrated **(Fig. 4)**.

The function of the cell can be compared with a factory, in which the DNA acts as the construction plan contained in the DNA and the necessary raw material. The DNA acts as the factory's master plan storing the genetic code and controlling the processes in the cell. The DNA is surrounded by the cytosol, a liquid containing thousands of different types of proteins constituting the raw material. The cytosol acts as the factory's assembly line allowing the raw material to be transported to the places where it is needed. The cell volume finally is enclosed by a cell membrane made from lipids and proteins.

To replicate, the cell goes through a process called mitosis. First, the entire DNA of the cell is replicated. Then, the original and its copy move to different parts of the cytosol. Then, the cell begins to divide, separating properly the two DNA-containing regions. The result of this process are two identical daughter cells.

… to Darwinian evolution

If the replication process would work perfectly and would produce identical replicas exclusively, no evolution would be possible. The term evolution designates random mutation of the genetic properties of the individuals coupled with natural selection. In order to be selected by the environment, the genetic mutation must translate into a phenological advantage of the individual. In his epochal oeuvre "On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life" Charles Darwin saw the natural selection process driven for example by climate change or by competition with other species for limited resources. As a further important selection mechanism he identified the sexual selection: Failing to mate prevents the individual to contribute to the evolution of the population's gene pool which is equivalent to eliminating its genetic properties from the gene pool. While these selection processes act on the level of the individual, they cause the population's collective gene pool to drift continuously beyond the lifetime of the single individual.

The Evolution of Life on Earth

The Sun and its planets are the result of the gravitational collapse of an interstellar cloud some 4'600 million years ago. The planets were formed from a debris disk that circulated the early Sun through accretion of dust grains on relatively short time scales of some tens of million years. The exact environmental conditions on the early Earth at that time, called the Hadean, are not known, but it is thought that the surface of the Earth cooled down to a level where water is liquid after about 200 million years. The heavily cratered surface of the Moon, lunar samples and meteorites have provided evidence for the early impact history of the Earth–Moon system. While impacts of comets and asteroids may have brought water and organic compounds to the early Earth, these impacts were devastating events that may have sterilized the whole surface of the planet and even boiled away the water on Earth. Eventually, though, the inventory of large bodies crossing the Earth's orbit declined, so that the frequency of these impacts decreased paving the way for the evolution of life on Earth.

Fig. 5: These living stromatolites have an age of around 3'500 years. The warm, shallow waters of Western Australian's Shark Bay favour the growth of micro-organisms, particularly cyanobacteria. Fossil stromatolites – literally layered rocks – also represent the oldest evidence for life on Earth. (Credit: University of South Carolina, Geology Department)

The conditions under which prebiotic chemistry on the early Earth occurred that would have eventually led to the first self-replicating organisms are not known. It is assumed that sandy beaches and volcanic mud pools may have provided the required milieu that protected prebiotically synthesized organic compounds from ultraviolet radiation and oxidation. Superheated columns of mineral-rich water that gush from vents in the seafloor, the so-called black smokers, may have been another potential environment for life to emerge.

The oldest traces of life on Earth are fossil cyano-bacteria that were found in 3.5 billion years old rocks in Western Australia, although there is a strong controversy about this topic in the scientific community. More secure morphological traces appear in the form of stromatolites in around 2.5 billion years old rocks. Stromatolites are the solidified ("lithified") remnants of microbial mats, in particular of cyano-bacteria. Active stromatolites can be found today in hyper-saline lakes and marine lagoons (Fig. 5). It can be said that these modern forms of cyano-bacteria constitute the most ancient organisms of life on Earth.

The history of life on Earth is also a history of mass extinctions. Life has been endangered or even nearly extinct many times since it emerged by massive climate changes, cosmic impacts or major volcanism events. It is estimated that more than 99.9% of all species that have ever lived are now extinct. An especially critical phase seems to have been the period between 750 and 580 million years ago when the Earth was completely covered by heavy ice sheets with surface temperatures as low as –50 °C according to the "Snowball Earth Theory". Such environmental conditions would have erased all but a small fraction of existing life deep on the floors of the oceans. Because the ice cover could not stop the continued output of the greenhouse gases by volcanoes, the Earth regained a moderate climate in a very short timeframe leading to hot environments according to the "Hot House Theory", now endangering the existing living systems which had successfully adapted to low temperature climates before. However, the subsequent phase of a warm, relatively stable climate allowed virtually every major phylum of animals to evolve within a relatively short time period of some 40 million years (the Cambrian explosion).

Mass extinctions have led to the loss of many species, but they stimulated also the evolution of life by creating unoccupied ecological niches where the remaining organisms could proliferate. The last known mass extinction occurred some 65 million years ago. It is thought to be caused by the impact of an asteroid of the size of 10 kilometres near what is now the north-western part of Mexico's Yucatan peninsula. This impact and the subsequent dust in the atmosphere led to a decade-long winter and to the extinction of 60 to 80% of all then living species, including the dinosaurs. Interestingly, some animal classes were less affected, like for example the amphibians and the mammals. The latter were present in the form of rat-like animals. After the cease of the dinosaurs' dominance, they were able to proliferate and to develop into the most complex animals on this planet, including men.

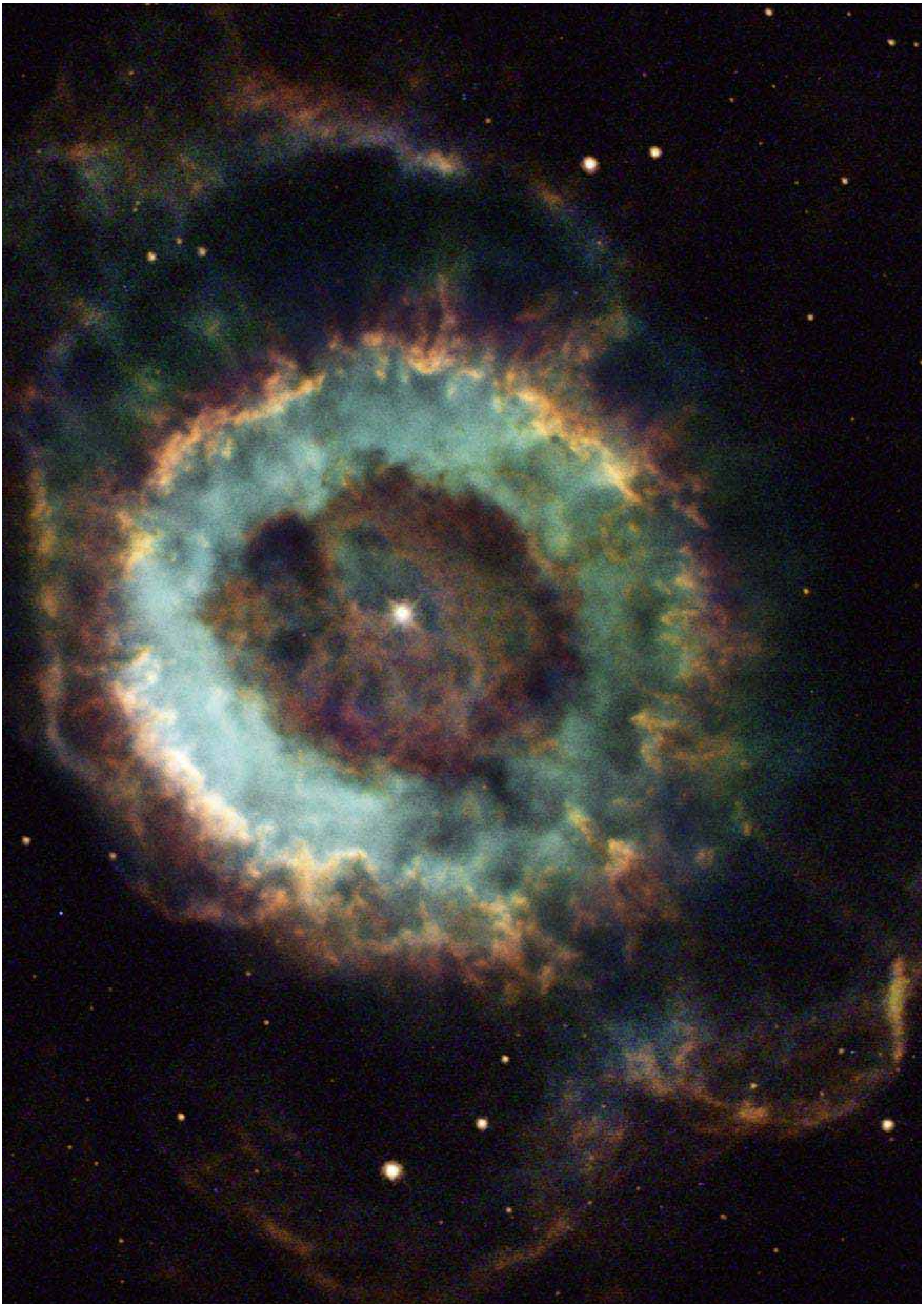
Co-evolution of Life and its Environment

According to the operational definition cited above, life is characterized as a self-sustaining chemical system with the ability to undergo Darwinian evolution. Through its chemical processes life is in constant exchange of matter and energy with its environment causing life to alter the environmental conditions leading to a co-evolution with its environment. This aspect is gaining central importance when it comes to searching for life in the universe as specific atmospheric biosignatures can be used as a proxy for biological activities on a planet.

The Earth's early atmosphere is supposed to have been dominated by carbon dioxide, nitrogen and water vapour. At that time, oxygen was present in the atmosphere only in trace amounts namely as a product of the breakdown of water vapour by the Sun's ultraviolet radiation or out-gassing from the interior. Therefore, no ozone layer was there to shield the surface of the Earth against incoming high energy radiation. This is one of the arguments in support of the hypothesis that the first living systems evolved in more sheltered environments such as sediments or deep waters.

The atmospheric concentration of oxygen rose to modern levels only about 2 billion years ago. However, oxygen was produced by cyanobacteria on the Earth through photosynthesis supposedly since 3.5 billion years. This leaves about 1.5 billion years of time on the Earth (the Archean epoch) when life was present, but the atmosphere was still anoxic. In fact, one of the key questions associated with the evolution of life on Earth is why it took so long for oxygen to reach these higher levels. Only this rise of oxygen in the atmosphere provided enough harvestable energy required by multicellular organisms.

Fig. 6: The white dwarf star in the centre of the planetary nebula NGC 6369 radiates strongly at ultraviolet wavelengths and powers the expanding nebula's glow. The nebula's main ring structure is about a light-year across. It contains the three biogenic elements oxygen, hydrogen, and nitrogen, which are coloured blue, green, and red respectively. In a remote future some of these elements eventually may help to constitute a living system. (Credits: Hubble Heritage Team, NASA)



Life from Space, Life into Space

The possibility that the emergence of life on Earth was based on organic molecules delivered to the Earth from space has become a subject of increasing interest. The analysis of carbonaceous meteorites has demonstrated that organic molecules, including amino acids, are formed in extraterrestrial environments. Astronomical observations have also shown that some of these molecules were formed already before the sun even started to form, and were incorporated into planets and smaller solar system objects such as comets and asteroids. Due to the extremely high temperatures on the planets during their formation, no organic molecules survived this epoch. The organic compounds necessary for the origin of life were formed either later in the atmosphere (e.g. through Miller-Urey-type reactions), or were delivered to the early Earth by large asteroid and comet impacts (as discussed above) or by their fragments, the meteorites and interplanetary dust particles. The study of meteorites, particularly the carbonaceous chondrites that contain organic matter up to 5% by weight, has allowed close examination of extraterrestrial organic material. One of the most famous sources of extraterrestrial material is the Murchison meteorite named after the Australian town where it reached the Earth in 1969 **(Fig. 8)**. Among other classes of organic compounds, more than eighty different amino acids have been identified in the Murchison meteorite at abundances of more than 1 part-per-million (ppm), eight of which are identical to those used in life on Earth as building blocks of proteins and enzymes. Also, the presence of sugar-related molecules and nucleobases was confirmed. Another source of information is the collection of micrometeorites that have been extracted from Antarctic ice. Many of these tiny objects in the 50–100 µm size

Fig. 7: This chicken embryo has been bred for seven days. Like its remote ancestors and like the majority of fossil and living tetrapodes, it will develop feet with five digits in the next few days, of which one however will disappear again to produce an individual with only four digits on its feet like all the currently living birds. This is an example of a vestigial trait, showing that the evolution of this individual (its ontogeny) mirrors the evolution of the birds in general (its phylogeny). (Credit: Hansjörg Schlaepfer)

Fig. 8: The Murchison meteorite. This is the most intensively investigated carbonaceous chondrite. It fell near the town of Murchison in Australia on 28 September 1969, only three months after the first lunar landing of Apollo 11. Fortunately, all the techniques that were developed at NASA to investigate the lunar samples for organic compounds could be applied to the meteorite as well. This allowed publishing the first report about the extraterrestrial amino acids in this meteorite shortly after its impact.

are unmelted chondritic micrometeorites, indicating that they had crossed the terrestrial atmosphere without suffering drastic thermal shock. In February 2006, the NASA Stardust spacecraft has returned, for the first time, samples from beyond the Earth-Moon system in the form of cometary dust particles, which will give scientists a new window on the composition of comets.

It is thought that the various extraterrestrial sources may have delivered about 10^{20} g of carbon to the Earth during the first 600 million years (the Hadean period). This delivery represents more carbon than the amount engaged in the current biomass, which is estimated at 10^{18} g. Whether or not these extraterrestrial compounds were the major source of organic molecules on the early Earth cannot be established even qualitatively since there are currently no estimates of the contributions from the other, indigenous sources.

If meteorites can transport organic matter to the Earth, the next logical question relates to whether living systems could have been delivered to the Earth. Of course, such specimens would experience the severe environmental conditions of space. Still, understanding survival in extreme conditions is essential for evaluating the potential for the interplanetary transfer of viable micro-organisms, and thus the potential that any life elsewhere in the solar system might share a common origin with life on Earth (or vice versa). Conditions in space and on other worlds are much more extreme than those encountered by any of the habitable extreme environments on Earth. Therefore, studies of survivorship beyond the Earth belong to the important tests of the resilience of Earth-originated life towards extreme conditions and its potential for dissemination in space.

Another option for life to leave its home planet is that of a civilized society using advanced technology to travel into space. This is what we call manned spaceflight **(Fig. 9).** The fact that all the niches on Earth are occupied by some forms of life suggests that it might be the ultimate destiny of life to leave its cradle and to proliferate into space.

Fig. 9: Launch of the NASA Space Shuttle Columbia on STS-109 on 1 March 2002. Currently, rockets with their inefficient chemical fuels are our only way to leave the Earth and to travel into space. If this journey is the destiny of mankind, we have only achieved the very first steps so far. A major leap forward, comparable with the introduction of jet engines in aeronautics that revolutionized air-travel, is required to travel anywhere beyond Mars. (Credit: NASA)

Strategies for Life Detection

When it comes to finding strategies for the detection of extraterrestrial life, two cases have to be distinguished depending on where the target is located:

■ *Within the solar system,* detection of living systems may be done by sending a spacecraft there with instruments to take *in-situ* measurements. Another option is remote sensing like for instance mapping the methane distribution in the Martian atmosphere from an orbiting spacecraft. The remote sensing approach allows only the detection of an active biosphere by finding gases in the atmosphere in concentrations that are beyond any abiotic processes. In contrast, the *in-situ* strategy allows the search for biosignatures that are either based on morphological, molecular or isotopic patterns from not necessarily currently living systems. There is not one single biomarker that can be used to unambiguously identify a signature of life. Therefore, the recognition of patterns, for example in the form of an overabundant presence of a certain group of molecules, becomes the favourite approach since life as we know it makes use of building blocks that are used repetitively to form larger structures. This approach is attractive also because it allows the recognition of patterns that do not necessarily have to be similar to terrestrial life.

■ *Beyond the solar system,* the distances are by far beyond the reach of any spacecraft. The quest for life, therefore, must be based on remote sensing techniques exclusively. Of primary interest are exosolar planets. So far, over 200 extrasolar planets have been found, but most of them are very different from what we expect to be a habitable planet. The test whether such a planet may harbour life will be based on so-called atmospheric biosignatures, similar to the spectroscopic approach mentioned above.

The European Space Agency (ESA) is currently planning a mission called Darwin **(Fig. 10)** to search for gases in the atmosphere of exosolar planets such as ozone, a proxy for the presence of large amounts of molecular oxygen, ozone, water, carbon dioxide, and methane. This approach, however, is extremely challenging since the luminosity of the star is orders of magnitudes higher than that of the planet. The analyses can be done by means of very sensitive long based spectrometers, which split the light reflected by a planet's atmosphere into its spectrum, where the content of these trace gases can be observed. The important clue here would come from the fact that some of these gases can be present in detectable amounts in the atmosphere of an exosolar planet only if there is a constant re-supply maintained by living systems. Any gas mixture that is in disequilibrium in an atmosphere to such an extent that it can not be explained by abiotic reactions would indicate biological activity on the planet.

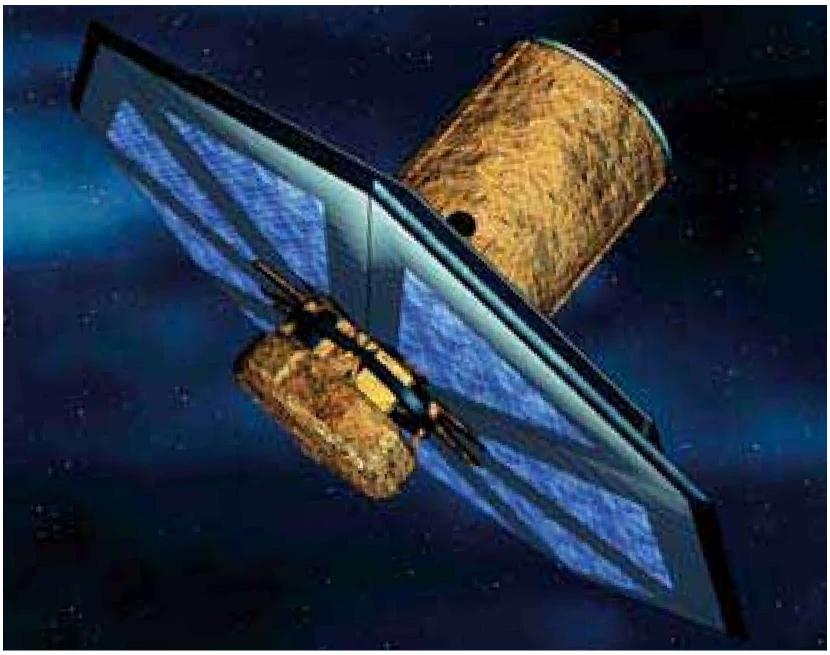


Fig. 10: One of the three space telescopes of the European Space Agency's Darwin mission. (Credit: ESA)

Life in the Universe

So far we have discussed the properties of life, how life evolved on Earth and how we can maximize the chance to detect life in the universe. Now, we address the question where we should search for life.

The Prerequisites for Life

Although no convincing signs of life have been found outside the Earth so far, a set of conditions can be compiled which must be fulfilled by a star and its planetary system in order to allow the development of life as we know it:

■ The central star must not be too massive (no more than about 1.5 times the mass of the Sun). More massive stars burn their hydrogen faster, and the time span in which the star is in a stable state is probably too short for life to originate and evolve on its planets. On the other hand, stars with less than about 10% of the solar mass will tidally lock planets into synchronous orbits. Such a planet would have a very hot side towards the star and a very cold side in the opposite direction, similar to Mercury in our own solar system. Provided the planet does not have the mass range of Jupiter-type planets, it will probably not be able to keep an atmosphere over periods of hundreds of millions of years. It is also highly unlikely that such a planet contains liquid water. In addition, the planet would be exposed to high energy solar flares providing a deadly environment for prebiotic chemistry and life.

■ For each star it is possible to determine the so-called habitable zone, which depends of course on how one defines the term "habitable". For example one could correlate "habitability" with the presence of liquid water on the surface of a planet with an appreciable atmosphere. If such an atmosphere contains a certain amount of greenhouse gases, such as methane or carbon dioxide, the habitable zone is widened towards the outer range of the stellar system. Since the luminosity of a star tends to increase during its lifetime, the habitable zone is being shifted outwards slowly.

■ Finally, the size of the planet is critical as well. The mass of a potentially habitable planet should not be too small; otherwise its gravitational field would be too weak to maintain a dense atmosphere over hundreds of millions of years. Also the planet should be massive enough to maintain active plate tectonics or at least some form of volcanism throughout its lifetime. Plate tectonics continually recycle oceanic crust back into the mantle at subduction zones and continually regenerate it at ocean ridges by the solidification of fresh magma. Without such a process, the planet does not have a feedback mechanism that would allow it to stabilize its climate and the physical conditions on the surface.

While these prerequisites apply to planets on which life could potentially develop, there are places in our solar system that are considered as potential habitats for some forms of life, but do not fulfil all these criteria. Of particular interest are the moons of the giant planets Jupiter and Saturn, where the lack of solar irradiation may be compensated by other mechanisms such as tidal heating to provide the required energy. Spacecraft data from the Galileo and Cassini missions have supported the notion that liquid water, and perhaps oceans, could be present under the icy surfaces of these objects.

Astrobiological Exploration of the Solar System

The solar system is of principle interest in the quest for extraterrestrial life. Most importantly, the distances to the other planets are such that, with our current technology, we have the capability to explore these worlds with robotic missions to perform *in-situ* measurements and eventually bring samples back for analysis here on Earth. Although we have set foot on the moon almost 40 years ago, we have not yet made the next step, which is the manned mission to Mars. And it seems that it will take at least another 30 years for this to happen. Of all the planets and small objects in our solar system, most are outside the habitable zone (as defined above). For example, Mercury, the innermost planet in the solar system, is far too close to the Sun and therefore too hot to harbour life. The surface temperatures on Mercury range from about –180 °C at the bottom of craters near the poles to about 400 °C at the sub-solar point. In combination with the absence of a dense atmosphere on Mercury, life as we know it could not develop there.

Venus

Venus is the most similar planet to the Earth in the solar system. It is often called the Earth's sister, since it has almost identical diameter and mass. Very early in its history, its environmental conditions may have also been similar to those of the Earth, and perhaps favourable for the emergence of life. However, due to its closer distance to the Sun, Venus has developed surface conditions, including temperatures of around 460 °C and permanent thick clouds of H_2SO_4 droplets, which make the current presence of any form of life highly unlikely.

The Earth's Moon

The Moon is thought to have been formed as a result of a collision between a very early, semi-molten Earth and a planet-like object with the size of Mars. The samples returned from the Apollo and Luna missions have shown that there are no signs of life and not even organic compounds present on the Moon's surface. The sterilizing UV and particle radiation it is exposed to would prevent any organic molecule to build up. However, the Moon may have been a crucial element in the development of life on the Earth. With the exception of Pluto and its moon Charon, Earth is the only planet in the solar system that has such a massive satellite relative to its own mass (the mass of the Moon is

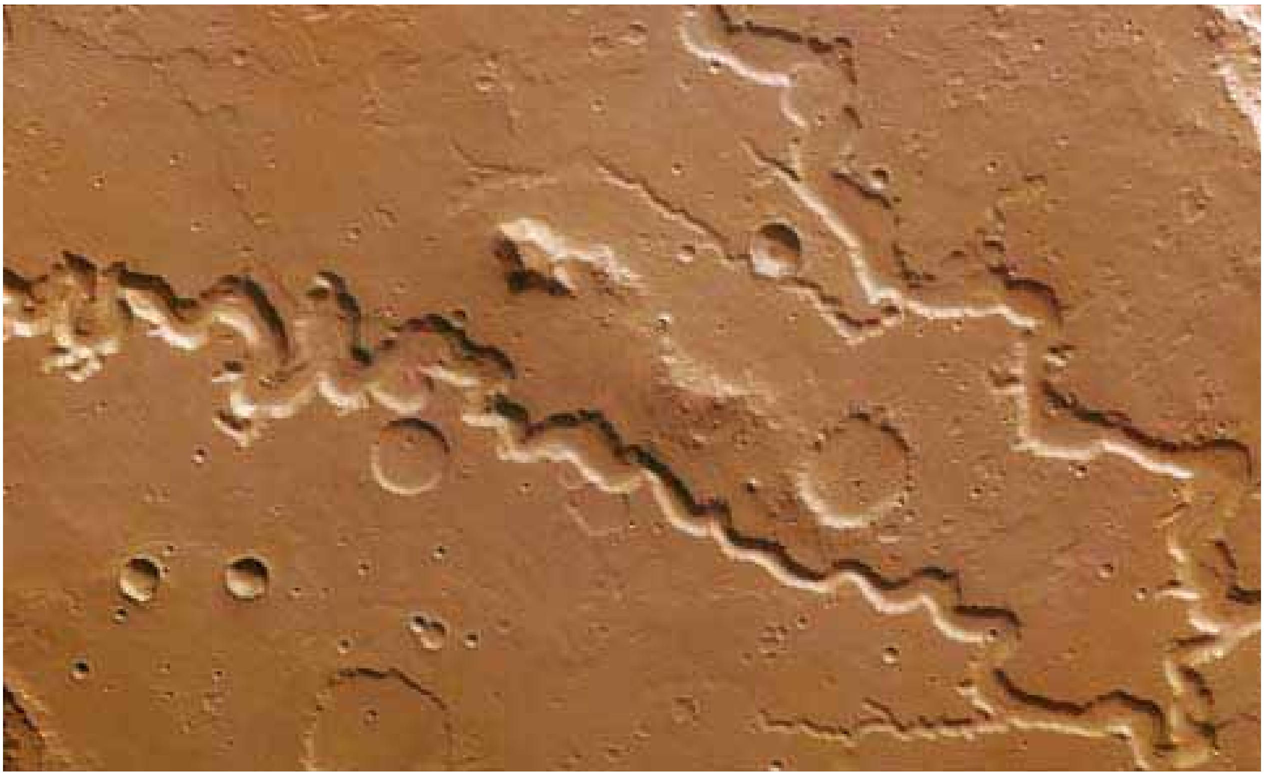


Fig. 11: Nanedi Valles, a roughly 800-kilometre valley extending southwest–northeast and lying in the region of Xanthe Terra, southwest of Chryse Planitia of Mars. In this view, Nanedi Valles ranges from approximately 0.8- to 5.0-kilometre wide and extends to a maximum of about 500 metres below the surrounding plains. The valley's origins remain unclear, with scientists debating whether erosion caused by ground-water outflow, flow of liquid beneath an ice cover or collapse of the surface in association with liquid flow is the responsible mechanism. This image was captured by the High-Resolution Stereo Camera (HRSC) onboard ESA's Mars Express. (Credit: ESA).

about $^1/_7$ of the Earth's mass). This high mass ratio has the effect that the Earth's rotation axis has always been relatively stable at its current inclination of 23.5°, which in turn provided a stabilized climate on the surface over billions of years.

Mars

Mars is the primary target for the search of traces of past or present life in the solar system. From the many Mars exploring spacecrafts and rovers there is convincing evidence that water existed in substantial quantities on its surface at some earlier epoch. We do not know, however, whether it was present on the surface for a sufficiently long period of time and if it is perhaps still present in subsurface aquifers. Based on the morphological evidence for the presence of surface water in the earliest epochs, including large outflow regions **(Fig. 11)** and layered sedimentary records, Mars initially had a dense atmosphere. However, due to the smaller mass of Mars as compared to the Earth, it lost its atmospheric gases to space, leaving it with the current thin CO_2-rich atmosphere.

The two Viking landers, which touched down on the Martian surface in 1976, carried biological life detection instruments as well as a gas chromatograph-mass spectrometer. These instruments were used to analyze soil samples in the immediate vicinity of the landers **(Fig. 12)**, but they failed to detect organic molecules in any of these samples. This finding is the primary evidence that points to an absence of life on the surface and immediate subsurface of Mars today. However, given the recent discovery of flourishing biospheres at 1'000 metres below the Earth's surface in South African gold mines, it is conceivable that a similar microbial community might be present in the deeper subsurface of Mars.

The Giant Planets Jupiter and Saturn

The atmospheres of Jupiter and Saturn (as well as Uranus) are mainly composed of hydrogen and helium, with a noticeable fraction of methane and a lower contribution of ammonia. The astrobiological interest of these planets is limited as they have no solid surface. In contrast, the interest in the giant planets' satellites is rapidly growing currently.

The Jovian Moons

The Galilean moons of Jupiter were visited several times by the Galileo spacecraft while orbiting Jupiter for almost 10 years. Magnetometer data have provided evidence for an ocean of liquid water beneath the icy crust of Europa, **Fig. 13.** Similar conditions may also exist on Ganymede and Callisto. While their distances from the Sun prevents these objects to receive sufficient solar radiation to maintain liquid water on their surfaces, their orbital geometries give rise to tidal forcing of their interior that provides a heat source which is probably sufficient to melt some of the subsurface ice layers. Although liquid water is supposed to be present on Europa, the chances for life to have originated and de-

Fig. 12: A Martian landscape acquired by the Viking 2 camera in 1976. While some parts of the lander are visible in the foreground, a rocky environment is seen in the background. (Credit: Edward A. Guinness, Washington University, St. Louis, USA)

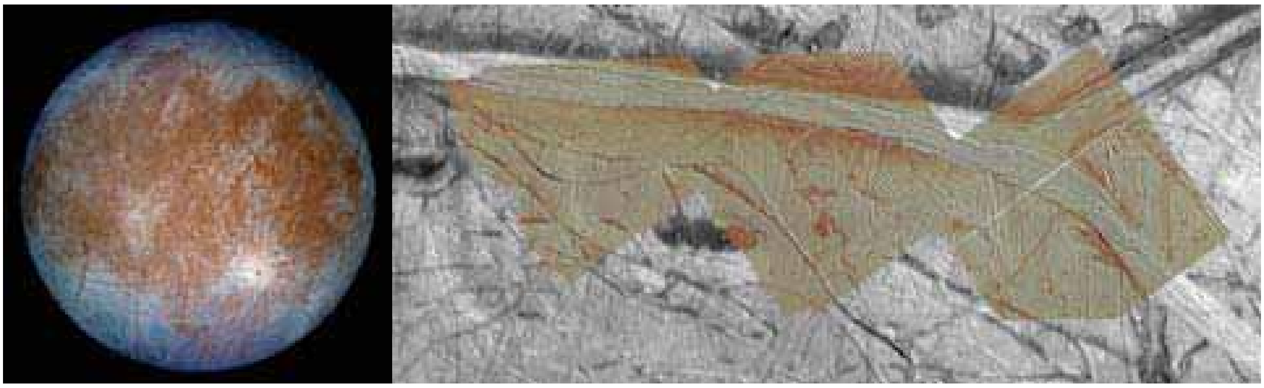


Fig. 13: Jupiter's icy moon Europa (left) and Agenor Linea (right), a bright white band on its surface obtained by NASA's Galileo spacecraft. Along this portion of Agenor is a "triple band," flanked by dark, reddish material of uncertain origin. On the right side of this image, Agenor splits into two sections. While these and other details of Europa's surface formations remain mysterious, the general results of Galileo's exploration of Europa have supported the idea that an ocean of liquid water lies beneath the cracked and frozen crust. (Credit: NASA).

veloped in this subsurface ocean are remote because it is not clear where organic compounds would have come from. Although some organics may have survived an impact on Europa, the transport of this material through the ice shell is not at all understood. Because of these uncertainties, Europa is still a prime target for future missions, which could include a ground penetrating radar instrument or even a sample return mission.

The Saturnian Moons

Since 1944, when Gerhard Kuiper detected methane in Titan's atmosphere, this moon was suspected to harbour life, as on Earth methane is a product of organic metabolism. Titan has a dense N_2 atmosphere rich in organics in both gas and aerosol phases. It represents, therefore, a natural laboratory for studying the formation of complex organic molecules on a planetary scale and over geological times. Despite the fact that Titan's surface temperatures are much lower than those found at the Earth's surface and that liquid water is totally absent therefore, the satellite provides a unique milieu to study the products of the fundamental physical and chemical interactions driving a planetary organic chemistry. However, neither the Cassini orbiter nor the Huygens lander have found any sign of life so far.

Recently, another moon of Saturn, the tiny Enceladus, has made scientific headlines when NASA announced the detection of huge water geysers on Enceladus by the Cassini orbiter **(Fig. 14).** Plumes of icy material extend far above its southern polar region. It is believed that the plumes stem from geysers erupting from pressurized subsurface reservoirs, potentially containing liquid water, while the surface cover is ice at a temperature of around –200 °C.

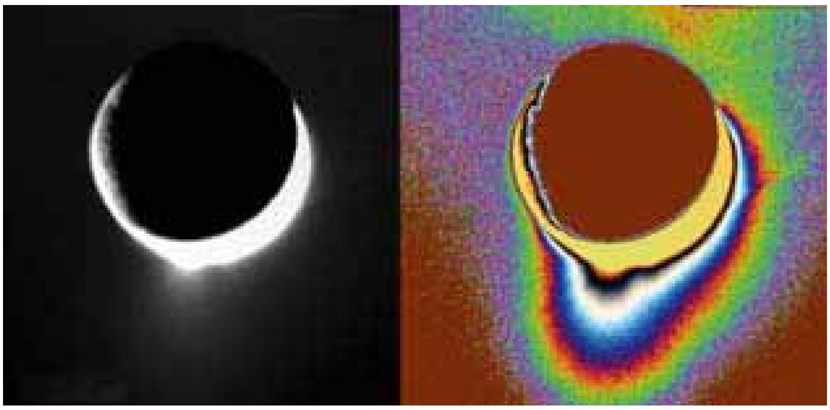


Fig. 14: Plumes of icy material extend above the southern polar region of Enceladus. (Credit: NASA/JPL/Space Science Institute)

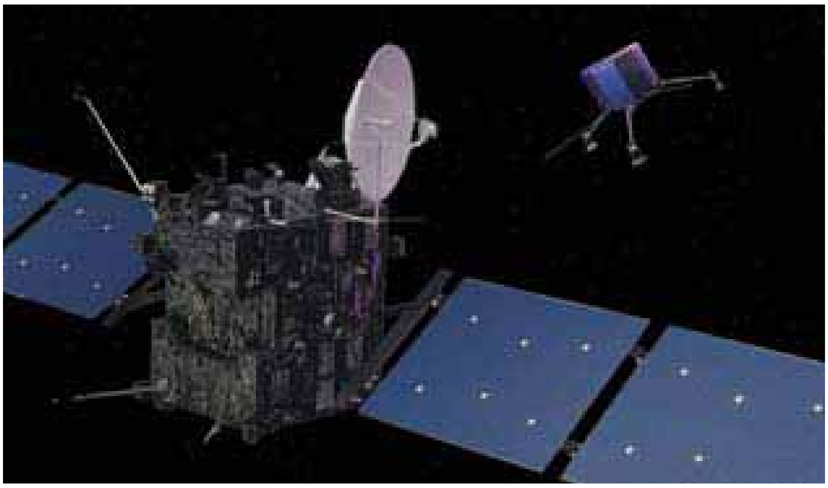


Fig. 15: An artist's view of the Rosetta mission and its lander Philae heading for the comet 67 P/Churyumov-Gerasimenko. Rosetta is under way since spring 2004 and will reach the orbit of its target comet in spring 2014. After close inspection of the comet, the probe will deliver the vehicle Philae, which is to touch down on the comet some days later. The probe itself will continue to circle around the comet for another year during its closes approach to the Sun. (Credit: ESA)

Objects on the Outer Rim of the Solar System

The icy giants Uranus and Neptune as well as Pluto and the Kuiper belt objects are too far away from the Sun to have liquid water on their surface; however, the exploration of these planets is immensely important to our general understanding of the formation of planetary systems.

Comets

Comets are believed to be Kuiper belt or Oort cloud objects whose orbit has been changed by gravitational pull from the outer planets bringing them into the inner solar system. As stated above, comets contain large amounts of water[6]. For example, two thirds of the material in Comet Halley's nucleus are water ice, and the rest is made out of silicate grains and organic particles. Substantial amounts of organic molecules including for example formaldehyde, methanol, etc. were detected in the comas of several comets. Improving detection capabilities of scientific probes allowed for the identification of many more organic substances in cometary comas, like for example ammonia, methane and acetylene, up to more complex molecules like cyano-acetylene. Unfortunately, we have no direct measurements about the composition of the nucleus yet. The European Space Agency's Rosetta mission with its lander Philae is on its way to comet 67 P/Churyumov-Gerasimenko and will address this question with its scientific payload **(Fig. 15).**

Outlook

Searching for new worlds has been one of mankind's very basic endeavours since its earliest days. Astrobiology is nothing else than the continuation of this effort into new dimensions. But at the same time it provides us with a deepened knowledge of the conditions of life on our own planet. In that sense, astrobiology may help us also to become aware of the unique value of life prompting us to safeguard the richness of species on our Earth.

[6] Spatium 4: Kometen by Kathrin Altwegg, October 1999

The Author

Oliver Botta was born in Liestal, Kanton Baselland, where he visited the elementary and secondary schools. Fascinated by the overwhelming wealth of molecules that can be composed from a limited set of atoms he decided to study chemistry at the University of Basel. He concluded the studies with the Ph. D. degree in organic chemistry in 1999. Then he moved to the Scripps Institution of Oceanography at the University of California at San Diego, USA, where he was engaged in organic trace analysis. This activity brought him in contact with the analysis of meteorites, thus combining his chemical background with his fascination of space. He also was involved in the early development of an analytical instrument for the robotic exploration of solar system bodies and that is now part of the science package of the European ExoMars mission.

From 2002 to 2004 Oliver Botta was engaged by the European Space Agency ESA as an external Post Doctoral Fellow at the University of Leiden in the Netherlands, where he not only continued his activities in the area of organic trace analysis of meteorites, but was also involved in astronomical observations of organic molecules in space. From April 1st 2004 onwards he was engaged at the International Space Science Institute with the primary task to organize an international forum on Astrobiology, leading the way for the Institute to increase its activities in this interdisciplinary scientific field. Last year, Oliver Botta then moved to the United States again, where he was engaged at the NASA Goddard Space Flight Center in Maryland to support the development of the Sample Analysis at Mars (SAM) instrument, a gas chromatograph-mass spectrometer for the future NASA Mars Science Laboratory rover mission. On July 1st 2006 he returned to ISSI to lead the Institute's future activities in Astrobiology.

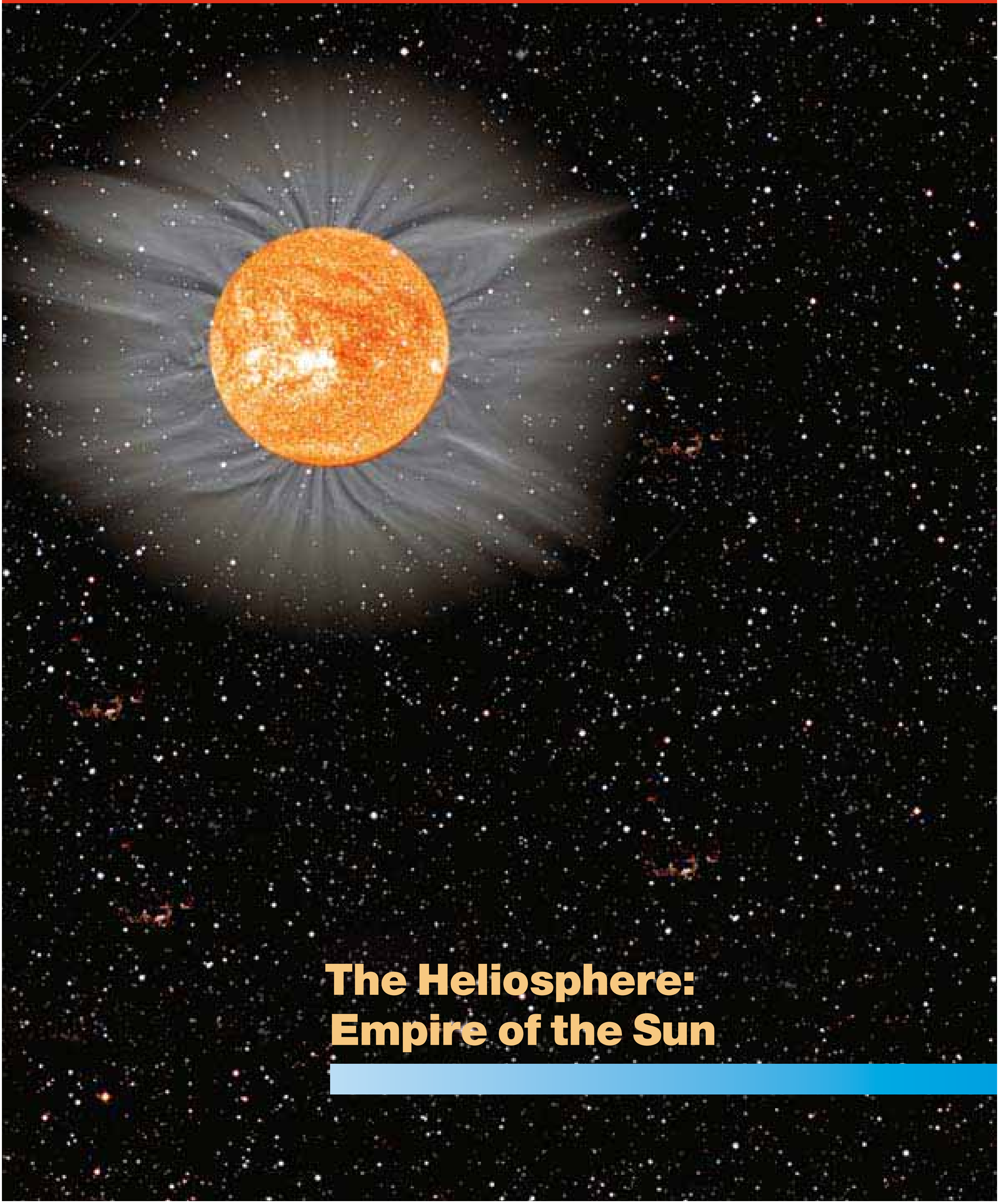
Astrobiology is not really a new scientific field, but more a rigorous attempt to combine the knowledge and the methodologies of the classical branches of science with the goal to lead to answers to some of mankind's most fundamental questions relating to the emergence of life on the Earth and the possibilities for life at other places in the solar system and beyond. This interdisciplinary approach requires scientists to communicate in a common language, and that is where Oliver Botta sees his role in his future engagement at ISSI. Together with his wife, Oliver Botta has two sons and a daughter, with whom he likes to play Lego, a concept that is comparable to nature's to build up complex systems by repeating and re-arranging (relatively) simple building blocks.

SPATIUM

INTERNATIONAL SPACE SCIENCE INSTITUTE

Published by the Association Pro ISSI

No 17, October 2006



The Heliosphere: Empire of the Sun

Editorial

In a landscape rich in lakes and forests in what was then eastern Pomerania, nowadays a part of Poland, there was a large farm with the wonderful name "Kornburg". It was into that environment that a boy was borne on 4 September 1926. The child grew up on his parent's estate and at that time no one could foresee that this youngster eventually would become one of the leading space scientists of the twentieth century.

He visited the schools in the nearby village and at the age of ten he was sent to the Gymnasium in Stolp (now Slupsk), the town near the Baltic coast. Mathematics and natural sciences were the subjects fascinating him most; so it was a logical choice to study physics at the University of Göttingen, one of the intellectual centres of Germany just recovering from World War II. There, he received the diploma and the Ph. D. in physics, but soon Carmen, a pretty girl just about to finishing her Gymnasium, attracted his interest just as well and the two married a few years later. Many stations marked their common further way to the Universities of Chicago, Miami and Toulouse and to the NASA centres at New York, Maryland, Houston and Pasadena; but during all these years, Bern was their favourite city.

It is with gratitude that Pro ISSI devotes the present issue of Spatium to Johannes Geiss, the boy from Pomerania, at the occasion of his eightieth birthday and to Carmen, the charming wife who accompanied Johannes through all the busy years.

The heliosphere is the subject of the present issue of Spatium, the empire of the Sun, or, to be more specific, that part of the universe where the Sun dominates over the surrounding interstellar medium. We are very thankful to Professor André Balogh, Director at ISSI, for his kind permission to publish herewith his fascinating talk held for the Pro-ISSI audience on 28 March 2006.

It is the Sun, from which Johannes Geiss learned about the composition of the pre-solar cloud that created our central star and the planets 4.6 billion years ago. His solar wind experiment, mounted five times on the Moon by the Apollo astronauts, allowed him and his co-investigators at the University of Bern to get deep insights not only into the solar system but into the secrets of the entire universe.

We are deeply indebted to you, Johannes, for all what you have done for science, for Bern and Switzerland and for the cultural progress of our society towards understanding this wonderful world!

Hansjörg Schlaepfer
Neerach, October 2006

Impressum

SPATIUM
Published by the
Association Pro ISSI
twice a year

INTERNATIONAL SPACE SCIENCE INSTITUTE

Association Pro ISSI
Hallerstrasse 6, CH-3012 Bern
Phone +41 (0)31 631 48 96
see
www.issibern.ch/pro-issi.html
for the whole Spatium series

President
Prof. Heinrich Leutwyler,
University of Bern
Publisher
Dr. Hansjörg Schlaepfer,
legenda schläpfer wort & bild,
CH-8173 Neerach
Layout
Marcel Künzi, marketing · kommunikation, CH-8483 Kollbrunn
Printing
Schellenberg Druck AG
CH-8330 Pfäffikon

Parts of the Sun's outer atmosphere are jettisoned into space in the form of the solar wind. This artist's view shows an image of the Sun by the SOHO extreme ultraviolet imaging telescope together with the corona as seen from the Earth during a total eclipse. (Credit: Hansjörg Schlaepfer)

The Heliosphere: Empire of the Sun[1]

Introduction

The Sun dominates, through its expanding atmosphere in the form of the solar wind, a large volume of space that extends to a distance over twice the orbital distance of Pluto or more than 80 Astronomical Units (AU)[2]. This is the heliosphere, the empire of the Sun. At its outer boundary, the solar wind is stopped by the surrounding interstellar medium. The structure of the heliosphere follows the variations of the solar activity cycle; it is formed by the ongoing interaction between solar wind streams of different speeds and the magnetic field of the Sun that is carried in the solar wind. At times of high solar activity, solar outbursts lead to very disturbed conditions that affect even the immediate neighbourhood of the Earth. Space missions have explored the properties in the different regions of the heliosphere, from the orbit of Mercury to its outer reaches. The study of the heliosphere is immediately relevant to a better understanding of our own space neighbourhood, but it is also a case study of a much larger class of stellar environments.

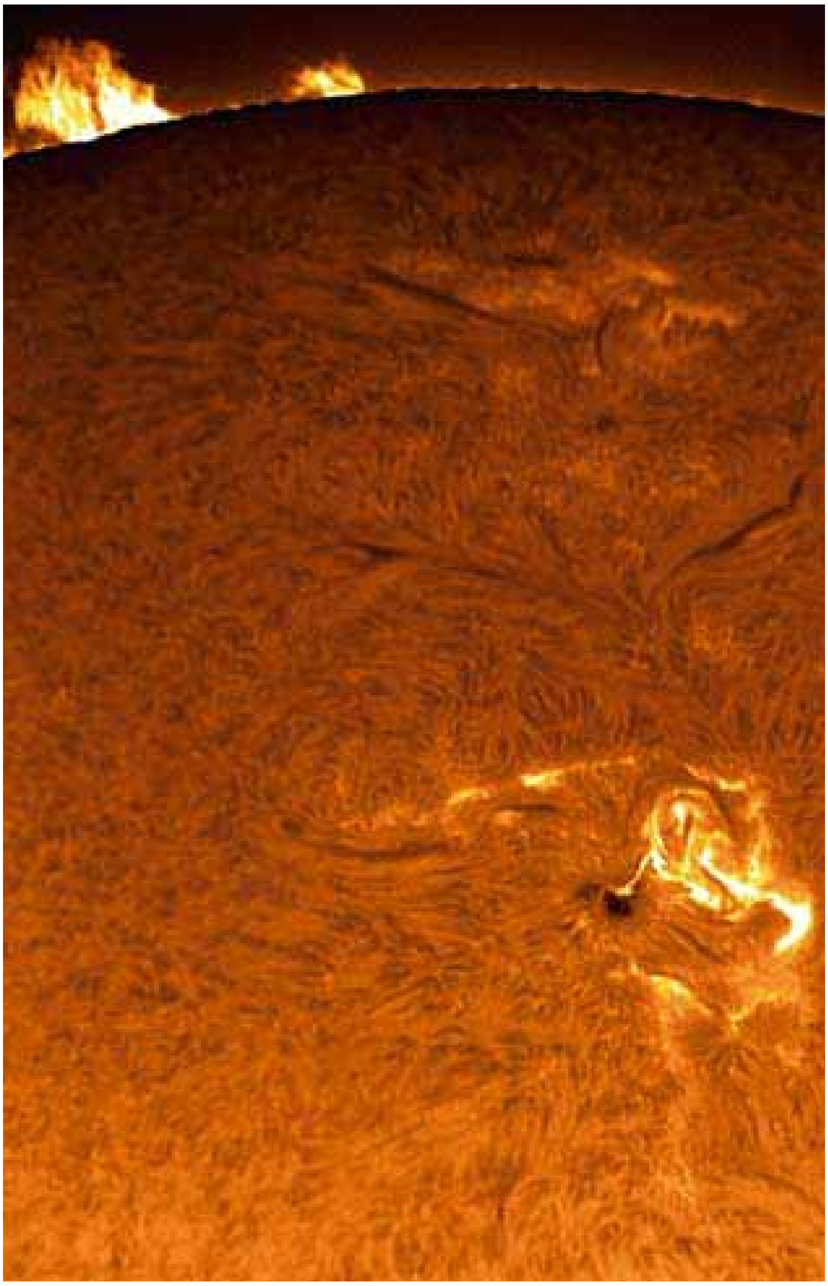


Figure 1: This image shows part of the Sun's surface in the Hα wavelength that is at the emission line of hydrogen. The sunspot has a diameter exceeding the size of the Earth. Relatively cool regions appear dark while hot regions appear bright. On the far left, solar prominences hover far above the Sun's surface. (Credit: Greg Piepol, sungazer.net)

[1] This text is based on a lecture by Prof. André Balogh for the Pro ISSI Association on 28 March 2006.
[2] The Astronomical Unit is defined as (approximately) the mean distance of the Earth from the Sun. The currently accepted value is 149,597,870 kilometres.

The Solar Wind and the Existence of the Heliosphere

The large volume of space around the Sun that we call the heliosphere is filled with the expanding atmosphere of the Sun. The upper solar atmosphere, the solar corona, visible from the Earth only at the time of solar eclipses **(see Figure 2)**, continuously expands into space at high speeds, varying between about 300 km/s and 1,000 km/s. This is the solar wind. The corona is a rarefied, very hot gas, with temperatures well in excess of one million degrees. At these temperatures, many of the electrons around the atomic nuclei are stripped away; leaving positively charged ions and negatively charged electrons in the gas. This is plasma.

Why the corona is so hot, when the surface of the Sun, the photosphere, is only about 6,000 degrees, has remained a mystery since the high coronal temperatures were identified in the 1940s. There is enough energy emerging from the Sun in the form of convective motions (mass transport from the interior to the surface) to supply the energy needed to heat the corona, but the way that this convective mechanical energy is transmitted to the gas in the corona is not really understood. An important factor in this is the magnetic flux that is also transported with the material to the solar surface. In the lower corona (the layers closest to the solar surface) magnetic fields emerging from the solar surface form very complex magnetic loops that can be observed by space-based solar telescopes or even from the ground at the times of total solar eclipses. It is likely that some form of waves and some form of magnetic dissipation are the main contributors to the heating of the corona. In any case, the corona is not only heated to these very high temperatures, but also, a part of it is expelled into space at high, supersonic speeds in the form of the solar wind.

The solar wind streams away from the Sun in all directions. Its composition is identical to the Sun's corona, that is approximately 95% protons (hydrogen cores) by particle densities, 4% alpha particles (helium cores) and 1% of carbon, nitrogen, oxygen, neon, magnesium, silicon and iron are the most abundant. These components are present as a plasma. The first detailed composition measurements were performed by Johannes Geiss's Solar Wind Composition (SWC) experiment on the Moon in the frame of the first five US Apollo missions 1969–1972. The solar wind was collected using a specially prepared metal-foil and then brought back for analysis **(see Figure 3)**. Today, the exact composition of the solar wind is routinely measured by instruments on Ulysses and ACE, two spacecraft carrying a Solar Wind Ion Composition Spectrometer.

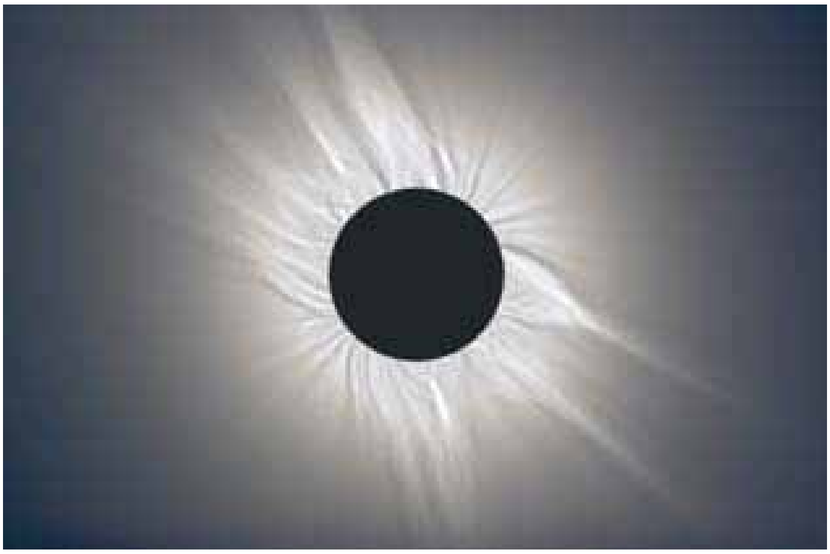


Figure 2: During a total solar eclipse, the Sun's corona is a marvellous sight. The subtle shades and features span a brightness range of over 10,000 to 1, making them difficult to capture in a single picture. But this composite of 33 digital images ranging in exposure time from $^1/_{8000}$ to $^1/_5$ second comes very close to revealing the solar corona in all its details. (Credit: Koen van Gorp)

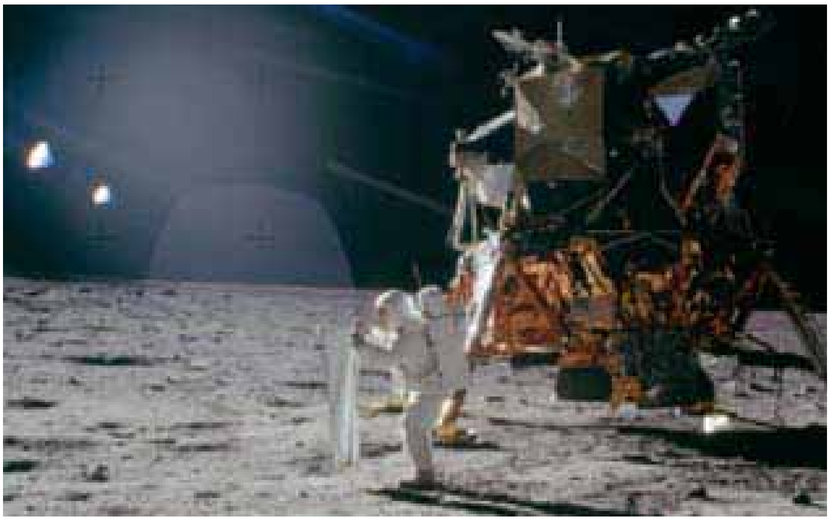


Figure 3: The US astronaut Edwin E. Aldrin unfurling the Solar Wind Composition (SWC) experiment of Johannes Geiss in the Mare Tranquilitatis on the Moon on 20 July 1969. This experiment was the first attempt to quantitatively analyse the composition of the solar wind. (NASA Photo S11-40-5872)

After more than four decades of observations, the properties of the solar wind have been well documented. The solar wind has an average density of 7 particles/cm^3 at the orbit of the Earth (but really highly variable from about 1 to 100 particles/cm^3), with a speed that varies from less than 300 km/s up to about 1000 km/s. The solar wind causes the Sun every second to loose up to 10^9 kg of matter. The density of the solar wind decreases as the inverse square of the distance to the Sun (it is a spherically expanding gas), but the speed of the solar wind varies very little all the way out to the outer boundary of the heliosphere. Both, the solar wind and the heliosphere have been postulated before space missions, since the early 1960s, have brought conclusive evidence about their existence. The two are really related, because if there is a continuously flowing solar wind, it is more than likely that it will be powerful enough to create the space for a "heliosphere" around the Sun, from which the interstellar medium (an even more rarefied gas than the solar corona and the solar wind) is excluded.

Manifestations of the Solar Wind

Perturbations of the Earth's Magnetic Field

Four kinds of observational evidence had been interpreted in terms of a solar wind and the heliosphere. First of all, periodic perturbations in the Earth's magnetic field were noted that matched the periods of the solar rotation (about 27 days) and the solar sunspot cycle (about 11 years), **see Figure 4.** The shorter, 27-day periodicities were associated with the passage, as the Sun rotates, with specific regions on the Sun that appeared to cause a higher level of auroral activity and other manifestations of geomagnetic disturbances. These regions were called M-regions that were thought to be somehow more magnetically active than other parts of the Sun. Their occurrence also changed with the 11-year solar cycle, and such M-regions appeared to be more active (caused

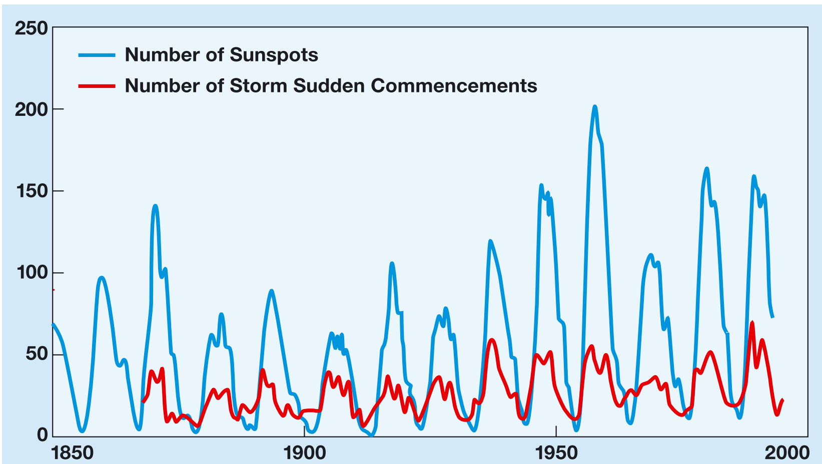


Figure 4: The periodicities observed in geomagnetic phenomena follow solar periodicities: the 11-year solar sunspot cycle and the 27-day solar rotation period. (Credit: Ellis, Maunder, Forbush, Bartels, Chapman)

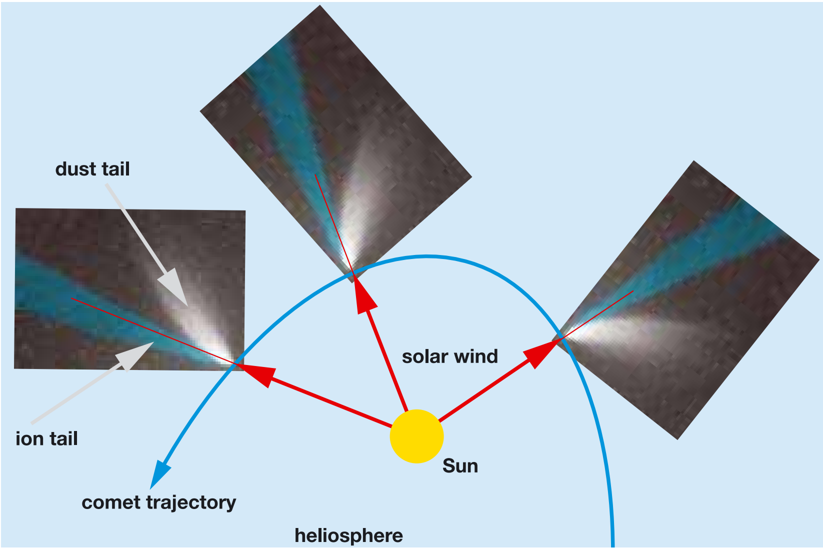


Figure 5: L.Biermann proposed that ion tails arise from atomic particles in the coma that are ionised by solar ultraviolet radiation which are then "entrained" by a "continuously flowing" corpuscular radiation from the Sun (now known to be the solar wind).

more terrestrial disturbances) not at the time of maximum in solar activity, but rather away from it. It is now known that such periodic disturbances are really caused by Corotating Interaction Regions which, as described in more detail below, are large scale structures in the solar wind caused by the collision of fast and slow solar wind streams. As the solar regions with which these structures are associated remain stable over several solar rotations, the interaction regions in the solar wind that cause geomagnetic disturbances also recur at the same time in each solar rotation. This is then observed as a 27-day periodicity in the terrestrial effects that was known well before the space age.

In addition, generally following large solar flares, large geomagnetic storms were observed. These geomagnetic storms are seen in the records as large, fast depressions, followed by a recovery over many hours or even a few days in the magnetic field measured in geomagnetic monitoring stations. This can be explained (and this explanation has proved valid even in the latest state of our knowledge) by the compression of the geomagnetic field by some large scale wave-front that travels from the Sun after large solar flares. These are now known to be Coronal Mass Ejections, described in more detail below. Both the matching periodicities at the rate of the solar rotation and the nature of the geomagnetic disturbances pointed to some agent that brought solar phenomena and disturbances to the vicinity of the Earth. Of course, this is the solar wind and the various structures that it carries in response to events on the Sun.

The second line of reasoning that indicated a continuous emission of particles from the Sun was developed in the 1950s. Ludwig Biermann worked out that the bluish tails of comets (now called the plasma tail) that always point radially away from the Sun can only be caused by particles constantly steaming also away from the Sun **(see Figure 5).** Even though the orientation of comet tails had been known for centuries, for long it was thought that the pressure of radiation (the visible light) from the Sun was responsible. Biermann's contribution was to demonstrate that radiation pressure was not enough and that particles travelling from the Sun at hundreds of km/s were necessary to create the plasma tails of comets. The other tail, the dust tail, curves away from the comet, still in a generally anti-sunward direction, is in fact generated by solar pressure, photons from the Sun striking the micron-sized dust particles emanating from the comet. Both tails are very visible in the photograph of comet Hale-Bopp in **Figure 6.**

Parker's Supersonic Flow Theory

In 1958, a highly controversial idea was put forward by Eugene Parker, then a young researcher in the University of Chicago. He had been influenced by the ideas of Biermann, but he set out to calculate the consequences of the

Figure 6: This image shows comet Hale-Bopp with its two tails: the dust tail arises from mass loss in the form of small particles shed by the comet, while the ion tail, a fainter feature, arises from ionized atoms and molecules. (Credit: J. C. Casado)

million-degree solar corona on what happens to this very hot part of the Sun's atmosphere. His theoretical solution to the problem included many simplifications, but still provided a sophisticated mathematical model for a solar wind that would escape from the solar atmosphere at supersonic speeds. We need to explain what supersonic speed means in the context of the solar wind. In a plasma, unlike in an ordinary gas like air, three kinds of waves can propagate: the so-called Alfvén wave that is just the wave which propagates along a magnetic field line as along a stretched string, and two kinds of sound waves (longitudinal compressional waves), the slow- and the fast-mode waves. Parker demonstrated that the solar wind, as it escapes the Sun's corona, travels at speeds in excess of the speed of the fast-mode sound or acoustic wave. This makes the solar wind a supersonic flow of plasma in interplanetary space.

Parker's idea was controversial at the time, as the scientific establishment favoured a different solution for the solar atmosphere, the so-called solar breeze that just evaporated at the outer edges of the corona. Nevertheless, Parker's theory found very soon, by 1962, uncontroversial proof from the first space missions in interplanetary space, in particular from NASA's Mariner 2 probe to Venus, that the solar wind was continuously observed at speeds of a few hundred km/s, close to what was predicted by Parker's theory. The density of the solar wind close to the Earth's orbit was found to be much less, about 7 particles per cubic centimetre, compared to 30 to 50 than in Parker's original theory. This discrepancy was mainly due to the simplifying assumption made by Parker that the temperature of the corona is constant. Since then, much has been learned about the solar wind. We know that it is a more complex phenomenon than treated by Parker, however, many of his conclusions remain valid, and in any case his thoughts have shaped the way we think about the solar wind and the heliosphere.

The Boundaries of the Heliosphere

If the solar wind flows from the Sun all the time in all directions, even if at speeds and densities that vary, its pressure can hold away the interstellar medium from the space around the Sun. Information about the properties of the interstellar medium are difficult to obtain, as no space mission has up to now reached it to provide direct information. We rely on indirect inferences about such parameters as the density, temperature, composition and magnetic field in the interstellar medium. Very sophisticated remote sensing techniques

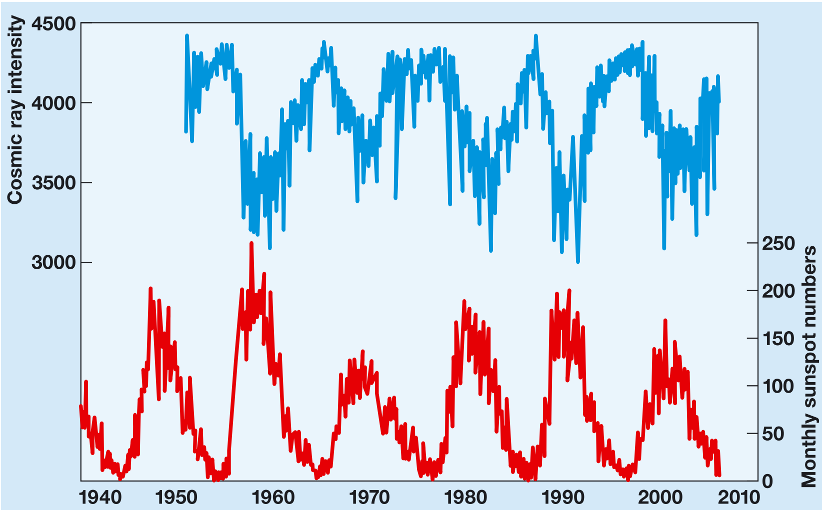


Figure 7: The 11-year variation in the intensity of cosmic radiation at Earth, in anti-phase with the sunspot cycle that implies a large volume of space around the Sun to which the access of galactic cosmic rays is somehow remotely controlled by the Sun.

have shown that even in the distant neighbourhood of the Sun the medium is not uniform but rather lumpy on very large scales. This means that calculating the size of the heliosphere from a balance of pressures between the solar wind and the interstellar medium is not very easy. Further below, the size of the heliospheric cavity will be estimated based on the best currently available estimates of the parameters of the interstellar medium.

Modulating the Cosmic Rays

The fourth and last early inference that a large volume of space was under the control of the Sun was made by Leverett Davis in 1956. His suggestion was based on the observation that the intensity of high-energy cosmic rays at Earth (see Spatium Nr. 11) increased when the number of sunspots was low and decreased when the number of sunspots was high. Sunspot numbers follow an 11-year cycle and are related to how "active" the Sun is. Activity on the Sun is measured in many different ways, but the way they change is closely related to the oldest indicator, the number of sunspots. There are generally more energetic outbursts from the Sun when the sunspot numbers are high, even though the outbursts are not directly related to sunspots. Almost certainly, underlying all measures of solar activity is the magnetic field of the Sun and the way the solar dynamo in the interior of the Sun is changing periodically.

The way cosmic ray intensities and sunspot numbers change in opposite directions is shown in **Figure 7.** How does this lead to the concept of the heliosphere? Cosmic rays are generated in the far reaches of the galaxy, in cataclysmic events such as supernova explosions; cosmic rays exist everywhere in the galaxy. As they reach the volume of space around the Sun, their propagation is impeded by the outward flowing solar wind and the diverse magnetic structures that are carried in the solar wind. If the solar wind and its structures change according the level of activity on the Sun (as they indeed do in response to the 11-year activity cycle), then cosmic rays will be more or less impeded in reaching the Earth. This means that when the Sun is in its more active state, with the largest number of sunspots, fewer cosmic rays can reach the Earth than when solar activity is low. Another way of looking at this is that at high solar activity a larger amount of energy needs to be expended by cosmic rays to reach the Earth, but as there are fewer cosmic rays of such higher energies, the number detected at Earth at such times is lower. Conversely, during low activity levels, lower energy cosmic rays can reach the Earth and as they are more numerous, their intensity increases.

This modulation of the intensity of cosmic rays in synchronism with the solar cycle can be explained by the existence of a large volume of space around the Sun in which the solar wind and its magnetic structures can impede the penetration of the cosmic rays that arrive from all directions in the galaxy. This is the volume of space around the Sun that Leverett Davis was the first to call the heliosphere.

The Size of the Heliosphere

The best indication of the size of the heliosphere was obtained when NASA's Voyager 1 spacecraft, launched in 1977, crossed the Termination Shock, one of its key outer boundaries, in December 2004, **see Figure 8.** The distance of Voyager from the Sun was then 94 Astronomic Units or about 14,100,000,000 km. This was a very important event in the history of heliospheric research, because after four or more decades of theoretical speculation, the measurement of the distance to this outer boundary provided a firm foundation to the calculations and modelling of not just the size, but also of our understanding of the heliosphere. Voyager 1 in fact made a set of important observations that will be described below; some of these are providing new questions because they don't fit the previously accepted theoretical mould.

Early estimates simply assumed that the flow of the solar wind, as it becomes rarefied the further we go from the Sun, will simply drop to a sufficiently low level of dynamic pressure which will balance the low pressure that the interstellar medium exercises to contain the solar wind. This is not only an oversimplified picture, but it is not conform to the important physical principles that describe the collision of a supersonic flow with another medium.

The Properties of the Interstellar Medium

In any case, the parameters of the interstellar medium in the vicinity of our Sun are not very well known. There have been many indirect estimates of the relevant parameters, but interstellar space is not at all uniform and has very large variations in its density, temperature, as well as the relative speed between different regions, **see the table below.** Primarily by examining starlight from different stars located in directions all around the Sun, it has been deduced that our distant neighbourhood in space, beyond the heliosphere, is rather emptier than interstellar space in general. This large and very rarefied volume, but with very high temperatures, is about 100 parsec in dimension (where 1 parsec, a unit normally used to measure astronomical distances, is 206,264.8062 Astronomic Units, or 3.1×10^{13} km, or yet 3.3 light years). But in fact, inside this bubble there are small irregular regions that are considerably cooler and denser than the average of the bubble. Even then, the Sun's immediate neighbourhood is described as a "warm, partially ionized diffuse interstellar cloud", this is the Local Interstellar Cloud or LIC whose properties define, together with those of the solar wind, the size of the heliosphere and the nature of its boundaries. The key parameters of the LIC are the density of neutral hydrogen atoms ($0.24/cm^3$), the density of electrons ($0.09/cm^3$), the ratio of ionized hydrogen (or number of electronless protons) to hydrogen

Interstellar Medium (ISM) Phases

Component	*Fractional volume (%)*	*Temperature (K)*	*Density (atoms/cm^3)*	*State*
Cold Neutral Medium (CNM)	1– 5%	50– 100	$1–10^3$	neutral hydrogen atoms
Warm Neutral Medium (WNM)	10–20%	1000–6000	$4 \cdot 10^{-4}$	neutral hydrogen atoms
Warm Ionized Medium (WIM)	20–50%	$10^3–10^4$	10^{-2}	ionized hydrogen
Hot Ionized Medium (HIM)	30–70%	$10^6–10^7$	$10^{-4}–10^{-2}$	highly ionized

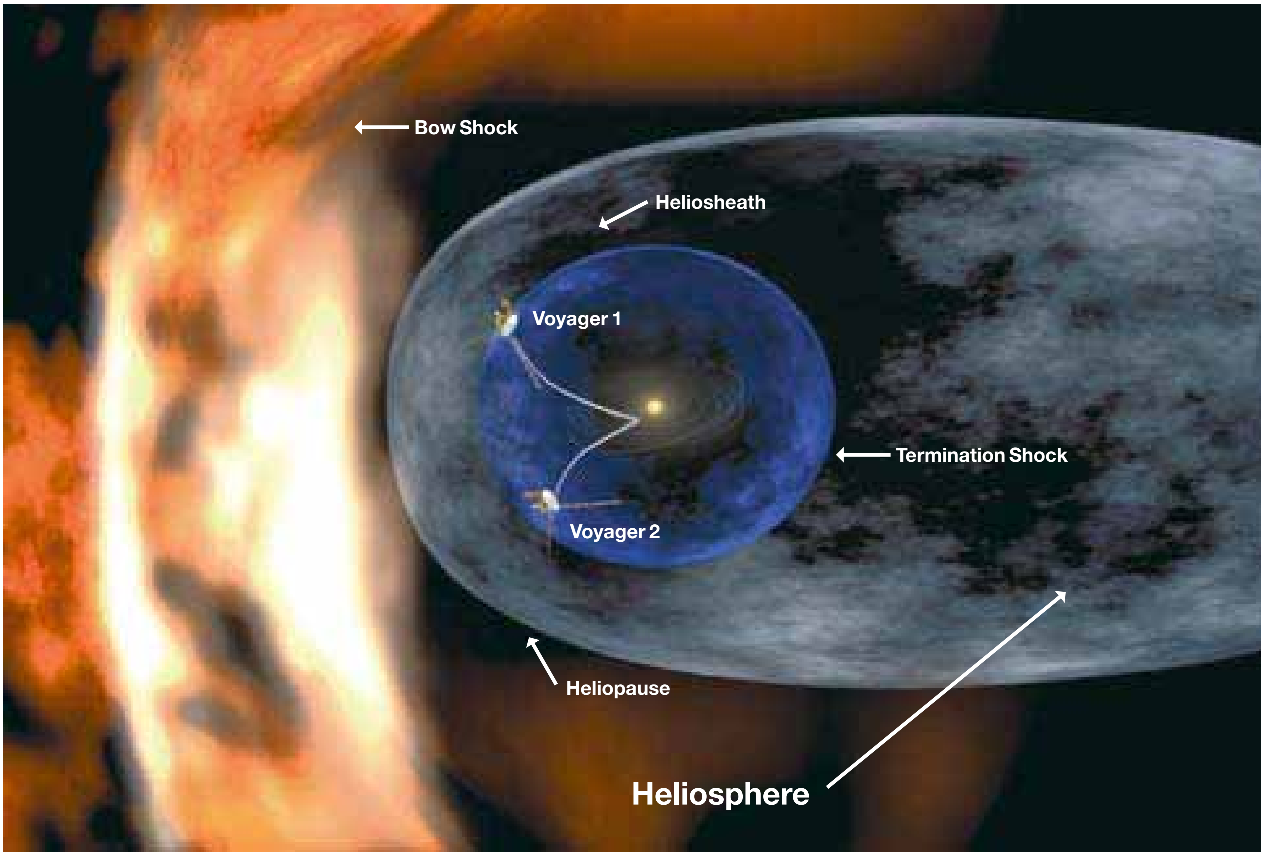


Figure 8: The boundaries of the heliosphere consist of the Termination Shock (where the solar wind slows down to subsonic speeds) and the heliopause (where the solar wind reaches the interstellar medium). In front of the heliopause there is the bow shock, where the interstellar medium is compressed by the heliosphere.

atoms (about 23%), the density of helium atoms (0.014/cm^3), the ratio of ionized to neutral helium (about 45%), the temperature (about 6,400 K, very similar to the temperature of the Sun's surface, the photosphere, but there are many orders of magnitude difference in their respective densities). Just for comparison, it is estimated that the temperature of the large local bubble is about a million degrees, but it consists mostly of very low-density ionized hydrogen, about 0.005 particles/cm^3.

Defining the Boundaries

For estimating the size of the heliosphere, using these parameters, we also need the physical principles of the interaction between the solar wind and the LIC. It has already been said that the solar wind is highly supersonic. As this flow runs into a near stationary medium, it needs to slow to subsonic speeds. In all kinds of flows in nature, the transition from supersonic to subsonic flow can only be a shock wave. There is a very large variety of shock waves, depending on the flows and the media in which the slowdown happens. Frequently quoted examples are supersonic aircraft, explosions, even the cracking of whips! In the astrophysical context, there are shock waves around planetary magnetospheres (where the solar wind slows down to flow around the obstacle) and extremely large shock waves generated by supernova explosions. In the case of the solar wind and the heliosphere, the shock wave is called the Termination Shock (this is the boundary that was crossed by Voyager 1 in late 2004) where the solar wind becomes subsonic, **see Figure 9.**

The Boundaries of the Heliosphere

Beyond the Termination Shock, the medium still consists of the now slowed down solar wind, until we reach the heliopause which is the boundary between the solar wind and the LIC. In space plasma physics, it is shown that plasmas of different origin cannot really mix (except in very special circumstances), so the LIC and solar wind plasmas have a distinct boundary that separates them. But there are complications. The LIC (unlike the solar wind) is not a fully ionized plasma, it contains neutral hydrogen and helium which are not subject to the laws of plasma physics. Many of these neutral atoms from the LIC penetrate into the heliosphere, from its distant boundaries all the way to the Earth's orbit where they have been observed, providing important information on the LIC. However, many neutral atoms are ionized when they encounter the solar wind (mostly by a process called charge exchange with the solar wind protons). Once ionized, these particles are "picked up" by the solar wind and can, if they are numerous, influence the local properties of the solar wind. Furthermore, such pick-up ions can be energized at the Termination Shock and, at high energies, become part of the cosmic ray population; as they are generally quite recognizable as a different population from the more generally observed galactic cosmic rays, they are called the "anomalous"

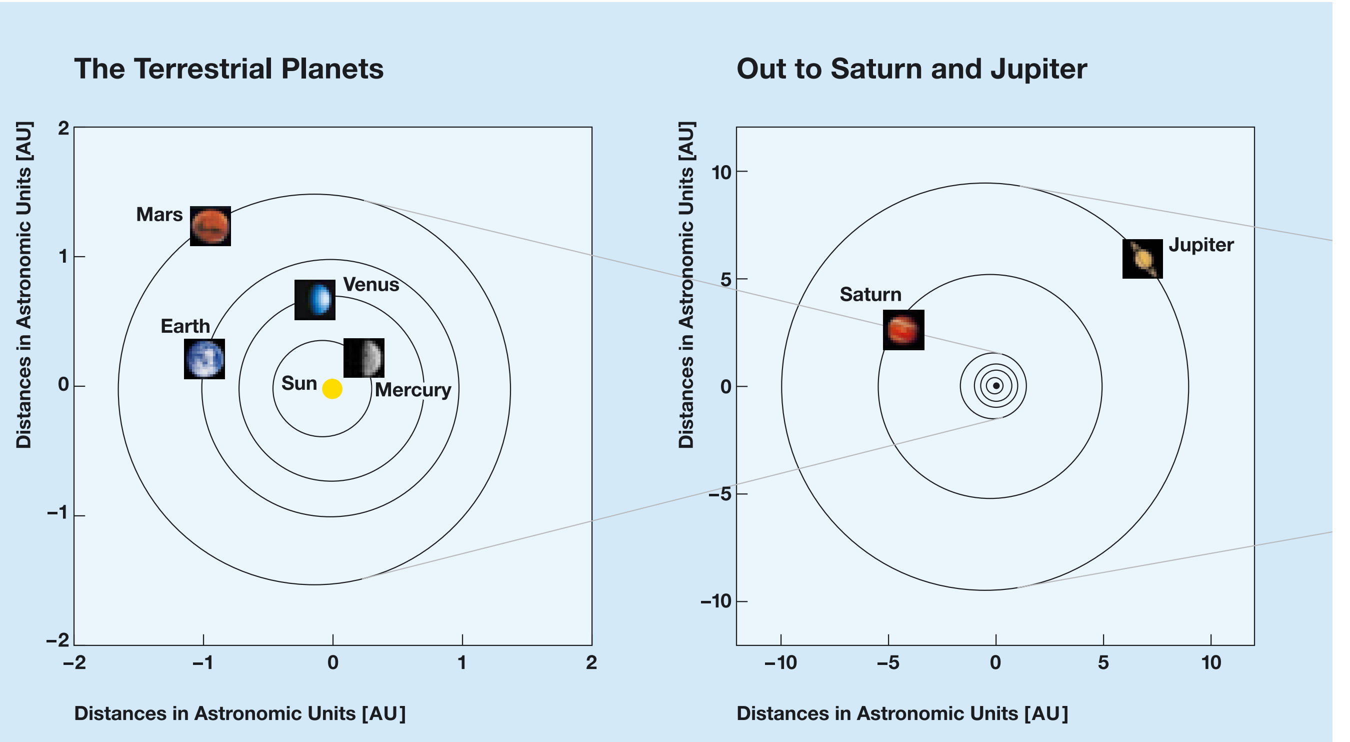
Figure 9: NASA's Voyager 1 spacecraft crossed the historical 100 AU milestone (equalling 15 billion kilometres) on 17 August 2006. (Credit: NASA / JPL)

cosmic rays. Just where these anomalous cosmic rays are energized has become one of the puzzles of the Voyager 1 observations after it crossed the Termination Shock, as this process does not appear to happen where expected from theory.

There have been many estimates in the past of the distance to the Termination Shock, all with many very tentative parameters. The very early estimate of Leverett Davis who could not know about the solar wind at the time (in 1955), but just made based on the modulation of cosmic rays, was 100–200 AU. Since then models with ever increased sophistication, but still with a limited certainty of the parameters, have suggested a distance of about 80 to 120 AU. The Voyager 1 observation at 94 AU falls well within the estimated range, thus justifying a posteriori many of the assumptions and parameters that have been used in the models.

Of course, the distance to the Termination Shock cannot be a constant, as the properties of the solar wind vary significantly, in particular with the 11-year solar activity cycle. As a result, the Termination Shock, as all other shock waves in space that we know, is in constant motion, moving in and out at speeds probably about 100 km/s. When Voyager 1 observed it, it was moving inward with about that velocity. The amplitude of the motion of the Termination Shock probably varies, depending on timescales: it makes large excursions (maybe as much as 10 AU or more) in response to solar cycle variations, but smaller ones in response to the always changing conditions in both the solar wind and, presumably, in the LIC as well.

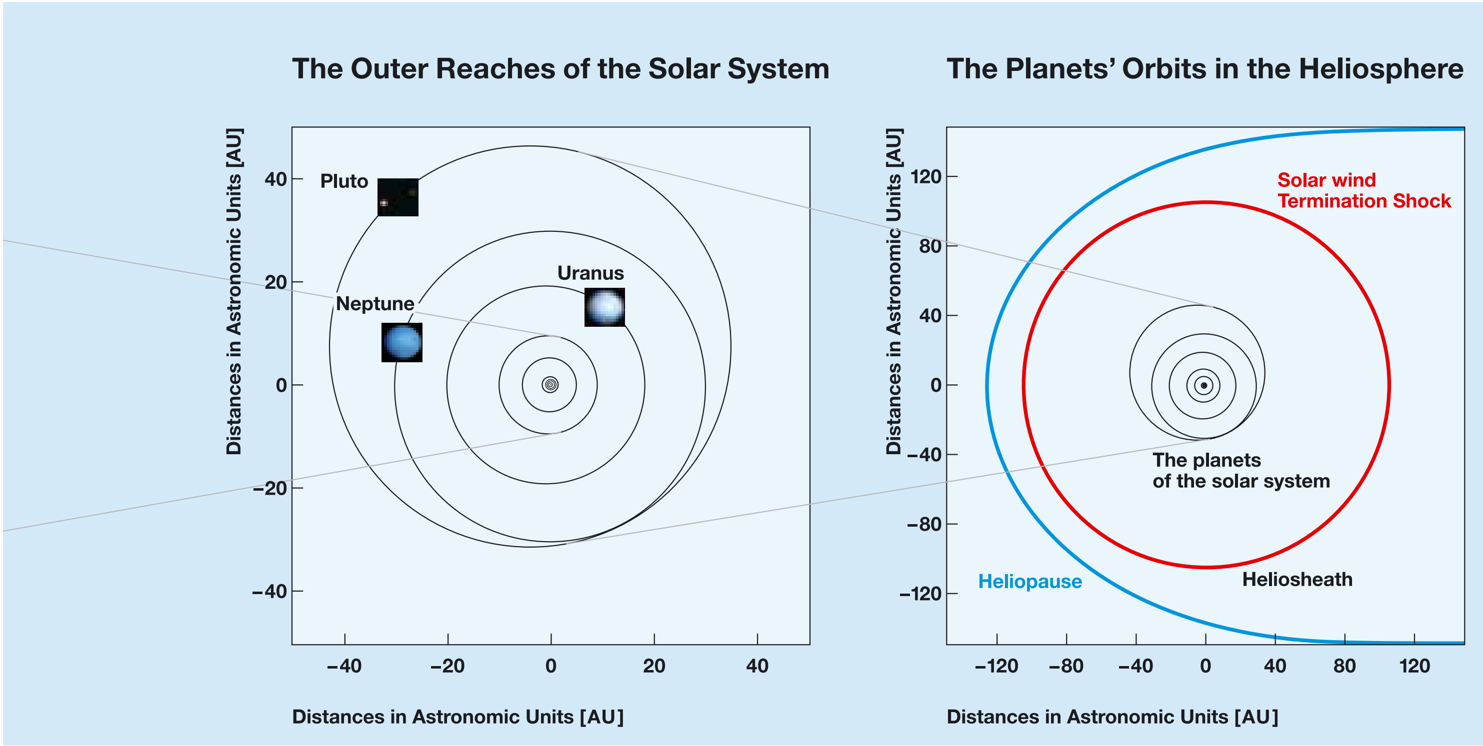
There are question marks concerning the existence of an outer shock wave, outside the heliopause and surrounding the whole heliosphere. This depends partly on the relative velocities of the Sun and the LIC, measured to be about 25 km/s, but partly also on the other parameters such as the density and temperature of the two colliding media. At this stage, scientific opinion is divided on this, while awaiting better observations of the relevant parameters.



By studying the relative motion of the Sun and the heliosphere in the local interstellar medium, it is estimated that occasionally, many times during the lifetime of the solar system, conditions in the LIC change significantly as large, cool and dense molecular clouds are encountered (with presumed densities of up to 10 particles/cm^3, or maybe 30 to 40 times greater than the present LIC). When that happens, the heliosphere is significantly compressed to perhaps less than a half or a third of its current size. How such conditions affect the Earth and its immediate environment, has not been really investigated, but clearly there will be noticeable consequences. A smaller heliosphere would lead to a weakened capacity to shield cosmic rays and therefore to a higher level of irradiation of the Earth by cosmic rays, which in turn would have important consequences on the Earth's climate and the evolution of life on the Earth. However, such changes are unlikely to happen on timescales measured by human generations; rather it is assumed that the heliosphere has been immersed in this cloud for the last perhaps 10,000 to 100,000 years.

It is interesting to compare the size of the heliosphere to the size of the solar system occupied by the planets. The Earth is well in the inner heliosphere, but really so are Jupiter (at about 5.5 AU) and Saturn (at about 10 AU). Even the gas giants Uranus and Neptune are well within the inner half of the heliosphere, while Pluto (lately defined a dwarf planet) is, when furthest away from the Sun, just about half-way out towards the Termination Shock. It can be safely said that the heliosphere, the empire of the Sun, is, on our human, solar-system-bound scales truly enormous. But then, on astrophysical scales, whether in the local bubble or in our galaxy, it is dwarfed by the vastness of the universe, **see Figure 10.**

Figure 10: This schematical image shows on successively increasing scales the inner solar system with the terrestrial planets, the gas giants Jupiter and Saturn, the outer solar system with Uranus, Neptune and Pluto, and finally the solar system in the heliosphere.



The Dynamic Structures of the Heliosphere

Having visited the outer reaches of the heliosphere and visited its boundaries, we now turn to the inner half of the heliosphere; perhaps even just the innermost tenth of it, out to about the orbit of Saturn. This is the part of the heliosphere that we know best and one that is most relevant to Earth-dwellers. The structure of the heliosphere in this inner region is wholly dependent on the Sun and the solar wind. There is a whole range of dynamic phenomena, mostly dependent on the 11-year solar activity cycle that characterizes the inner heliosphere.

Probing the Heliosphere

Thanks to a large number of interplanetary spacecraft, since Mariner 2 in 1962, the solar wind is a well known and well understood medium. Questions of course remain, but the key processes and phenomena, their dependences as a function of position in the heliosphere and of time have been well observed and described. Among the key space missions dedicated to heliospheric studies, special mention must be made of the two USA-German Helios probes in the mid 1970s, the Pioneer 10 and 11 probes, and the two Voyagers, 1 and 2. Helios remains to date the first spacecraft ever to have observed the solar wind in the innermost heliosphere, well inside the Earth's orbit. The Pioneers and Voyagers have explored the middle and outer heliosphere. Several spacecraft have observed the solar wind at the orbit of the Earth, from the early IMP series through ISEE-3 and now Wind and ACE. The Ulysses mission launched in 1990 and scheduled to operate at least until mid-2008 has been unique in its orbit over the poles of the Sun to provide observations of the three-dimensional heliosphere. These missions and the many scientific teams that have worked with them now over two generations have brought the knowledge of the heliosphere to its current advanced level.

The Solar Wind through the Solar Activity Cycle

The principal changes in the solar wind, as a consequence of changing solar activity levels, take place in three dimensions, mainly out of the ecliptic plane which contains the Earth's orbit and is within 7.25° of the solar equatorial plane. Ground- and space-based solar observations have shown that there are many important changes in the Sun and its corona that accompany the sunspot cycle. One of the

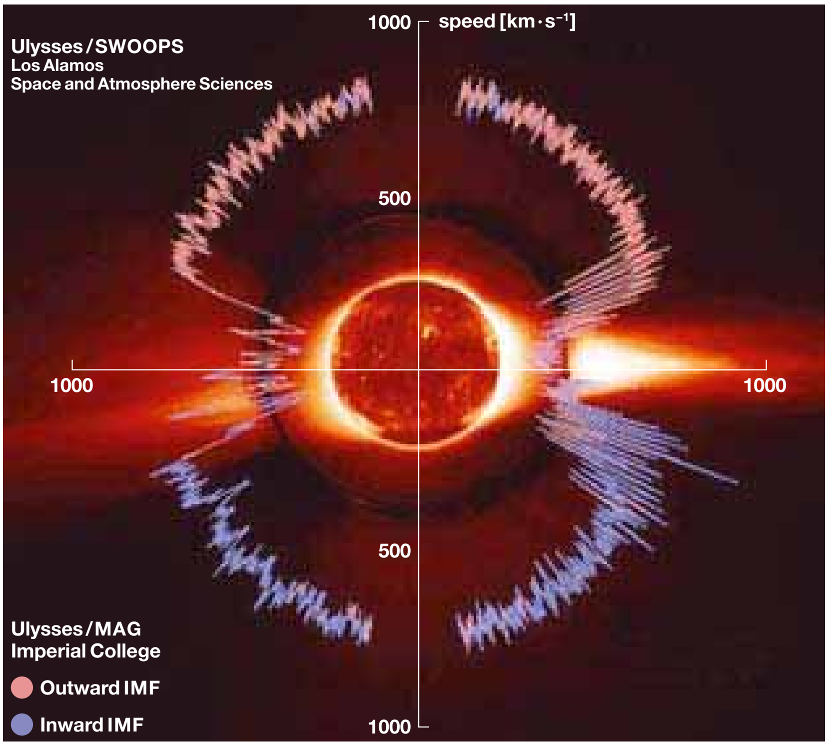


Figure 11: The solar wind profile as a function of solar latitude probed by the Ulysses spacecraft around solar minimum activity. The low speed portion of the solar wind is found mainly in the Sun's equatorial plain, whereas the high speed portion covers the higher solar latitudes. (Credit: Ulysses SWOOPS team, PI: D.J. McComas)

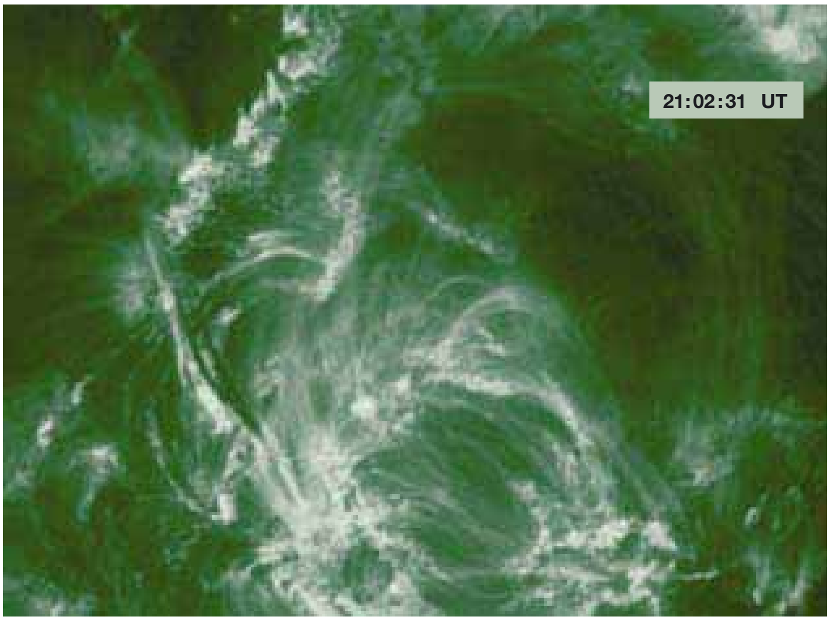


Figure 12: NASA's Transition Region and Coronal Explorer (TRACE) spacecraft discovered new features of the Sun's surface, termed "solar moss". The solar moss consists of hot gas at about two million degrees centigrade which emits extreme ultraviolet light. It occurs in large patches, about 6,000–12,000 miles in extent, and appears between 1,000–1,500 miles above the Sun's visible surface, sometimes reaching more than 3,000 miles high.

main discoveries concerning the solar wind in the 1970s was that there are two kinds of solar wind, fast streams (at speeds in excess of about 650 km/s) and slow solar wind (with speeds less than about 550 km/s), **see Figure 11.** Most of the parameters that describe the solar wind, such as its density, temperature and composition, are significantly different in the two kinds of solar wind, not just its speed. Also in the 1970s, solar observers identified dark, relatively cool regions in the corona as the origin of the fast solar wind streams, these regions are the coronal holes. One of the characteristics of coronal holes is that the solar magnetic field that they contain is open to interplanetary space, so that the magnetic field lines dragged into the heliosphere by the solar wind originate mostly from coronal holes.

The origin of the slow solar wind is less well understood, it is far more variable in its parameters than the fast solar wind, and it is likely to be generated near or at the edges of hot regions in the solar corona where the magnetic field is tightly held in complex loop systems, **see Figure 12.**

The magnetic field of the Sun that threads the solar corona is carried into space by the solar wind from the open-field regions, the coronal holes. The Sun, unlike the Earth, for instance, does not have simple magnetic North and South Poles. Around the minimum activity phase, there are large coronal holes covering the heliographic poles of the Sun, and the magnetic fields that emanate from these are of opposite polarities in the North and the South. This is the closest the Sun ever gets to showing a magnetic dipole such as the magnetized planets. As solar activity increases, these polar coronal holes shrink and fragment, so that there is apparently much less open magnetic flux and the polarities are much more mixed in the corona.

As the magnetic field is carried out in the solar wind, the magnetic polarities are separated by a so-called neutral line, separating the inward and outward pointing magnetic fields. In interplanetary space, the surface that separates the polarities is called the Heliospheric Current Sheet (HCS). Near solar minimum, this vast surface is close to the equatorial plane of the Sun, while near solar maximum, it becomes very complex and highly inclined with respect to the solar equator. Pictures of the solar corona, taken at the time of total eclipses **(see Figure 2)** or rather more routinely from a space-based observatory such as SOHO **(Figure 13),**

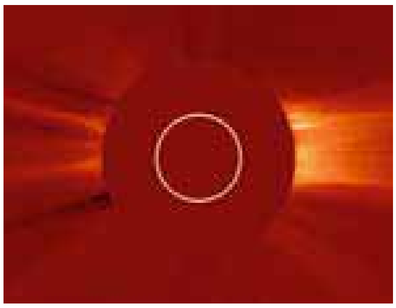


Figure 13: During solar minimum, the solar corona consists of bright streamers along the Sun's equator. Image taken by SOHO's EIT instrument. (Credit ESA)

show long, bright streamers that are confined near the solar equator near solar minimum, but point at high heliographic latitudes when the sunspot number is high **(Figure 14).**

The two kinds of solar wind are well delineated around solar minimum: all the fast solar wind comes from the large polar coronal holes and the slow wind from above the equatorial regions where the magnetic field remains in the form of loop systems. The HCS is always embedded in slow solar wind, just as the streamers are seen to originate above the hot loop system in the equatorial corona. A further effect that shapes the heliospheric medium is what is called the over-expansion of the fast wind: even though coronal holes above the poles have an angular extent of only about 30 degrees away from the poles of the Sun, the fast wind fills the inner heliosphere to much lower heliolatitudes, it is as though at least at the edges of coronal holes, the solar wind flows bend in a direction away from the solar poles. Around solar minimum, the polar coronal holes remain approximately the same for many periods of solar rotation.

At solar maximum, coronal holes are small and can be found everywhere on the Sun, not just near the poles. They are also generally short lived, often appearing and disappearing within a single solar rotation that is within a few days. As a result, the solar wind is rarely fast, but generally mixed, some not-so-fast wind and slow wind streams mingle at all heliolatitudes. The contrast between the solar wind at solar minimum and solar maximum is well illustrated by the observations of Ulysses.

Closer to Solar Minimum Activity: Corotating Interaction Regions

The formation of polar coronal holes, following solar maximum activity, is not a simple process, but involves the migration of magnetic regions of opposite polarities towards the heliographic poles. While these large coronal holes form, parts of them reach to low latitudes, towards the equator. This means that, at equatorial latitudes, both fast and slow solar

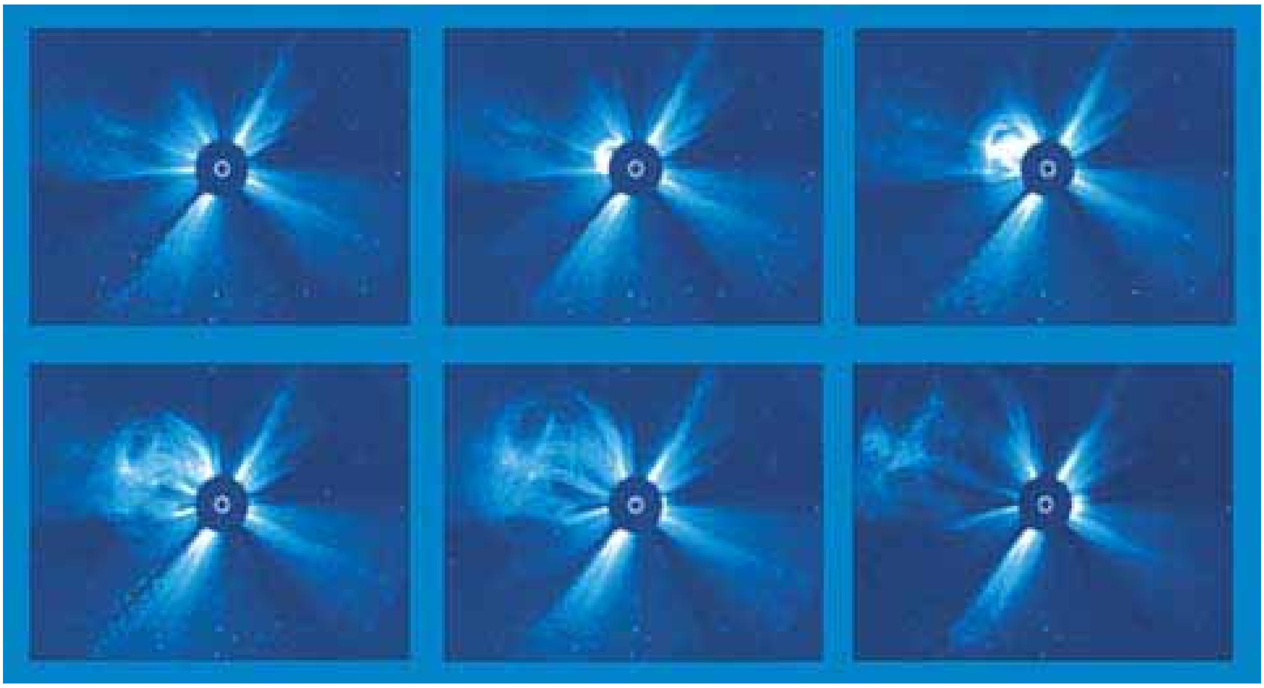


Figure 14: During solar maximum, the coronal streamers tend to point to higher elevations. In addition, large amounts of material are often ejected into space in the form of Coronal Mass Ejections. (Credit: ESA)

wind streams are generated in successive solar longitude ranges. As the Sun rotates, fast and slow solar wind streams are emitted alternately in a given radial direction. But then fast wind catches up with the slow stream ahead of it. In the interaction between fast and slow wind, the solar wind plasma is compressed and heated. As the compressed plasma travels out away from the Sun, it even forms travelling shock waves at its leading and trailing edges. If, as happens close to solar minimum, the flow pattern of fast and slow streams remains the same over successive solar rotation periods (each solar rotation is ~ 26 days long), the outward travelling interaction regions appear to rotate with the Sun, hence their name "Corotating Interaction Regions" or CIRs. Such CIRs persist for a year or more, usually in the period between solar maximum and minimum activity and constitute the major structuring process in the inner heliosphere. In addition, the shock waves that form at their leading and trailing edges constitute an important source of energetic particles in the heliosphere, not competing with cosmic rays, but still providing an important example of how energetic particles can be produced in the universe.

Around Solar Maximum Activity: Coronal Mass Ejections

Solar maximum activity has been historically measured by the number of sunspots. This remains a useful criterion, although we now know that sunspots themselves are only a symptom of considerable changes in the Sun's upper layer, the convection zone. Solar activity in fact increases in response to the complexity of magnetic fields emerging to the surface of the Sun; the increased complexity restructures the magnetic fields in the solar corona. In the course of this restructuring process, very large and occasionally explosive amounts of magnetic energy are transformed into sudden heating of the corona that expel very large amounts of coronal material into space, embedded in the solar wind. These events are Coronal Mass Ejections (CMEs); their number increases from close to zero at solar minimum to several per day close to solar maximum.

In many cases, CMEs appear to be closed magnetic structures, unlike the ordinary solar wind that has open magnetic fields embedded in it. CMEs, when directed towards the Earth, often carry enough hot plasma and strong magnetic fields to cause major disturbances in the terrestrial magnetic field; these events are the magnetic storms. The strongest of these storms, usually before or after solar activity maximum, depress the Earth's magnetic field and cause an increase in the intensity of radiation near the Earth in such a way that can damage spacecraft and even affect large sections of terrestrial power supply grids. The very large-scale CME "bubbles" occupy large volumes in the heliosphere around solar maximum activity (when the most powerful CMEs occur) and modify significantly the structure of the heliosphere. As the CMEs propagate out towards the outer heliosphere, they often amalgamate and the largest among them form what has been called a Global Merged Interaction Region (GMIR) which acts as a barrier between the inner and outer heliospheric regions and significantly impedes the access of cosmic rays to the vicinity of the Earth.

One of the effects of large-scale, probably merged CMEs is the effect they have on the outer boundaries of the heliosphere. Some intriguing radio noise observations by the Voyager spacecraft in 1983 and 1992/93 have been interpreted as the signs of radio emission from the heliopause when the large-scale CMEs reach it and disturb it. This provides yet another tentative measure of the distance of the heliopause: estimating the travel time of the CMEs that cause the radio emissions gives a distance of about 150 AU to the heliopause, a figure that is broadly in line with theoretical expectations.

Large-Scale Structures and the Modulation of Cosmic Rays

The cause of cosmic ray modulation is the increase in complexity of heliospheric structures from solar minimum to solar maximum activity. However, the precise processes that control the access of cosmic rays into the inner heliosphere are not fully understood. Almost certainly, several factors jointly play a role in this process. One such process is the formation of GMIRs; it has been noted that cosmic ray intensity decreases significantly, almost in a step-like way, after the passage of a GMIR. But the modulation process as measured at the Earth is smoother and almost certainly involves higher levels of turbulence (or simply disorder) in the heliospheric magnetic field around maximum activity. The access of cosmic rays is controlled by a combination of GMIRs, turbulence and the heliospheric boundary; the modulation of cosmic rays is caused by changes in these factors associated with the activity level through the solar cycle.

Models of the Heliosphere and its Boundaries

There are many models of the heliosphere that attempt to take into account the known characteristics of the solar wind as it propagates to large distances from the Sun and the characteristics of the Local Interstellar Cloud that are deduced from remote and indirect observations. The models tend to agree on the basic parameters, such as the approximate distance of the outer boundaries, the Termination Shock and the heliopause. However, the problem of modelling in detail is quite difficult, as there are several components of the solar wind and the interstellar medium to be brought into the model. Current modelling work is greatly helped by the increase in computer power that enables multiple parametric studies to be carried out. The nature of the boundaries and their effects remain uncertain, as was discovered by Voyager 1 recently, as described below.

One aspect seems generally accepted: this the Hydrogen Wall, a region in front of the heliopause in which there is a significant increase in the density of interstellar neutral hydrogen atoms. First predicted by models, the existence of such a wall has found strong support in the observation of an increase in the absorption of radiation, selectively in the hydrogen spectrum, from nearby stars, such as Alpha Centauri and Sirius. This observation could also be used to try and detect stellar winds similar to that of our Sun around other stars.

The Termination Shock and Beyond: Voyager 1 Results

Most models of the heliosphere estimated the distance to the first boundary, the Termination Shock, to be around 100 AU or perhaps somewhat less, dependent on the activity cycle of the Sun. The two Voyager spacecraft have been approaching such distances in the past few years, so the expectation has been strong that at least Voyager 1 (the furthest away) would meet the Termination Shock sooner rather than later. This significant event occurred in December 2004, at a distance of 94 AU from the Sun, when the instruments on board Voyager 1 unanimously indicated the crossing of a shock front, clearly the Termination Shock wave where the solar wind becomes subsonic. At first sight, the observations seemed to match the predictions, but as Voyager 1 moved further out into the heliosheath, the expected increase in the so-called Anomalous Cosmic Ray (ACR) population was not observed. The ACRs are interstellar atoms that have been ionized and raised to high energies, somewhere at or near the heliospheric boundaries; their existence has long been considered as a strong indication of the strength of the Termination Shock.

Previously, ACRs had been expected to be energized at the Termination Shock, in fact close to the part of the shock front which directly faces the flow of interstellar matter. A recent suggestion to explain these observations is that the geometry of the Termination Shock is not suitable for the energization process at the expected location, but that ACRs originate around the flanks of the heliosphere, where the geometry of the Termination Shock enables the energization to be carried out.

This result has shown the importance of observations to verify the theoretical models. Voyager's great achievement is that it has lasted the long voyage to the edge of the heliosphere: the 27 years it took from Earth to travel to this first boundary illustrates the scale of the heliosphere when measured against what our space missions can achieve.

The Future of Heliospheric Research

Interplanetary spacecraft have explored the solar wind and the heliosphere since the early 1960s. Of these many were in Earth orbit and some just ahead of the Earth in the direction of the Sun near a Lagrange point in space, a point where the Earth's and the Sun's gravity balance. These spacecraft have measured in great detail the properties of the solar wind and monitored the changes in these properties on all time scales. As a result, a very large data base is now available on the solar wind.

But the properties of the solar wind change as a function of distance from the Sun and as a function of solar latitude in ways that are usually not easily predictable. Space missions that have travelled towards the Sun or away from the Sun have contributed to our understanding of how the solar wind and the heliosphere changes as a function of distance away from the Sun. The still unique Ulysses mission, a joint undertaking between ESA and NASA, has been in its polar orbit around the Sun since early 1992, opening a completely new, and in many ways unexpected perspective to observe changes in the structure of the heliosphere with solar latitude.

Continued monitoring of the near-Earth solar wind as well as remote solar observations will extend the data base for heliospheric research. A few space missions are dedicated to such research. NASA's Stereo mission, to be launched in autumn 2006, consists of two identical spacecraft that will accompany the Earth around the Sun. They will follow the same orbit as the Earth, but one ahead, the other behind along the orbit. These two spacecraft will make observations of CMEs from two observation points, to build up a stereoscopic image of these events as they propagate away from the Sun, towards the Earth. This will allow three-dimensional models of the shapes and structures of CMEs to be verified and new models to be developed. Another mission that is under construction by NASA is the Interstellar Boundary Explorer (or IBEX) that will use a novel technique to study the heliosphere's Termination Shock and the heliosheath. In fact, IBEX will remain in Earth orbit, but will monitor energetic neutral atoms that are generated at the heliospheric boundaries and can travel unimpeded through the solar wind. By building up images of the intensities of such neutral atoms, the dynamics of the heliospheric boundary can be better understood. A more ambitious mission, ESA's Solar Orbiter, is also being planned that will combine a relatively close approach to the Sun (maybe a third of the distance between the Sun and the Earth) and moving out of the Sun's equatorial plane, following the example of Ulysses, but not aiming to reach polar latitudes. The objective of the Solar Orbiter, currently planned for a launch in 2015, is to combine remote observations of the Sun and the solar corona with in situ observations of the solar wind to link in more detail than before the corona with the heliospheric medium.

Possibly the most ambitious space mission in the planning is NASA's Interstellar Probe. No launch date has been set and there are some key technical issues to be resolved. The Interstellar Probe is attempting to shorten the time needed to reach the boundaries of the heliosphere and go beyond them, into interstellar space. In order to achieve this aim, a large solar sail, using the radiation pressure of the Sun's photons will be used to accelerate the spacecraft away from the Sun. The speed expected to be achieved is about 70 km/s, or four times faster than was achieved by the Voyager craft. This speed would allow the Interstellar Probe to reach a distance of 200 AU in about 15 years. Voyager reached the first boundary, the Termination Shock at nearly 100 AU, almost 30 years after launch, clearly the Interstellar Probe would bring a better, quicker answer to the questions concerning the boundaries of the heliosphere and just what lies beyond.

The Author

André Balogh is one of the many Hungarians who had to leave their country in the wake of the Hungarian revolution against the Soviet occupation in 1956. He first completed his secondary education in France and then went to Paris to study telecom engineering at the École Nationale Supérieure des Télécommunications. In 1964 he moved to England where he found an employment at the Imperial College in London as a Research Fellow of the then European Space Research Organization ESRO. There he first specialized in research of the Earth's magnetosphere and the nearby interplanetary medium. Capitalizing on his knowledge as a telecom engineer he soon became involved in the development of scientific instruments for space research, which led to his appointment as Principal Investigator for the magnetometers onboard Ulysses (launched 1990 and still working perfectly) and on the four Cluster spacecraft (launched in 2000). In parallel he became a professor for Space Physics in the Physics Department of Imperial College, an appointment which he held until his retirement in 2005 when he was appointed Emeritus Professor of Space Physics. André Balogh knew ISSI from his former visits in Bern participating in Workshops, International Teams and as a visitor. When he was offered the task of a part-time Director of ISSI he gladly accepted as this new appointment allows him to actively shape the space research programmes in his favourite field but also beyond.

Space has attracted the interest of the very young André already. Later, after his studies of engineering sciences, he decided to realize his youth's dreams and to become a space scientist. In the vastness of the universe, the heliosphere is of primary interest to André Balogh as this is the area which is accessible to space missions and can be explored by in situ observations. The complexity of physical processes involved in the creation of the heliosphere is another element challenging space scientists like A. Balogh. And last but not least, it is the sheer dimension of the heliosphere which makes this part of our environment so amazing: yet it is our extended home, still imaginable on the human scale.

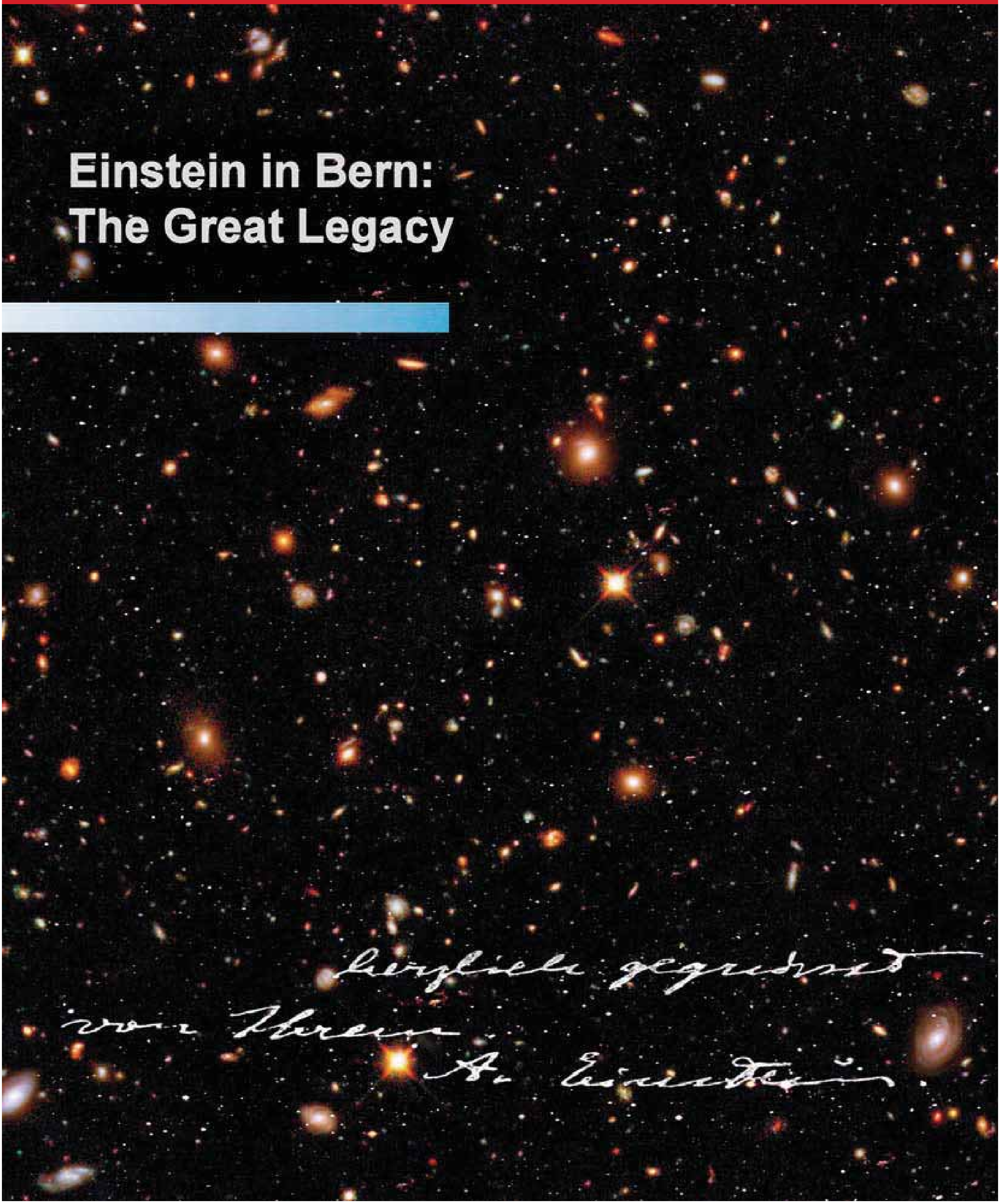
INTERNATIONAL SPACE SCIENCE INSTITUTE

SPATIUM

Published by the Association Pro ISSI

No. 18, February 2007

Einstein in Bern: The Great Legacy



Editorial

Omnia rerum principia parva sunt. The beginnings of all things are small.

This famous quote of Marcus Tullius Cicero is more than true for the middle drawer of an ordinary desk at the Kramgasse 49 in Bern. This drawer was called the office for theoretical physics by its owner, clearly a euphemism initially, but more than appropriate by the time when its contents prompted nothing less than a revolution of theoretical physics.

The desk belonged to the patent clerk of third rank Albert Einstein, who during the office hours had to treat the more or less ingenious inventions filed to the Patent Office, while in his spare time had set out to invent a new physics.

One might expect that such high-flying studies could at best occupy a few scientists in their laboratories. It was worse off: even the brightest representatives of the worldwide science community needed at least twenty years to fully grasp the epochal power of the patent clerk's ideas while for the great public Einstein's theories continue to stand for the inaccessibility of science.

Nevertheless, much has been said about Albert Einstein on the occasion of the hundredth anniversary of the annus mirabilis 1905 in Bern, the year, when the middle drawer of his desk became really *the* office for theoretical physics. Rudolf von Steiger, Professor at the University of Bern and Director of the International Space Science Institute together with Thomas H. Zurbuchen, Associate Professor at the University of Michigan have endeavoured successfully to translate the fascinating ideas of Albert Einstein for a larger audience, not just in Bern, but in many stations all over the world and to highlight some of the traces he continues to leave in our daily life. We are greatly indebted to the authors for their kind permission to publish herewith a revised version of their multi-media presentation.

Hansjörg Schlaepfer
Brissago, January 2007

Impressum

SPATIUM
Published by the
Association Pro ISSI

INTERNATIONAL SPACE SCIENCE INSTITUTE

Association Pro ISSI
Hallerstrasse 6, CH-3012 Bern
Phone +41 (0)31 631 48 96
see
www.issibern.ch/pro-issi.html
for the whole Spatium series

President
Prof. Heinrich Leutwyler
University of Bern
Layout and Publisher
Dr. Hansjörg Schlaepfer
CH-8173 Neerach
Printing
Stämpfli Publikationen AG
CH-3001 Bern

Front Cover
This composite image provides a glimpse to the remotest regions of the universe explored by the Hubble Space Telescope so far. The evolution of the cosmos is certainly a topic that Albert Einstein discussed with his two colleagues of the Akademie Olympia in Bern, Conrad Habicht and Maurice Solovine. The overlay image shows an excerpt of a message he wrote to the latter. (Credit: ESA, NASA, Hansjörg Schlaepfer)

Einstein in Bern: The Great Legacy[1]

Thomas H. Zurbuchen, University of Michigan, Ann Arbor, MI, USA
Rudolf von Steiger, International Space Science Institute, Bern, Switzerland
Brian Grimm, Paper Cardinal Design, Oakland, CA, USA

Introduction

The French philosopher and mathematician Henri Poincaré[2], in 1902, published a book entitled "La Science et l'Hypothèse". In this book he identified what were – in his opinion – the most important unsolved problems in science. One problem concerned the way light interacts with metal surfaces and obviously is able to eject electrons out of these surfaces. The second problem had to do with the apparently random zig-zag motion of pollen observed under the microscope, called Brownian Motion. The third problem was the failure of experimental physics to detect how light propagates, for example, from stars to astronomical telescopes. The assumption of a thin, nearly massless ether had been questioned by an experiment by Michelson and Morley some 25 years earlier that failed to find any evidence for such an ether.

Only a few years later a young patent clerk by the name of Albert Einstein had solved all three of these problems in a convincing manner.

Figure 1: Beautiful Bern. Looking over the roofs of downtown Bern to the hills of the Bernese Mittelland to the permanently snow-covered High Alps with the Blüemlisalp in the centre. (Credit: Bern Tourism, Bern)

[1] The present text follows a lecture by R. von Steiger in the Historisches Museum in Bern, 22 August 2006. Similar talks were held by T. Zurbuchen and R. von Steiger in over 20 locations world-wide

[2] Jules Henri Poincaré, 1854 Nancy, France – 1912 Paris, French mathematician, physicist and philosopher.

Figure 2: Bern, Kramgasse 49. It is here that the Einsteins lived between 1903 and 1905. From this house Albert Einstein revolutionized physics by his publications in 1905. (Credit: Einsteinhaus Bern)

Setting the Stage

The story plays in Bern, the capital of Switzerland, in the very heart of Europe. The historical roots of Bern date back to the La Tène time, the 5th to 1st centuries B.C. Modern Bern was founded by Duke Berchtold V von Zähringen in 1191. A legend tells us that he decided to name the new city after the first animal he would catch on a hunt; this was a bear, prompting him to name the place Bern. Bern is considered one of the most beautiful cities in Europe. It is located close to the Aare, a river originating in the Swiss Alps that brings clean water from the mountains to the city. Naturally, the U-shaped river bend was an attractive location for the city in medieval times, providing protection from three directions. Today, Bern houses approximately 150,000 people. To outsiders, Bern is known for its history, spanning many centuries, its bear pit housing the animal also found in the Bern flag, and its ambiance that is certainly unrivalled. To its visitors, Bern is often described as "gemütlich" – you immediately feel the warmth of its people, and its beauty.

In the old part of the city, in a house whose origins date back to the first city expansion in 1218, a story unravelled that was so groundbreaking and new that it still has effects today. In this house at Kramgasse 49[3] **(Figure 2)** Albert Einstein lived from October 1903 to May

1905. It was one of the seven different locations where he resided during his seven year stay in Bern.

For all practical purposes, things did not go very well during that time. He had finished his exams, allowing him to submit a doctorate at the University of Zurich. Out of the three graduating students applying for a doctorate, two were hired as teaching assistants, but Einstein was not. He was married to Mileva Maric and they had a child, so the sheer economical necessities forced him to accept a job at the "Amt für geistiges Eigentum" – the Swiss office for intellectual property, or, as we now call it, the Patent Office. He was hired as a clerk of third rank. The office hours were from 7 a.m. to 6 p.m. every working day and from 7 a.m. to 5 p.m. on Saturdays. Only Sundays were off. The Einsteins shared the kitchen and the bathroom with four other families, while their own apartment contained two rooms and a foyer. Nothing of this simple setting revealed that here one of the most ingenious human spirits ever silently was at work. And his objective was nothing less than to answer the three major open issues identified by Henri Poincaré three years before while dealing with such mundane items as shower heads or refrigerators, during his office hours. He did not even have unlimited access to libraries as he generally worked in the patent office during the time when the library was open.

We will now concentrate on what Einstein wrote in three of his most important papers, but let's not forget how, under what conditions he lived when writing these papers.

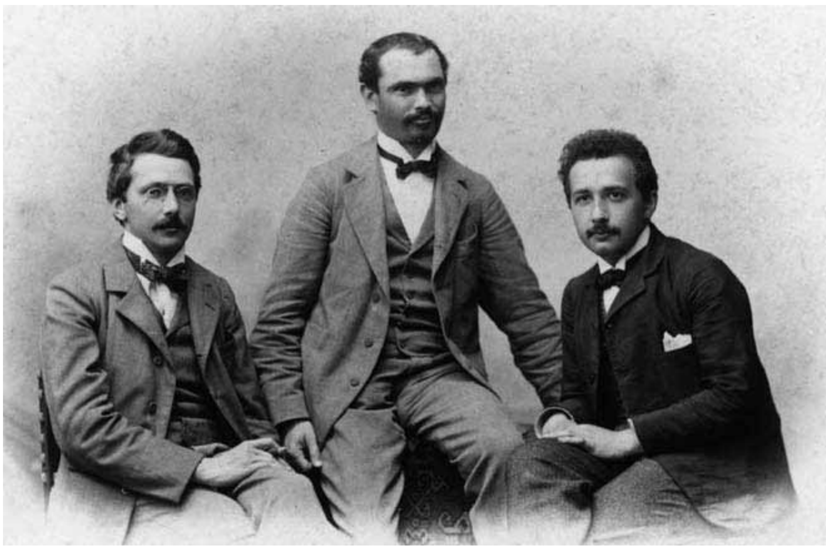


Figure 3: The members of the Akademie Olympia. From left to right: Conrad Habicht, Maurice Solovine and Albert Einstein. (Credit: ETH-Bibliothek, Zürich)

[3] Today, this is the location of the Einstein house.

The Great Scientific Trilogy

The genius could not be halted by the measly setting. Rather he discussed the ardent problems together with his colleagues of the Akademie Olympia, Conrad Habicht and Maurice Solovine **(Figure 3)**, and then went back to his desk, the middle drawer of which he called his office for theoretical physics, and put his thoughts to paper.

The Photoelectric Effect

In March 1905 Albert Einstein published a short article in the *Annalen der Physik* entitled *Über einen die Erzeugung und Verwandlung des Lichts betreffenden heuristischen Gesichtspunkt,* a paper that in 1921 earned him the Nobel Prize. At that time, all physicists knew what light was. Whether from stars, the Sun, or from radio antennas, light clearly propagated as a wave. Just like sound, light can propagate around the corner – you can easily see this for yourself. Just drop a coin in a mug, then back away until you lose sight of the coin. Next fill the mug with water. Voilà, the coin reappears though it still rests on the bottom of the mug. Light propagates very much like sound, bending, adding, and subtracting, and very much behaving like liquid waves on a lake or sound-waves in air. Just like sound-waves in air, light-waves were assumed to propagate in a medium, the so-called ether, which spanned the universe.

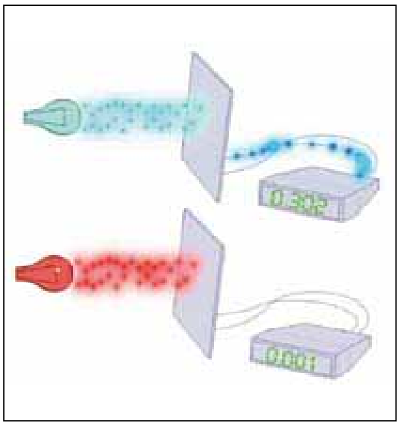


Figure 4: The photoelectric effect. The experimental setup demonstrating the different properties of blue and red light respectively.

There was, however, a problem with this notion as illustrated in **Figure 4.** The experimental arrangement consists of two lamps, one red and one blue, shining on a metal plate. The metal plate is hooked up with a wire, so that electric current from this plate can be measured. Electric current basically is made out of a bunch of moving electrons, negatively charged particles. Now, without any lamp on, there is a very slight current, almost nothing at all. Then, the red lamp is turned on. The current increases only slightly, but does not become really strong at all. The blue lamp has exactly the same wattage and therefore emits exactly the same power as the red lamp. But when this blue lamp is turned on, the current really jumps. What happened? Although we irradiated the plate with exactly the same power, the blue light causes the electrons to leave the metal plate, while the red light has practically no effect.

Einstein's explanation shook the understanding of light at its basic roots. He argued that light is made of particles, called photons. Each photon has an energy content which is determined only by its colour. So, with that new concept, the explanation of this intriguing experiment becomes quite simple: when the photons of the light sources hit the metal surface, they interact with the electrons in the metal. A certain minimum energy is required for an electron to be freed from the metal and to contribute to the electric current. The red light is made from lower-energy photons in larger quantity whose energy content is not sufficient to free the electrons in the metal. This is why the red light does not cause a significant increase of current. In contrast, the blue photons, when hitting the surface, cause an electric current to flow, as they possess enough energy to free electrons in the metal plate. Note that this effect works no matter how dim the light may be. Blue light particles do simply have sufficient energy, and increasing the brightness just brings more of them out of the metal and thus further increases the current.

Einstein's interpretation was highly controversial at that time. Based on

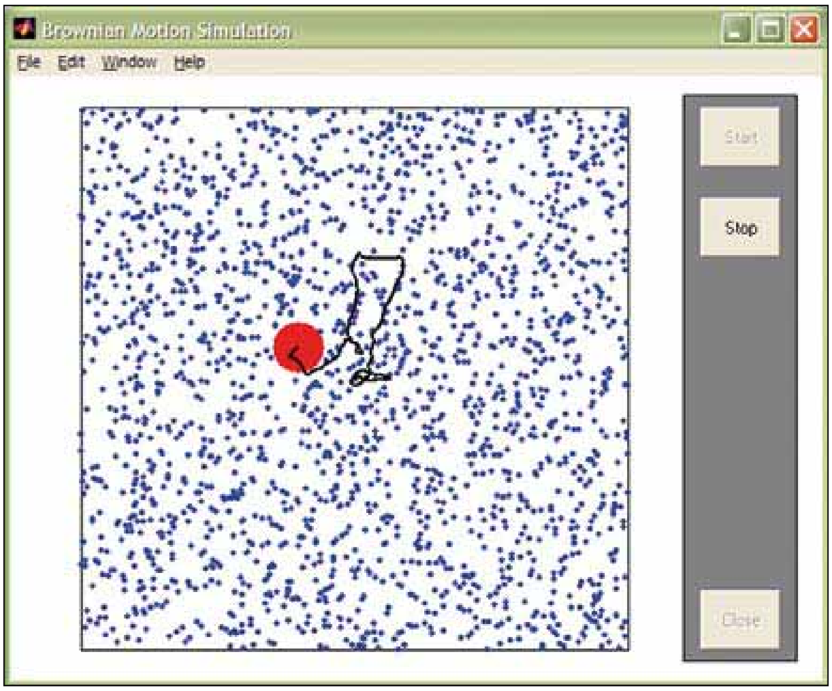


Figure 5: Brownian Motion. The British botanist Robert Brown observed in 1827 that the pollen of plants suspended in water was moving around in a random way. This sketch shows what might have been the path of one such pollen, bumped by billions of collisions with the smaller and hence invisible water molecules in the solution.

[4] Robert Andrews Millikan, 1868 Morrison, Illinois – 1953 San Marino, California, American physicist, Nobel Prize 1923.

a series of experiments, Robert Millikan[4] validated all predictions made by Einstein relative to the photoelectric -effect, but he came to the conclusion that *despite the apparently complete success of the Einstein equation for the photoelectric effect, the physical theory on which it was designed to be the symbolic expression is found to be untenable.* Einstein's theories continued to be controversial and only through the wise foresight of the Nobel committee did Einstein get the ultimate recognition for this paper in 1921.

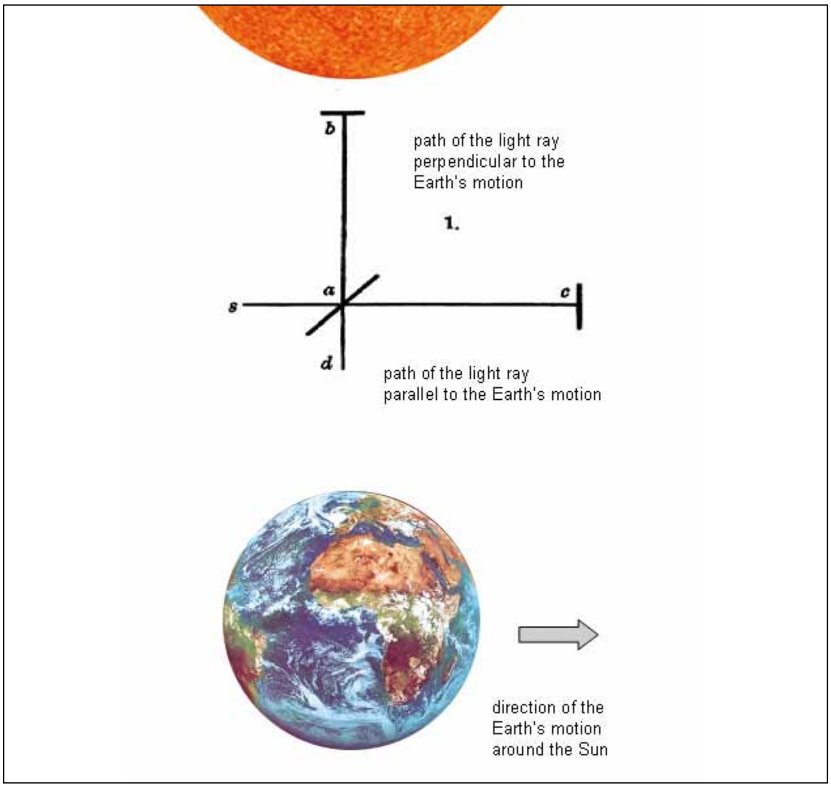


Figure 6: The Michelson-Morley experiment Before Einstein, it was assumed that light propagates in the ether at rest in the universe. Therefore, the light ray parallel to the Earth's motion (a-c) should return faster to the semi-reflective mirror as compared to the perpendicular light beam (a-b). But the experiment proved otherwise.

Brownian Motion

Only two months later, in May 1905, Albert Einstein solved the second of Henri Poincaré's great challenges. The Scottish botanist Robert Brown[5], during extensive work with his microscope some seventy years earlier, had made a strange observation. The pollen of *Equisetum* suspended in a water solution untiringly rushed around in a random zig-zag motion, **Figure 5**. Such observations had been made prior to Brown, but were generally interpreted as showing that the pollen grains were really little animals swimming in the water. Brown was aware of this and therefore performed a series of experiments that involved sprinkles of glass, rock and other inorganic material in the water. No matter what type of small particles he used, Brown observed the very same motion in all experiments. Therefore he had to exclude the earlier interpretation and progress to a notion that the motion was not of biological origin, but had its roots in the physical properties of water and its interaction with these grains.

Einstein's view of Brownian Motion is surprisingly straightforward: the suspended small particles are constantly bombarded by the water molecules which in turn are much smaller and therefore not visible under the microscope. If a water molecule had the diameter of a tennis ball, the size of the pollen would be approximately 100 metres. These tiny water molecules restlessly strike the particle and push it continuously. Billions of such collisions occur every second. Random fluctuations may cause a whole convoy of particles to hit the pollen knocking it this way and some time later it may experience a push from another direction and so on, as shown in **Figure 5**.

Einstein solved the problem by interpreting water to be made of small particles and not, as was assumed before, a jelly-like, continuous liquid. It was already known at the time that many issues in thermodynamics, the part of physics that describes how heat is created, ex-

[5] Robert Brown, 1773–1858, Scottish botanist, discoverer of Brownian Motion.

changed and transported, could be explained in the context of average properties of particles, bumping into each other and moving around. On the average, the model of these interacting particles describes quite well how heat behaves. However, this picture was largely assumed to be a model only and leading scientists refused to assume that there would be any departures from that average behaviour on the level of the individual particle. Einstein's interpretation established two very important new aspects. Firstly, the fact that water consisted of small, individual particles was established firmly. While this idea was brought up long before Einstein, his interpretation of Brownian Motion established this new notion without any doubt. Secondly, and much more importantly, his interpretation of the statistical nature of fluids opened the door to the modern, statistical description of particles. On the average, things are pretty predictable, but for an individual particle, things become random and therefore unpredictable.

Special Relativity

One month later, in June 1905, Albert Einstein addressed the theory of relativity, which of the three problems is perhaps farthest away from our everyday experience. In contrast to the first two topics, Einstein based his reasoning not on a particular experiment, but rather he was motivated by the problems scientists had with light and its propagation. At the time, physicists thought that light, just like sound in air, propagated in a medium, generally called ether, as we have seen above. That ether would presumably sit at rest in the universe. This hypothesis was tested by Albert Michelson[6] and Edward Morley[7] in Cleveland in 1887, **Figure 6**. The basic idea of their experiment is quite straightforward: the Earth orbits around the Sun at a speed of about 30 kilometres per second, or 1/10,000 of the speed of light. The light beam in a laboratory directed along the path of the Earth's mo-

Figure 7: The heart of downtown Bern. The Kramgasse, where the Einsteins lived between 1903 and 1905, seen from the Münster tower (Photo: Frank Rutschmann).

tion should speed up or slow down in the ether at that rate depending on the actual direction while its speed should not be affected by the Earth's velocity when it is directed perpendicular to the Earth's motion. The outcome of this experiment was very disappointing: the speed of the light was found to be independent of the direction relative to the Earth's motion. While this experiment was a failure in the eyes of many scientists, it was an inspiration for Einstein.

To understand Einstein's reasoning, we have to introduce the concept of a co-ordinate system. A co-ordinate system is a three-dimensional map allowing points in space to be brought in relation to each other. A co-ordinate system attached to the Earth can be used, for example, to measure distances between two cities. We may centre the co-ordinate system, for example, in Bern. Then, Zurich is some 80 km to the north-east. As we are free to centre our co-ordinate system where we like, we might centre it at the Sun constantly oriented relative to the stars. In this co-ordinate system, the Earth is roughly 150 million kilometres away circling the Sun approximately in one year. So far, so good and simple. But now we can add time to the picture by labelling each point of space with time by letting a clock run there. By doing so, our co-ordinate system has become four-dimensional: three space dimensions and one time dimension, but it continues to be quite simple. The question Einstein asked himself is how events in one system can be transferred into another co-ordinate system, or more specifically, how physical experiments in one co-ordinate system relate to physical experiments in another system which is moving relative (this is why we speak of relativity!) to the first one with a constant speed. Such systems are called inertial co-ordinate systems. Of course, the results of the ex-

[6] Albert Abraham Michelson, 1852 Strelno, Poland – 1931 Pasadena, California, German-American physicist, Nobel Prize 1907.
[7] Edward Williams Morley, 1838 Newark, New Jersey – 1923, American scientist.

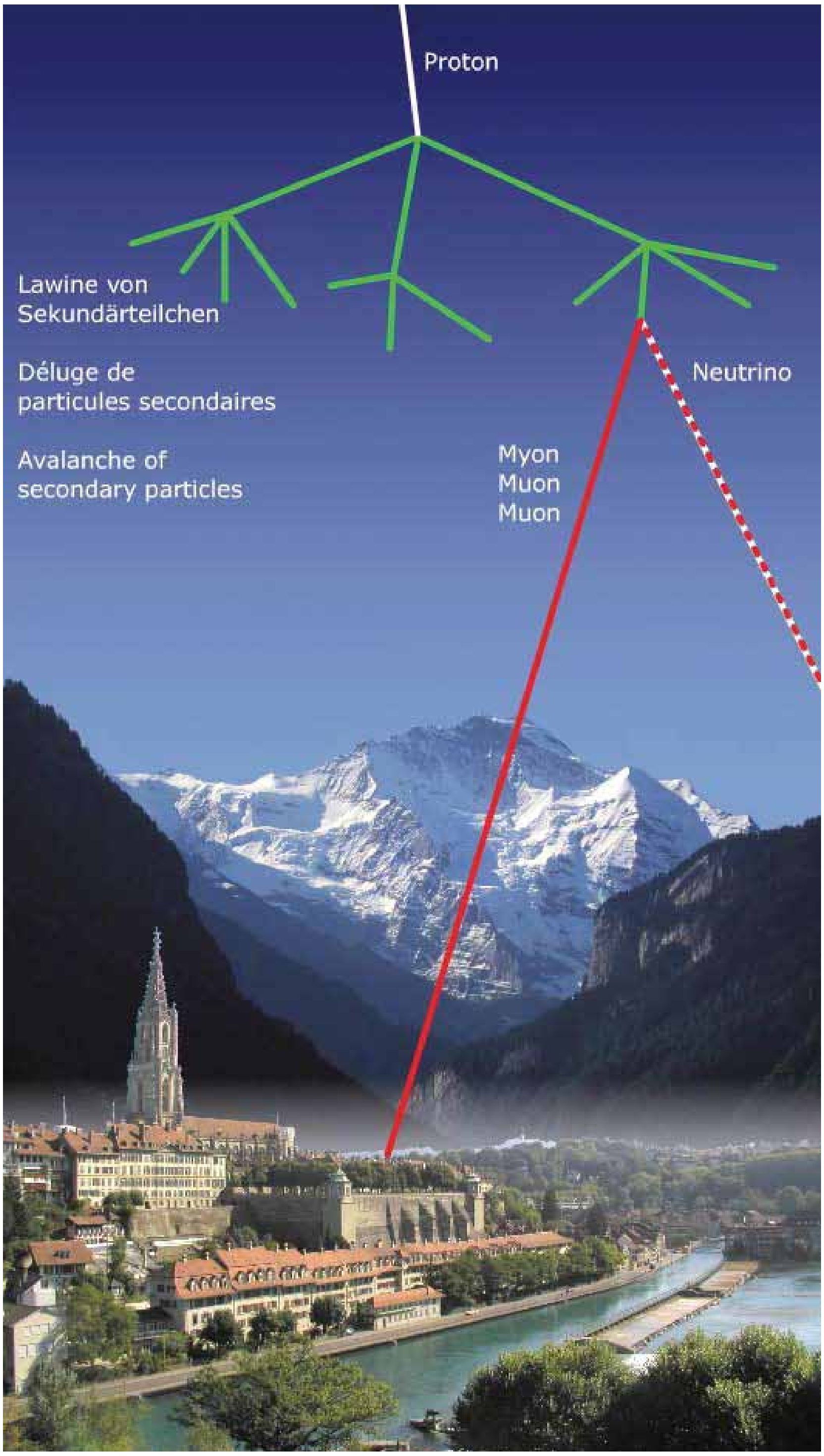


Figure 8: Demonstrating the relativity of space and time. The fast moving muons experience the dilatation of time that allows them to travel farther than their limited lifetime would suggest. (Credit: BHM)

periments should be independent of the co-ordinate system much like the distance between Bern and Zurich must be independent of the co-ordinate grid we use. This, however, is particularly tricky for a field of physics called electrodynamics, the science that describes the propagation of light as well as the electrical and magnetic phenomena, like for example the processes on the surface of the Sun or the behaviour of the European electrical network.

Einstein focused his analysis on two principles and their consequences. The first principle is the Principle of Relativity: the laws of physics are the same in each co-ordinate system, or the other way round: one cannot distinguish between (inertial) co-ordinate systems based on physical experiments. The second principle is a direct consequence of the negative result of the Michelson-Morley experiment. It simply states that the speed of light is absolute and constant in all co-ordinate systems.
When defining these principles, Einstein was firmly guided by observable phenomena: time is what you can read off a clock. Clocks can be synchronized only by exchanging signals at finite speed such as the speed of light. The conclusions drawn by Einstein are truly amazing: if the speed of light is absolute, time and space must be relative!

We can verify this striking conclusion by an experiment, see **Figure 8**. It starts in the depths of space when an exploding star accelerates particles to very high speeds, nearly to the speed of light. Some of these so-called cosmic rays enter the upper atmosphere of the Earth and collide with air particles. These collisions create a whole shower of new and mysterious particles; most of them collide with other air molecules or decay very quickly so they don't make it very far. Some particles in the shower, however, the so-called muons, interact only weakly with the air molecules. They can travel therefore all the way to the ground. Muons are the heavy cousins of electrons, but while electrons have an unlimited lifetime, the lifetime of a muon is a mere 2.2 microseconds. So, we expect no muon to make it from the upper atmosphere down to the ground, as it will decay before getting there, even when propagating at the speed of light. More specifically, we expect the muon to travel no more than 660 metres before decaying. But our experiment proves otherwise!

In order to identify the muons, we use spark chambers. These are stacks of metal plates separated by thin gaps with a high voltage across. When a muon passes through, it leaves a trace of visible sparks in the chamber, which allows us to notice its presence. Now, we set up one such spark chamber at the Jungfraujoch at an altitude of 3,580 metres and the other in Bern at 560 metres. While we expect the spark chamber on the Jungfraujoch to see many muon counts thanks to its high elevation, no counts in Bern are expected as it is deeper in the atmosphere than the muon's expected travelling distance of 660 metres before decaying. In the experiment both chambers happily count muons, the one in Bern a bit fewer than the one high up on Jungfraujoch. So, apparently about half the muons counted at the altitude of Jungfraujoch make it to Bern without any problems and are seen as small sparks in this chamber.

What happened? We argued before that no muon can travel further than 660 metres even at the speed of light before decaying. And still they travel through the altitude difference of 3,020 metres, five times as much. Einstein's notion of relativity helps us to explain our findings. We can do so from two different points of view, either in a co-ordinate system fixed to the Earth, or in a co-ordinate system fixed to the flying muon. In the Earth-fixed co-ordinate system, we know that the distance from Jungfraujoch down to Bern is 3,020 metres, and the particle's speed is roughly the speed of light. Due to the muon's high speed, Einstein argues that the time as experienced by the muon slows down dramatically. That means, that, as seen by our clocks, the muon's life is extended or, in Einstein's terminology, time is relative. By translating from the muon's co-ordinate system to our coordinate system attached to the Earth, time is dilated.

On the other hand, we can look at this experiment as if we were sitting on the muon. Again, we know that the muon lives a mere 2.2 mi-

Figure 9: The world-famous Zytglogge tower. The tower itself was part of the original city walls at the beginning of the 13th century. The great bell was cast in 1405. The clock is more recent, its mechanism dates back to 1530. (Photo Hansjörg Schlaepfer)

croseconds on average. Its speed is nearly the speed of light. But now, what happens with the distance? Einstein again explains this. As seen by the muon, the distance experienced between Jungfraujoch and Bern contracts. As Einstein puts it: space is relative also. Thus, neither space nor time remain absolute, but rather they are relative to their state of motion. What remains absolute is the speed of light.

Einstein and Us Today

As stated above, Albert Einstein's interpretations were highly controversial at his time. So it is no wonder that it took many years to fully grasp their content. Eventually though, the strengths of his ideas became obvious and scientists and engineers began to exploit their potential in favour of our everyday world. Each one of the three papers discussed here marked the beginning of a whole new field of physics. We will now look at some examples where his contributions are fundamental.

Photo Effect → The Laser

Einstein's explanation of the photoelectric effect prompted scientists over the next two decades to develop quantum theory, and engineers endeavoured to manipulate these tiny components of light. The laser is just one device which exploits the notion of light quanta. Incidentally, Einstein himself laid the groundwork for the laser through his later discovery of stimulated emission in 1918. The term laser is an abbreviation of *Light Amplification by Stimulated Emission of Radiation*. This type of light – which does not exist in nature – has a wealth of applications today, such as for example, transmitting billions of bits per second between spacecraft, or as scalpels that make extremely sharp cuts with much less bleeding than conventional scalpels. They can also be used to burn and evaporate tumours. The laser seen in **Figure 10** is a carbon dioxide laser operating at a wavelength of 10.6 micrometres. This is in the invisible infrared – the red ray is merely a small conventional laser to guide the operator. In the picture you can see Dr. Berchtold von Fischer from the Lindenhof hospital in Bern operating on an apple. He normally operates on breast cancer and has co-developed a new scheme that allows sparing as many lymph nodes as possible, thus reducing post-operation complications considerably. This is only one of countless applications today that can trace their source back to Einstein's photon hypothesis.

Brownian Motion → Nanotechnology

Brownian Motion gave Einstein the hint that matter is indeed composed of atoms. An atom is extremely small. Most of the things we use in our everyday life have sizes of metres, like our bodies, or a fraction of millimetre like our hair. For the much smaller world at Brownian Motion, microscopes allow us to ac-

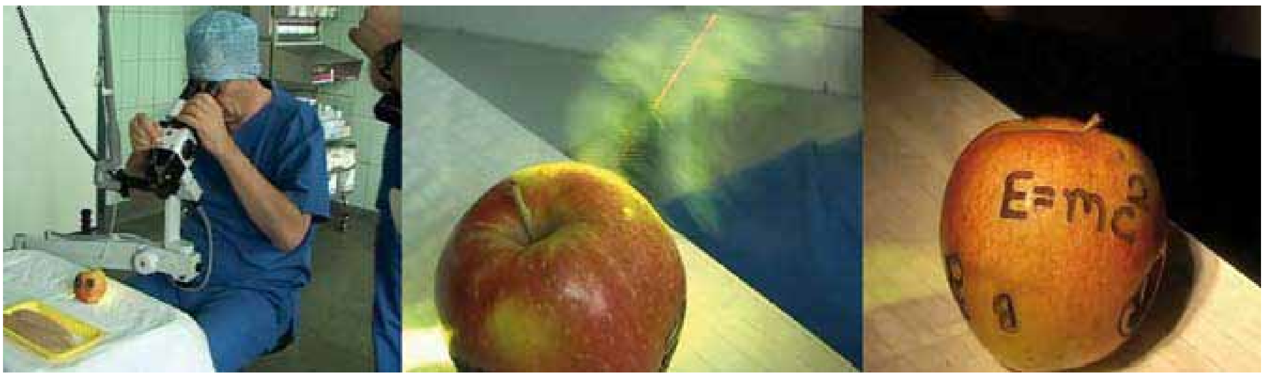


Figure 10: Einstein in the operating theatre. The coherent light of lasers finds an overwhelming variety of applications today. This sequence shows a surgeon demonstrating a carbon dioxide laser writing Einstein's famous formula on the skin of an apple.

tually see how cells evolve and how viruses attack; but the most amazing microscope ever developed is the Scanning Tunnelling Microscope (STM) which won the Nobel Prize for Gerd Binnig[8] and Heinrich Rohrer[9] in 1986. A very small tip, made only of a few atoms, is moved very close to a material. Then, some electrons can jump (or "tunnel") from the material to this tip and are detected as a current, see **Figure 11**. The closer the tip, the more electrons can jump the gap. Atoms are like bumps in materials and the tip, moved very accurately, allows one to measure a landscape this extremely small. Let's just quickly get a sense of the dimensions we are dealing with here: Washington, D.C. is roughly 4,500 km away from Los Angeles. Stretching a platinum wedding ring from coast to coast, from Los Angeles to Washington, would lead to a distance between two neighbouring atoms of just one centimetre. On this scale the thickness of a hair would be roughly 3,000 metres.

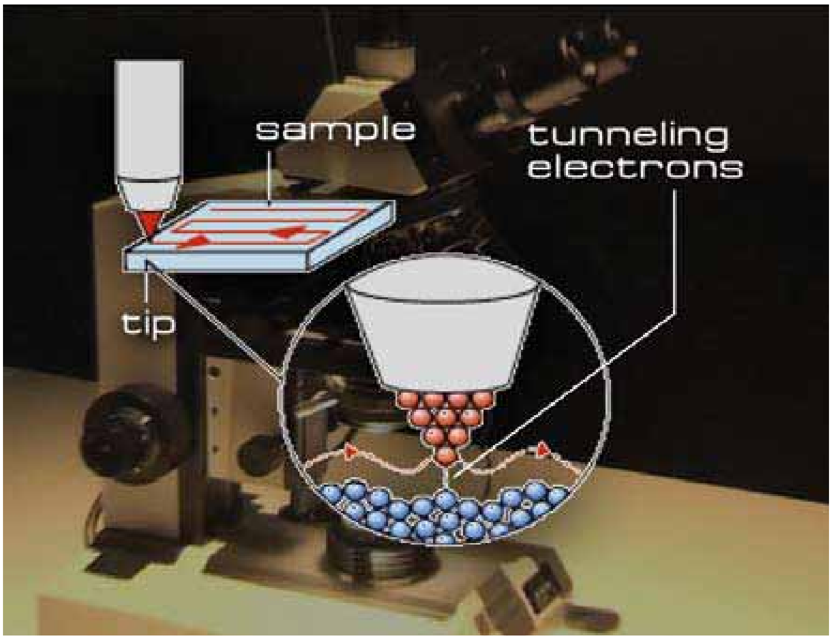


Figure 11: The principle of the Scanning Tunnelling Microscope. A very small tip only a few atoms wide is moved closely over the surface of a sample. With decreasing distance of the tip from the surface an increasing number of electrons become able to jump to the tip and be detected in the form of an electric current. This allows generating an image of the sample's surface.

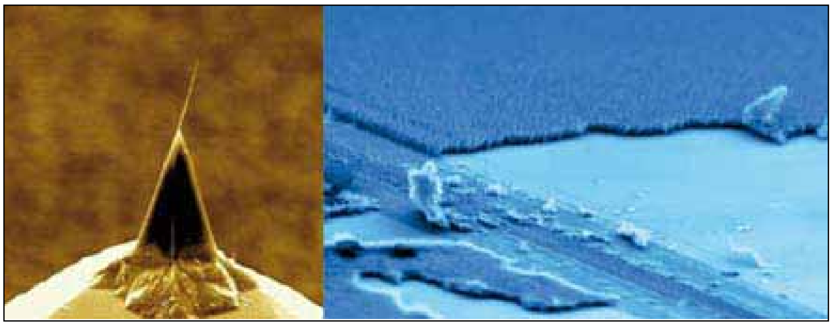


Figure 12: Glimpses into the nanoworld. On the left side, a carbon nanotube is depicted. Its thickness is a mere 10 atoms. The image to the right shows the structure of a carpet made up of such nanotubes.

These nanoworlds have the most amazing structures. **Figure 12** to the left shows a nanostructure that has been getting a lot of press recently: a carbon nanotube. These tubes have a thickness of only 10 atoms or so, but are unbelievably strong, about a hundred times stronger than steel at only a sixth of the weight. We are currently learning to grow these structures and to use them. The trick is to grow such tubes en masse, leading to materials that have almost miraculous properties. There are already prototypes of a new display type allowing for monitors that are a lot sharper than what we use today. Also, nanotubes will certainly have countless applications for air and space travel.

[8] Gerd Binnig, 1947 Frankfurt am Main, German physicist, Nobel Prize 1986.
[9] Heinrich Rohrer, 1933 Buchs, Switzerland, Swiss physicist, Nobel Prize 1986.

Special Relativity → Cosmology

Finally, let's look at the consequences of the paper on special relativity, written in June 1905, with its three-page addendum of September 1905. In this short text, Einstein came to his famous equation $E=mc^2$. This notion states that energy and mass are equivalent and thus may be exchangeable. Simple as it sounds it changed our view of the world. Together with his paper on general relativity, which included gravity, Einstein described all the processes around us, how the Sun creates its energy, how planets orbit the way they do. When we explore these worlds, we take advantage of his theories in many different ways. We use them in building rockets and in getting spacecraft into orbit.

One of the most popular applications using Einstein's theory of relativity is the Global Positioning System, GPS for short. It consists of a constellation of 24 spacecraft orbiting the Earth at a distance of 20,183 kilometres. There, the Earth's gravity is slightly weaker than here on the surface. According to the theory of general relativity, this causes the clocks aboard the satellites to tick a bit slower than ours. While the difference is minute, it would cause the GPS to provide position information with an error of some 100 metres, which would make the system useless, if it weren't adjusted for Einstein's equations. So, when you next take a taxi in the city of Bern, remember that the taxi driver, guided by a GPS receiver, tacitly applies the laws found here by Albert Einstein a century ago.

Relativity perhaps has the most important consequences for our understanding of the universe and the diverse objects we find in it. One consequence was observed first in 1919 by Sir Arthur Eddington[10] on his famous scientific expedition (see also Spatium 14). The Sun seems to deflect light, like a giant lens, moving the apparent location of stars ever so slightly. Light from a distant star may be distorted just like a lens distorts it. Einstein's theory of general relativity predicts this, and the theory of relativity triumphs. The reason light gets deflected is because gravity curves space itself. Light looks for the fastest path, which is not necessarily straight.

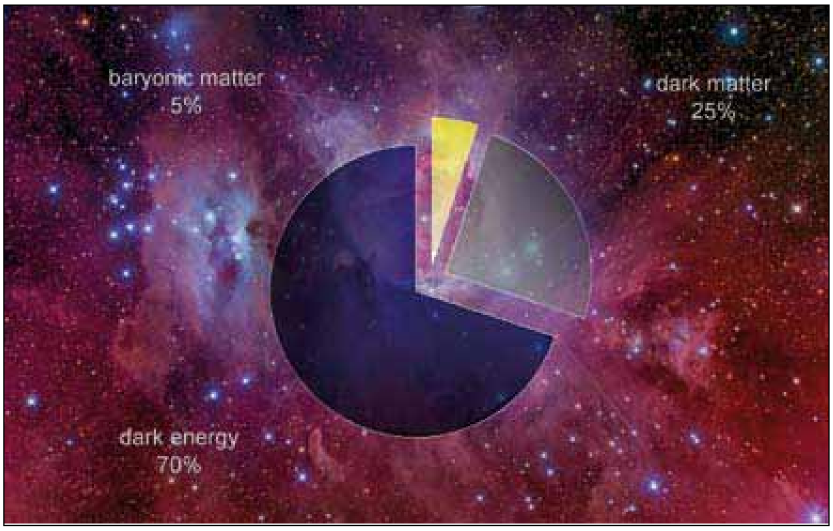


Figure 13: Mysterious cosmos. Baryonic matter (including neutrinos, etc.) forms the stars and the galaxies. It is this type of matter which we experience in our daily life. But this is only a small fraction of what constitutes the universe. The vast majority, namely dark matter and dark energy, is a theoretical assumption required to explain astronomical observations.

Einstein's theories have their most important consequences when applied to the history, the current state, and the future of the universe. Those belong to the most important questions humans have struggled with over the last millennia: What happened to the universe before the solar system was formed? This question has an amazing and unexpected answer, which actually was derived from Einstein's equations: The universe, we observe, started roughly 14 billion years ago and is now ex-

[10] Sir Arthur Stanley Eddington, 1882 Kendal, UK – 1944 Cambridge, British astrophysicist.
[11] Abbé Georges Henri Lemaître, 1894 Charleroi, Belgium – 1966 Löwen, Belgium, Belgian priest and physicist.
[12] Alexander Alexandrovich Friedman, 1888 St. Petersburg – 1925 Leningrad, Russian cosmologist and mathematician.

panding at a rapid speed. In 1927, a Belgiam catholic, Montsignor Le Maitre[11], concluded that such an expanding universe was a solution that came directly out of Einstein's equations. A Russian physicist, Alexander Friedman[12], had come to the same conclusion even before, but he died of typhus fever during WWI, before his solution was widely known. This expansion was first observed by Edwin Hubble[13], an astronomer working at the Californian Institute for Technology, in 1929. Hubble found that every galaxy seemed to move away from us, at a speed that increased with increasing distance – similar to a huge explosion. But not an ordinary one at that: It's not that galaxies explode into a previously existing empty space, but space itself is created in the "explosion", and it looks to everyone anywhere in the universe, not just us, that we are at its centre. The Hubble Space Telescope has continued these observations and has managed to even go beyond that. Looking at the largest distances and comparing to observations at the smaller distances, the Hubble telescope provided one of the most exciting and puzzling results of the last few years: In contrast to what we would expect, the expansion of the universe appears to be accelerating: the explosion is still powered today by a mysterious force. Cosmologists found an appropriate name for it: dark energy. It turns out that Einstein had even put this effect into his equations – even though he personally never believed in the actual existence of dark energy. In order to solve his equations in the way he considered reasonable, and only a static, eternal universe appeared reasonable back then, he had to put in an additional term called the cosmological constant. When he realized that the universe is indeed expanding, he dropped the term and called it his "biggest blunder" (grösste Eselei), because it made him fail to predict that the universe cannot be static and eternal. And yet this very term is now used by cosmologists when describing the accelerating expansion of the universe.

This adds to the puzzle we have today, summarized in **Figure 13**. In fact, we have an almost embarrassing situation right now: We know that the universe is made of multiple constituents. We observe stars and galaxies made of atoms. This constituent is called baryonic matter, the ordinary matter we know from our daily life. But there is another class of matter, called dark matter, as it does not directly interact with photons, therefore we cannot see it, but we can measure its gravitational force on ordinary matter. Both ordinary matter and dark matter together are still only a minor part of the universe. The major part is made up of the mysterious dark energy that drives the expansion. Neither dark matter nor dark energy, however, is fully understood today.

Conclusions

Does this sound familiar? We are in a situation again where somebody, like Henri Poincaré, could write a paper summarizing the most important problems we do not currently understand, and a successor to Albert Einstein should come and solve these open issues. Certainly, for most physicists, dark matter and dark energy are such problems: What is the nature of our universe, and what is it made of? Or, more specifically, what is the nature of dark matter and of dark energy?

Our story started in Bern. This city, with its beauty, and its charm, affected the world in one of the most profound ways thanks to one of its most important inhabitants ever: Albert Einstein. On 31 December 1999 *Time* magazine selected him as the Person of the Century – a lonely patent clerk who was asking important questions and did not back off when people did not agree with him. For those of us who have been affected by Einstein throughout our professional and personal lives, the Bern year of 1905 will always remain a miracle.

But the most important effect Einstein's work has on our world is that it became more beautiful. Einstein pointed out the underlying texture of nature that people did not see before he came along. Science can do that, and will do that again in the future.

[13] Edwin Powell Hubble, 1889 Marshfield, Missouri – 1953 San Marino, California, U.S. American astronomer.

SPATIUM

The Project and its Authors

In late 2004, when preparations for the centennial of Einstein's annus mirabilis got intense, the two first authors independently received invitations to give a talk about that singular event in the history of science. Even though neither of us is a theoretical physicist or a historian of science we decided to team up and jointly develop such a talk, together with video specialist Brian Grimm with whom we had worked before.

Our principal qualification for giving such a talk is that we had both studied in Bern, the city, where Einstein lived 100 years ago. So we decided to begin our presentation with some impressions of this beautiful town. The second message is that Einstein did not only develop relativity, but also gave theoretical explanations of Brownian Motion and of the photoelectric effect, each of which can rightly be considered as of equally fundamental importance. In the third part of the talk we advance time by 100 years to see what came to be of these three seminal topics by giving examples of the numerous applications today.

The talk has been given by both of us numerous times in 2005 and 2006 in the USA, in Asia and Europe. It is a pleasure to thank here our supporters and hosts: the Swiss Department of the Exterior, the Swiss Embassies in Washington, D.C., Seoul, Beijing, and Tokyo, the Swiss Centers in Boston and San Francisco, the Adler Planetarium in Chicago, the Historical Museum in Bern and more.

Rudolf von Steiger was born in Bern. He completed his Diploma (1984) and Ph.D. (1988) in Physics also at the University of Bern. After a postdoctoral year at the University of Maryland he returned to Bern, where he joined the newly founded International Space Science Institute (ISSI) in 1995. Today he works as a director of ISSI with a guest professorship at the University of Bern.

Thomas H. Zurbuchen was born in Heiligenschwendi. He completed his Diploma (1992) and Ph.D. (1996) in Physics at the University of Bern. Then he moved to the University of Michigan in Ann Arbor, where he works today as an Associate Professor for Space Science and for Aerospace Engineering.

Brian Grimm is a video artist and the owner of Paper Cardinal Design based in Oakland, California. He specializes in technical and scientific animations for facilitating efficient and effective communication. Together with his partner Tanja Andrews he is also the author of the award-winning video blog on www.freshtopia.net.

INTERNATIONAL SPACE SCIENCE INSTITUTE

SPATIUM

Published by the Association Pro ISSI

No. 19, August 2007

4440: A Secret Number in Astronomy

Editorial

This issue of Spatium is devoted to the history of astronomy, more precisely to its evolution in Europe and the area adjacent to the Mediterranean Sea. It is a play on different stages: In the early days of astronomy, the observation of heavenly process was on the one hand a vital necessity to tune agricultural activities to the changing seasons and on the other hand the Sun was worshipped as the inexhaustible source of light and life. Early astronomers were priests therefore linking the art of objective observations with the richness of pious inspirations.

The initially united stages splitted off when scientists found that the wandering of the Sun, the Moon and the other celestial bodies is ruled by nothing else than a set of soulless mathematical functions. On the religious stage, emerging dogmas incorporated suitable world models that received the status of eternal truth irrespective of the continued progress made on the stage of science. Galileo Galilei contributed to further deepening the ditches between the two stages by stipulating that a hypothesis, in order to be scientifically acceptable, must be provable everywhere at any time. This of course was provocative news for the clerical establishment which did not hesitate to denounce him of heresy and to prohibit him to further teach his findings: the two stages had entered a state of fierce war for the next centuries.

Galilei even added a third stage that is the use of technical means to overcome the limitations of the human senses to observe nature. Although his means were nothing more than a pair of simple lenses packed in a tube of cardboard called telescope, the shiny objects that now suddenly became visible challenged the human spirit enough to let him make every effort to improve the miraculous device even further. In that sense, the history of astronomy is also the history of the evolution of technology and today's extensive use of space-borne instruments for astronomical research is nothing else than a continuation of Galiei's use of telescopes some 400 years ago.

All these different stages – and many more – were the subject of a most fascinating lecture Prof. Giovanni F. Bignami, then at the University of Pavia, now President of the Italian Space Agency, gave the Pro ISSI audience in November 2006. The present text follows loosely the lines of his presentation; may it provide our readers the same wealth of inspring thoughts as Prof. Bignami's talk offered his attentive audience…

Hansjörg Schlaepfer
Neerach, August 2007

Impressum

SPATIUM
Published by the
Association Pro ISSI

INTERNATIONAL SPACE SCIENCE INSTITUTE

Association Pro ISSI
Hallerstrasse 6, CH-3012 Bern
Phone +41 (0)31 631 48 96
see
www.issibern.ch/pro-issi.html
for the whole Spatium series

President
Prof. Klaus Pretzl
University of Bern
Layout and Publisher
Dr. Hansjörg Schlaepfer
CH-8173 Neerach
Printing
Stämpfli Publikationen AG
CH-3001 Bern

Front Cover
The recently found sky disk of Nebra, Central Germany, is the most ancient image of the sky. Dating back to 1600 B.C., it represents the Sun, the Moon and – probably - the Pleiades. It is assumed to have served determining the beginnings of new seasons for agricultural purposes, but it certainly served religious traditions as well. (Credit: Landesamt für Archäologie Sachsen-Anhalt, Germany, Hansjörg Schlaepfer)

4440: A Secret Number in Astronomy[1]

Dr. Hansjörg Schlaepfer, Neerach

Introduction

Mankind has been looking at stars with naked eyes for the last 4,000 years collecting data and understanding important astronomical facts. Since Galileo Galilei gave us the telescope, 400 years ago, astronomy from the ground has reached a profound comprehension of our universe; infinitely greater than had been possible in the previous millennia. In a spectacular acceleration, from millennia over centuries to decades, the last 40 years have given us astronomy from space, yielding a view of celestial objects over the whole electromagnetic spectrum, i.e. the most important channel through which the sky sends us information about itself. In decoding such information, gathered from telescopes in space, we have, once again, vastly surpassed the work done over the previous centuries and created entirely new scientific disciplines such as planetology or cosmology.

4,000 Years of Naked Eye Astronomy

From Early Testimonies...

There was a time – not too far back – when the night sky was really dark and crowded with uncountable stars. The Milky Way was the wonderful lane that leads from the ho-

Fig. 1: Summer solstice over Stonehenge. The great stone circle was erected between 3000 B.C. and 1600 B.C. It consists of a circular structure, exactly aligned with the rising Sun at the solstice. (Credit: Andrew Dunn)

[1] The present text follows a lecture by Professor Giovanni F. Bignami, President, Italian Space Agency, Rome, for the Pro-ISSI audience on 13 November 2006 and is freely retold by Hansjörg Schlaepfer.

rizon to the zenith and down again to the opposite side. Today, atmospheric pollution absorbs most of this faint radiation and the scattering from the ubiquitous artificial light sources hides the beauties of the night sky to most of us. Fortunately, mankind has evolved under clear night skies and has been impressed by the silently wandering stars. We do not know the beginnings of their observations; the eons since have erased most, if not all, of the traces of early astronomical endeavours. Interestingly though, these seem to have evolved independently(?) at many different places all over the world some 4,000 years ago. One of these places is Stonehenge, UK **(Fig. 1)**, of which the oldest traces date back as far as 3000 B.C. The megalith's precise orientation towards the rising Sun at summer solstice is a clear indication, that astronomy played an important role in the life of their constructors.

On the Continent, more precisely in Central Germany, another milestone of early astronomical endeavours from that time period was found recently: the sky disk of Nebra, see the **front cover.** Dating back to 1600 B.C., it represents mankind's very first known image of the sky. The latest analytical technologies were used to unravel some of the disk's secrets: it is thought that its designers had contacts to the eastern Mediterranean area, to Mycenae and even further to Mesopotamia, where in the fertile half-moon between the rivers Euphrates and Tigris the Sumerian culture prospered at that very time. The Sumerians are known to have developed not only what could be called the first "professional" astronomers but also the means to bequeath much of their broad knowledge. They invented a form of writing in clay with a wedge-shaped stylus thereby producing documents that survived the many millennia since. Thanks to these documents, we know that the Sumerian priests studied the stars and knew to correlate astronomical alignments with the changing seasons. Based on their observations they set up a calendar consisting of a year with twelve months, each of which began at the first sighting of the new moon. Sometimes, based on a respective royal order, the eleventh months had to be repeated, in order to bring the lunar calendar in line again with the Sun's. A Sumerian day consisted of six watches, each being two double hours long. Each hour was divided into 60 minutes, and each minute was divided into 60 seconds. The year had nominally 360 days, and so their heavens were divided into 360 degrees. Many of these ancient concepts have survived thousands of years and constitute today our counting system for time and angles, while in the other areas the decimal system, presumably of Indian origin, has become the standard system.

The cuneiforms, as the Sumerian clay tablets are called, tell us also about the Sumerian endeavours to understand the processes in the sky in order to sow and to harvest at the appropriate time. No doubt, the need for sufficient food and water supply was at the root of their interest in astronomy and, since fertility of the fields was a gift from the gods at that time, religious and scientific endeavours could co-evolve in untroubled harmony…

While the Euphrates and the Tigris were the source of wealth for the Sumerians and the later Babylonians, another large river, the Nile, was the cradle for the Egyptian civilization. The Nil's periodic inundations, lasting from July to October, required the mastering of the calendar to tune agricultural activities to the river's habits: the Egyptians broke a year down into 365 days, which was largely sufficient for practical purposes, and it seems that further efforts to understand the mechanics of the heavens were not made, at least until the later Hellenistic period.

Not a great river, but a long sea shore was the key asset of the classic Greek civilization that evolved approximately at the same time on the northern shores of the Mediterranean Sea. The great Greek philosophers Socrates[2] and Plato[3] laid the foundations of our western philosophy. Socrates' thinking and Plato's dialogues are characterized by mythic-allegoric symbols for describing the world: their real world is that of the ideas, while the human perception

[2] Socrates, 470 B.C. – 399 B.C., ancient Greek philosopher.
[3] Plato, 427 B.C. – 347 B.C., ancient Greek philosopher.

of the physical world is at best a blurred image. In contrast, for Aristotle[4], a pupil of Plato, the visible appearances became real and the world of the ideas an imperfect approximation of the empiric reality: this was the dawn of science.

Regarding astronomy, it was Aristarchos of Samos[5], who approached nature under this new paradigm. He estimated the diameter of the Sun. The result was a factor of twenty too short, but nevertheless, Aristarchos' Sun had a volume equalling 250 Earth volumes making it much greater than the Earth. It was nothing else than a logical consequence to assume the Earth to circle the Sun (and not vice versa): the modern heliocentric world model was born.

Alexander the Great[6], the most successful military commander of the ancient world, was a contemporary of Aristotle. He enlarged the accessible world towards Central Asia and India thereby boosting not only trade and economic evolution, but cultural exchange with the newly discovered civilisations as well. He founded many cities, for example Alexandria in Egypt, which soon became the centre of the known world. Here, a famous Museion was erected, that is a temple for the Muses, or – in modern terminology – an academy of sciences. Its library became the world's largest counting up to 900,000 scrolls. Of course, such an impressive collection of knowledge attracted the very best scientists of the time. Eratosthenes of Kyrene[7], one of the first directors of the library, is known for having found the Earth's perimeter around 200 B.C. He observed a draw well in Syene, Upper Egypt, illuminated by the Sun at summer solstice right down to its floor, while at the same time a well in Alexandria was illuminated only partially at an angle of 7.2°. The distance between the two cities was known from the royal couriers to be 5,000 stadia. From these data he was able to estimate the Earth's perimeter at 39,250 km, just within 2% of the exact value.

... to the First Great Astronomers...

On the northern shores of the Mediterranean Sea, most probably in Nicaea, now Izmir, Turkey, Hipparchos[8], the greatest astronomer in ancient times, saw the light of the day in 190 B.C. He was the first to derive accurate quantitative models for the motion of the Sun and the Moon. To this end, he made use of observations and knowledge accumulated over centuries before. He was also the first to compile a trigonometric table allowing him to solve any triangle. With his solar and lunar theories together with his numerical trigonometry, he was probably the first to develop a reliable method to predict solar eclipses.

In contrast to Aristarchos, Claudius Ptolemy[9] found the Earth at the centre of the universe three hundred years later. Ptolemaeus used Hipparchos' system to explain the motions of the Sun, the Moon, and the five planets known at that time. It was accurate enough to predict the position of the planets for naked-eye observations. Ptolemaeus authored three major oeuvres which strongly influenced astronomy, of which one is the *Mathematical Syntaxis* (in Greek called ΜΕΓΙΣΤΗ ΣΨΝΤΑΞΙΣ, translated in Arabic as Al Magiste and later in Latin as Almagest). The Almagest is our most important source of information on ancient Greek astronomy, see **Fig. 2.** It includes a star catalogue containing 48 constellations with the same names as we still use today.

...to the Decline of European Astronomy...

Ptolemaeus's death in 168 and the destruction of the Alexandrian library that occurred presumably at approximately that time brought the evolution of western science to a sudden end. In addition, the philosophic schools of the late Roman Empire were oriented backward to

[4] Aristoteles (English: Aristotle), 384 B.C., Stageira, Greece – 322 B.C., Chalkis, ancient Greek philosopher.
[5] Aristarchos of Samos, 310 B.C., Samos, Greece – 230 B.C., Alexandria, Egypt, ancient Greek astronomer and mathematician.
[6] Alexander the Great, 356 B.C., Pella, Northern Greece - 323 B.C., Babylon (today Iraq), King of Macedon.
[7] Eratosthenes of Kyrene, 276 B.C. – 195 B.C., ancient Greek mathematician, geographer and astromer.
[8] Hipparchos, 190 B.C., Nicaea, Asia Minor – 120 B.C., Rhodes, Greece, ancient Greek astronomer, geographer, and mathematician.
[9] Claudius Ptolemaeus (English: Ptolemy), 90, Alexandria (?), Egypt – 168, Alexandria, Hellenistic mathematician, geographer, astronomer, and astrologer.

Fig. 2: A late copy of the Almagest by Ptolemy translated in Latin around 1451.
The drawings illustrate the computation of the duration of solar and lunar eclipses.

Plato, thereby creating an intellectual substrate that did not aim at scientific progress. Even worse, the initially suppressed Christianity became state religion. It was a fundamentalist movement that addressed mainly the "poors in spirit", social strata that were far away from any interest in science. The Ptolemaic system of a geocentric world was formulated just at the right time to enter the Christian dogmatic system, where it firmly stood for the next 1,400 years. In 529, Iustinianus I[10] closed the academy of Athens, the last Hellenistic school: the long dark night of the medieval time in Europe had begun. Although some early Christian philosophers, like for example Origen[11], tried to reconcile the evolving Christian doctrines with Hellenistic thinking, the new movement began to extinct everything not literally in line with the Bible. That became the destiny for the vast majority of the great testimonies of ancient human thinking, including for example the Almagest. A chance for survival had only those works that had been compiled or copied elsewhere by late-Roman librarians.

... to Arabian Astronomy ...

Fortunately though, science found a new culture medium on the eastern side of the Mediterranean Sea: In 622, Muhammed[12] founded a new religious and political movement unifying elements from Judaism and Christianity. The new religion, the Islam, rapidly developed an impressive unifying power: its influence eventually reached Andalusia in the west and Central Asia and China in the east. Like some 900 years before, under Alexander the Great, economy, handcraft and trade re-flourished. Damascus, later Baghdad, became the new world's capitals. And in contrast to the parochial Christianity, the evolving Islamic culture was open to new ideas and science was seen as an important cultural accomplishment. Under Kalif Al Mansur, around 760, sages from all over the world convened to make up the world's intellectual centre in Baghdad. The works of the ancient Greek thinkers were translated meticulously into Arabic and the classic Almagest of Ptolemy was not only studied but further developed also, see **Fig.3.** It was enriched by the Hindu-Arabic numeral system, now called the Arabic numerals. Kalif Al Mamun ordered the construction of a new great observatory near Baghdad in 829 and Muhammad Al Fagan was one of the leading astronomers there. He authored a book entitled "Elements of Astronomy", which served as a standard text book in Europe until the end of the 16th century.

Towards the year 1000, the centre of Arabic astronomy moved slowly to the west, first to Egypt, then to the Moorish Castile in Spain. Under the Kalifs of Cordoba, astron-

[10] Flavius Petrus Sabbatius Iustinianus, 482, Iustiniana Prima (today Serbia) – 565, East Roman Emperor.
[11] Origenes, (English: Origen), 185, Alexandria, Egypt (?) –254, Tyros, today Leban (?), one of the early fathers of the Christian Church.
[12] Muhammed, 570, Mecca (today Saudi Arabia) – 632, Medina, founder of the Islam.

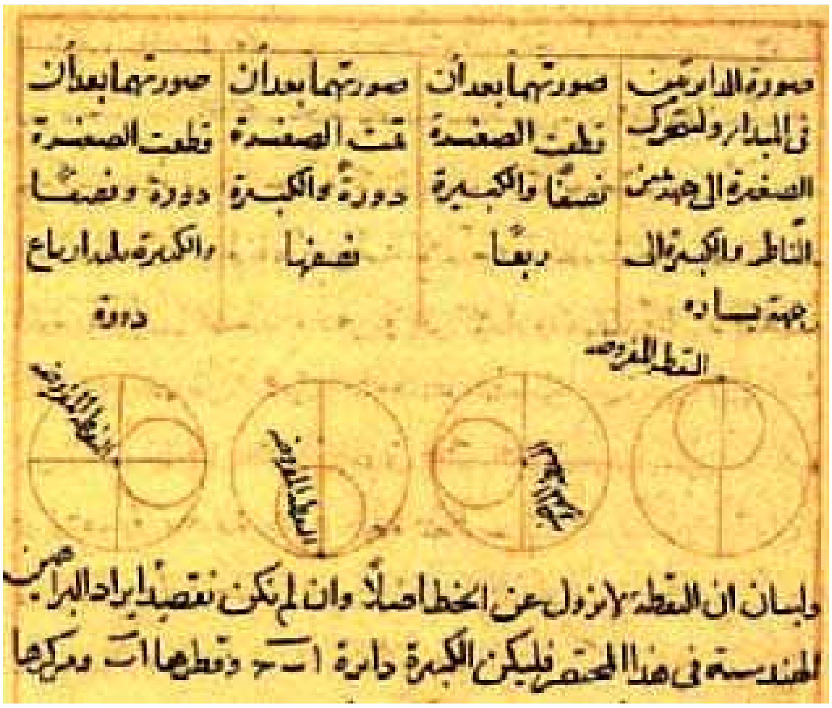


Fig. 3: An example of a late Arabic book based on Ptolemy's Almagest. The Picture shows a page of Nasir ad-Din at-Tusi's "Memoirs on Astronomy", written in the fourteenth century.

omy saw a first, albeit transitory, revival in Europe in the 11th and 12th century, but the "reconquista" by the Castilian Christians not only brought the Arabic presence in Europe, but also their scientific endeavours to an end. A land mark of occidental philosophy of that time is the Divina Comedia by Dante[13], one of the masterpieces of world literature, firmly rooting in the classic scholastic tradition.

Fortunately, astronomy sparked off again in the eastern parts of the Arabic empire, more precisely in Baghdad and Megara, where new astronomical institutions were established.

In the vicinity of Tabriz, in today's north-western Iran, an observatory was founded in 1259 that had a large library counting as much as 400,000 manuscripts. Some 200 years later, in 1420, the Mongol Ulugh Begh ordered the construction of an observatory on the Silk Road in Samarkand, today Uzbekistan. He personally developed the first star catalogue that was not simply an improvement of Ptolemy's oeuvre, but contained many new elements. This catalogue eventually became known in Europe and went into press in 1655, at a time, when it was already outdated by insights of European scientists.

...to the Revival of European Astronomy

Under the increasing Mongolic pressure the Arabian astronomy came to an end, while the occident was fortunately about to leave the dark medieval times and to become open for science. At that time, many precious goods were imported over the Silk Road from the east. It is probably on that path that the art of printing reached Europe around 1300. Initially, it was used to decorate clothes for religious purposes exclusively. The art of paper making, originally invented in China also, was imported during the Moorish conquest of Andalusia. The first paper mill was established in 1120 near Valencia; later mills were founded in Italy and Germany around 1400. In 1452, the German goldsmith Johannes Gutenberg invented the moveable type printing in Mainz. This new technology was quicker and more durable than the previous laborious woodblock printing, which required the different letters to be carved out from a solid block of wood. This new technology rapidly spread through Europe: in 1469, the first print press was established in Venice, 1470 in Paris, 1473 in Krakow, Poland, and it took more than 150 years for the first printing press to start operating in North America (1638).

Of course, the Church knew to put the promising technology rapidly into its service; but also scientists began to exploit printing for their

[13] Dante Alighieri, 1265, Florence, Italy – 1321, Ravenna, Italy, Italian poet and philosopher.

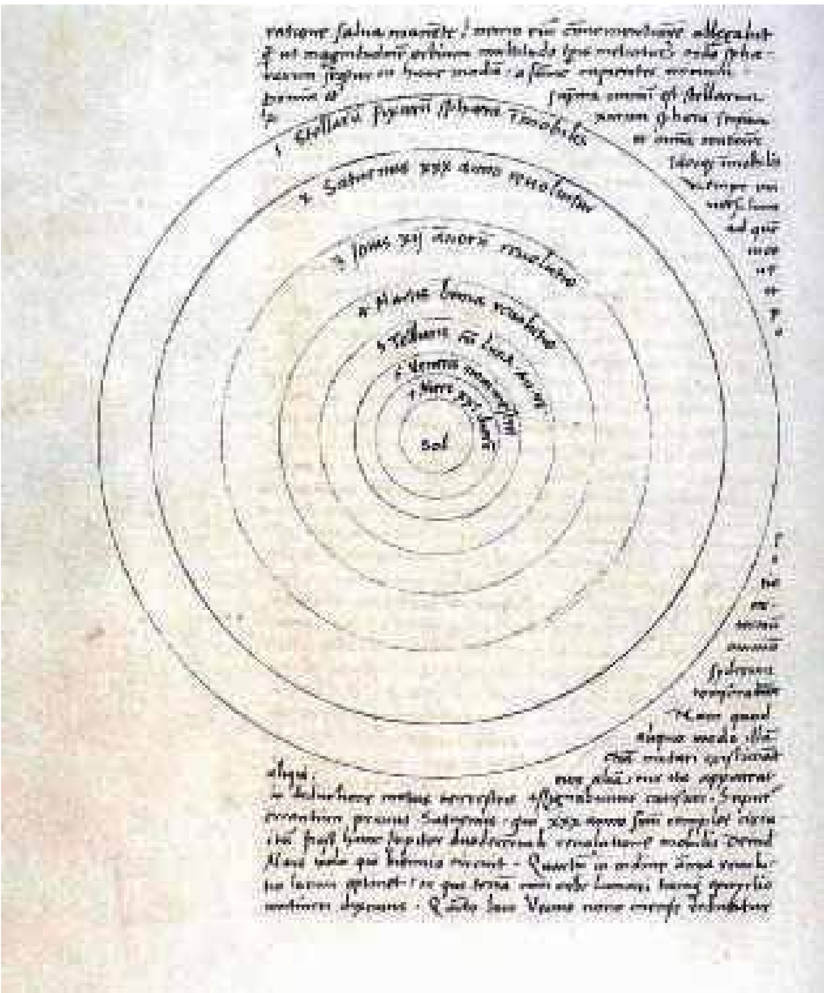


Fig. 4: Nicolas Copernicus: De revolutionibus orbium coelestium (1543).

needs. One of earliest great European scientists, who used the art of printing, was Nicolas Copernicus[14]. His Commentariolus (Little Commentary) was initially nothing more than a handwritten text outlining some of his ideas on a heliocentric world model. But around 1515 his friends urged him to publish and even the clerical establishment invited him to go into press. It was in 1543 only, the year of Copernicus' death, that his seminal oeuvre "De revolutionibus orbium coelestium" was printed, see **Fig. 4.** At first, the "De revolutionibus" did not cause any worries even though its contents were in contradiction to the official dogmas that adhered to the Ptolemaic world model. Rather, it took some 60 years for the clerical establishment to recognize the explosive force of that book and to put it rapidly on the "Index Librorum Prohibitorum", the list of the prohibited books considered a danger to the true faith.

At about the same time, Giordano Bruno[15], another adherer of the heliocentric world model and an early proponent of an infinite and homogeneous universe, was accused of heresy by the Roman inquisition. As he refused to recantate, he was burned at the stake in Rome and all his works were placed on the Index, where they remained for the subsequent 400 years.

Fortunately though, ideas cannot be extirpated by burning their proponents. Far to the north, in safe distance from the Roman inquisition, the Dane Tycho Brahe[16] was about to study law together with a variety of other subjects that attracted the youngster's interest. The solar eclipse in 1560 impressed him so deeply that he stopped his studies and began to devote himself to astronomy. As an intended lawyer he was used to work precisely and he quickly became aware of the need to improve and enlarge the available

[14] Nicolas Copernicus, 1473, Thorùn (Prussia) – 1543, Frombork, Poland, mathematician, astronomer, jurist, physician, diplomat, and soldier.
[15] Giordano Bruno, 1548, Nola, Italy – 1600, Rome, Italian philosopher, priest, cosmologist, and occultist.
[16] Tycho Brahe, 1546, Knudstrup, Denmark – 1601, Prague, Danish astronomer.

astronomical instruments, and to construct entirely new ones, **Fig 5.** Assisted by his diligent sister Sophia he was able to make astronomical measurements of unprecedented precision.

Unfortunately for Sophia and Tycho, King Christian IV succeeded King Frederic II, who had been a strong supporter of their astronomical research. Brahe decided to leave Denmark and to move to Prague in 1599. There, during the last two years of his life, he was assisted by the young Johannes Kepler.

Thanks to its outstanding accuracy, Brahe's work marks the climax of naked eye astronomy, but also its end, as new means for observing the sky had silently emerged in the meantime that bore the potential of a quantum leap in astronomy...

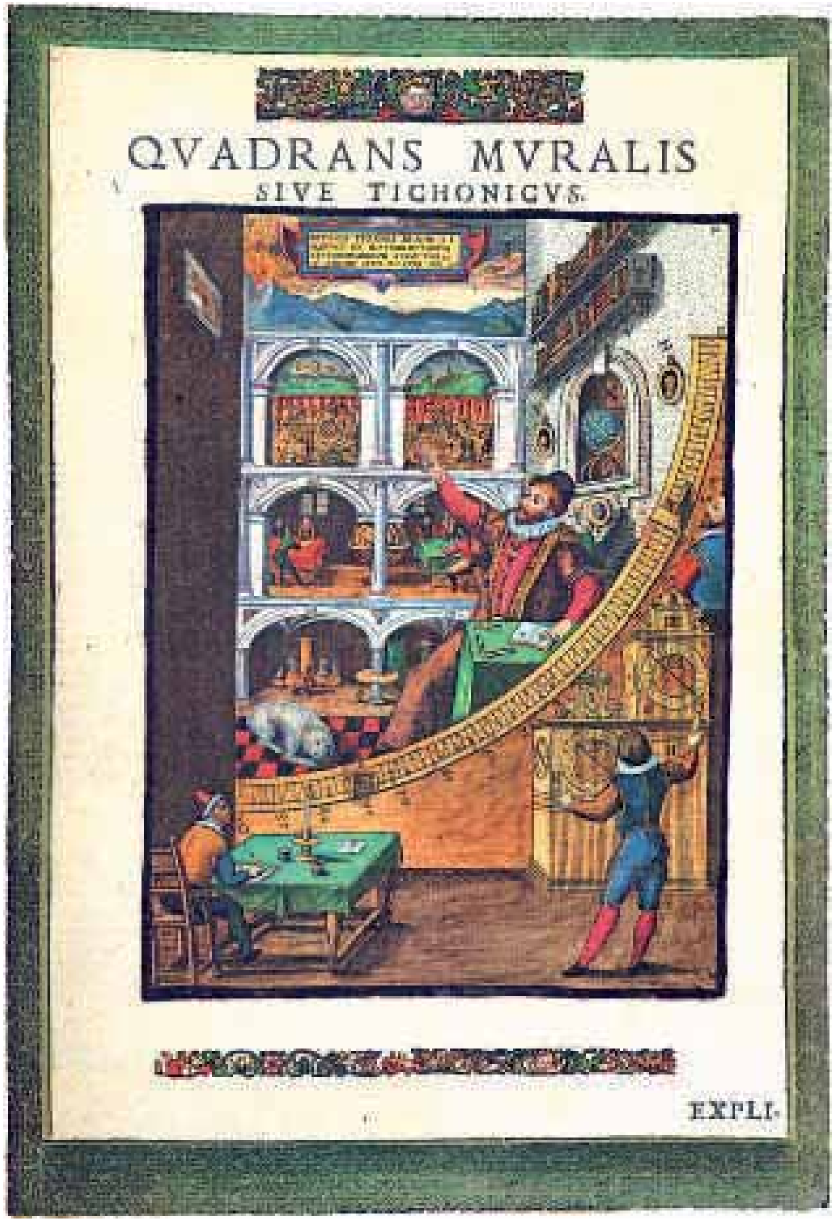


Fig. 5: Tycho Brahe at work. Brahe is depicted in the centre of the image together with his assistant reading the scale on the right, the assistant responsible for the time reading on the right bottom and the assistant taking the notes on the left bottom.

400 Years of Astronomy with Telescopes

Small Things Change the World

Murano, a small town on an island in the Venetian lagoon, had been a commercial port since many centuries when in 1291, the Venetian Republic decreed its glassmakers to move to Murano as the glassworks represented a severe fire danger to the city of Venice. These glassmakers had developed and refined technologies to produce a variety of precious glass types amongst which were reading stones of the best then available quality **(Fig. 6)**. Such reading stones were a necessity not only in the *Serenissima*[17], but found their customers far abroad, for example in the Netherlands. The legend tells that on a sunny summer morning some children played in the shop of the spectacles maker Hans Lippershey[18] in Middelburg. Incidentally, they found that the combination of two such lenses brought a new quality to the image as compared to that produced by a single lens. Lippershey, a clever businessman, immediately recognized the potential of the children's invention and successfully started building and selling pairs of lenses as telescopes all over Europe.

Fig. 6: A reading stone. (Credit: Carl Zeiss)

Such a novel device came to Pisa in Italy, specifically to Galileo Galilei[19]. The first of six children, young Galileo wanted to become a priest. At his father's urging, however, he enrolled for a medical degree at the University. But soon, Galilei recognized that medicine was not his favourite field. He then switched to mathematics, a decision for which humanity remains deeply indebted as it was the prerequisite for Galilei to become the father of modern physics. At the age of 25 already, he was appointed the chair of mathematics at the University of Pisa. A mere three years later, Galilei moved to the University of Padua to teach geometry, mechanics, and astronomy. During these fertile years as a professor, Galilei made his significant discoveries in both pure and applied science. He pioneered the use of quantitative experiments whose results could be analyzed subsequently with mathematical methods. Just three years after Lippershey's invention, in 1611, Galilei built his first own telescope that he presented to the then wealthy Venetians. Those welcomed Galilei's device enthusiastically for various purposes such as for military information gathering or for merchants who found it useful for their shipping business. Galilei on his part was more interested in gaining scientific insights. Those came along very quickly: in January 1610 already he

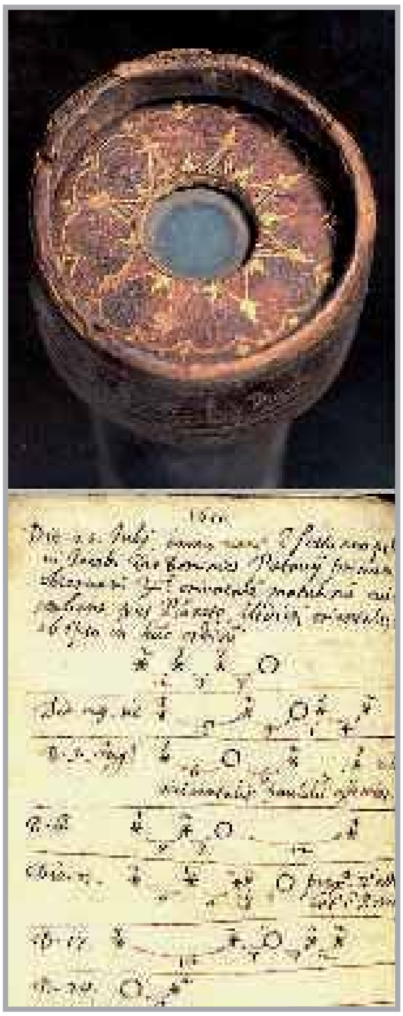


Fig. 7: A telescope used by Galileo Galilei (upper table) and an excerpt of his astronomical diary of 1610.

---

[17] La Serenissima, a name for the Republic of Venice, from the title Serenissimo, literally meaning "the most serene".

[18] Hans Lippershey, 1570, Wesel, Germany – 1619, Middelburg, the Netherlands, Dutch lensmaker.

[19] Galileo Galilei, 1564, Pisa – 1642, Arcetri (Italy), Italian physicist, astronomer, and philosopher.

discovered three of Jupiter's four largest moons: Io, Europa, and Callisto. A mere four nights later, he discovered Ganymede, Jupiter's fourth moon. In his diary **(Fig. 7),** he noted that the moons appeared and disappeared periodically, which he attributed to their movement behind Jupiter. From this, he concluded that they were orbiting the planet. Unsuspectingly, Galilei published his discoveries in a small treatise called "Sidereus Nuncius" in 1610. The oeuvre was devoted to the Medici for the housing they offered him at their court. But the generous patrons did not realize what worries the booklet was about to cause: Galilei's finding of a planet that is orbited by smaller bodies was revolutionary, to say the least, as the officially still accepted geocentric world model required all celestial bodies to safely circle the Earth. Equally important from a scientific point of view was Galilei's trendsetting approach of making his claims visible for everybody, in contrast to the dogmatic approach to defining the truth "ex cathedra". The "Sidereus Nuncius" made Galilei a dangerous subject for the clerical establishment which did not hesitate to denounce him of heresy. Galilei went to Rome to defend himself against the accusations, but, after lengthy trials he was finally forbidden in 1616 to either advocate or teach Copernican astronomy furthermore. In addition, he was confined in Siena. Totally blind, he was allowed to retire to his villa in Arcetri, where he died quietly in 1642.

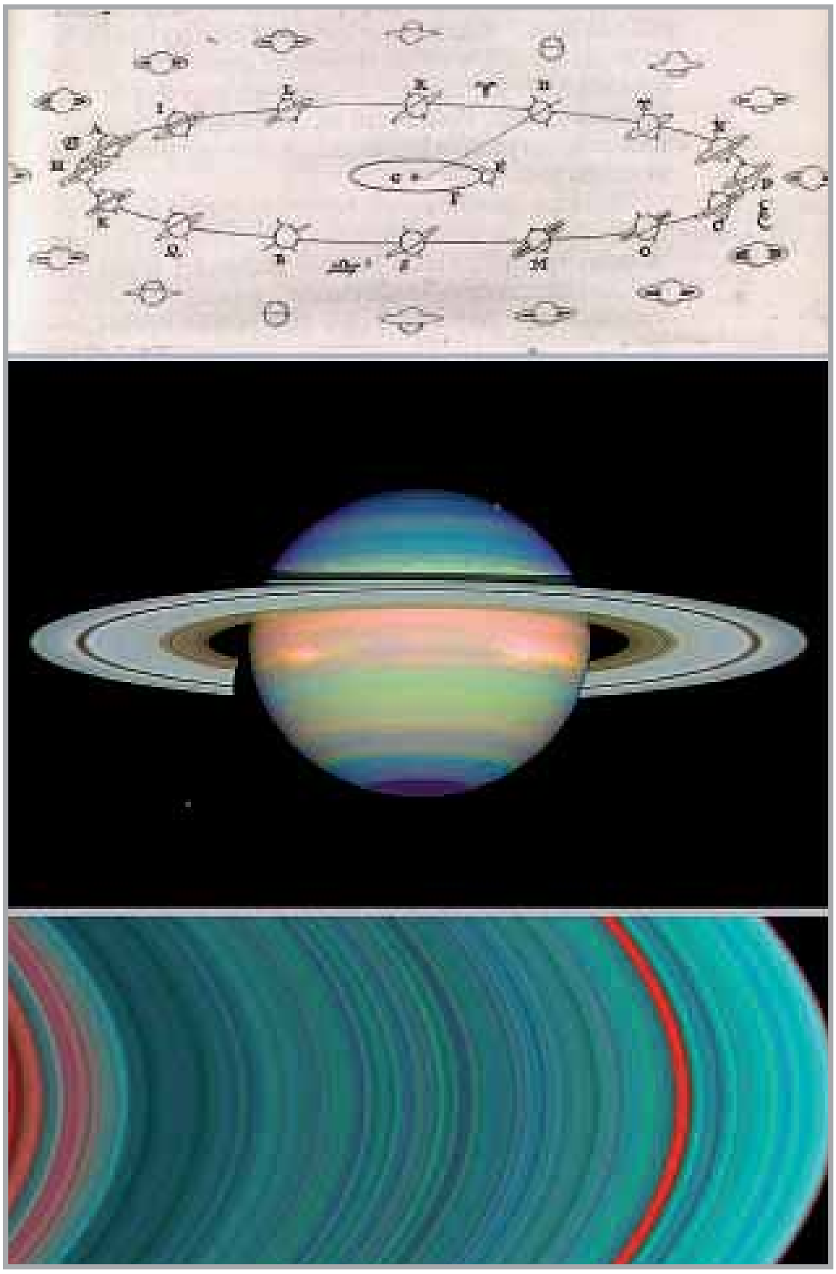


Fig. 8: Saturnian impressions: The upper table shows an excerpt of Christiaan Huygens' observations of Saturn of 1655. The middle table presents a false colour image of Saturn in the infrared acquired by the Hubble Space Telescope in 1998. The table below shows the Saturnian ring system observed by the Cassini spacecraft in 2004.

Again in a safe distance from the inquisition, more precisely in the Imperial Free City of Weil der Stadt, now in the German state of Baden-Württemberg a boy was born in 1571. Nothing in the early life of Johannes Kepler[20] would have indicated that this youngster of fragile health was later to become one of mankind's most prominent scientists. His father Heinrich died when Johannes was five, and his mother was tried for witchcraft. Still, she showed him a comet in 1577 that left a long trace not only on the sky but in the youngster's mind as well. He began studying theology, but soon it turned out that he was an excellent mathematician, so he switched and became a teacher of mathematics and astronomy in Graz, Austria. In early 1600, Kepler met Tycho Brahe near Prague, where he stayed as a guest. He was allowed to analyze some of Brahe's observations of Mars. Brahe was impressed by Kepler's theoretical abilities and soon allowed him further access to his data. Two days after Brahe's unexpected death in late 1601, Kepler was appointed his successor as imperial mathematician: he inherited his master's records as well as the responsibility to complete the unfinished work. During the following years and after careful analysis of Brahe's data, Kepler realized that the planetary orbits follow three relatively simple mathematical laws. Two of them are derived in the "Astronomia nova", 1609, while the third is contained in the later "Harmonices mundi", 1619. When Kepler heard of Galileo's telescopic discoveries, he started theoretical and experimental investigations of telescope optics as well. He published the results in "Dioptrice", 1611, where he described an improved telescope also, now known as the astronomical or Keplerian telescope, using two convex lenses that can produce higher magnification than Galileo's combination of convex and concave lenses.

Kepler's oeuvre was a seminal success and found enthusiastic readers all over Europe. One of those was the Dutchman Christiaan Huygens[21] who was building his own telescopes at that time. He managed to improve their imaging performance to such an extent as to allow him the discovery of Saturn's moon Titan in 1655. He also examined Saturn's planetary rings, and found that they must consist of rocks. But Huygens was not only a diligent observer, but also a gifted physicist: he derived the wave nature of the light and developed the probability theory as well. And his hobby, as to say in modern terms, was the existence of life on other planets: he published his reasoning in the book "Cosmotheoros", presumably one of the very first oeuvres on exobiology.

The Quest for Astronomical Discoveries

The seventeenth century saw Europe in a deep crisis; the Thirty Years' War was fought between 1618 and 1648, basically a conflict between Protestants and Catholics, but also a dispute on the role of the religion in the modern state. Devastating famine and disease all over Europe was one of its terrible consequences. Nevertheless, in some isolated spots, science was able to survive. Far to the south of Europe, Giovanni Domenico Cassini[22], a contemporary of Christiaan Huygens, studied mathematics and astronomy and became a professor for astronomy at the University of Bologna. In 1665, he determined both the rotation periods of Jupiter and Mars, and was one of the first to observe the polar caps of Mars. On invitation by the *Roi Soleil,* Louis XIV, Cassini moved to Paris. Based on his collaboration with Christiaan Huygens, he developed and used very long air telescopes. In 1672, he was able to measure the distance of Mars by triangulation and thereby to refine the dimensions of the solar system. Specifically, he determined the value of the astronomical unit (AU), just 7% short of the currently accepted value of 149,597,870,691 metres. Cassini discovered also Saturn's moons Iapetus, Rhea, Tethys, and Dione and the gap separating the two parts of the Saturnian rings that later received the name Cassini Division in his honour.

[20] Johannes Kepler, 1571, Weil der Stadt, Germany – 1630, Regensburg, German mathematician, astronomer and astrologer.
[21] Christiaan Huygens, 1629, The Hague (the Netherlands) – 1695, The Hague, Dutch mathematician, astronomer and physicist.
[22] Giovanni Domenico Cassini, 1625, San Remo, France – 1712, Paris, Italian/French astronomer, engineer, and astrologer.

In England, struck also by religious quarrels, another genius was about to revolutionize physics and the observational means in the hands of astronomers. Sir Isaac Newton[23] published the "Philosophiae Naturalis Principia Mathematica" in 1687, where he described the universal law of gravitation and the three laws of motion, laying the groundwork for classical mechanics. This allowed him to derive Kepler's laws of planetary motion from a purely theoretical point of view. He showed that the motion of objects on Earth and of celestial bodies is governed by the same set of natural laws. The unifying and predictive power of his laws became central to the advancement of the heliocentric system at the time. Based on his studies of the properties of light, he concluded that any refracting telescope (based on lenses) must suffer from the dispersion of light into the different colours. This led him to invent the reflecting telescope today known as the Newtonian telescope.

While Newton's mirror telescope constituted a major progress in observational astronomy, another star in mankind's history did not even need to observe the universe with his eyes, but rather used the overwhelming power of his rational thinking to develop theories on the evolution of the solar system: Immanuel Kant[24]. Besides an enormous wealth of philosophical oeuvres, he wrote the "Allgemeine Naturgeschichte und Theorie des Himmels", 1755, where he correctly derived the theory of the evolution of planetary systems based on Newtonian mechanics. According to Kant, our solar system has been formed out from a form of nebula where the different planets concretized independently from each other. In addition, he saw our solar system as being merely a smaller version of the fixed star systems, such as the Milky Way and other galaxies.

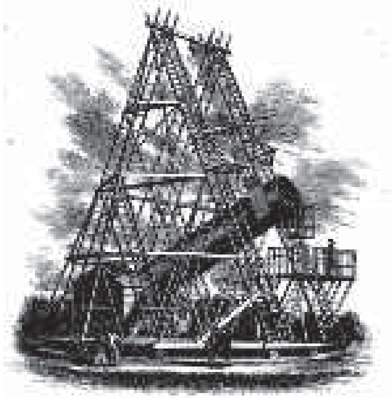


Fig. 9: The 12 m telescope built by William Herschel.

The early eighteenth century allowed Europe to recover from its sufferings finally and art and sciences were reaching new highs. It was the time of the baroque, the time of Johann Sebastian Bach and Georg Friedrich Händel. Music was the favourite subject of the young William Herschel[25] in Hanover, Germany. Born around 1738 as one of ten children, young Friedrich Wilhelm came to the Hanoverian Guards regiment, which was ordered to England in 1755, when the crowns of England and Hanover were united under George II. A gifted youngster, Friedrich quickly learned English and at the age of nineteen, he changed his name to Frederick William Herschel. He became a successful music teacher, bandleader and composer like his father, but eventually he got interested in mathematics and astronomy also. He started polishing mirrors and building telescopes of which he constructed over 400 of various and increasing size during his lifetime. The largest and most famous one is a reflecting telescope with a 12 m focal length and an aperture of 126 cm, see **Fig. 9.** This instrument allowed him to discover planet Uranus in 1781, which prompted him to give up his career as a musician and to become a great astronomer. Later, Herschel discovered Mimas and Enceladus, two Saturnian moons, and Titania and Oberon, two moons of Uranus. From studying the proper motion of stars, he was the first to realize that the solar system is moving through space, and he determined the approximate direction of that movement. He also studied the structure of the Milky Way and concluded correctly that it has the shape of a disk.

While the Northern Italian Valtellina valley is well known for its fine wine made of the Nebbiolo grape, it is certainly not so for scientific importance. But in the rural town

[23] Sir Isaac Newton, 1643, Woolstrope, Lincoln (UK) – 1727, Kensington, founder of classical mechanics.
[24] Immanuel Kant, 1724, Königsberg (Prussia) – 1804, Königsberg, German philosopher.
[25] Sir Frederick William Herschel, 1738, Hanover (Germany) – 1822, Slough (UK), German-British astronomer and composer.

of Ponte in Valtellina, Giuseppe Piazzi[26] saw the light of the day in 1746. He began his career with theological studies and finished his novitiate at the convent of San Antonio, Milan. Later, he studied at various Italian universities and became a professor for dogmatic theology in Rome. But he was not only a gifted clergyman, but an excellent mathematician as well. In 1780, one year after the French revolution, he was called to the chair of higher mathematics at the academy of Palermo, Sicily, where the Bourbon King Ferdinand I of the Two Sicilies was on the throne. A fervent supporter of astronomy he gave Piazzi a major grant for building an observatory. In order to acquire the best available instruments, Piazzi went on a shopping tour throughout Europe, specifically to Paris and to England. The equipment he had acquired was placed on top of a tower of the Royal Palace in Palermo. Observations began in 1791, and the first report was published just one year later. The excellent equipment – and the clear sky of Sicily – allowed him to make scientific contributions soon, such as to estimate more precisely the obliquity of the ecliptic, the length of the tropical year, and the parallax of the fixed stars. He saw the necessity for a revision of the existing star catalogues and for the exact determination of their positions. In 1803, he published a list of 6,784 stars and in 1814 a second catalogue containing no less than 7,646 stars. While he was at his regular nightly work in early 1801, Piazzi made the discovery of his life: a heavenly body he identified as a fixed star at first. But upon repeating the measurements during the following nights, he found that this star had shifted slightly. Therefore it would rather have been a planet, or as he called it cautiously, *a new star.* Unfortunately, he was not able to observe this new star any longer as it was lost in the glare of the Sun. Even worse, he was unable to compute its orbit with the then available methods for regaining it after passing the Sun. Fortunately, the German mathematician Carl Friedrich Gauss[27], another star in the science firmament, read about Piazzi's discovery and difficulties and, after some weeks of intense calculations, developed a new method of orbit determination that allowed Piazzi to successfully locate the new star again. Gauss published his results in 1809 under the name of "Theoria motus corporum coelestium in sectionibus conicis solem ambientum" (theory of motion of the celestial bodies moving in conic sections around the Sun), which was to become a major cornerstone of the art of astronomical computation. Then only, it became clear that Piazzi's assumption was correct and this object was not a comet but much more like a small planet. Piazzi named it *Ceres Ferdinandea,* honouring the Roman and Sicilian goddess of grain Ceres and his sponsor, King Ferdinand I. The Ferdinandea part of the name had to be dropped later for political reasons, while the reference to the immortal goddess could be maintained without causing worries. Ceres turned out to be the first, and largest, of the asteroids in the belt between Mars and Jupiter. Under the terms of a 2006 International Astronomical Union resolution, Ceres is now called a dwarf planet, just like the former planet Pluto.

Giuseppe Piazzi had the fortune to find a wealthy patron willing to support his scientific endeavours. Others are wealthy enough to devote themselves to their passion like for instance the British William Lassell[28], who made a fortune by brewing beer. But he was not only interested in fine beverages, but also in astronomy; so he decided to build an observatory near Liverpool which he equipped with a respectable 24-inch reflector telescope. Like many others, he ground the mirrors himself to an excellent quality, but in contrast to others he pioneered an equatorial mount of his telescope for easy tracking of celestial objects as the Earth rotates. Et voilà: spectacular discoveries did not wait for long. Invited by William Herschel to observe Neptune, he discovered Triton, its largest moon in 1846, in 1848 he independently co-discovered Hyperion, a moon of Saturn, and in 1851 Ariel and Umbriel, two

---

[26] Giuseppe Piazzi, 1746, Ponte in Valtellina (Italy) – 1826, Naples, Italian monk, mathematician, and astronomer.
[27] Carl Friedrich Gauss, 1777, Brunswick, Lower Saxony (Germany) – 1855, Göttingen, Hanover, German mathematician and scientist.
[28] William Lassell, 1799, Bolton, Lancashire (UK) – 1880, Maidenhead, British astronomer.

new moons of Uranus. Lassell realized that the atmospheric conditions in Liverpool were unfavourable for astronomy at that time already and, thriving for new discoveries, he decided to move to Malta, where he made further observations until his death in 1880.

The Martian Saga

The next to silently enter the astronomical scene is Giovanni Virginio Schiaparelli[29]. As hydraulic engineer and civil architect he received an appointment as a teacher of mathematics in an elementary school in Turin in 1856, which however did not fully satisfy him. "Without taking into account my almost absolute poverty", as he wrote later, he decided to devote himself to astronomy. In 1857, a small stipend terminated his poverty at least partially, and enabled him to receive training in astronomy in Berlin. Upon his return to Italy he assumed the post as *secondo astronomo* at the famous observatory at the Brera Palace in Milan. Although the instruments were outdated, Schiaparelli made the best use of the available resources, and in April 1861 he succeeded in making a major discovery: Hesperia, the 69th asteroid. However, political turmoil suddenly changed his life. In 1860, Giuseppe Garibaldi dispossessed the King of the Two Sicilies, and in 1861 Victor Emmanuel II became the first King of the united Italy. The new sovereign provided Schiaparelli with a telescope of far superior quality that allowed him to study the planets in more detail. During the fall of 1877 Mars was close to the Earth. This constellation prompted Schiaparelli to study this neighbouring planet carefully for the purpose of drawing up a new map. Thanks to his meticulous observations, the result was a tremendous advance over anything that had appeared before. This forced him, however, to set up a new nomenclature of the Martian topography that was consistent with his observations. Exploiting his intimate knowledge of classical literature and the Bible, he chose the names for the topology of Mars that are still valid today, **Fig. 10.** His map was the first to include surface features called "canali" by Schiaparelli. This term was later mistranslated to the English "canals", a failing that played an important role in the planet's subsequent mystification.

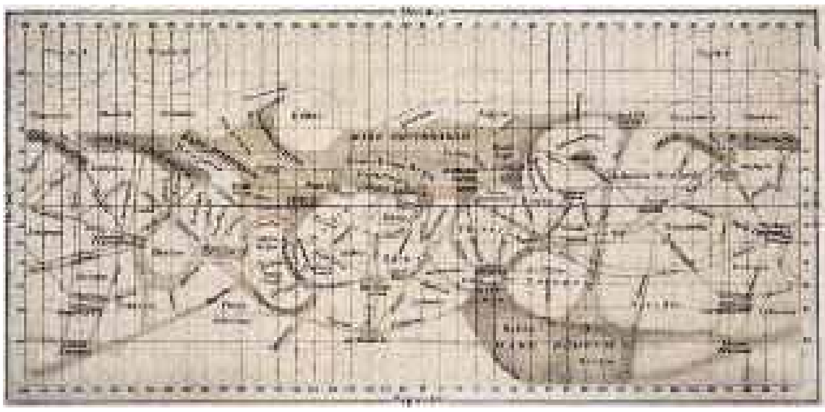


Fig. 10: The map of Mars by Giovanni Virginio Schiaparelli, 1890.

Unfortunately, Schiaparelli lost eye sight in his older years. This in turn was recognized as the chance of his life by one of his diligent readers, Percival Lowell[30]. The seminal mistranslation of the "canali" to the English "canals" triggered the businessman, a descendant of the distinguished and wealthy Boston Lowell family, to set up his own observatory. In search of the best-suited site, he travelled throughout the United States to finally choose Flagstaff, Arizona, at an altitude of over 2,500 m, where he built an excellent observatory. Lowell's belief was that the Martians were keeping their planet alive via a global channel network that brings water from the Martian polar caps to the equatorial zones. He published this hypotheses in three books including detailed maps of the Martian surface: this was the beginning of the Martian saga and the intelligent Martians.

[29] Giovanni Virginio Schiaparelli, 1835, Savigliano (Italy) – 1910 Milan, Italian astronomer.
[30] Percival Lowell, 1855, Boston (USA) – 1916, Flagstaff, Arizona, US American astronomer.

The Birth of Modern Astronomy

While Lowell's work strongly fostered the general public's interest in space research, it had only a limited impact on space science. The contrary is true for another American, who in his younger years studied law in Oxford, but was fascinated by astronomy also: Edwin Hubble[31]. After studies at the University of Chicago he was offered a position at the Mount Wilson Observatory in California, where the world's largest telescope was just completed at that time. His main interest was focussed on the cepheids, a class of stars that periodically brighten and dim again with a frequency which is a uniform function of their brightness. Therefore, the intrinsic brightness can be inferred from the period, independently from their actual distance from the Earth, qualifying them as the ideal yard sticks to measure distances in the universe. Based on these findings he discovered the proportionality between the distances of the cepheids and their velocity away from the Earth, more precisely with their red shift. He was able to plot a trend line that showed that the universe is rapidly expanding: the Big Bang was born. This in turn was bad news for Albert Einstein[32] on the other side of the Atlantic, who, some years before, had formulated his cosmological equation: in line with the general consensus at that time he had assumed that the universe is static and eternal. To this end he had to tune the equation appropriately by introducing an artificial term, which became known as the cosmological constant. When he learnt about the findings of Hubble, Einstein quickly cancelled this term and qualified it as the biggest blunder of his life.

Fig. 11: Aerial view of the European Southern Observatory Paranal site with the Visible and Infrared Survey Telescope in the foreground and the four Very Large Telescopes in the background (Credit: Gerhard Hüdepohl, ESO)

That was around 1929, when not only scientific breakthroughs shocked the world, but also the collapse of the Wall Street stock market followed by a global economic crisis. In Germany, Hitler came to power some years later. World War II broke out. Science and engineering were forced to serve the development of new and more efficient weapon systems. This was the case also for a group around Wernher von Braun[33] in the German Peenemünde who was developing liquid-fuel rocket engines for aircraft and jet-assisted take-offs as well as for long-range ballistic missiles. After WWII, his group was transferred to Fort Bliss, Texas, where these *"Prisoners of Peace"* continued their work in rocketry and helped the United States to build up an own rocket programme. In 1946, the first US-built V2 rocket was launched at White Sands Missile Range. The American government, however, was not really interested in their work at that time and only embarked on a modest rocket-building programme. That, however, changed dramatically from one day to the other, when in autumn 1957, the Soviet Sputnik 1 spacecraft beeped around the Earth

[31] Edwin Powell Hubble, 1889, Marshfield, Missouri (USA) – 1953, San Marino, California, US American astronomer.
[32] Albert Einstein, 1879, Ulm (Germany) – 1955, Princeton (USA), theoretical physicist, Nobel Prize 1921.
[33] Wernher Magnus Maximilian Freiherr von Braun, 1912, Wirsitz (Prussia) – 1977, Alexandria, Virginia (USA), German scientist.

as the first artificial satellite. Although it was not equipped with a science payload, it marked the dawn of a new era in astronomy: from now on, the technical means were available to transport scientific instruments to space, where they could observe the universe unhampered by the partially opaque atmosphere, and even more, to visit and to probe celestial objects in situ, at least within our solar system.

It is, however, interesting to note here that the space age did not terminate the research with Earth-borne systems. Rather, rapid technological progress made numerous discoveries with Earth-bound telescopes possible. One of the most famous sites of astronomy is the La Silla **(Fig. 11)**, one of the driest places on Earth, where the European Southern Observatory (ESO) international organization operates a suite of the most advanced telescopes from its headquarters at Garching near Munich, Germany. As an example, just lately, ESO announced that Prof. Mayor of the University of Geneva and his colleagues have discovered there the first extra-solar star that has an Earth-like planet just in the right distance to the central star to possibly harbour liquid water. This extra-solar planet will be a hot candidate for the search for life beyond the boundaries of the solar system with future space missions.

40 Years of Space-Based Astronomy

There is no doubt: space technology has allowed astronomers to make quantum steps in the understanding of the universe in a way that exceeds all what has been done in the millennia before. The sheer number of past, current and planned space missions requires us to concentrate on some of the most important milestones in the history of space-based astronomy.

The Moon

The Moon with its distance of about 384,000 km from the Earth offers itself as a natural first object for space missions. Indeed, only two years after Sputnik 1, i.e. in 1959, the history of space-based astronomy began with the launch of the Russian Luna 2 spacecraft, the very first space mission. On its journey to the Moon, it discovered the solar wind. It then splashed onto the surface of the Earth's companion thereby becoming the first artificial object there. In the same year, in late 1959, the Luna 3 spacecraft reached a Moon orbit that allowed for the exploration of the far side that had never been seen before, **Fig. 12.**

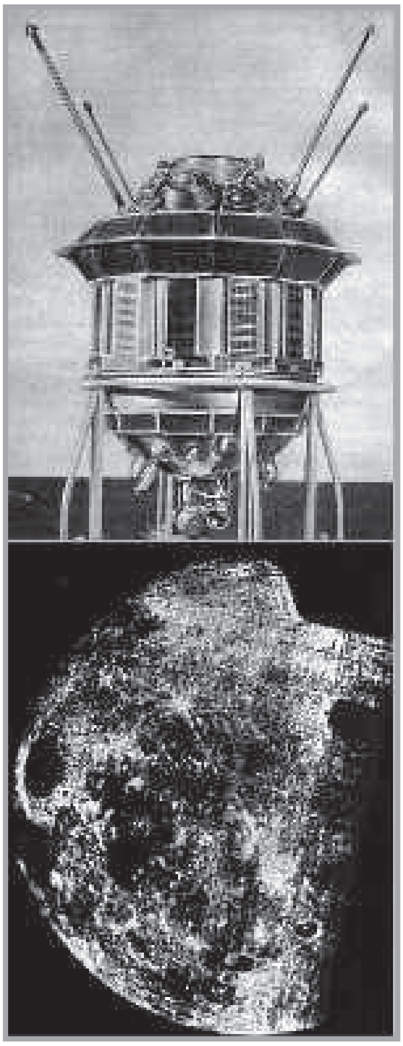


Fig. 12: The Soviet Luna 3 spacecraft. It was the first space probe to reach an orbit around the Moon and to explore its far side (table below).

During the subsequent years, the Soviets rapidly improved their mastering of space technology and went on with the Luna programme. It was in 1966 when the Luna 9 spacecraft performed a spectacular first soft landing on the surface of the Moon and returned the first detailed images from its surface, **Fig. 13.**

The unexpected advent of the Soviet Sputnik 1 spacecraft caused an outright shock in the western world: the United States multiplied their space effort to catch up with the Soviet Union. On 25 May 1961, President John F. Kennedy announced the objective of landing an American astronaut on the Moon before the end of the decade. To this end, NASA was created, and the financial backing of space programmes was dramatically increased. A co-ordinated countrywide approach in such different areas as large rocket motors, computer and telecommunication technology, and many others, allowed NASA to realize President Kennedy's vision on 20 July 1969: Neil A. Armstrong, mission commander of Apollo 11, became the first human to set his foot on the surface of the Earth's companion. "That's one small step for a man, one giant leap for mankind", he reported to ground control. The mission was a complete success, not only with regard to human exploration of space, but also in the field of space science: a few minutes after stepping onto the dusty Moon surface, Armstrong rolled out the solar sail experiment developed by Professor Johannes Geiss of the University of Bern **(Fig. 14)**, thereby executing the first man-tended space experiment. It consisted of an aluminium foil that trapped the particles arriving from the Sun. Upon its return to the laboratories in Bern, the trapped particles were out-gassed from the foil and analyzed to unravel the secrets about the composition of the Sun and the history of the solar system. The experiment was successfully repeated three times during subsequent Apollo missions.

Fig. 14: US Astronaut Neil A. Armstrong. The picture shows him unfurling the solar wind experiment of the University of Bern. (Credit: NASA)

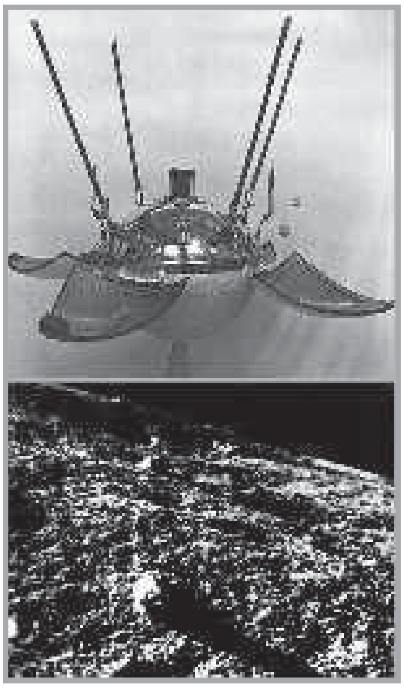


Fig.13: The Soviet Luna 9 spacecraft. It was the first probe that landed on the Moon and returned images from its landing site (table below).

Since 1972, when the last humans explored the Moon, it was visited by automated probes only, such as the US Clementine spacecraft in 1994 and ESA's first mission to the Moon, SMART-1, in 2005. Current plans of NASA as well as ESA intend to exploit the Moon as an attractive first outpost when it comes to travelling to Mars. This vision, however, may turn into reality in the 2030 time frame only.

The Planets

The success of Soviet spacecraft prompted NASA to devise space science missions also. The first successful US spacecraft was Mariner 2, that reached an orbit around Venus in 1962. Its instruments confirmed Venus to be a very strange world covered with cool clouds over an extremely hot surface. The subsequent Mariner 4 spacecraft performed the first close-by trajectory with Mars in 1964 and returned the first detailed pictures of its surface, **Fig. 15.**

Within the class of planets, Mars and Venus are the prime objects of space missions thanks to their orbits that bring them relatively close to Earth. To reach the outer solar system, very powerful launchers are required that provide the spacecraft with sufficient energy, while a journey to the innermost planet Mercury requires complex orbital manoeuvres to bring the spacecraft down to the appropriate orbit. The first space probe to reach Mercury was the US Mariner 10 vehicle. Launched in 1973, it became the first spacecraft successfully exploiting the gravity field of planets to change its trajectory: it flew by Venus and used its gravitational field to bring its perihelion down to the level of Mercury's orbit. This manoeuvre was originally inspired by the orbital mechanics calculations of the Italian scientist Giuseppe Colombo[34]. The images transmitted by Mariner 10 in 1974 are still the best available information on Mercury's surface, as no other spacecraft has visited this planet since.

Breathtaking successes in the advancement of space technology took place on the Soviet side also. Their Venera 13 was the first spacecraft to successfully land on the hostile surface of Venus: the lander survived for two hours in an environment characterized by a temperature of 450 °C and a pressure equalling 84 times that of the Earth's atmosphere, **Fig. 16.** Its instruments witnessed a loud thunder: Venera 13 became the first space probe to record sound from another world.

It was in the mid-nineties that Mars re-entered into the focus of planetary research. This planet is the closest to the Earth and not too dissimilar to our home planet. This qualifies Mars as the preferred object for human exploration. In addition, Mars is suspected to possess liquid water underneath its surface, which, at least theoretically, could be the basis for some actual or fossil form of life. In 1996, NASA launched the Mars Global Surveyor to successfully map the planet's surface paving the way for subsequent exploration missions. One year later, the Mars Pathfinder arrived at Mars together with the rover Sojourner that explored the Martian surface. This mission returned huge amounts of detailed information on the Martian surface. End of 2003, ESA's Mars Express mission entered a Martian orbit from which it continues to transmit detailed three-dimensional images of the planet's surface as well as information on its subsurface structure thanks to its surface-penetrating radar sensor. In 2004, Mars received the visit of the US twin Spirit and Opportunity, two remotely controlled rovers. Although their expected life time was some 90 days they continue to defy the hostile environment and to successfully explore the Martian surface.

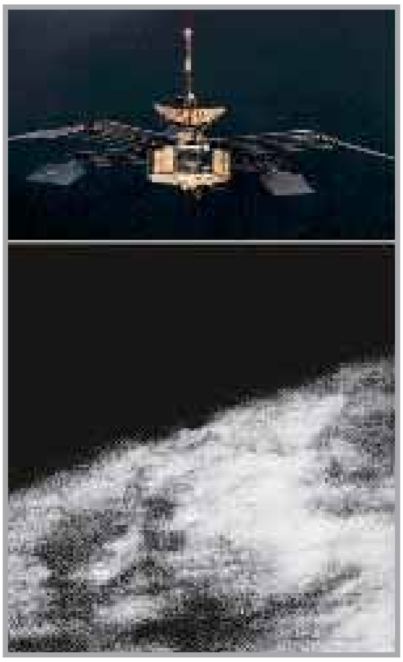


Fig. 15: The US Mariner 4 spacecraft (upper table) and the first close-up picture of Mars (below). (Credit: NASA)

Fig. 16: The Soviet Venera 13 spacecraft (above) and a detail of the Venus surface (below).

[34] Giuseppe Colombo, 1920, Padova (Italy) – 1984, Padova, Italian mathematician and engineer.

The Sun

Mastering complex orbital manoeuvres was one of the key requisites for another milestone in space-based astronomy: Ulysses, a joint ESA/NASA mission that was launched from the US Space Shuttle in 1990. Its objective was to explore the Sun from high latitudes for the first time. The first part of Ulysses' orbit brought the spacecraft away from the Sun towards Jupiter whose enormous gravitational field was exploited to jettison it out of the ecliptic plane on an orbit over the polar regions of the Sun **(Fig. 17, middle table).** After 17 years of operation, its suite of instruments continues to provide first-class scientific data and to contribute to the understanding of our most important star, the Sun[35].

The Outer Solar System

One of the very landmarks in the outer solar system exploration is the US Pioneer 10 spacecraft. Launched in 1972, it was the first to visit the planet Jupiter, after which it followed an escape trajectory out of the solar system. Its last weak signal was received in early 2003, when it was 14 billion km away from Earth, but it presumably continues its journey into the realm of interstellar space.

Just like Ulysses, the NASA Galileo spacecraft was launched from a Space Shuttle flight also. Its main objective was to explore Jupiter and its four moons. On its way, it executed the first asteroid fly-by and arrived at Jupiter in 1995, becoming the first spacecraft to orbit the giant gas planet. One of the stunning results of the Galileo mission is that the moon Europa most probably harbours a water ocean beneath its icy surface which might allow some form of life to develop, as liquid water is thought to be one of the key ingredients for life[36].

The dimensions of the universe by far exceed the capabilities of our imagination. Even the solar system, nothing more than a tiny part of the Milky Way which itself is only one of billions of galaxies in the universe, has dimensions which make its exploration lengthy endeavours that often exceed the active life of the scientists who invented the mission. The US Voyager 1 spacecraft **(Fig. 18)** is no exception thereof: launched in 1977, thirty years ago, it visited Jupiter and Saturn and was the first probe to provide detailed images of the moons of these planets. It is now the farthest man-made object from Earth, still rushing away from the Sun at a faster speed than any other space probe so far. In August 2006, it reached the milestone of 100 Astronomical Units, that is a distance of 15 billion km from the Earth. It is assumed to have entered now the heliosheath, the region between the solar system and the interstellar space[37]. At this distance, the signals from Voyager 1 take more

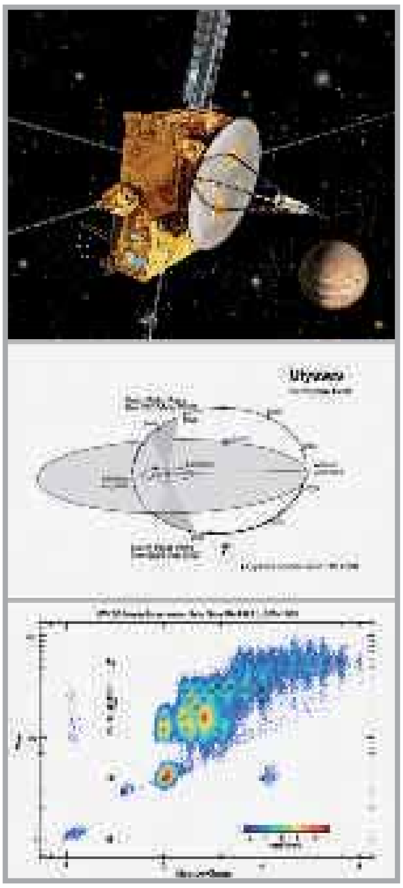


Fig: 17: The Ulysses spacecraft. The current third orbit brings it from Jupiter towards the southern polar regions of the Sun, then to the perihelion in August 2007, to the northern polar regions and back to Jupiter's orbit again. The table below shows example data returned by the Solar Wind Ion Composition Spectrometer (SWICS) instrument that was developed at the University of Bern. It displays the results of solar wind measurements, more precisely the particle counts as a function of particle mass over particle mass per charge ratio. (Credit: Physikalisches Institut, Universität Bern)

[35] See also Spatium 2: Das neue Bild der Sonne.
[36] See also Spatium 16: Astrobiology.
[37] See also Spatium 17: The Heliosphere.

Fig. 18: NASA's Voyager 1 spacecraft. Launched in 1977 it returned the first detailed pictures of the giant planet's atmosphere, of which a false colour image is shown in the table below. (Credit: NASA)

than thirteen hours to reach the control centre on Earth. The spacecraft has reached escape velocity; together with Pioneer 10, Pioneer 11, and Voyager 2 it is now an interstellar space probe forever rushing away from the solar system.

From the Voyager 1 mission, it was known that the Saturnian moon Titan is covered by a dense methane atmosphere. As on Earth methane is always the by-product of biological processes, Titan began to challenge the international scientific community. It was in 1997 that the joint ESA/NASA mission Huygens/Cassini was launched that reached Saturn after a seven years journey in 2004. On Christmas Day 2004, the Cassini spacecraft jettisoned its twin, the Huygens probe, towards Titan, where it arrived on 14 January 2005. A complex and fully automatic sequence of re-entry manoeuvres was needed to reduce the probe's relative speed from some 40,000 km per hour to a velocity that allowed the use of parachutes. During descent, the probe explored the properties of the moon's enigmatic atmosphere, and transmitted the data via the Cassini orbiter to ground control on Earth. Then it performed successfully the first landing in the outer solar system. The probe remained active for a further 70 minutes at an ambient temperature as low as –180°C, gathering the first images from the surface of Titan, while Cassini continues to orbit Saturn and to explore its many moons[38].

A further mission to the outer solar system is NASA's New Horizons spacecraft currently heading towards Pluto to visit the dwarf planet in 2015 for the first time and to explore its moons Charon, Nix and Hydra. Launched in early 2006 by the powerful Atlas V-500 vehicle, the New Horizons spacecraft could be inserted directly into an Earth- and solar-escape trajectory with an Earth-relative speed in excess of 16 km per second qualifying it as the fastest spacecraft launch up to now.

Comets and Interplanetary Dust

So far, space missions concentrated on the exploration of the Moon and of the planets, as their gravity field helps to guide spacecraft on their final course to the target object. Perfect mastering of celestial mechanics is required when it comes to explore smaller objects like comets[39]. It was ESA's Giotto probe that succeeded in 1986 to execute the first close fly-by of a comet and to transmit the first detailed images of its cometary nucleus, **Fig. 19.** The

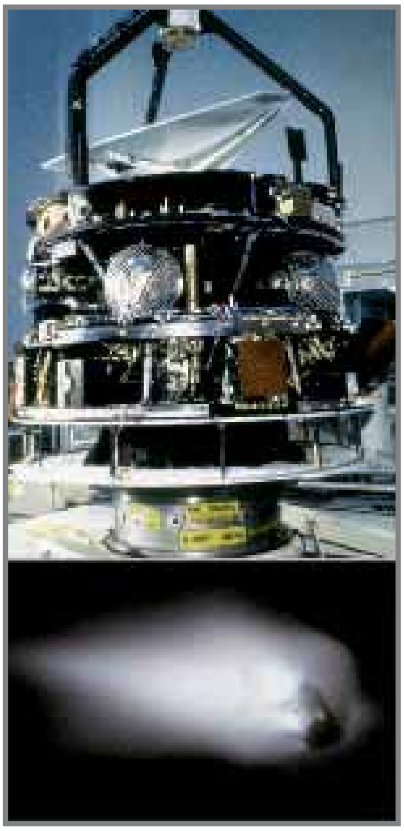


Fig. 19 shows the Giotto spacecraft (upper table). Giotto executed the first visit to a comet at close quarters by flying by the comet Halley in 1986. The lower table shows the first ever image of a comet's nucleus. (Credit: ESA)

[38] See also Spatium 15: Titan and the Huygens Mission.
[39] See also Spatium 4: Kometen.

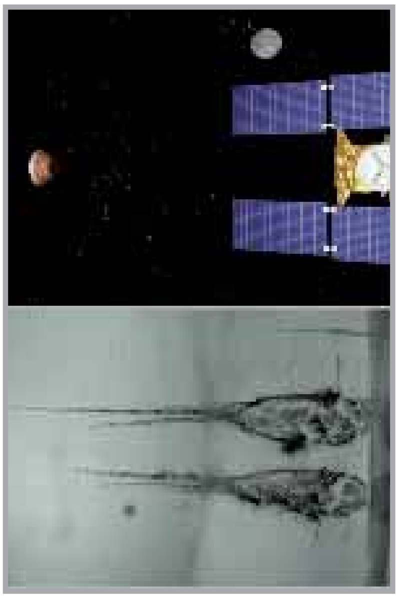


Fig. 20: The Stardust spacecraft. The upper table shows the spacecraft jettisoning the dust sample container back to the Earth, while on the lower table two impacts of dust particles can be seen, that penetrated the storage material from the right. (Credit: NASA)

comet Halley explored by Giotto is a notable comet in the sense that it is the first that was recognized as periodic. Historical records show that Chinese astronomers observed the comet's appearance as early as in 240 B.C. and regular observations after 240 B.C. are recorded by Japanese, Babylonian, Persian, and other astronomers.

It is obvious that celestial bodies like planets, moons and even comets are scientifically attractive objects as they contain much information on their own evolution and that of the solar system as a whole. It is, however, less obvious that the interplanetary dust, that are the particles in the interplanetary space, is a promising scientific object as well. The dust may stem from comets, which in the vicinity of the Sun receive enough thermal energy to evaporate some of their matter into space. But the dust may be the in-fall from interstellar space as well, making it even more promising. The objective of the US Stardust mission **(Fig. 20)** was to collect samples of such dust and to return them back to Earth. More specifically, it was intended to capture both interstellar dust and cometary dust samples from comet Wild 2 and its coma. The dust was captured by a specially developed medium called aerogel, a silicon-based solid with a porous, sponge-like structure, in which 99.8 percent of the volume is empty space. This medium served to decelerate the high velocity dust particles without altering their chemical composition and to conserve them safely during the rest of the flight. After a journey of more than 5 billion km, the sample container landed on Earth in 2006. In an unprecedented effort, the analysis of the dust particles is currently executed by a great number of scientists all over the world.

The Universe

While all the previously mentioned missions were focused on a specific target object such as a moon, a planet, a comet or interplanetary dust, a general purpose space-borne telescope is doubtlessly an attractive further application of space technology. The Earth's atmosphere is transparent in some wavelengths, such as for instance in the visible region, where the Sun by chance has its maximum radiative intensity, but it blocks many other wavelengths preventing Earth-bound sensors to receive any information from the universe. An observatory outside the atmosphere could easily overcome these problems, and it is no wonder that such ideas had been put forward long before the required technological assets were at hand. This exactly is the case for the Hubble Space Telescope (HST), **Fig. 21,** for which the preparative actions had been initiated as early as in 1946. But great things require their time: beset by delays and budget problems, it went into orbit in 1990 as a joint NASA/ESA endeavour only, and when it was there, its main mirror was found to be defective. Fortunately though, the HST was the first spacecraft conceived as an astronaut-tended instrument that periodically would receive updates as the technologies advance. Therefore, NASA decided to execute a repair mission with the objective of upgrading the defective mirror by a specifically designed lens that had to be installed on the orbiting telescope by an astronaut. This was the great chance for our Swiss astronaut Claude Nicollier, who, together with his US colleagues, managed to perfectly execute the challenging repair work in a breathtaking extravehicular activity in late 1993. When the telescope was switched on again after the repair and the now bright and sharp images reached the control centre on the Earth, it became quickly clear that the HST is one of the most important milestones in space-borne astronomy[40].

[40] See also Spatium 12: Ten Years Hubble Space Telescope.

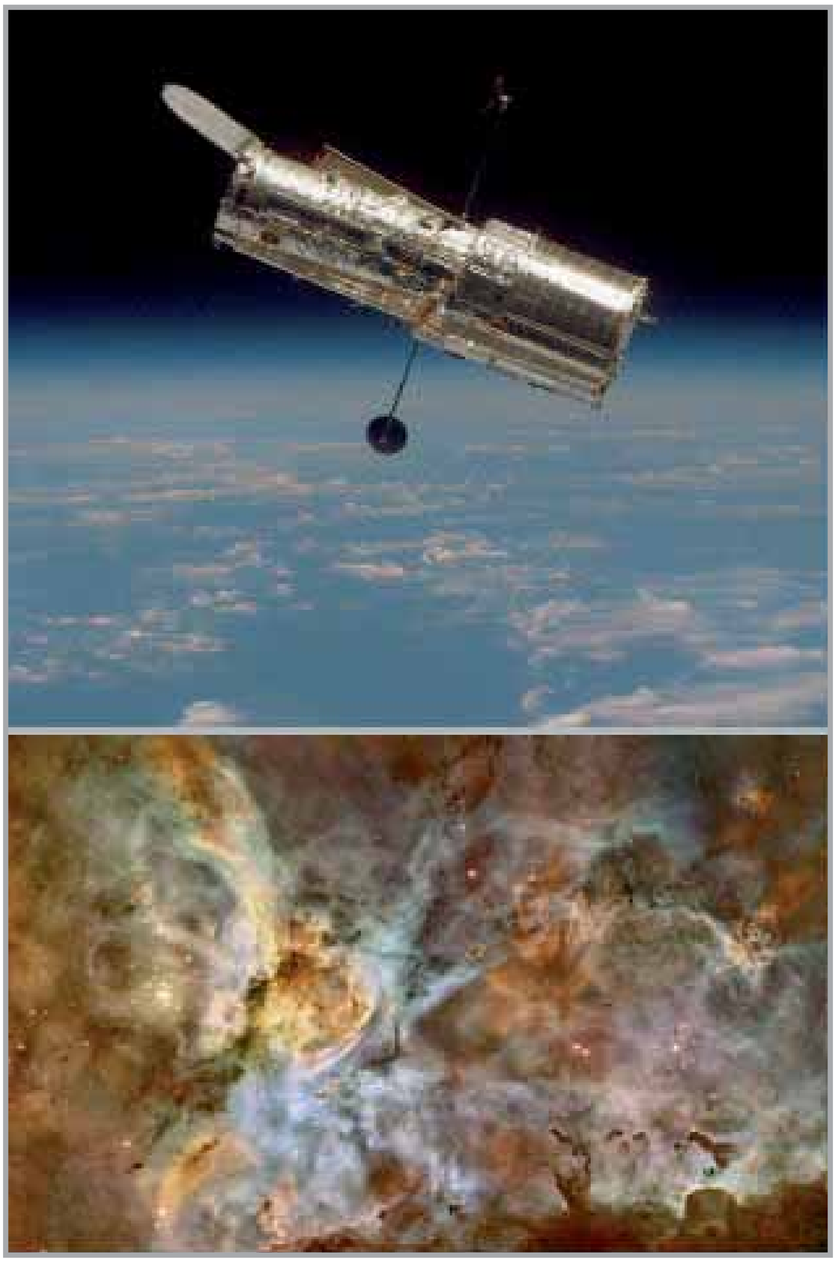


Fig. 21: The Hubble Space Telescope. Since its launch 17 years ago, the spacecraft has transmitted almost 500,000 images of more than 25,000 celestial objects. This information allowed scientists all over the world to publish about 7,000 scientific papers. In celebration of its 17th anniversary a team of astronomers has assembled one of the largest panoramic images ever taken with HST: the lower table shows the central region of the Carina Nebula spanning over 50 light years. (Credit: ESA/NASA)

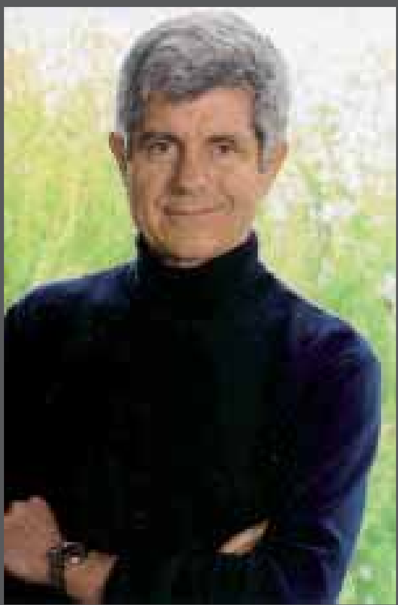
Outlook

Past and ongoing space missions have revealed an incredible amount of information about the solar system and the universe. But, the fascination of science is that an answered question always rises many new ones. Space research makes no exception thereof. One of the key open issues refers to the nature of the dark matter[41]. In the midforties already, the Swiss astronomer Fritz Zwicky became aware of the problem that the dynamic behaviour of galaxies cannot be explained by the universal laws of Kepler taking into account the visible mass of the galaxies only. He therefore stipulated an additional source of gravitational pull to keep the stars together. This missing mass is called now dark matter, but its very nature remains unknown. Another puzzling discovery was made in the last few years: according to our expectation, the expansion of the universe should be continuously decelerating. But in contrast, it seems to be accelerated by a so far unknown force termed dark energy.

These are just two examples that challenge new generations of space scientists. Many problems remain open and may become the subject of fascinating discoveries in the years to come. And beyond all attempts to understand the mechanics of the universe there remains the question of whether the human spirit is really alone in this incredible vast universe.

[41] See also Spatium 7: In Search of the Dark Matter in the Universe.

The Author



Giovanni Fabrizio Bignami received his laureate in physics in 1968 at the University of Milan. Here, he worked as an assistant to the Chair of Advanced Physics the next years. In 1971 he began his career in science politics when he became research scientist at the Italian Consiglio Nazionale delle Ricerche (Italian National Research Council). A fellowship allowed him to work at the NASA Goddard Space Flight Center from 1973 to 1975, where he was involved in the analysis and interpretation of gamma ray astronomy data. During this research activity he discovered one of the brightest neutron stars, the Geminga in the Gemini constellation. Its name is a contraction of "Gemini gamma-ray source", and coincidentally means "it's not there" in the Italian Milanese dialect. This discovery won him 1993 the "Bruno Rossi" Prize of the American Astronomical Society together with J. Halpern.

From 1974 to 1975 G. F. Bignami was a visiting assistant professor at the Catholic University of America in Washington DC, in 1970 he was a visiting scientist at the Max Planck Institute für Kernphysik in Heidelberg, Germany. Later, he served the Italian National Space Agency (ASI) and the European Space Agency (ESA) in several functions. He participated in ESA's Astronomy Working Group (1984–88), in ESA's Space Science Advisory Committee (1994–98), he was an Italian Delegate to ESA's Science Programme Committee (1998–2002), then Vice-Chairman of ESA's Science Programme Committee (1999–2002) and Chairman of ESA's Space Science Advisory Committee (from 2002), just to give an incomplete list of his activities for ESA. On national grounds, he served ASI as Director of Science (1997–2002) in a phase when his budget doubled from 50 M€ to 110 M€. In early 2007, G. F. Bignami's career culminated by the nomination as President of the Italian Space Agency. On the academic side, G. F. Bignami served as Professor in Physics at the University of Cassino, Italy, and Professor of Astronomy at the University of Pavia (from 1997).

One of the key concerns of G. F. Bignami is the co-operation between the two neighbouring countries Italy and France. This is why he served as a member of the Scientific Council of the French Centre National de Recherche Spatiale (CNRS) and later as a Director of the Centre d'Etude Spatiale des Rayonnements at Toulouse.

G. F. Bignami's scientific activities are linked to the majority of space science missions. He held major responsibilities for instance in ESA's XMM-Newton mission as a Principal Investigator for the European Photon Imaging Camera. Quite naturally, such an impressive scientific career led to a number of honours and awards amongst which only the Officier de l'Ordre National du Mérite de la République Française can be mentioned here.

An important endeavour of Bignami concerns public outreach and science popularization activities. He currently gives about 30 public talks per year, writes regularly for major Italian as well as international newspapers and magazines. And last but not least, as a diligent student of old languages, he published a book containing the first English translation (in iambic pentameters) of Galileo Galilei's longest poem.

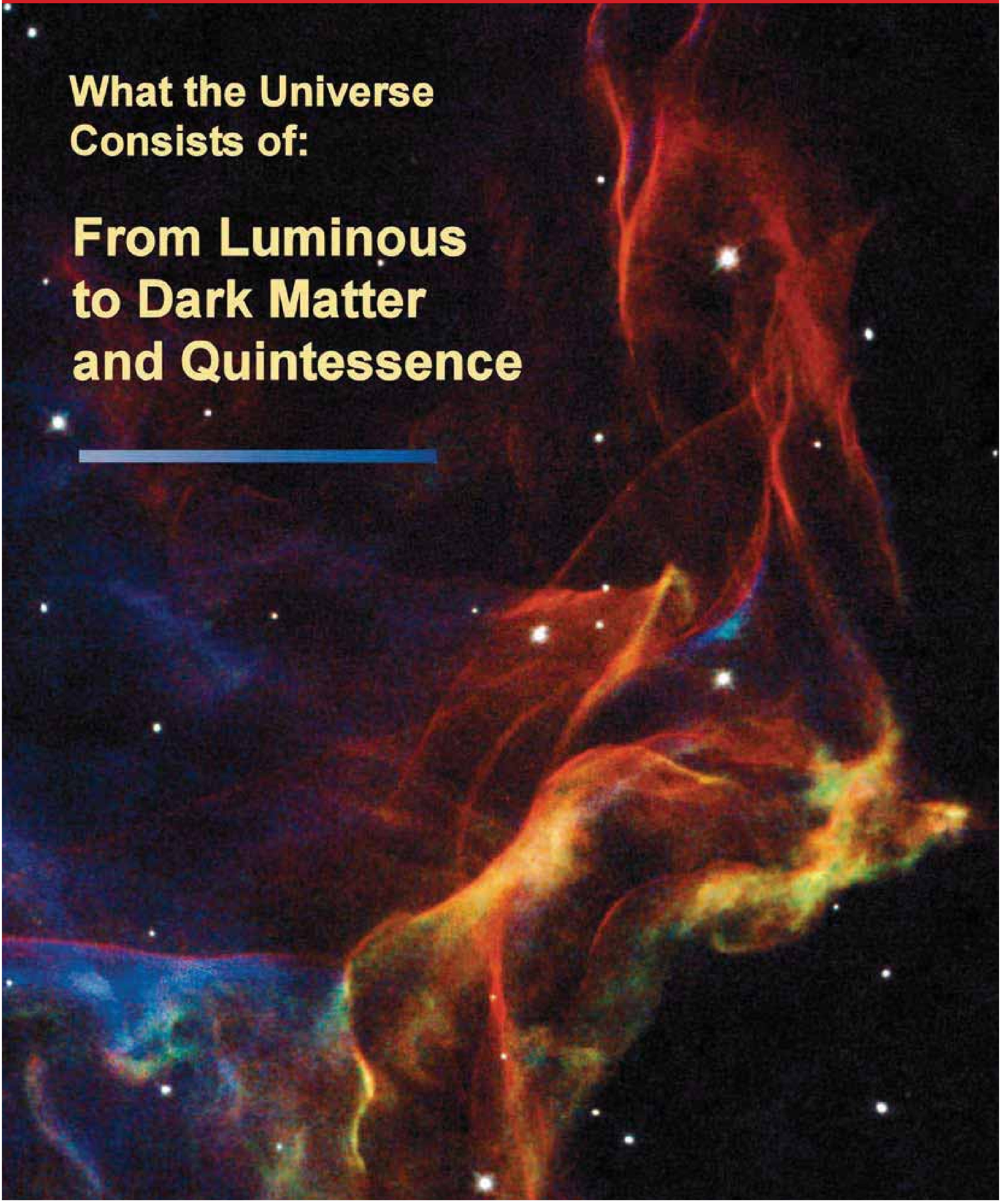
INTERNATIONAL SPACE SCIENCE INSTITUTE

SPATIUM

Published by the Association Pro ISSI
No. 20, November 2007

What the Universe Consists of:

From Luminous to Dark Matter and Quintessence



Editorial

Initially you were the most important entity all around. Later you began discovering your environment that quite naturally constituted the middle of your growing Universe. At school the grasp of your mind expanded rapidly. You learned about the Sun that is so important for your life that it henceforth quietly occupied the centre of your world. Later, you came to know that your beloved Sun is just a fairly small star at an arbitrary spot in a galaxy called the Milky Way that is by far not the centre of the Universe (which has none indeed). Even worse: you had to become accustomed to the idea of a Universe that has been expanding for billions of years at an unimaginable pace and the only open question was whether there is enough mass around for the big crunch to eventually happen or alternatively whether the Universe was to expand forever. Not enough, just a couple of decades ago scientists disclosed that beyond the stars we can see there must be much more hidden mass of unknown nature that by its sheer quantity rules the fate of our Universe. Tellingly, they called it dark matter. And recently they found that the Universe has started rushing out even more quickly than it had done before, driven by an unknown energy. Again, cosmologists found an appropriate designation for their discovery: dark energy, a term that does not hide our ignorance about its true nature.

The present issue of Spatium brings you gradually along this evolution and when you finally close its last page you will have reached that ultimate state of humbleness that Sokrates described so well: I know nothing except the fact of my ignorance. Fortunately though, this purgatory process is not without great reward: your mind will be purged and readied for all the fascinating news science will bring along in the years to come.

We owe much gratitude to Prof. Uwe-Jens Wiese, Director of the Institute for Theoretical Physics at the University of Bern, who authored the present text after his intriguing lecture for the Pro ISSI audience on 27 March 2007. And we congratulate you, dear reader, for the pertinacity to go with us the thorny way that leads to a refined degree of ignorance…

Hansjörg Schlaepfer
Brissago, November 2007

Impressum

SPATIUM
Published by the
Association Pro ISSI

INTERNATIONAL SPACE SCIENCE INSTITUTE

Association Pro ISSI
Hallerstrasse 6, CH-3012 Bern
Phone +41 (0)31 631 48 96
see
www.issibern.ch/pro-issi.html
for the whole Spatium series

President
Prof. Klaus Pretzl
University of Bern
Layout and Publisher
Dr. Hansjörg Schlaepfer
CH-8173 Neerach
Printing
Stämpfli Publikationen AG
CH-3001 Bern

Front Cover
The Veil Nebula, one of the most spectacular supernova remnants in the sky, is seen here by the NASA/ESA Hubble Space Telescope. The supernova explosion presumably occurred some 5,000 to 10,000 years ago and could have been witnessed by ancient civilizations. (Credit: NASA / ESA)

What the Universe Consists of: From Luminous to Dark Matter and Quintessence[1]

Preface

In contrast to the saying that "there is no free lunch", according to Alan Guth from the Massachusetts Institute of Technology (MIT), the Universe is the ultimate free lunch. Luckily for us, the Universe has reached the rather old age of 13.7 billion years which gave us enough time to develop. Also it contains the type of matter that forms the material basis of life. Endowed with great curiosity, we urge to understand the deep meaning of it all. In particular, the curious mind asks questions such as: "Why did the Universe become so old?" and "What does the Universe consist of?". This article is an attempt to explain some of our current knowledge concerning these questions to interested laymen and laywomen.

Having stated this, it is obvious that no mathematical background should be required to understand this text. However, physics is a quantitative science, which is fascinating, in particular, because its results can be expressed in mathematical terms. Indeed, mathematics is the universal language that Nature has chosen to express herself in. Our human language, on the other hand, is often inadequate to accurately describe physical reality. For that reason, the author has decided not to avoid mathematical formulas completely. The formulas require no more mathematical apparatus than multiplication and addition, and are meant to deepen the understanding of some of the problems discussed here. Still, it is not necessary to understand the formulas in order to appreciate the rest of the article. In any case, the author likes to encourage the reader to try to understand the equations. Mathematics is such a beautiful language – even used by Nature herself – that we should not completely forget how to use it. Just as we cultivate the use of foreign languages, we may aim at "speaking" some mathematics as well.

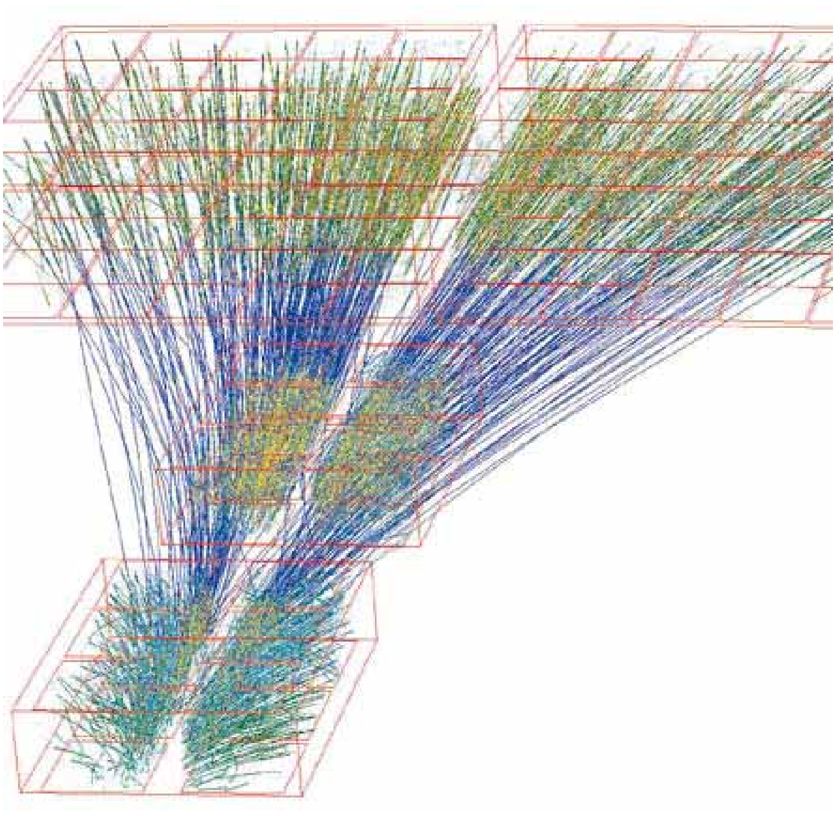


Fig. 1: One of the most fascinating results of modern physics is the fact that the fate of the Universe in its largest dimensions is intimately connected with the properties of elementary particles. Hence, the exploration of the smallest dimensions can provide us with indications regarding the future of the Universe. This picture portrays the result of the collision of a lead projectile on a lead nucleus at CERN, Geneva. (Credit: CERN)

[1] The present issue of Spatium contains a summary of the lecture by Prof. Uwe-Jens Wiese, Institute for Theoretical Physics, University of Bern, held for the Pro ISSI audience on 27 March 2007.

Introduction

The matter that we find on Earth has been investigated by humans ever since the dawn of scientific thinking. Ancient classifications into four essences – air, water, earth, and fire – have long been superseded by our modern understanding of atomic, nuclear, and particle physics. As we have known for about one hundred years, ordinary matter consists of atoms in which negatively charged electrons circle around a tiny positively charged nucleus. The electrons are bound to the nucleus by electric Coulomb forces. These forces are mediated by the electromagnetic field whose smallest quanta are photons – the particles that light consists of. Photons are coupled to any charged particle by the fundamental force of electromagnetism. Since they consist of charged particles, atoms can emit or absorb light. Also the matter in the Sun can emit light only because it contains charged particles. In particular, the luminous matter in the Sun and in other stars is made of the same stuff (namely electrons and nuclei) as the matter that we find on Earth.

Basics of Particle Physics I

Baryons are the family of subatomic particles which are made of three quarks. The family notably includes the proton and neutron, which make up the atomic nucleus, but many other unstable baryons exist as well. The term "baryon" is derived from the Greek barys (heavy), because at the time of their naming it was believed that baryons were characterized by having greater mass than other particles.

The **Large Hadron Collider (LHC)** is a new particle accelerator at the European Organization for Nuclear Research (CERN) in Geneva. It is due to start operations in 2008 and to probe deeper into matter than ever before. It will collide beams of protons at ultra-high energy which is about the equivalent of a flying mosquito, but the protons' energy will be squeezed into a space about a million million times smaller than a mosquito.

MACHOs (Massive Astrophysical Compact Halo Objects), candidates for baryonic cold dark matter.

WIMPs (Weakly Interacting Massive Particles), potential candidates for non-baryonic cold dark matter.

Atomic nuclei consist of positively charged protons and electrically neutral neutrons which in turn consist of quarks and gluons. Unlike electrons or photons, quarks and gluons have never been observed in isolation. They are subject to the strongest fundamental force in Nature – the so-called strong interaction – which permanently confines quarks and gluons inside protons and neutrons. Strongly interacting particles such as protons and neutrons are known as baryons. Since protons and neutrons are much heavier than electrons, the mass of atoms is dominated by baryons. For that reason, the ordinary matter that stars and planets consist of is known as baryonic matter.

Obviously, not all of the baryonic matter is actually luminous. For example, unlike stars, the Earth and the other planets are not sufficiently heavy to ignite nuclear fusion in their cores. Ordinary matter that does not actively shine is known as baryonic dark matter. Since the Sun is much bigger than the planets, most of the matter in the solar system is luminous. The biggest piece of baryonic dark matter in the solar system is the giant planet Jupiter. Dark matter objects such as Jupiter are classified as MACHOs (Massive Astrophysical Compact Halo Objects). Observations of distant galaxies imply that there must be a lot of dark matter in the Universe. In fact, there must be more dark matter than what can be attributed to MACHOs. Potential candidates for non-baryonic dark matter are so-called WIMPs (weakly interacting massive particles). Since these particles are electrically neutral, they cannot emit light and thus they are a form of dark matter. Also they do not participate in the strong interactions, and hence they are non-baryonic. Today there is a lot of indirect evidence for WIMPs, but they have never been detected directly. This may change when the Large Hadron Collider (LHC) at CERN starts operating in 2008.

As first observed by Edwin Hubble in the 1920s, the Universe is expanding, i.e. all galaxies are moving away from each other. Originally it was expected that the expansion would be slowed down due to gravitational attraction between the different galaxies. However, as we learned in 1998 from the observation of very distant supernova explosions, the expansion is actually accelerating. What counteracts the gravitational pull due to the luminous and dark matter inside the different galaxies? Currently, the best candidate is vacuum energy, i.e. energy that fills all of "empty" space, and that is not clustered in galaxies. The vacuum energy density is far too small to be directly detectable in a terrestrial experiment. However, since vacuum energy fills the entire Universe, it adds up to a large amount. Indeed, detailed observations of the cosmic microwave background radiation – a remnant of the hot big bang – indicate that vacuum energy dominates the Universe. The nature of the vacuum energy (which is sometimes also called dark energy) is a big puzzle. A static form of vacuum energy is a cosmological constant, as first introduced by Albert Einstein in general relativity. A dynamical form of vacuum energy is known as quintessence. Since we have learned a lot about air, water, earth, and fire, we may hope to also learn more about the nature of this fifth essence.

In the rest of this article we will take a closer look at luminous and dark matter as well as at quintessence. In particular, we will see that elementary particle physics – i.e. physics on the smallest length scales – has a big impact on cosmology – the physics on the largest scales. The composition of the Universe determines its evolution and is thus also essential for understanding our own existence.

Luminous Matter

In this section we will be concerned with the ordinary baryonic matter that we find on Earth and that has been studied extensively in countless atomic, nuclear, and particle physics experiments. We know that this form of matter exists everywhere in the observable Universe because we see the light that it emits. The light emitted from distant galaxies reaches us after a long journey through the expanding Universe. During this process, the wavelengths of the light are stretched because space itself is expanding. As a consequence, the frequencies are red-shifted. Still, the red-shifted spectra consist of the same spectral lines that are characteristic for the luminous matter here on Earth. In order to understand the structure of ordinary matter, we need to learn something about its constituents, electrons as well as protons and neutrons, which in turn consist of quarks and gluons.

Fundamental Forces

We distinguish four fundamental forces:

- the strong interactions
- the electromagnetic interactions
- the weak interactions
- and gravity.

The electromagnetic interactions, responsible for the binding of atoms, are mediated by the exchange of photons between charged particles such as electrons and nuclei. The strong interactions, responsible for the internal binding of nuclei and of the protons and neutrons themselves, are mediated by the exchange of gluons between quarks. Finally, the weak interactions, responsible for radioactive decay, are mediated by the exchange of heavy W- and Z-bosons **(see Basics II** on the next page). One expects that gravity is mediated by gravitons. However, since gravity is an extremely weak force, its quanta have never been detected. The four fundamental forces act on three types of matter particles – quarks, electrons, and neutrinos. The interactions that the different matter particles participate in are listed in **Basics II.**

The Structure of Matter

The matter that we find on Earth consists of atoms in which electrons orbit an atomic nucleus. The nucleus consists of protons and neutrons, which in turn consist of quarks confined together by gluons. Protons consist of two u-quarks and one d-quark, while neutrons consist of one u-quark and two d-quarks. The electric charge of a u-quark is $Q_u = \frac{2}{3}$ and that of a d-quark is $Q_d = -\frac{1}{3}$, such that the charges of proton and neutron result as

$$Q_p = 2Q_u + Q_d = 2\,\tfrac{2}{3} - \tfrac{1}{3} = 1$$
$$Q_n = Q_u + 2Q_d = \tfrac{2}{3} - 2\tfrac{1}{3} = 0. \quad (1)$$

Protons and neutrons are tiny objects of a size of about 10^{-15} metres.

Basics II

Fundamental forces and matter particles. The following table indicates which fundamental force acts on which matter particle:

Forces	Strong	Electro-magnetic	Weak	Gravity
Mediators	Gluon	Photon	W- and Z-Boson	Graviton
Quarks	yes	yes	yes	yes
Electrons	no	yes	yes	yes
Neutrinos	no	no	yes	yes

The **quarks** are basic constituents of matter, which participate in all fundamental forces including the strong interactions. The other constituents are electrons and neutrinos, which are not strongly interacting. It is quarks that make up protons and neutrons, with three quarks within each of these particles. Protons and neutrons as well as other particles consisting of three quarks are known as baryons.

The **gluon** is an elementary particle that causes quarks to interact, and is indirectly responsible for the binding of protons and neutrons together in atomic nuclei.

Quantum chromodynamics (QCD) is the theory of the strong interaction, one of the four fundamental forces. It describes the interactions of the quarks and gluons found in protons and neutrons. It is an important part of the standard model of particle physics.

The **W- and Z-bosons** are carrier particles that mediate the weak nuclear interactions, much like the photon is the carrier particle for the electromagnetic force.

The **neutrino** is an elementary particle travelling close to the speed of light. It lacks an electric charge, is able to pass through ordinary matter almost undisturbed, and is thus extremely difficult to detect. Neutrinos have a minuscule, but non-zero, mass too small to be measured currently.

Electrons have the electric charge $Q_e = -1$ and are, as far as we know, even point-like. The simplest atom – hydrogen – consists of a single proton that represents the atomic nucleus surrounded by a single electron, **see Fig. 2.** A hydrogen atom is hence electrically neutral because

$$Q_H = Q_p + Q_e = 1 - 1 = 0. \qquad (2)$$

A helium atom consists of two protons and two neutrons forming the atomic nucleus surrounded by two electrons. Again, just as any other atom, the helium atom is neutral because

$$Q_{He} = 2Q_p + 2Q_n + 2Q_e = 2 + 0 - 2 = 0. \qquad (3)$$

The luminous matter in the Sun is ionized, i.e. electrons are no longer bound to the nuclei. The light emitted from the Sun originates from the nuclear fusion of protons and neutrons to helium. When two protons and two neutrons form a helium nucleus, the mass of the nucleus is a bit smaller than the sum of the masses of protons and neutrons, i.e.

$$M_{He} = 2M_p + 2M_n - m. \qquad (4)$$

According to Einstein's famous equation, the small mass deficit m corresponds to a binding energy $E = mc^2$ (where c is the velocity of light) which is liberated in the nuclear fusion process. This energy, which reaches us in the form of sunlight, is the source of life on Earth.

The Origin of Baryonic Mass

The mass of the baryonic matter (either luminous or dark) contributes to the gravitational pull between the galaxies that counteracts the cosmic expansion. Hence, to understand the evolution of the Universe, it is helpful to understand the origin of mass. As we know from Sir Isaac Newton, the gravitational force that attracts us to the planet is proportional to the mass of the Earth. Where does that mass originate from? Well, the Earth consists of baryonic matter whose mass is dominated by atomic nuclei. Since nuclei consist of protons and neutrons, we should ask where their mass comes from. In fact, pro-

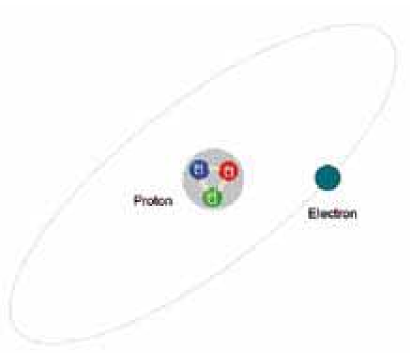


Fig. 2: The hydrogen atom consists of a proton that forms the atomic nucleus and an electron. The proton itself consists of two up-quarks (u), each with a charge of 2/3 and one down-quark (d) with a charge of −1/3. This results in a charge of 1 for the proton. The electron with its charge of −1 makes the hydrogen atom electrically neutral. The quarks in the proton are bound together by the gluons, represented here by the wave symbols between the quarks. The gluons mediate the strong interaction as detailed by the theory of quantum chromodynamics (QCD). This drawing is not to scale; rather the proton has a diameter of 10^{-15} m, while the orbit of the point-like electron is 10^5 times the proton's diameter.

tons and neutrons consist of almost massless quarks and gluons. As we have already seen for helium, the mass of a composite object may differ from the sum of the masses of its constituents due to binding energy. Since quarks and gluons are confined inside protons and neutrons by the strong force, binding energy is the main source of mass. The energy of bound quarks and gluons is described by quantum chromodynamics (QCD, **see Basics II**) – the theory of the strong interactions, which was developed by Harald Fritzsch from the Ludwig Maximilians University in Munich, Murray Gell-Mann from the Santa Fe Institute, Heiri Leutwyler from Bern University, Yoichiro Nambu from the University of Chicago, and by other physicists. Indeed, the strong binding energy of quarks and gluons inside protons and neutrons is the origin of baryonic mass.

The Feebleness of Gravity

We experience gravity as a rather strong force, just because we have the gigantic body of the Earth underneath us. Yet, gravity is much weaker than the other fundamental forces. To illustrate the feebleness of gravity, let us compare the electrostatic repulsive Coulomb force:

$$F_e = \frac{e^2}{r^2} \tag{5}$$

between two protons at a distance r with the attractive gravitational force

$$F_g = \frac{GM_p^2}{r^2} \tag{6}$$

between these particles. Here e is the basic electric charge unit and G is Newton's gravitational constant. In natural units of Geheimrat Max Planck's quantum $\hbar$ and the velocity of light c, the strength of electromagnetism is characterized by the so-called fine-structure constant **(see Basics III)**

$$\alpha = \frac{e^2}{\hbar c} \approx \frac{1}{137}\ . \tag{7}$$

Similarly, the strength of the strong interactions is characterized by a coupling α_s which is close to 1 at low energies. Since α is quite a bit smaller than 1, electromagnetism is a relatively weak force. In natural units of $\hbar$ and c, Newton's constant

$$G = \frac{\hbar c}{M^2_{Planck}} \tag{8}$$

can be expressed through the Planck scale M_{Planck} **(see Basics III and V)**. Combining the equations (5), (6), (7), and (8), the ratio of the gravitational and the electrostatic force between two protons thus takes the form

$$\frac{F_g}{F_e} = \frac{GM_p^2}{e^2} = \frac{\hbar c}{e^2}\left(\frac{M_p}{M_{Planck}}\right)^2 \approx 137\left(\frac{M_p}{M_{Planck}}\right)^2. \tag{9}$$

Hence, if the proton mass M_p would be of the order of the Planck scale M_{Planck}, gravity would be as strong as the electromagnetic interactions. However, the proton mass is much smaller than the Planck scale, namely $M_p/M_{Planck} \approx 10^{-19}$ such that gravity is, in fact, an incredibly weak force with

$$\frac{F_g}{F_e} \approx 10^{-36}. \tag{10}$$

If we want to understand why gravity is so weak, we must hence understand why the proton is so much lighter than the Planck scale. As we discussed before, the proton mass can be derived from the theory of the strong interactions. As was pointed out by Frank Wilczek from MIT, the feebleness of gravity is a consequence of asymptotic freedom **(see Basics III)**, the fact that quarks and gluons interact only weakly at high energies. David Gross from the Kavli Institute in Santa Barbara, David Politzer from the California Institute of Technology (Caltech), and Wilczek won the Nobel prize in 2004 for explaining this property of QCD. Without asymptotic freedom, the proton and thus all baryonic matter in galaxies would be much heavier. Hence, the gravitational pull between the galaxies would be much stronger. In that case, the expansion of the Universe could come to a halt, and the Universe would re-collapse in a big crunch. This underscores how the physics of elementary particles determines the fate of the Universe.

Dark Matter

Based on the motion of galaxies in galaxy clusters, already in the 1930s the Swiss astronomer Fritz Zwicky concluded that besides the visible luminous matter there must be large amounts of hidden dark matter inside galaxies. In the 1970s, the existence of dark matter was confirmed by the measurement of the velocity of stars rotating around the centre of a galaxy. The velocity of a star can be inferred from the emitted light spectrum that is shifted due to the Doppler effect. The galaxy rotation curves show that stars at the edge of a galaxy have larger velocities than one would expect based on the luminous matter inside the galaxy, **see Fig. 3.** According to Newton's law of gravity, the velocity v of a star of mass m rotating around the centre of a galaxy at a distance r is determined by

$$F_g = \frac{GMm}{r^2} = m\frac{v^2}{r} \Rightarrow v^2 = \frac{GM}{r} \qquad (11)$$

Here M is the mass of the matter within the distance r from the centre of the galaxy. Hence, by measuring v through the Doppler effect, one can infer the mass M. If M is larger than the mass of the visible luminous matter, one will conclude that there is also dark matter – unless one concludes that Newton's law of gravity should be modified.

Basics III

Asymptotic freedom is a property of some classes of physical theories in which the interaction between the particles, such as quarks, becomes arbitrarily weak at ever shorter distances, i.e. length scales that asymptotically converge to zero (or, equivalently, energy scales that become arbitrarily large).

The **fine-structure constant** is the fundamental physical constant characterizing the strength of the electromagnetic interaction. It is a dimensionless quantity, and thus its numerical value is independent of the system of units used.

The term **Planck scale** refers to a length scale in the neighbourhood of the Planck length 1.616×10^{-35} m, or a time scale in the neighbourhood of the Planck time 5.390×10^{-44} sec. At this scale, the usual concepts of space and time are expected to break down, as quantum indeterminacy becomes virtually absolute.

Dark Matter versus Alternative Theories of Gravity

As was pointed out by Sean Carroll from Caltech in one of his inspiring popular talks, the interplay between dark matter and theories of gravity has a long history. It had been known since the 1820s that the orbit of the planet Uranus was not completely accounted for by Newton's law applied to the Sun and the seven planets known at that time. In 1846 the mathematician Urbain Le Verrier explained the irregularities in the orbit of Uranus by predicting the existence of "dark matter": an eighth planet. It was a great triumph of Newton's law that Neptune was discovered almost exactly in the predicted position later in the same year. According to our modern classification, Neptune is a form of baryonic dark matter. In fact, the sunlight reflected from its surface is so dim that, in contrast to Uranus, Neptune cannot be seen with the naked eye. Inspired by his success, Le Verrier also investigated the irregularities in the orbit of Mercury and predicted another planet which he called Vulcan. The many false claims of Vulcan's discovery ended only after Einstein had explained Mercury's perihelion shift as an effect of general relativity. In that case, there was indeed no hidden dark matter to be found. Instead, Newton's law of gravity had to be modified. One should keep in mind that, after the construction of general relativity, Newtonian gravity is not wrong – it is only incomplete. In fact, Newton's theory of gravity re-emerges from Einstein's general relativity in the limit of small velocities and weak gravitational fields.

After this brief discourse on the history of science, we might ask if the rotation curves of galaxies indeed imply the existence of dark matter. Moti Milgrom from the Weizmann Institute in Rehovot has challenged this standard view in his model of modified Newtonian dynamics (MOND). In MOND, Newton's law is modified such that the rotation curves of galaxies are described without the need for dark matter. However, unlike general relativity, MOND is not a fully satisfactory theory of gravity. It is designed to avoid dark matter, but it does not explain all the rest of gravitational physics. Indirect evidence for dark matter also comes from the analysis of the cosmic microwave background radiation to which we will turn later. Let us therefore adopt the generally accepted conclusion that, besides luminous matter, there must be large amounts of dark matter within galaxies.

Dark Matter Candidates

As we said before, massive planets like Jupiter qualify as baryonic dark matter. By gravitational lensing, similar objects have been identified in the halo of our galaxy. MACHOs are definitely a viable dark matter candidate. However, their number is too small to account for all of the dark matter. Other dark matter candidates are WIMPs, **see Basics I**. In contrast to MACHOs, WIMPs are individual weakly interacting massive elementary particles. They should not participate in the strong or electromagnetic interactions. If they did carry an electric charge, they could scatter light and would thus be visible. Hence, WIMPs must be electrically neutral. If they did participate in the strong interactions, they would be bound inside exotic baryons, which have never been detected. Hence, WIMPs must be a non-baryonic form of dark matter. According to **Basics II**, neutrinos indeed do not interact strongly or electromagnetically. They only participate in the weak interactions and in gravity. Until recently, it was not known if neutrinos carry mass. However, due to tremendous recent progress in neutrino physics, we now know that neutrinos indeed have a small but non-zero mass. Indeed, besides the cosmic microwave background radiation of photons, there must be a similar background radiation consisting of an enormous number of neutrinos. Due to their large number, neutrinos are an important dark matter component. Since they are not very massive they are not classified as WIMPs. Because they are light and were produced with very high energies immediately after the big bang, neutrinos are a form of so-called hot dark matter. Hot dark matter alone would not lead to the structures observed in the galaxy distribution today. Hence, one expects that there are, in addition, large amounts of cold dark matter consisting of very massive weakly interacting elementary particles.

The standard model of the known elementary particles does not include a massive WIMP candidate. However, some extensions of the standard model, for example, those with supersymmetry (SUSY) predict the existence of such particles. If it is stable, the lightest SUSY particle is a WIMP and thus a good cold dark matter candidate. Using cryogenic detectors (including the ORPHEUS detector in Bern shown in **Fig. 5**[2]) one has tried to detect WIMPs which should be present everywhere in the galaxy. However, their interactions are so weak that none have been observed until now. In the near future, the Large Hadron Collider at CERN promises to provide enough energy to produce and detect WIMPs in a controlled collider experiment **(see Fig. 6).** Again, it is particle physics that holds the promise to teach us more about the composition of the Universe.

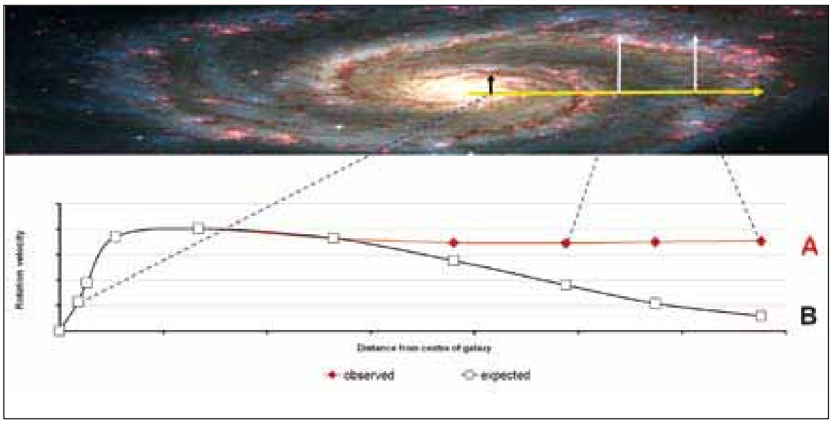
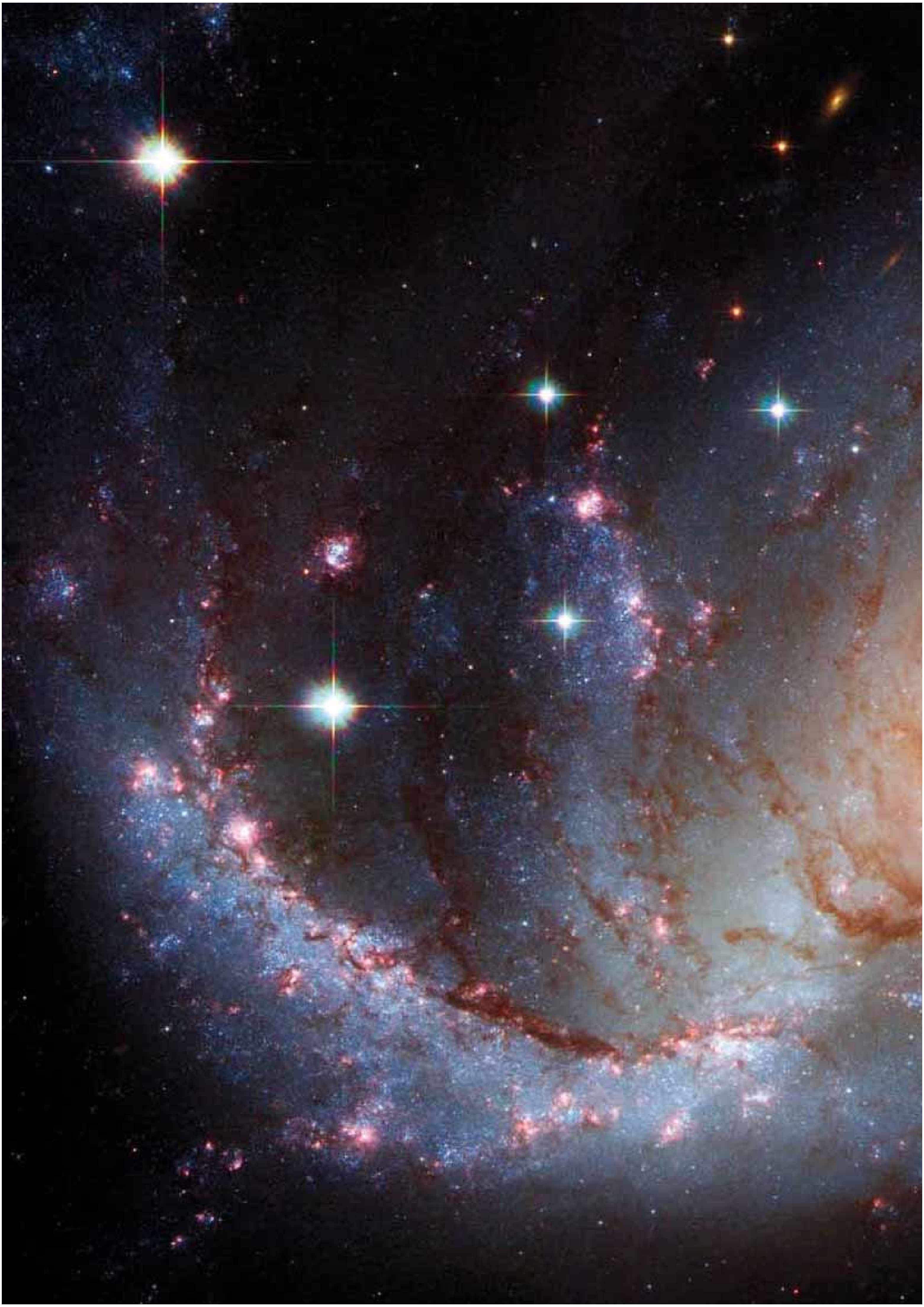
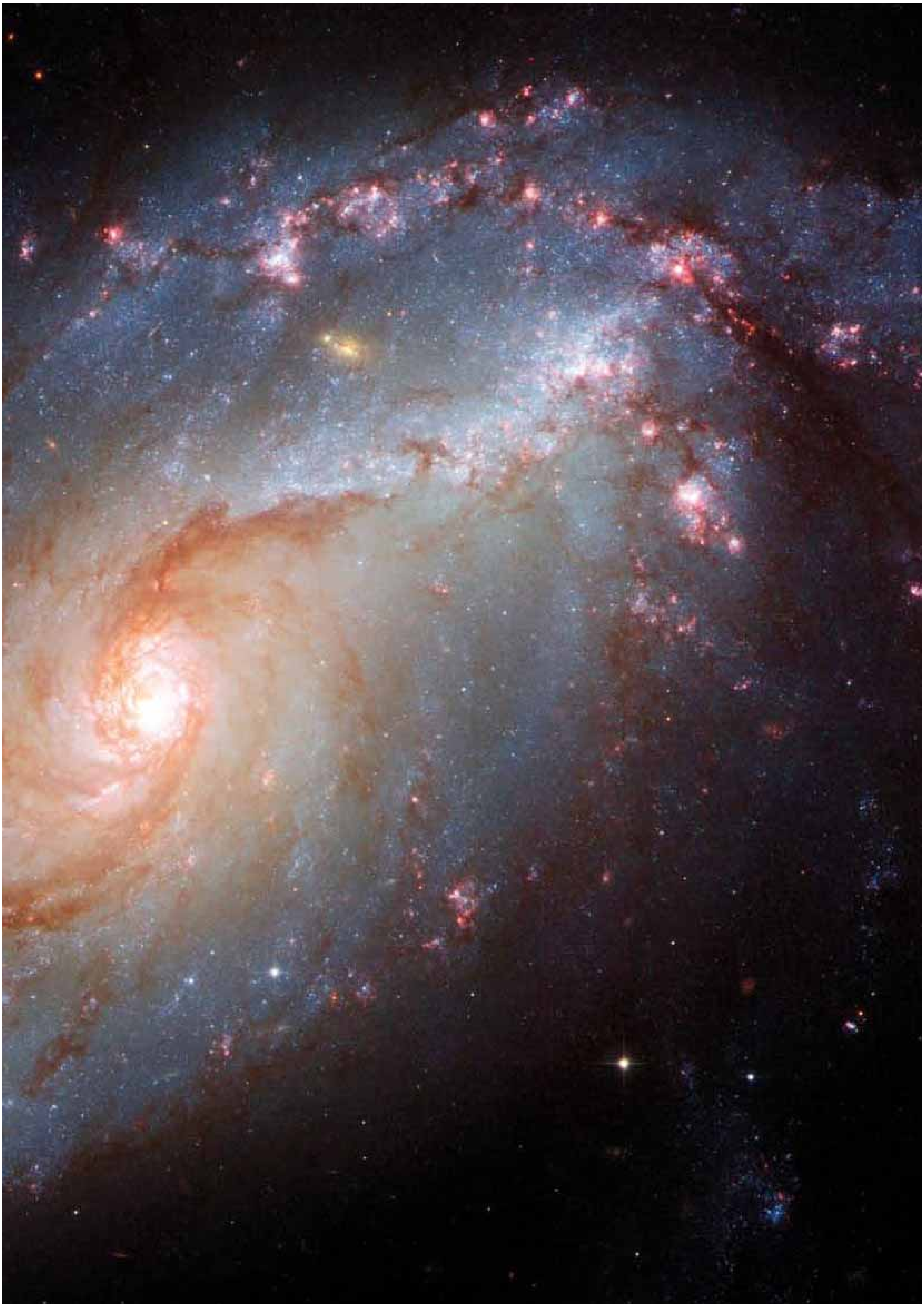


Fig. 3: Rotation curve of a spiral galaxy. The observed rotation velocity of stars at a large distance from the centre of the galaxy (A) is larger than one would expect based on the luminous matter (B). The discrepancy can be explained by the presence of dark matter.

Fig. 4 (on the next pages): ***Beautiful Universe.*** This Hubble Space Telescope view of the spiral galaxy NGC 1672 unveils details in the galaxy's star-forming clouds. Dust lanes extend away from the nucleus and follow the inner edges of the galaxy's spiral arms, where clusters of hot young blue stars form and ionize surrounding clouds of hydrogen gas that glow red. (Credit: NASA, ESA, and The Hubble Heritage Team)

[2] See also Spatium 7: In Search of Dark Matter in the Universe





Basics IV

The **Higgs field,** named after the British physicist Peter Higgs, is a postulated quantum field, which is believed to permeate the entire Universe. Its presence is required in order to explain the large mass of those particles which mediate the weak interactions (the W- and Z-bosons). The photon, which mediates the electromagnetic interactions, on the other hand, is massless.

The **cosmic microwave background radiation** is a form of electromagnetic radiation discovered in 1965 that fills the entire Universe. It is interpreted as the best evidence for the big bang model that stipulates the early Universe to be made up of a hot plasma of photons, electrons and baryons. As the Universe expanded, adiabatic cooling caused the plasma to cool until it became favourable for electrons to combine with protons and form hydrogen atoms. This happened at around 3,000 K or when the Universe was approximately 380,000 years old. At this point, the photons did not scatter off the now neutral atoms and began to travel freely through space. The photons have continued cooling ever since; they have now reached 2.725 K and their temperature will continue to drop as long as the universe continues expanding. Accordingly, the radiation from the sky we measure today comes from a spherical surface, called the surface of last scattering, from which the photons that decoupled from interaction with matter in the early Universe, 13.7 billion years ago, are just now reaching observers on Earth.

The inflaton is the generic name of the unidentified scalar field that may be responsible for an episode of inflation in the very early Universe. According to inflation theory, the inflaton field provided the mechanism to drive a period of rapid initial expansion that shaped the Universe immediately after the big bang. The name inflaton follows the convention of field names, such as photon field and gluon field, which end with „-on".

The Origin of Non-Baryonic Mass

As we have seen, the origin of the mass of baryonic dark matter is well understood in terms of strong interaction energy of quarks and gluons inside protons and neutrons. We have also seen that neutrinos and WIMPs are candidates for hot and cold non-baryonic dark matter. What is the origin of this non-baryonic mass? For example, what is the origin of the neutrino masses? In the original minimal version of the standard model of elementary particle physics, neutrino masses were even zero. After the discovery of non-zero neutrino masses, we know that we must use an extended version of the standard model. In this model the neutrino masses are free parameters, which are at least naturally small due to a mass hierarchy mixing mechanism (known as the see-saw mechanism) first discussed by Peter Minkowski from Bern University. Also the masses of electrons and quarks are free parameters not predicted by the theory. These parameters are related to couplings to the so-called Higgs field **(see Basics IV)**, which has a non-zero value in the vacuum as a result of electroweak symmetry breaking. The dynamics of the Higgs field thus affects the masses of the other elementary particles. One major goal of the Large Hadron Collider is to produce a quantized oscillation of the Higgs field – the so-called Higgs particle – which is the last still unobserved particle in the standard model. The search for the Higgs particle may also shed some light on the dynamics of electroweak symmetry breaking and thus on the origin of the masses of the other particles. If SUSY WIMPs will

Fig. 5: The cryogenic ORPHEUS detector in the underground laboratory at Bern University has been used in the search for WIMPs. Due to their very weak interactions, the detection of WIMPs is extremely difficult and has not yet been achieved. See also Spatium-7: In Search of Dark Matter in the Universe.

be produced at the LHC, their mass should result from supersymmetry breaking which is a highly speculative subject. Hence, in contrast to baryonic matter, the origin of the mass of non-baryonic dark matter is presently much less well understood. It will require an enormous amount of experimentation at very high energies to make progress on these exciting questions. The LHC is a first giant leap in this direction.

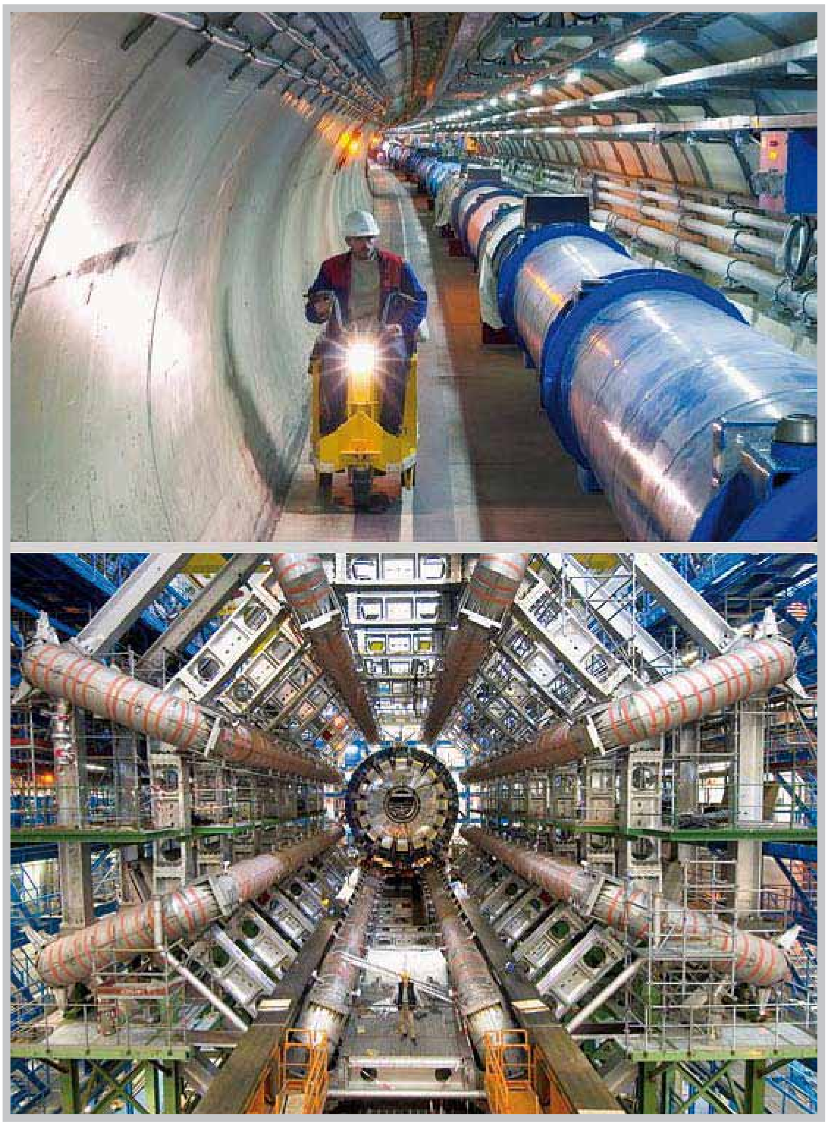


Fig 6: The Large Hadron Collider (LHC) experiment at CERN: The upper plate shows a section of the 27 km long underground ring-tunnel in which protons will be accelerated and collided with each other at the highest energies ever reached in any collider experiment. The lower table shows the gigantic ATLAS (A Toroidal LHC ApparatuS) detector built by an international collaboration including the Laboratory for High-Energy Physics at Bern University. The detector will be used to search for Higgs particles, WIMPs, and other as yet undiscovered elementary particles.

Vacuum Energy

As observations of very distant supernova explosions and detailed investigations of the cosmic microwave background radiation have shown, luminous and dark matter alone cannot explain the composition of the Universe. Indeed, there is evidence for vacuum energy which, unlike luminous or dark matter, does not cluster inside galaxies but fills all of space homogeneously. This form of energy counteracts the gravitational pull of the matter and leads to an accelerated expansion of the Universe. Vacuum energy may be represented by a static cosmological constant or by dynamical quintessence. In the framework of the inflationary Universe, vacuum energy also plays a central role immediately after the big bang.

Evidence for Vacuum Energy

Supernova explosions are among the most violent events in the Universe, **see front cover**. When a large star runs out of nuclear fuel, it collapses under its own gravity. The resulting shock wave expels most of the mass of the star in a gigantic explosion. One distinguishes different types of supernova explosions. Those of type Ia are events that emit a well-defined amount of energy and follow a standard pattern of luminosity change over time. When a distant supernova of type Ia is observed on Earth, it serves as a standard candle whose distance can be inferred from its apparent luminosity because the actual brightness is known. At the same time, the velocity of the emitting galaxy resulting from the expansion of the Universe can be deduced from the red-shift of the emitted light spectrum. In 1998 two international collaborations, the supernova cosmology project and the high-z supernova search team, found that distant supernovae are fainter than one would expect for a Universe whose expansion is decelerated by gravitational pull, **Fig. 7.** Consequently, very distant supernovae of type Ia provide evidence that the expansion of the Universe is actually accelerating. A sufficient amount of positive vacuum energy can indeed counteract the gravitational pull of the matter and give rise to an accelerated expansion of the Universe.

> **Basics V**
>
> **Man-made versus Natural Units**
>
> The most basic physical quantities – length, time, and mass – are measured in units of metres (m), seconds (sec), and kilogrammes (kg). Obviously, these are man-made units appropriate for the use at our human scales. For example, a metre is roughly the length of an arm, a second is about the duration of a heart beat, and a kilogramme comes close to the mass of a heavy meal. Expressed in man-made units, Nature's most fundamental constants are: the velocity of light:
>
> $$c = 2.9979 \times 10^{8}\, m/\text{sec},$$
>
> Planck's quantum
>
> $$\hbar = 1.0546 \times 10^{-34}\, kgm^2/\text{sec},$$
>
> and Newton's gravitational constant
>
> $$G = 6.6720 \times 10^{-11}\, m^3/kg\ \text{sec}^2,$$
>
> Appropriately combining these fundamental constants, Nature provides us with her own natural units (also known as Planck units): the Planck length
>
> $$l_{Planck} = \sqrt{G\hbar/c^3} = 1.6160 \times 10^{-35}\, m,$$
>
> the Planck time
>
> $$t_{Planck} = \sqrt{G\hbar/c^5} = 5.3904 \times 10^{-44}\, \text{sec},$$
>
> and the Planck mass
>
> $$M_{Planck} = \sqrt{\hbar c/G} = 2.1768 \times 10^{-8}\, kg.$$
>
> Planck units are not too practical in our everyday life. For example, an arm has a length of about 10^{35} l_{Planck}, a heartbeat lasts roughly 10^{44} t_{Planck}, and our body weighs about 10^{10} M_{Planck}. Still, l_{Planck}, t_{Planck}, and M_{Planck} are the most fundamental basic units that Nature provides us with. It is interesting to ask why we exist at scales so far removed from the Planck scale defined by these three fundamental units. For example, why does a kilogramme correspond to about 10^8 M_{Planck}? In some sense, this is a historical question. The amount of platinum-iridium alloy deposited near Paris about a hundred years ago, which defines the kilogramme, obviously is an arbitrarily chosen man-made unit. If we assume that the kilogramme was chosen because it is a reasonable fraction of our body weight, we may rephrase the question as a biological one: Why do intelligent creatures weigh about 10^{10} M_{Planck}? If biology could explain the number of cells in our body and (perhaps with some help of chemistry) could also explain the number of atoms necessary to form a cell, we can reduce the question to a physics problem. Since atoms get their mass from protons and neutrons, we are led to ask: Why is the proton mass given by $M_p \approx 10^{-19}$ M_{Planck}? This physics question is addressed in the main text.

The cosmic microwave background radiation consists of photons that originated from a mass extinction of matter and antimatter about one second after the big bang. The early Universe contained a gas of charged particles and was thus opaque. Only when the Universe had sufficiently expanded and cooled down could atoms form, thus neutralizing the matter. At that time, about 380,000 years after the big bang, the Universe became transparent, and the cosmic photons decoupled from the matter. Hence, the red-shifted cosmic photons that we observe today still carry detailed information about the Universe at the decoupling time. Remarkably, the temperature of the cosmic background radiation is uniform to a degree of 1 in 100,000. The small deviations from uniformity were measured with very high accuracy by the Wilkinson microwave anisotropy probe (WMAP) satellite in 2003, **Fig. 8,** upper table. The WMAP data allow us to reconstruct the composition of the Universe. The total amount of energy in the Universe determines if, on the largest scales, space is curved or flat. It is flat only if the energy density ρ assumes a critical value ρ_c. For a flat Universe one thus obtains the ratio

$$\Omega = \frac{\rho}{\rho_c} = 1. \qquad (12)$$

The WMAP data indicate that $\Omega = 1$ and hence that, on the largest scales, space indeed is flat. The matter fraction $\Omega_M = \Omega_b + \Omega_{nb}$ consists of a baryonic and a non-baryonic contribution, Ω_b and Ω_{nb}, and the vacuum contribution is denoted by Ω_Λ. As illustrated in **Fig. 8**, lower table, the best fit to the WMAP data corresponds to an energy cocktail of only about 5 percent of ordinary baryonic (luminous or dark) matter (i.e. $\Omega_b = 0.05$), about 20 percent of non-baryonic dark matter ($\Omega_{nb} = 0.20$), and 75 percent vacuum energy ($\Omega_\Lambda = 0.75$). Altogether, we thus obtain

$$\Omega = \Omega_M + \Omega_\Lambda = \Omega_b + \Omega_{nb} + \Omega_\Lambda = 0.05 + 0.20 + 0.75 = 1. \quad (13)$$

Interestingly, the amount of dark matter that one infers from the observation of galaxy clusters is consistent with $\Omega_M = \Omega_b + \Omega_{nb} = 0.25$. Furthermore, the observed abundances of light nuclei such as helium and lithium, which were produced in the primordial nucleosynthesis a few minutes after the big bang, also imply $\Omega_b = 0.05$. Finally, the amount of vacuum energy inferred from type Ia supernova explosions is again consistent with $\Omega_\Lambda = 0.75$.

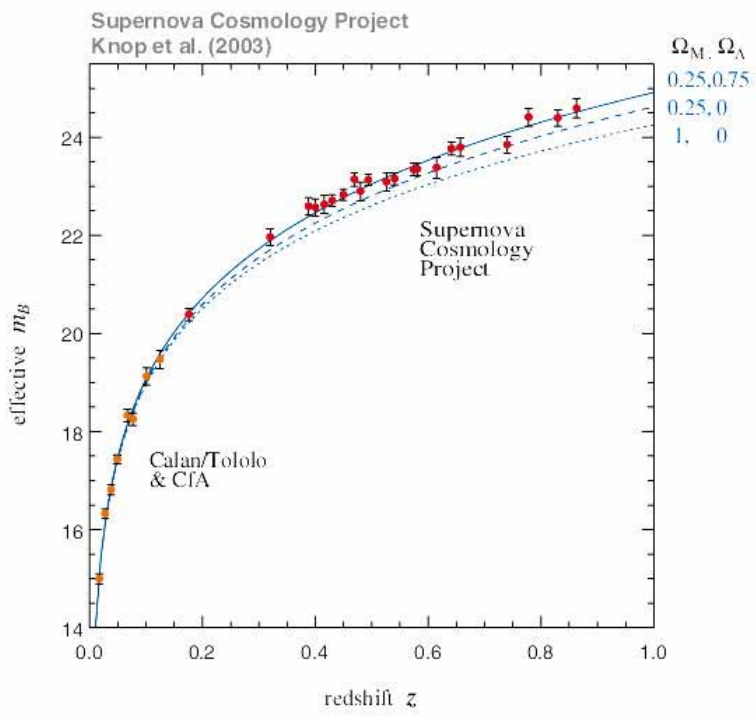


Fig. 7: The Hubble diagram based on very distant type Ia supernovae observed at high redshift obtained by the international supernova cosmology project. The observed luminosity is smaller than originally expected, which implies that the expansion of the Universe is accelerating. The best fit to the data is obtained for $\Omega_M = 0.25$ and $\Omega_\Lambda = 0.75$, indicating that 25 percent of the energy of the Universe is due to (luminous or dark) matter, while 75 percent is vacuum energy.

The Cosmological Constant

What is the nature of the vacuum energy? Unlike the energy of MACHOs or WIMPs that is clustered in galaxies, vacuum energy homogeneously fills all of space. In order to obtain a static Universe, Einstein had included a cosmological constant Λ in the equations of general relativity. Once Hubble had observed the cosmic expansion, Einstein is quoted as considering the introduction of Λ his biggest blunder. In quantum field theory the vacuum fluctuations of the fields do give rise to vacuum energy. However, that energy is formally divergent. Naive attempts to make it finite lead to the estimate $\Lambda = M^4_{Planck} c^5/h^3$ which is a factor of 10^{120} bigger than the amount of vacuum energy necessary to explain the observed accelerated cosmic expansion. This shows drastically that we have presently no idea how to compute the tiny but non-zero vacuum energy density. Indeed, the cosmological constant problem is one of the greatest puzzles in physics today. At present we do not have an established theory of quantum gravity, and hence we do not even know the rules from which one could perhaps derive the value of Λ. We have seen that the property of asymptotic freedom of quantum chromodynamics explains naturally why the proton mass is very much smaller than the Planck scale. Perhaps the correct theory of quantum gravity has a similar property that guarantees the same for Λ.

Inflation and Quintessence

A dynamical form of vacuum energy plays a central role in Guth's inflationary early Universe. Putting aside the cosmological constant problem, he postulates that Λ is indeed zero in the true vacuum – the state of absolute lowest energy. However, he then assumes that in the moment of the big bang the Universe started out in a false vacuum in which the quantum fields did not have their final vacuum values. In particular, a yet unidentified inflaton field **(see Basics IV** on page 12) is suggested to be slowly rolling towards its true vacuum value, **Fig. 10.** The energy of the inflaton field in the false vacuum then acts like vacuum energy. As a consequence, the inflationary Universe

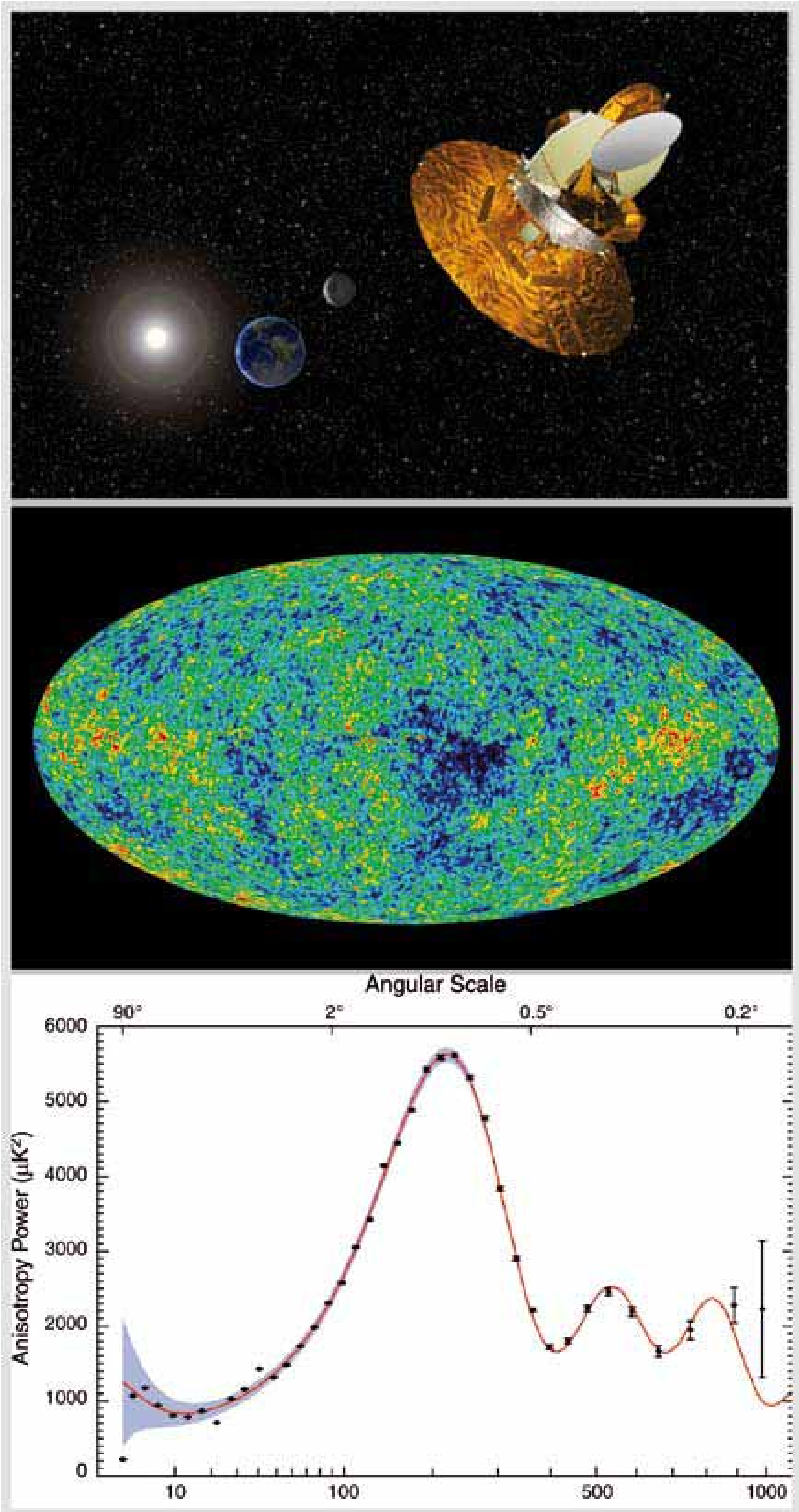


Fig. 8: The antennae of NASA's Wilkinson microwave anisotropy probe (WMAP) point away from the Sun and detect photons of the cosmic microwave background radiation (upper table). These photons were generated immediately after the big bang and have been travelling through the expanding Universe for about 13.7 billion years. The middle plate shows tiny 1 part in 100,000 variations in the temperature of the cosmic microwave background radiation as a function of the angular position in the sky. This radiation carries detailed information about the physical conditions about 380,000 years after the big bang. In the lower plate the intensity distribution of the cosmic microwave background radiation is analysed as a function of angular scale. It contains detailed information about the composition of the Universe. The best fit to the data is obtained for $\Omega_b = 0.05$, $\Omega_{nb} = 0.20$, and $\Omega_\Lambda = 0.75$, indicating that only 5 percent of the energy of the Universe is due to ordinary (luminous or dark) baryonic matter, while 20 percent is due to non-baryonic dark matter, and 75 percent is vacuum energy.

undergoes an accelerated exponential expansion. The idea of inflation solves many problems of the standard big bang cosmology. First, Guth predicted that $\Omega = 1$, long before this was confirmed by the WMAP data. In this way, inflation explains why the Universe is old and flat. Second, it also solves the horizon problem: Why do the cosmic photons coming from opposite ends of the Universe have almost exactly the same temperature? This is a puzzle because they originated from regions separated by two times the horizon distance and should thus not have been in causal contact. In the inflationary Universe, space is stretched so much that those regions indeed were in causal contact very early on. Third, inflation even explains the 1 part in 100,000 deviations from uniformity in the temperature of the cosmic background radiation, which formed the seeds for structure formation in the Universe. These initial fluctuations are attributed to the quantum fluctuations of the inflaton field. Finally, inflation also explains why magnetic monopoles (isolated magnetic north or south poles), which unavoidably arise in grand unified theories, are no longer to be found in the Universe today. They simply got extremely diluted by the inflationary expansion. Interestingly, inflation includes a mechanism by which it can naturally end. This happens when the inflaton field finally reaches its true vacuum value. While inflation has been a very attractive theoretical speculation for a long time, the WMAP data have recently confirmed it in great detail.

In 1987, long before there was observational evidence for vacuum energy, Christof Wetterich from Heidelberg University applied the idea of dynamical vacuum energy to today's Universe. Again, he assumed that some quantum field – in this case called a cosmon field – has not yet reached its true vacuum value. This provides us with dynamical vacuum energy – known as quintessence – which may drive the ac-

Fig. 9: The European Space Agency's Planck mission is scheduled for launch in 2008. It aims at probing the cosmic microwave background radiation with the highest resolution so far. The cosmic background was discovered accidentally by Penzias and Wilson in 1965 by means of a terrestrial microwave receiver. In 1992, the NASA cosmic background explorer COBE delivered the first detailed maps of the temperature distribution of the cosmic background. In 2002, NASA's Wilkinson microwave anisotropy probe (WMAP) charted the cosmic background even more precisely (Fig. 8). The Planck spacecraft will go a step forward by mapping the cosmic microwave background anisotropies with a temperature resolution of the order of 10^{-6}. The Swiss space industry led by Oerlikon Space of Zurich contributed the entire payload module structural subsystem, seen here in the company's clean room. (Credit: Oerlikon Space, Zurich)

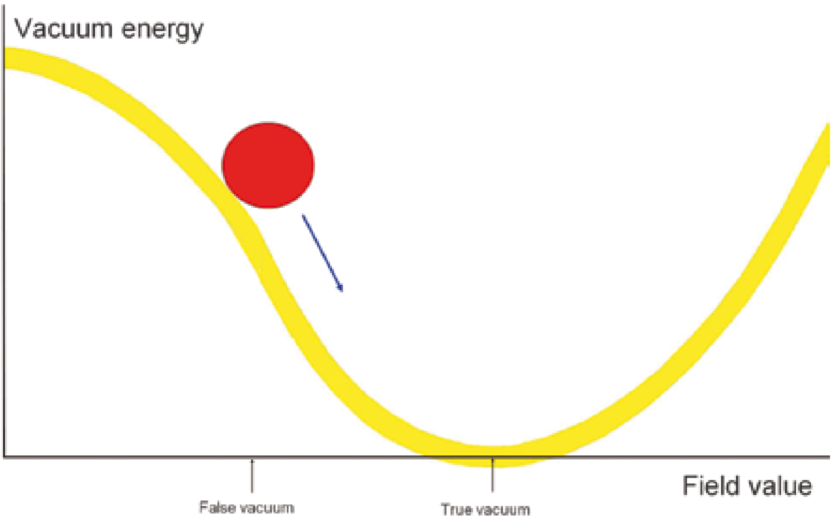


Fig. 10: Vacuum energy as a function of the value of an inflaton field. Immediately after the big bang, the field is displaced from its true vacuum value. The energy of the inflaton field in the false vacuum leads to an exponential expansion of the Universe. The inflationary epoch ends when the inflaton field rolls down to its true vacuum value.

celerated expansion of the Universe today. We then arrive at the following picture of the cosmic evolution. Immediately after the big bang, the Universe was dominated by dynamical vacuum energy in the form of an inflaton field rolling down to its true vacuum value. When this value is reached, inflation ends and the energy stored in the inflaton field is released in the form of an extremely hot gas of elementary particles. The Universe then becomes radiation dominated. It keeps expanding (although no longer at an exponential rate) and, consequently, the hot gas cools down. As a result, particles and antiparticles in the gas annihilate almost completely (at about 1 second after the big bang) thus producing the enormous number of photons and neutrinos in the cosmic background radiations. Only a tiny surplus of matter survives the mass extinction of particles and antiparticles and constitutes all the matter that we find in the Universe today. Once the gas has cooled so much that neutral atoms form (at about 380,000 years after the big bang) the cosmic photons decouple and the Universe becomes transparent. Around that time the Universe begins to be matter dominated. The atoms then form large structures and the first stars are born (at about 500 million years after the big bang). After about 10 billion years of expansion, the matter is diluted so much that vacuum energy (of a much smaller magnitude than during the inflationary epoch) again becomes noticeable. By now (at 13.7 billion years after the big bang) vacuum energy has begun to dominate over the matter. If the vacuum energy exists in the form of a cosmological constant, the Universe will from now on expand forever at an exponential rate. If the vacuum energy is dynamical and exists in the form of quintessence, i.e. if it is carried by a cosmon field, the exponential expansion may eventually come to an end. This will happen when the cosmon field finally reaches its true vacuum value. All this happens over time scales of billions of years. It is interesting to ask why we happen to exist around a time when matter dominance is being replaced by vacuum dominance.

Life in the Multiverse

Once we adopt inflation as a paradigm for cosmology, we must accept that it is a never ending process. In other regions of space the inflaton field may not yet have reached its true vacuum value and thus new exponentially expanding Universes are constantly being created at an astounding rate. This process will practically not be detectable from our own Universe, but the idea of inflation naturally leads to the concept of a very large Multiverse, containing our Universe as just one part.

If the value of the vacuum energy is determined by the value of a cosmon field, the same could be true for other physical "constants" such as the fine-structure constant α or the masses of quarks, electrons, and neutrinos. If so, there is no reason why these parameters should necessarily have the same values in other regions of the Multiverse. In our Universe, the quark masses seem to be fine-tuned so that nuclear physics happens in a very peculiar way. Similarly, chemistry would work totally differently if the fine-structure constant α or the electron mass were only slightly different. Why do we live in a Universe in which all those parameters are so well suited for the development of life? Similarly, we might ask why we are not living on the surface of the Sun. Obviously, the conditions there are inappropriate for the development of life. If we contemplate the existence of other Universes with other physical conditions, we should not be surprised that we developed in one that is hospitable to life. We may in turn use the anthropic principle, i.e. the fact of our own existence, to "explain" the particular values of various fundamental parameters: they are simply what they must be in order to allow life to develop. The question why we live in a transition period between matter and vacuum dominance may perhaps also be answered by the anthropic principle: at much earlier or much later times the world might not be hospitable for life.

As tempting as it may seem, the anthropic principle should only be used as a last resort when all other possible explanations fail. For example, we could have used the anthropic principle to argue that hydrogen or helium behave just the way they do, simply because this is necessary for our own existence. Such argumentation could, in fact, have stopped us from developing atomic or nuclear physics. Fortunately, as some interesting consequence of the laws of Nature, we are endowed with great curiosity and we can understand things at a very deep level. The anthropic principle should not stop us from figuring out everything that can possibly be understood.

Conclusions

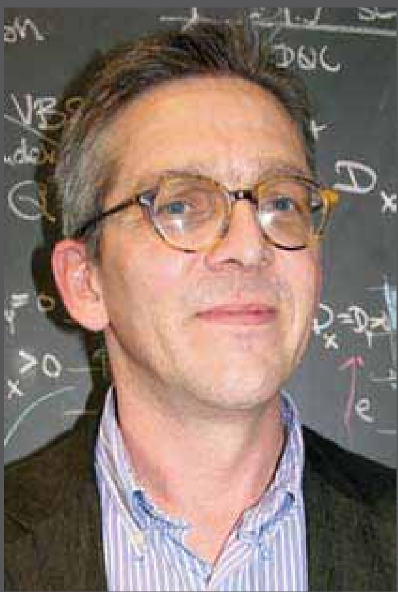
Consistent observations of galaxy clusters, the abundances of light nuclei, distant type Ia supernova explosions, and the cosmic microwave background radiation have revolutionized our understanding of the composition of the Universe and lead to the following conclusions. The ordinary baryonic matter that we find here on Earth and that we understand well constitutes only about 5 percent of the energy of the Universe. Other 20 percent are non-baryonic dark matter of a yet unidentified type. We may hope to soon produce this matter in the form of WIMPs at the LHC. Still, the majority of 75 percent of the energy in the Universe is vacuum energy filling space homogeneously. The nature of this energy, if it is a static cosmological constant or dynamical quintessence, is very much unclear. We live in a Universe that provides physical conditions suitable for our own existence, and that may be part of a much larger Multiverse. Other parts of the Multiverse may be inhospitable to life but are very far away and practically impossible to investigate. Although human endeavour may thus be naturally limited, and we may never find ultimate truth, there is no doubt that there are numerous open questions and many exciting things to explore in our Universe for current and future generations of curious minds.

Acknowledgements

I like to thank the Association Pro ISSI for inviting me to give the talk on which this text is based. Very special thanks go to Hansjörg Schlaepfer for numerous discussions about the main text, as well as for contributing beautiful figures and clear explanations in the boxes concerning the basics. I am also indebted to Heiri Leutwyler for a careful reading of the manuscript and for a number of very useful suggestions.

SPATIUM

The Author



Uwe-Jens Wiese received his diploma in physics at Hannover University in 1984 and the Ph.D. degree in Theoretical Particle Physics in 1986. He completed his studies during several post-doctoral positions at Hamburg University, the DESY Höchstleistungsrechenzentrum in Jülich and at the University of Bern. In 1993, he completed his habilitation at the Rheinisch Westphälische Technische Hochschule (RWTH) Aachen.

From 1994 to 1999 Uwe-Jens Wiese was an Assistant Professor at the Massachussetts Institute of Technology (MIT). In 1999 he was promoted to an Associate Professor and later to a Full Professor at MIT. In 2001 he followed a call from the University of Bern to succeed Hans Bebie. Currently he is the director of the Institute for Theoretical Physics at the University of Bern. Although he loved working at MIT, Uwe-Jens Wiese also appreciates the attractive academic environment in Bern. Even compared to the world-famous MIT, the quality of the work performed in Bern is compatible, only the variety of subfields that are being covered is smaller. His research interests focus on the dynamics of quarks and gluons in particle physics and of strongly correlated electron systems in condensed matter physics. He uses analytic methods as well as numerical simulations to gain insight into these fascinating and very active fields of research.

Uwe-Jens Wiese is actively engaged in the public outreach of physics; he is an active member of the organization committee for "Physik am Samstag", a series of public lectures given by leading scientists at the University of Bern that are addressed to high school students and their teachers.

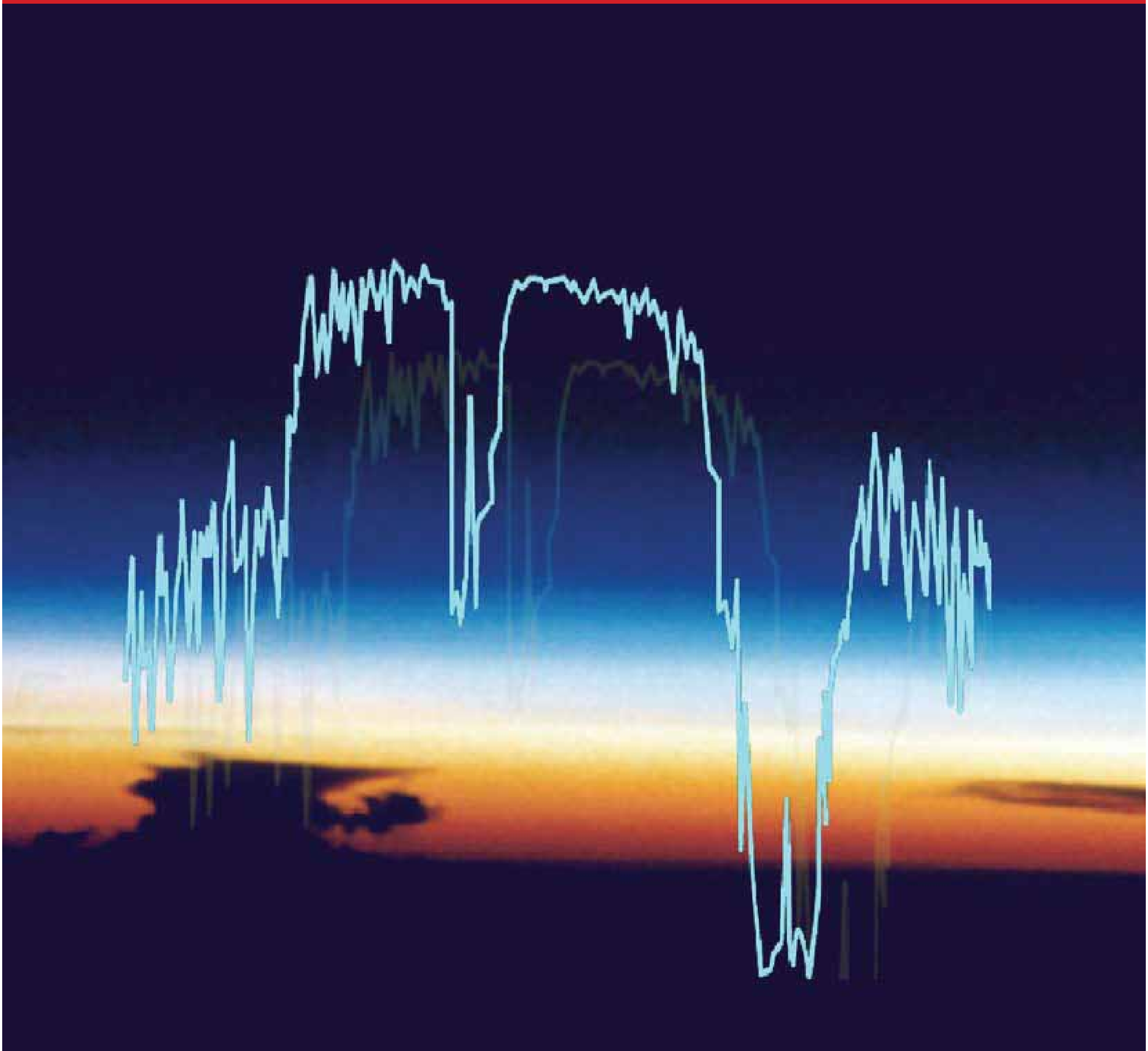
When asked what fascinates him about physics, he says: "I find it very intriguing that the basic laws of Nature can be formulated in mathematical terms. Indeed, the fundamental constituents of matter are an embodiment of mathematics – the universal language that Nature has chosen to express herself in. In some sense, by evolving the human subspecies of the theoretical physicist, the Universe is thinking about itself. Being part of this process together with colleagues all around the world is a fascinating and most rewarding experience."

INTERNATIONAL SPACE SCIENCE INSTITUTE

SPATIUM

Published by the Association Pro ISSI

No. 21, February 2008



The Earth's Ozone Layer

Editorial

Eritis sicut deus scientes bonum et malum (1.Mos. 3, 5).

God created heaven, Earth, plants and animals, and he wished to populate the land with man and wife. He created Adam and Eve, and he provided them with all goods necessary for life. They were, however, submitted to the temptation of grasping the fruit of the forbidden tree. Praising the tree of wisdom, the snake prophesied: you will be God alike, conscious of good and evil. Adam and Eve ate from the tree, lost their innocence and God expelled them from paradise.

Johann Wolfgang von Goethe lets Mephisto, disguised as a professor, write these very words into the scholar's register as a warning of the dangers of unmindful use of abilities. The scholar assumed to have met great Faust, but in fact, he encountered the Devil personally.

These magnificent ciphers are true not only for the pupil in his avid search for wisdom, but for our ambivalent use of technologies as well: wisdom can increase the quality of our life, while trespassing nature's boundaries leads us to the loss of paradise.

The story of the Earth's ozone layer is the modern version of the unfortunate experience Adam and Eve were to make. This is the message that Dr. Yasmine Calisesi, an expert in atmospheric chemistry and a former scientist at ISSI, conveys in this Spatium. Based on her lecture for the Pro ISSI audience on 31 October 2007 she tells us the ozone story. This story may have found a happy end, but at the same time it is a forerunner of other stories, such as for example that of global warming, a topic requiring similar decisive actions by the industrialized societies like those taken to safeguard the Earth's ozone layer.

May the present issue of Spatium contribute to our common wisdom on the dos and don'ts in full responsibility of *scientes bonum et malum*.

Hansjörg Schlaepfer
Brissago, February 2008

Impressum

SPATIUM
Published by the
Association Pro ISSI

INTERNATIONAL SPACE SCIENCE INSTITUTE

Association Pro ISSI
Hallerstrasse 6, CH-3012 Bern
Phone +41 (0)31 631 48 96
see
www.issibern.ch/pro-issi.html
for the whole Spatium series

President
Prof. Klaus Pretzl
University of Bern
Layout and Publisher
Dr. Hansjörg Schlaepfer
CH-8173 Neerach
Printing
Stämpfli Publikationen AG
CH-3001 Bern

The Earth's Ozone Layer[1]

Preface

The discovery of the ozone hole over the Antarctic in 1985 set a milestone in the history of environmental research. In addition to the profound meaning of such a finding, it also stirred the public awareness of ecological matters to an unprecedented extent. Suddenly, everyone became aware of the fact that unmindful use of "unclean" technologies can put humankind's sheer survival into question. From 1930 onwards, thousands of tons of chlorofluorocarbons (CFC's) came into use each year in a variety of applications and were finally released into the atmosphere. It took nearly half a century for science to persuade the governing instances of the threatening effects these chemicals have on the ozone layer, that shields us so efficiently from dangerous solar radiation. Fortunately, though, scientific knowledge at the time of the discovery of the antarctic ozone hole was such that, under the leadership of the United Nations, the international community was able to define and urgently implement the required concerted actions to prevent the complete destruction of the ozone layer. Today, we have observational data that indicate a decline of the CFC concentration in the stratosphere. Models predict a recovery of the ozone layer to its pre-1970's values in the next fifty to hundred years, based on the unrealistic condition, however, that the stratosphere remains untouched by the effects of climate change.

The ozone hole story provides an unprecedented example of humankind's ability to solve global environmental issues, the next challenge to be tackled being undoubtedly that of global warming.

The present issue of Spatium is devoted to the story of the ozone hole. It aims at familiarizing the reader with the crucial importance of the ozone layer for all living organisms on Earth, and to enhance his/her motivation to contribute to what we owe to future generations: a planet Earth that is capable of supporting the peaceful evolution of mankind.

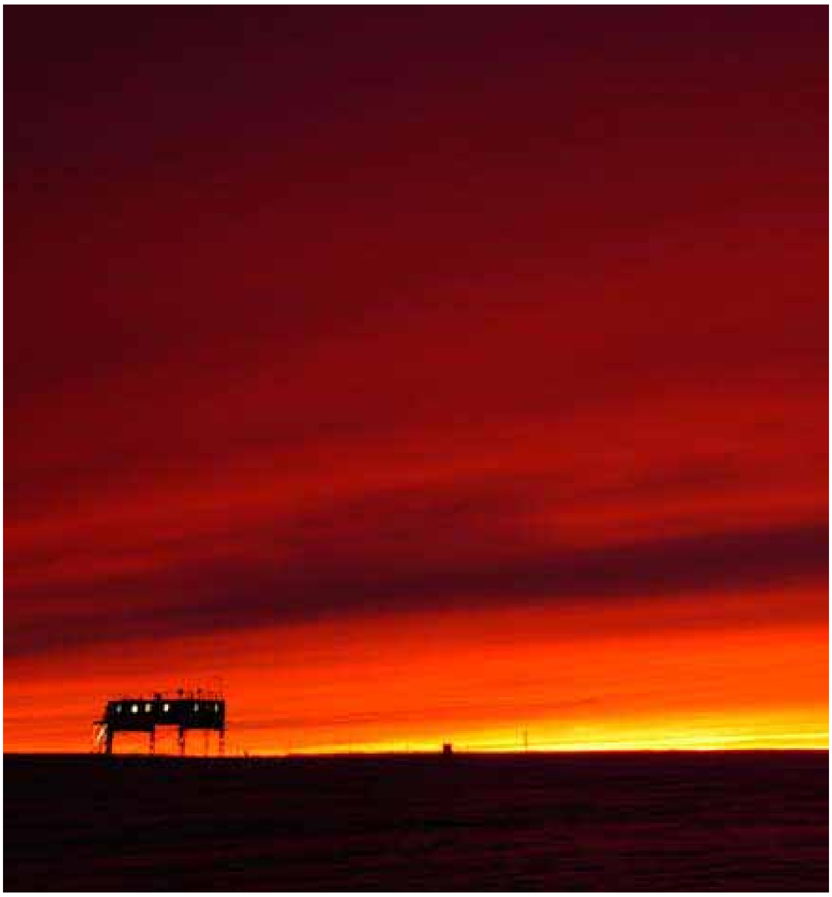


Fig. 1: Observations from Halley Bay in Antarctica provided the first conclusive data on the ozone hole in 1985. This image shows the British Simpson platform in Halley Bay at sunset. What it does not show, however, is the air temperature that might have been as low as -50 °C. (Credit: http://www.flickr.com)

[1] The text for the present isssue of Spatium was drafted by Dr. Hansjörg Schlaepfer and revised by Dr. Yasmine Calisiesi, who gave the Pro ISSI audience an introduction to the subject on 31 October 2007.

Setting the Stage

In this section, we will address the most important mechanisms governing the production and loss of ozone in the stratosphere. We will also highlight the reasons why the stratospheric ozone layer is so important for life on Earth.

The Light of the Sun

The Earth's ozone layer is the result of the combined influence of the Sun and Earth, more specifically the solar radiation and the Earth's atmosphere. Our planet circles the Sun at a distance of some 150 million km. This happens to be right within the narrow zone around the Sun, within which liquid water can exist on a planet's surface, allowing life to develop on Earth.

The amount of power that the Earth receives from the Sun can be compared to the output of some 100 million nuclear power stations. This enormous power meets the Earth in the form of electromagnetic radiation. More specifically, it consists mainly of visible light, associated to infrared and ultraviolet radiation.

As outlined in **Fig. 2**, the intensity of the radiation emitted by the Sun is not uniform at all wavelengths. Rather, its maximum is in the visible part of the spectrum. As a fascinating result, evolution led the sensitivity of the (human) eye to match exactly this maximum in the solar spectrum. The Sun's light is often termed "white", but actually, it is composed of a continuum of colours as represented in **Fig. 2.**

The sunlight enters the top of the atmosphere and passes through various layers down to the surface. In the atmosphere, though, not all the wavelengths are transmitted equally. Its constituents absorb preferentially specific portions of the spectrum; as we will see later, the part of the ultraviolet radiation denoted by UV-C in **Fig. 2** is totally absorbed by ozone in the atmosphere. The less energetic radiation, that is, parts of the UV-B and UV-A solar spectrum above 280 nm, are not completely absorbed by ozone and can penetrate through the atmosphere down to the Earth. Levels of UV-B and UV-A at the Earth's surface are thus dependent on the amount of ozone in the atmosphere. Note that, being the most energetic one, UV-C radiation is the most dangerous to human cells. As it is completely ab-

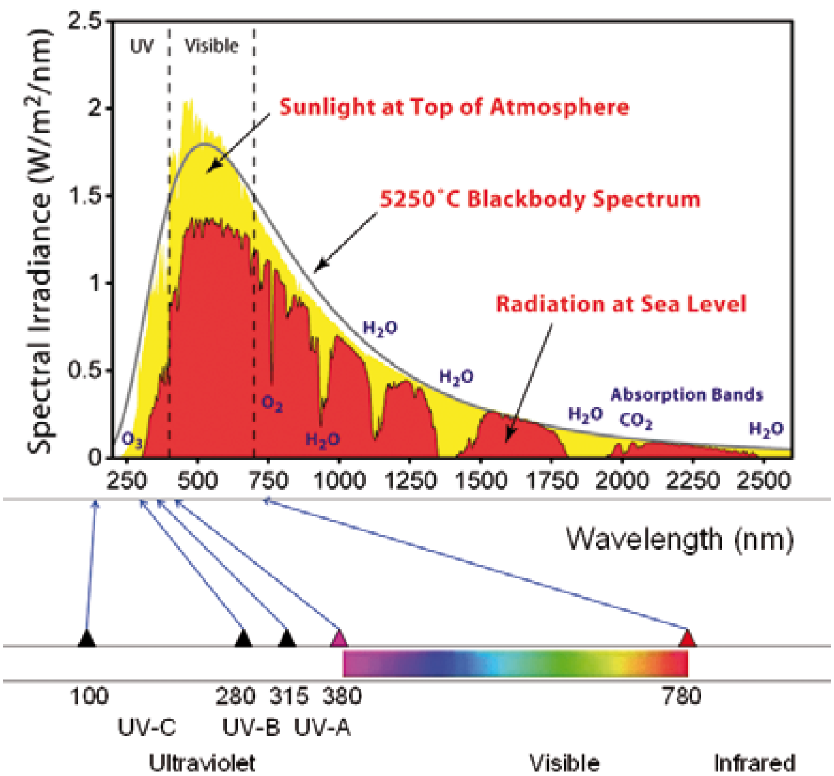


Fig. 2: The solar radiation spectrum. The upper part shows the solar radiation spectrum at the top of the atmosphere (yellow shaded area). Various components of the atmosphere absorb specific parts of the spectrum, resulting in reduced radiation strengths at sea level (red area). The lower part shows a magnified portion of the spectrum, centred on the visible wavelength range with ultraviolet to the left and infrared to the right. The ultraviolet is commonly divided into three parts, spanning from the UV-C range at the short wavelengths towards the longer wavelengths of UV-B and UV-A.

sorbed by the ozone layer, only the UV-B and UV-A parts of the solar spectrum are the target of sun-protecting lotions.

Biological Properties of Light

Light is one of the key requirements of life. Solar radiation provides cells with the energy required to support their functions. The absorption of light by a cell depends on both the wavelength of the radiation and on its intensity. In the case of the human skin, extended exposure to visible light may cause accelerated ageing and sunburn. Intensive exposure to infrared radiation may give rise to retinal damage of the eye and to skin burn. The least energetic UV-A radiation penetrates deep into the skin, and causes early aging. UV-B radiation, however, is the dangerous component of the sunlight on Earth's surface as its photons contain enough energy to destroy the desoxyribonucleic acid (DNA) in the skin cells, which is the carrier of the genetic code. Upon a certain dose of UV-B radiation, such cells can develop a malignant tumour, which is eventually fatal for the entire organism.

To summarize, light in adequate intensities and quality is a necessity for life. On the other hand, however, light at too high energies may also destroy it.

The Earth's Atmosphere

In this section, we shall take a closer look at the composition of the atmosphere in order to understand how the ozone layer is established.

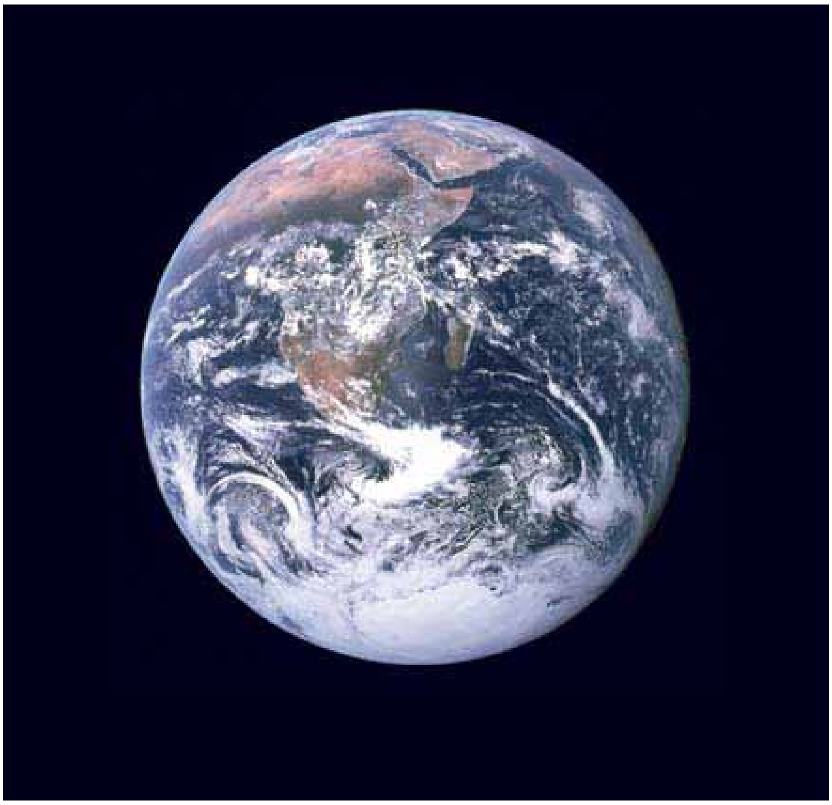


Fig. 3: The Earth seen by the Apollo 17 crew. In this representation and assuming a "thickness" of 100 km, the Earth's atmosphere would measure mere half a millimetre. (Credit: NASA)

The Earth's atmosphere acts like a blanket on the planet: it is the place of the global climate and responds to varying solar radiation regimes. Still, it is a very thin blanket, and its density decreases rapidly from the surface up to interplanetary space **(Fig. 3).**

The atmosphere is a mixture of a multitude of chemical constituents. The most abundant of them are nitrogen N_2 (78%) and oxygen O_2 (21%). These gases, as well as the noble gases (argon, neon, helium, krypton, xenon), possess very long lifetimes and hence are well mixed up to about 100 km altitude. Minor constituents, such as water vapour, carbon dioxide, ozone, and many others, show lower, more strongly varying concentrations with space and time. As mentioned above, these gases pile up in an atmosphere that exhibits a decreasing pressure over height. According to the laws of elementary thermodynamics, we would thus in a first guess expect the temperature to fall from the ambient value we experience at the Earth's surface to a very low value close to space. As we will see, this simplistic approach is far from reality!

Vertical Structure of the Atmosphere

Fig. 4 shows the complex temperature profile of the atmosphere. Some regions show a decreasing temperature with increasing height, which agrees with our everyday experience: climbing a mountain brings us in colder air. Some other regions, in turn, display an increasing air temperature with increasing height. It is precisely following these differing behaviours that the atmosphere is commonly divided into different layers, each being defined by the sign of its temperature gradient.

Let us take a closer look at the different layers of the atmosphere:

- The **troposphere** extends from Earth's surface to 6–8 km over the poles and 18 km over the equator. Solar heating of the surface causes the adjacent air masses to warm up. As warm air is lighter than cold air, it rises to higher elevations where pressure is lower, leading thus to an adiabatic cooling of the rising air. Hence, air temperature in the troposphere decreases with increasing height. The troposphere is the layer within which most of the atmospheric water vapour is contained. This vapour condenses when the air temperature falls under the water dew point, giving rise to clouds. The troposphere is hence the realm of weather. The upper boundary of the troposphere, where the temperature gradient changes from negative to positive is called the tropopause.

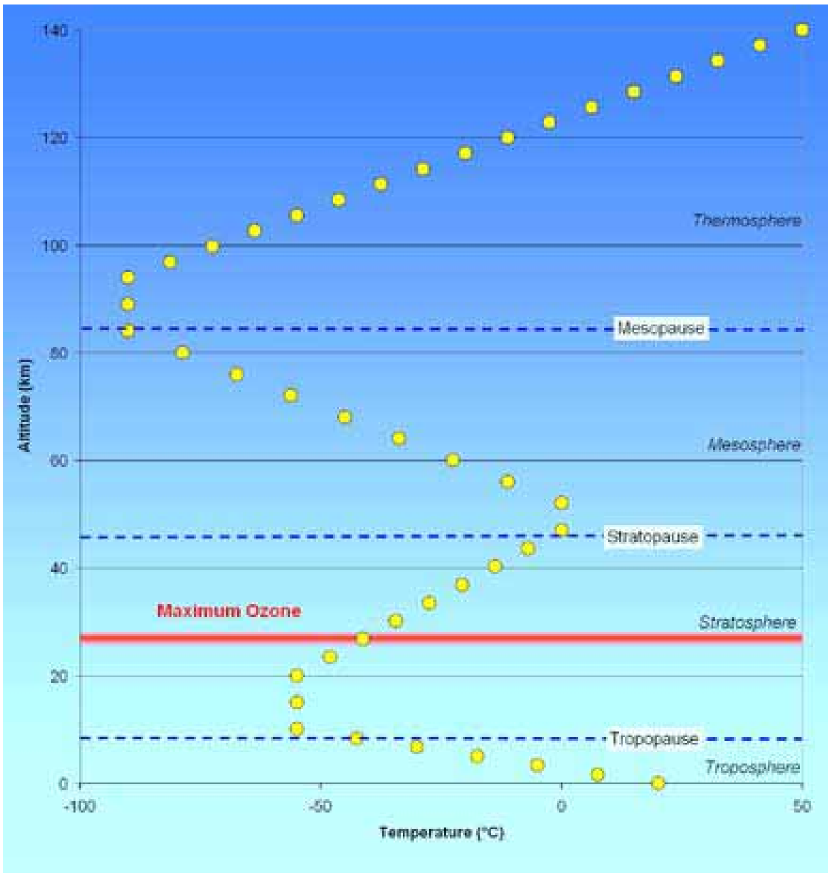


Fig. 4: The vertical structure of the atmosphere. The yellow dots mark the temperature profile compared to height. The layer boundaries are located at altitudes at which the temperature gradient changes sign.

- The **stratosphere** extends from the tropopause to an altitude of about 50 km. It is in the stratosphere that the ozone layer is found. More precisely, it is the place of the maximum ozone concentrations. As we have seen above, ozone molecules have the important property of absorbing ultraviolet radiation from the Sun and transforming it into heat. As solar radiation is stronger, and air density is lower at higher altitudes, air temperature increases here with increasing height. One of the consequences of the positive temperature gradient characterizing the stratosphere is the vertical stability of this atmospheric layer. That is, only very limited vertical mixing takes place within this layer, impeding among others the removal of long-lived pollutants and trace gases by precipitation, but also allowing the existence of local maxima in concentrations such as the one building the ozone layer.
The upper boundary of the stratosphere is called the strato-

pause. Above this boundary, the air temperature tends to fall again with increasing height.

- The **mesosphere** extends from the stratopause to some 85 km. At these altitudes, atmospheric pressure and thus molecule density is already very low. The mesosphere is the region within which most meteorites are consumed when entering Earth's atmosphere at very high velocity. The mesosphere is delimited at its upper boundary by the mesopause. At higher altitudes, air temperature again begins to rise with increasing height.

- The **thermosphere** ranges from the mesopause to some 500 to 1,000 km. At these high altitudes, the residual atmospheric gases sort into strata according to molecular mass. Here, temperature increases with altitude due to absorption of highly energetic solar radiation by the small amount of residual oxygen still present. Temperatures are highly dependent on solar activity, and can rise up to 2,000 °C. Radiation causes the air particles in this layer to become electrically charged, enabling radio waves to bounce off and be received at distant places on Earth. In addition, the thermosphere is also the place of the auroras, which occur when fast solar wind particles hit the air molecules causing them to emit light. The density of gas molecules in the thermosphere is already so low that the International Space Station can orbit the Earth in the upper part of the thermosphere, between 320 and 380 km. The thermosphere ends at the thermopause.

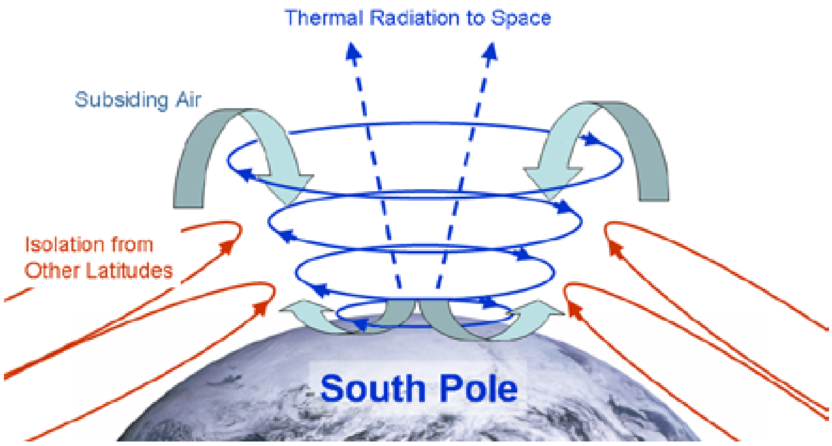


Fig. 5: The polar vortex. During the polar winter, the polar vortex leads to an isolation of the air masses. This prevents fresh air with high ozone content entering the polar region, while the ozone content within the vortex remains low due to the lack of solar radiation. The polar vortex is a natural feature taking place each year over the winter pole, independently of the existence of an ozone hole. It is, however, a key element of the development of an ozone hole in the presence of ozone depleting substances.

- The **exosphere** begins at the thermopause. There is no defined upper limit to this layer, as the concentration of gases diminishes continuously into space. It is from the exosphere that atmospheric gases, atoms, and molecules can escape into space.

Dynamics of the Atmosphere

The Sun not only causes a complex atmospheric temperature profile in the vertical plane, but it does so in the horizontal plane too. As water surfaces heat up much slower than land surfaces when the Sun is overhead, air masses over the oceans tend to be cooler than those over large land areas. This in turn sets up air pressure differences between land and sea that give rise to winds. As the Earth spins around its axis, it transmits acceleration to all objects placed within its gravity field. The resulting force causes the direction of any moving object to be bent. It is called the Coriolis force.

The horizontal pressure gradient and acceleration provided by the Coriolis force largely dominate the movement of air masses in the troposphere. They give rise to a great variety of global wind systems like for example the trade winds that were used to carry trade ships across the world. These winds cause a continuous exchange of air masses within the troposphere and therefore a stabilization of the air temperature over the Earth. This leads to a horizontal temperature profile that decreases from the equator to the poles, as the solar input is the highest when the Sun is right overhead. While this decrease is small in the tropical and subtropical zones, more pole-ward at some 40° north and south the temperature gradient becomes steeper finally reaching temperatures in the polar regions as low as -80°C.

The Polar Vortex

One of these wind systems is specifically important when it comes to understanding the development of the polar ozone holes. During the polar winter, surface heat is lost to space by radiation. This means that the surface cools down to very chilly temperatures. The adjacent air masses cool down similarly, and as cool air is denser than warm air, the typical polar high-pressure zone builds up. This causes polar surface air masses to flow towards the Equator. As the Coriolis force is very strong near the poles, the Equator-ward movement implies that air mass trajectories are bent towards the east. There, they come in contact with surfaces of higher temperature, causing them to warm up and subsequently to rise. From here, the air flows back to the pole at high altitudes, as indicated in **Fig. 5.** This pattern of circumpolar winds is called the polar vortex. Its main feature is the isolation of the polar air masses from the adjacent, lower latitudes air. As we will see later, this structure is an essential ingredient for an ozone hole to develop over the poles.

Ozone

We have seen earlier that the Earth's atmosphere consists mainly of nitrogen and oxygen. Oxygen is present in the air predominantly in the form of oxygen-oxygen molecules O_2. In contrast, an ozone molecule consists of three oxygen atoms O_3. Ozone is a bluish gas with a sharp, irritating odour.

In the following sections, we will explore the ozone's ambivalent role depending on its location: close to the Earth's surface in the ambient air, ozone is toxic when breathed in large quantities. In the stratosphere on the other hand, ozone is an indispensable shield that protects life against the high-energy part of the sunlight.

Discovery of Ozone

The earliest mention of ozone possibly dates back as far as to the eighth century B.C. The great ancient Greek poet Homer narrates the adventures of Odysseus and his friends on their long journey home to Ithaca, following the fall of Troy. In a terrible thunderstorm, the hero's crew finds the air filled with fire and brimstone as the lightning strikes their vessel. What the plagued seamen smelled as brimstone was nothing else than ozone produced by the lightning discharge.

As for much other wisdom, no other reference to ozone was made during the mediaeval times. It was only in the late 18th century that experiments to uncover the secrets of electricity provided new insights into this miraculous gas. Ozone was first observed in a laboratory when scientists experimented with high voltage discharges. Much like in the lightning of a thunderstorm, the energy released by an electrical discharge is able to produce ozone from the oxygen in the air. Martinus van Marum[2], a Dutch scientist, constructed a giant double plate-glass frictional electrostatic generator between 1780 and 1790. The high voltage field generated by the machine caused "persons within 10 feet of the plates to experience a sort of creeping sensation over them, as if surrounded by a spider's web". The stench of ozone gas, "the odour of electrical matter", was most apparent during operation of the huge machine. Still, the cause of the strange smell was not identified yet. Rather, it was up to Christian Friedrich Schönbein[3], a professor of chemistry at the University of Basel, to detect the gas in 1840 and to give it the appropriate name: ozone, from the Greek "οζειν", the smelling.

The second half of the 19th century marks the development of qualitative methods for measuring ozone. The so-called Schönbein paper was used, that showed a colour change under the influence of ozone (and unfortunately of other gases too).

[2] Martin von Marum, 1750, Delft (NL) - 1837, Harlem (NL), Dutch scientist.
[3] Christian Friedrich Schönbein, 1799–1868, German - Swiss chemist.

One of the first laboratories to set about routine surface ozone measurements was the Paris Municipal Observatory, located in Montsouris Park in Paris. These measurements began in 1876, and continued for 34 years. The quality of the collected ozone data is doubtful due to the imprecision of the method in use. Nevertheless, this series provides a unique source of information for pre-industrial levels of surface ozone, making it a precious data set for comparisons with present-day measurements.

In 1879, it was noted that only small parts of the Sun's UV-B and no UV-C radiation are able to penetrate Earth's atmosphere down to the surface. Shortly after this discovery, Cornu[4] and Hartley[5] deduced that this absorption was due to stratospheric ozone, as the measured surface ozone concentrations were far too low to explain the observed absorption spectra. It was thus concluded correctly that there must be much more ozone somewhere in the higher atmosphere. This reasoning was later verified by observations of the stratospheric ozone layer.

Physical Properties of Ozone

As mentioned above, an ozone molecule consists of three oxygen atoms. While the binding between the two oxygen atoms in an oxygen molecule (O_2) is very strong, the binding energy of the third oxygen atom in ozone (O_3) is rather weak. This implies that the third oxygen atom is easily lost, qualifying ozone as a potent oxidizer that is today used in many technical applications such as for example the purification of drinking water. In the present context, the most relevant are the spectroscopic properties of ozone: as outlined in **Fig. 6**, atmospheric ozone and molecular oxygen together block all incoming radiation at wavelengths shorter than 280 nm. As noted above, this is the portion of the solar spectrum that is most dangerous for living tissues in general and for the human skin specifically.

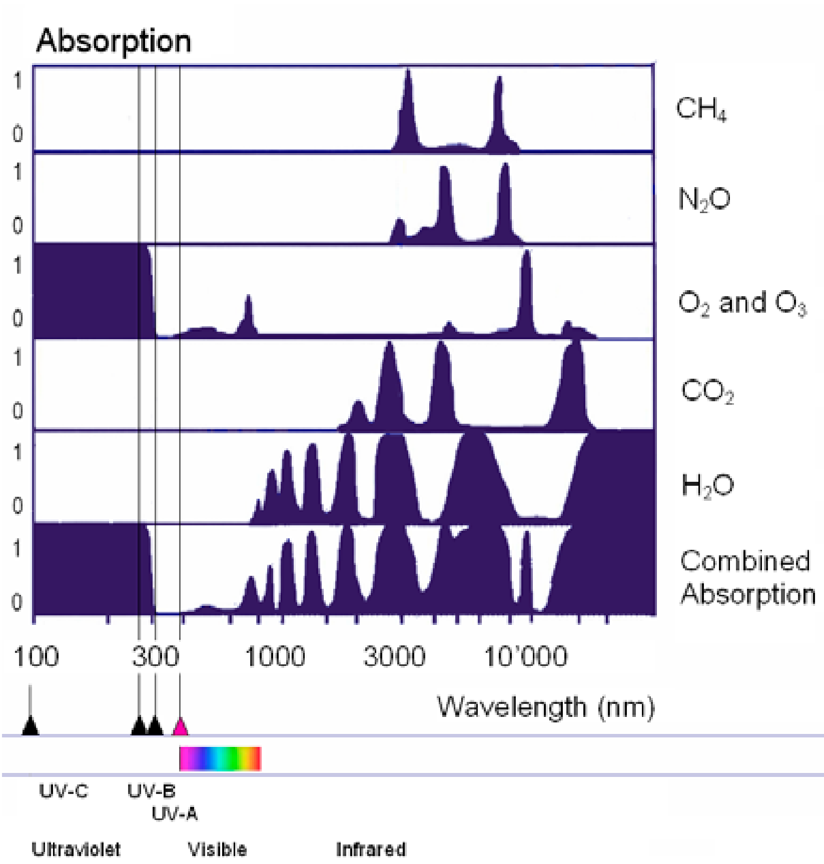


Fig. 6: Schematical representation of the absorption spectra of some constituents of the atmosphere. Note that ozone (O_3) and oxygen (O_2) are responsible for the blocking of all sunlight at wavelengths shorter than 280 nm.

---

4 Marie Alfred Cornu, 1841, Orléans, France - 1902, Romorontin, France, French physicist.
5 Sir Walter Noel Hartley, 1845, London - 1913, Braemar, Scotland, British physicist.

The Earth's Ozone Layer

There is no major natural source of ozone in the troposphere. In contrast, stratospheric ozone is generated by the light of the Sun. More specifically, the ultraviolet light emitted by the Sun is able to split O_2 molecules in the stratosphere into their components, which then react with other O_2 molecules to form ozone. More than 90% of the atmospheric ozone is contained in the stratosphere. The existence of an ozone layer was however not a permanent feature on Earth, as atmospheric oxygen built up only late in Earth's history as a result of early forms of life. It is fascinating to note here that primordial organisms produced a waste product that became a key prerequisite for later, more complex forms of life.

Evolution of the Ozone Layer

The Earth condensed from the gas and dust of the solar nebula about 4.6 billion years ago. The proto-Earth swept up the material of the gaseous solar nebula in a process of accretion over millions of years[6]. Many of the gases now present on Earth were collected via collisions with other celestial bodies. Eventually, the bombardment subsided as the growing Earth had swept up most of the smaller bodies and the remaining gases in its orbit around the Sun. Heavier gases were trapped in the congealing molten rocks, while lighter gases captured by Earth, such as methane, ammonia, and hydrogen, formed Earth's first atmosphere. This early atmosphere was completely free of oxygen.

Later, gases originally trapped below the surface began to enter the atmosphere through volcanic eruptions and other openings in the crust. The atmosphere began to fill with carbon dioxide, nitrogen, and water vapour. Oceans developed when the surface temperature of the early Earth had fallen enough to allow water to exist in liquid form. Large amounts of carbon dioxide were removed from the atmosphere and trapped in carbonate rocks and in seawater. Photolysis of carbon dioxide and water vapour yielded some atmospheric oxygen, but at that time, every oxygen atom reacted quickly with rocks on the surface and gases already present in the atmosphere. The atmosphere therefore continued to remain free of oxygen.

Finally, life developed, and began to influence its environment[7]. As there was still no oxygen in the atmosphere, no ozone layer existed either, which could have shielded land-based organisms from the Sun's ultraviolet radiation. We assume therefore life to have originated in niches in the oceans, where the shielding role was taken over by water. Initial forms of life were various types of bacteria fit for an anoxic environment. The photosynthetic metabolism of the first plants (types of algae) and cyanobacteria (blue-green algae) initiated to produce oxygen some 2.5 billion years ago. That caused the oxygen content of the atmosphere to rise steadily, allowing the ultraviolet part of the sunlight to produce ozone in the high atmosphere. It is thought that the atmosphere reached 50% of its present ozone content as early as about one billion years ago, while oxygen levels themselves were still only a small fraction of their present value. About 400 million years ago, however, oxygen and ozone had presumably both reached their today's values. At that very time, the first land animals also appeared. It is therefore probable that the development of higher evolved forms of life outside the oceans was closely linked to the development of the protective ozone layer.

Discovery of the Ozone Layer

In contrast to tropospheric ozone, which can be readily identified by means of simple chemical methods, the measurement of stratospheric ozone requires more complex remote sensing techniques. Hartley's discovery of the existence of stratospheric ozone in 1881 opened the floor for a scientist whose name is closely linked with the ozone layer: Gordon Miller

6 See Spatium-6: From Dust to Planets
7 See Spatium-16: Astrobiology

Fig 7: Gordon Miller Bourne Dobson around 1970. (Credit: photo presumably by Dr. A. Dziewulska-Losiowa 1970)

Bourne Dobson[8] **(Fig. 7)**. As a lecturer in meteorology in Oxford, he studied the trails of meteorites together with F. A. Lindemann, then Head of the Clarendon Laboratory at Oxford University. Their research led them to observe that the trails of meteorites did not disappear quickly at high altitudes. From this behaviour, they concluded that the temperature profile above the tropopause was probably not constant as was expected, but that rather it appeared to increase substantially with altitude. Dobson inferred correctly that the cause of the warm stratosphere was heating by the absorption of ultraviolet solar radiation by ozone, and he set out to make measurements of the amounts of ozone in the stratosphere and of their variability.

Dobson was not only an excellent theoretician; he was a gifted experimenter as well. In 1925, he developed the first ozone remote sensing instrument **(Fig. 8)**. The "Dobson spectrometer" working principle is based on the comparison of the sunlight intensity measured at two different wavelengths of the solar spectrum: one at which solar light is unaffected by ozone, and one at which the sunlight is strongly absorbed by ozone. By comparing the measured light intensity at the two wavelengths, Dobson was able to determine what is called the total ozone content of the atmosphere above the point of observation. This term relates to the amount of ozone integrated along a line from the observation level up to the top of the atmosphere. The first Dobson spectrometer was built in Dobson's laboratory in summer 1924. Extensive measurements performed during the following year yielded the main features of the seasonal variation of ozone, among others the springtime maximum and the autumn minimum at mid-latitudes. These results were so encouraging that Dobson decided to construct a further series of five instruments during the winter of 1925–26.

Dobson installed these instruments at various European locations, allowing him the discovery of the relationship between total ozone and synoptic-scale weather systems. It was the prospect of potential use of ozone information to improve weather forecast that spurred the initial interest in long-term ozone observations, and paved the way for the establishment of global ozone monitoring networks. Among the five instruments disposed by G.M.B. Dobson across Europe, one was placed at Arosa, in the Swiss Alps, at the Lichtklimatisches Observatorium then headed by F. W. Paul Götz. Here, routine measurements continued well after the first experi-

Fig. 8: One of the first Dobson spectrometers. Although emerging technologies have since allowed the development of a multitude of alternative ozone remote sensing methods, modern-day versions of the Dobson spectrometers still constitute the pillars of worldwide ground-based ozone monitoring networks. (Credit: University of Oxford, Atmospheric, Oceanic and Planetary Physics Dept.)

[8] Gordon Miller Bourne Dobson, 1889, Oxford – 1976, Oxford, British physicist and meteorologist.

ments of G.M.B. Dobson, qualifying today the Arosa time series as the longest data sequence of the total ozone content worldwide **(Fig.9).** It is one of the results of F. W. Goetz's research to show that the maximum in ozone concentrations, when described in absolute values (absolute molecule densities), was likely to be at an altitude of approx. 25 km.

Honouring G. M. B. Dobson, the total ozone value is now commonly measured in Dobson Units (DU). If the total ozone column over a certain point on Earth were to be compressed to a pressure of one atm. at a temperature of 0°C, the resulting slab would have a typical thickness of 3 mm. This by definition corresponds to a value of 300 DU **(Fig. 10).**

Quite naturally, the discovery of the stratospheric ozone layer spurred physicists and meteorologists to explain the relevant production and loss mechanisms. It was only in 1929 and 1930 that Sidney Chapman[9] published the theory of ozone formation and destruction. The Chapman Reactions today provide the basis of all ozone chemistry schemes.

Diary of A Stratospheric Ozone Molecule

The Chapman Reactions describe the production and loss mechanisms for ozone under the influence of sunlight. Let us have a closer look at the life cycle of a single ozone molecule. The first step happens when ultraviolet radiation breaks apart an oxygen molecule (O_2) into two free oxygen atoms O:

$$O_2 + hc/\lambda \rightarrow O + O \quad (\lambda < 242\,\text{nm}) \qquad (1)$$

where hc/λ represents the energy supplied by an ultraviolet photon, with h the Planck constant, c the speed of light, and λ the photon's wavelength. The dissociation of the oxygen molecule can take place if the wavelength lies below 242 nm.

The rate, at which ozone is formed, is slow, since (1) the sunlight intensity is low at these very short wavelengths and (2) the density of molecules in the stratosphere is low.

Free oxygen atoms are highly reactive. Typically, they react with other oxygen molecules to form an ozone molecule:

$$O + O_2 + M \rightarrow O_3 + M \qquad (2)$$

where M is an inert species, usually O_2 or N_2, which carries away the excess reaction energy. This

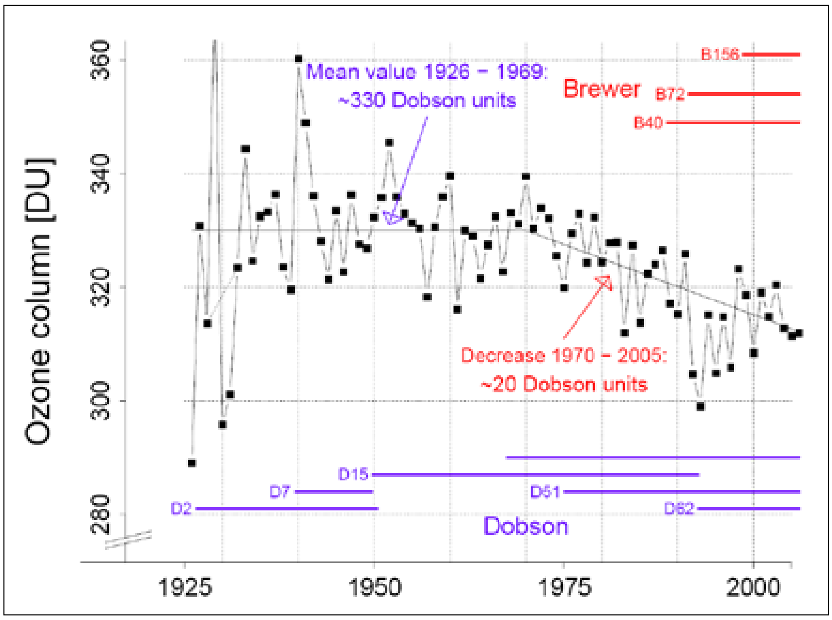


Fig. 9: The "Arosa total ozone time series" (1-year averages). One of Dobson's first instruments (the D2) was installed in Arosa in 1926. Since then, regular measurements have been made with further Dobson-(and Brewer-)instruments leading to the longest total ozone record worldwide. As shown above, total ozone amounts above Arosa were approximately constant (330 DU) until the mid 1970's. From then onwards, a significant decrease was observed. (Credit: Bundesamt für Meteorologie und Klimatologie MeteoSchweiz, 2007)

[9] Sydney Chapman, 1888, Manchester - 1970, British physicist and geophysicist.

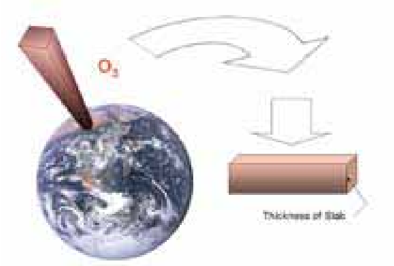


Fig. 10: Definition of the Dobson Unit. If all the ozone over a certain point on the Earth's surface were compressed to one atmosphere at a temperature of 0° C, the resulting slab would measure typically 3 mm. In this case, the total ozone amounts to 300 Dobson Units (DU).

production of heat is the reason for the higher temperatures prevailing in the stratosphere with respect to the lower troposphere.

In this process, a new ozone molecule is created. Ozone in turn also absorbs ultraviolet radiation:

$$O_3 + hc/\lambda \rightarrow O + O_2 \ (\lambda < 310\,\mathrm{nm}) \quad (3)$$

Here the ozone molecule is split into one oxygen atom and an oxygen-oxygen molecule that can react again as in (2) to produce ozone. In another step, ozone can also be destroyed by reaction with a single oxygen atom:

$$O + O_3 \rightarrow O_2 + O_2 \quad (4)$$

Here, two O_2 molecules are produced that are ready to participate in further sequences of the Chapman reactions. In the unpolluted stratosphere, this process may continue for a long time, and the ozone molecule may serve its important function many times.

The model described by Sidney Chapman led to predictions of ozone concentrations that are much higher than those actually found in the stratosphere. It was assumed, therefore, that additional mechanisms must exist for its destruction. In 1970, Paul Crutzen[10] suggested additional reactions involving nitrogen monoxide (NO) gas as a catalyst:

$$NO + O_3 \rightarrow NO_2 + O_2 \quad (5)$$

$$NO_2 + O \rightarrow NO + O_2 \quad (6)$$

Net:

$$O + O_3 \rightarrow 2\,O_2$$

In these reactions, the NO molecule combines with an ozone molecule to form nitrogen dioxide NO_2 and oxygen O_2. Then, NO_2 reacts with atomic oxygen to produce another NO molecule, and an oxygen molecule O_2. The crucial point of this reaction sequence is that the NO molecule enables the destruction of ozone but remains conserved allowing for further destructive reactions. NO is therefore called a catalyst, i.e. a chemical substance enabling a chemical reaction without being consumed itself. In other words, a single NO molecule can destroy ozone molecules as long as it is not washed out from the stratosphere by other mechanisms.

The next step towards understanding the evolution of the ozone layer was taken in 1974 when Mario Molina[11] and F. Sherwood Rowland[12] showed that chlorofluorocarbons (CFC's, **see box**) can reach the stratosphere, stay there for a long time and, under the effect of sunlight, release chlorine atoms that can destroy ozone in a catalytic reaction much like that involving NO. At that time, CFC's were widely used in many technical as well as consumer applications and thoughtlessly released into the atmosphere.

Reverting now to our ozone molecule, we conclude that its lifetime is dramatically shortened in the presence of a variety of chemicals such as chlorine, nitrogen or bromine, the products of CFC's in the stratosphere. These alarming discoveries were not a long time coming…

Chlorofluorocarbons (CFC's) are compounds containing chlorine, fluorine and carbon. Halons are compounds that contain, in addition, bromine. Compounds containing H, F and C are called hydro fluorocarbons (HFC's). During the 20th century, CFC's were marketed under the trade name Freon and other names, and became widely used as refrigerants, propellants in aerosol cans, inflating agents in foam materials, solvents and cleansing agents.

[10] Paul Josef Crutzen, 1933, Amsterdam, Dutch meteorologist, Nobel Prize laureate 1995 in chemistry.
[11] Mario José Molina, 1943, Mexico City, Mexican chemist, Nobel Prize laureate 1995 in chemistry.
[12] Frank Sherwood Rowland, 1927, Delaware, Ohio, US-American chemist, Nobel Prize laureate 1995 in chemistry.

Discovery of the Ozone Hole

The International Geophysical Year 1958 (IGY) allowed scientists to take part in a worldwide series of coordinated observations of various geophysical phenomena. At that time, no less than 44 Dobson spectrometers were already distributed around the world. These participated in, and later continued the monitoring programme by acquiring data on the total ozone content at various places on Earth. The most frightening news came in 1985 from Halley Bay, Antarctica. There, three British scientists, Joe Farman, Brian Gardiner, and Jonathan Shanklin, of the British Antarctic survey, had observed a 30% decrease in the October total ozone values between 1970 and 1984. These results were corroborated by another report by the Japanese scientist Shigeru Chubachi, based at the Syowa Antarctic station. Although this anomaly was at first not detected by satellite-based measurements, it was soon confirmed by these, as well as by in-situ ozone probes and airborne measurements. The ozone hole had been discovered.

This finding prompted the countries to act. In September 1987, representatives from around the world met in Montreal to sign a treaty setting up sharp limits on the use of

Fig. 11: Polar stratospheric clouds. A complex chemistry takes place on the surface of the ice particles that form polar stratospheric clouds involving chlorine and bromine released from CFC's. As soon as the solar radiation reaches sufficient intensity in early spring, the Cl- and Br-radicals are released, leading to the rapid catalytic loss of O_3 in the antarctic stratosphere (Credit: H. Berg, Forschungszentrum Karlsruhe)

CFC's and halons. The Montreal Protocol established a new way of viewing environmental problems. In the past, no such issue had ever been addressed with so much determination. The Montreal Protocol tackled the ozone issue early, at a moment when damage had already been proven, but hope still existed to avoid an even larger danger.

The Montreal Protocol called also for an increase in atmospheric research, which motivated many research laboratories to launch field expeditions to inhospitable Antarctica. These aimed at solving the mystery of ozone depletion there. In contrast to all expectations, it was found that the destruction of the ozone layer took place at a much lower altitude and to a much larger extent than predicted by the theory of Crutzen, Molina, and Rowland. Farman, Gardiner, and Shanklin already had noted in their 1985 Nature paper that the observed October ozone decrease went along with an increase in concentration of CFC's in the Southern hemisphere over the same time period.

Then, further alarming news came along quickly: besides the building-up of anthropogenic chlorine in the stratosphere, scientists reported that the ozone loss is accelerated by polar stratospheric clouds, see **Fig. 11.** Within the polar vortex, icy cloud particles in the stratosphere offer surfaces to chemical reactions that transform chlorines and bromines into efficient destructive chemicals to break ozone molecules. These reactions take place so quickly that practically all of the ozone over Antarctica is destroyed within a few weeks during the September and October (antarctic spring) period. In addition, the polar vortex effectively prevents air from mixing; therefore, no fresh air containing ozone can enter the antarctic stratosphere during this period.

As a result of intensive research activities at that time, it became clear that the Montreal Protocol would not go far enough to protect the fragile ozone layer. Even with the 50 percent cuts mandated by the treaty, levels of chlorine and bromine would still rise in the stratosphere, meaning that the ozone loss would only worsen with time. In June 1990, delegations met again in London and voted to strengthen the Montreal Protocol significantly.

Given the longevity of CFC molecules, recovery times are measured in decades: according to recent estimates, a CFC molecule takes an average of 15 years to reach the stratosphere from the ground. There, it can stay for up to a century, destroying many thousands of ozone molecules during its lifetime. We have, therefore, to face the fact that even if the release of CFC's into the atmosphere were to be stopped instantly, CFC densities in the stratosphere would continue to increase and the ozone destruction would continue. There is some hope, however: as a result of the ban on CFC's imposed by the Montreal Protocol and its amendments, the latest observational data indicate a global stabilization of the CFC concentration in the stratosphere.

The Earth's Ozone Layer Today

Today, the ozone layer is one of the Earth's most closely watched environmental parameters. It is fortunate that space technology has provided unprecedented observational means covering the remote polar zones, which are so important for understanding the ozone issue but so difficult to access from ground-based stations.

Several institutions regularly publish global ozone surveys, such as for instance the World Meteorological Organization (WMO), which issues regular bulletins with up-to-date information on the state of the ozone layer. The European Space Agency (ESA) has commissioned the Koninklijk Nederlands Meteorologisch Instituut (KNMI) to implement an Internet site under the Agency's Data User Programme (DUP). The TEMIS (Tropospheric Emission Monitoring Internet Service) aims at computing and delivering global concentrations not only of ozone but also of other tropospheric trace gases and aerosols derived from data gathered by space-borne instruments such as the Global Ozone Monitoring Experiment (GOME), the Scanning Imaging Absorption Spectrometer for Atmospheric CHartographY (SCIAMACHY) and the Along Track Scanning Radiometer (ATSR) onboard ESA's Envisat spacecraft.

In the following chapters, we will look first at some global results published on the TEMIS server. We will concentrate on two specific days in the year 2007: 23 March, at the end of the antarctic summer, and 3 October, at the end of the antarctic winter.

Global Total Ozone Maps

In **Fig. 12**, the global total ozone values are shown for the two days selected. Based on our former reasoning, we would expect the highest values to be reached in the tropical zones, where the Sun is overhead and hence delivers the strongest irradiation to produce ozone.

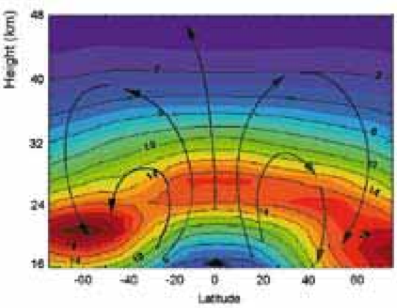


Fig. 13: The Brewer-Dobson circulation illustrated by a schematic, seasonally averaged cross-section through the global atmosphere from pole to pole. At the Equator, tropospheric air masses are lifted up to the stratosphere. At an altitude of some 40 km, they separate in two streams towards the poles. At about 60° latitude, these air masses flow down towards the troposphere and then back to the equator. The result is the so-called Brewer-Dobson circulation. The background colours indicate the local mean ozone density in Dobson Units per vertical kilometre. It can be seen that the highest concentrations are reached in the lower stratosphere between some 80° and 60° North and South latitudes. (Credit: Old Dominion University, Center for Coastal Physical Oceanography, Virginia, USA)

Both upper panels, however, prove the contrary: the largest ozone amounts are rather found at the high latitudes. We conclude therefore that an additional process must take place that reduces ozone values at the tropics and increases them at the winter/spring pole. It was again G. M. B. Dobson, preceded by A. W. Brewer, who was able to find a slow circulation bringing air masses from the equatorial troposphere up into the stratosphere. There, it is enriched in ozone by the strong solar radiation. These air masses then split and flow pole-wards, whereby the ozone enrichment of the middle and polar latitudes takes place. Then, the ozone-rich air sinks down to the troposphere and flows back to the tropics. We now expect the highest values of total ozone at the polar latitudes just outside of the polar vortex. **Fig. 12** illustrates the results of this mechanism, confirming the existence of the so-called Brewer-Dobson circulation as shown in **Fig. 13.**

Arctic Total Ozone Maps

The situation in the northern hemisphere is shown in **Fig. 12** (middle tables) for the same two days as above. At the end of the arctic winter, on 29 March 2007, the total ozone values reached very high values (up to 500 DU, left table). Due to its much lower dynamic stability, the northern polar vortex often gives rise to an attenuated polar ozone hole as compared with its southern counterpart. However, lower ozone concentrations corresponding to the northern ozone hole can still be observed in the left-side panel at about 60°N and longitudes between 30 and 120°E. Obviously, the northern polar vortex was weak and perturbed enough during the winter 2006/2007 to allow the supply of tropical, ozone-rich air to the polar stratosphere even within the vortex. At the end of the arctic summer, on 3 October 2007, the values were between 300 and 350 DU, which is close to normal.

As the majority of humans live under the tropical and northern middle latitudes, the danger for the world's population caused by the depletion of the arctic polar ozone layer has been relatively low over the past 40 years. However, excursions of the polar vortexes towards the equator, induced by perturbed atmospheric dynamics, could lead to direct threats to the populations, such as for instance during winter 1996/97. At that time, an outright arctic vortex together with a large ozone hole developed. The latter spread down well towards densely inhabited parts of Europe and the United States, as indicated by **Fig. 14.** The question of the stability of the northern (and to a lesser extent the southern) polar vortex is still the object of investigations. Relationships have been demonstrated with other global-scale dynamic patterns, such as the equatorial quasi-biennial oscillation (QBO).

Antarctic Total Ozone Maps

In **Fig. 12,** the antarctic total ozone concentration is shown again for the same two days. The left table presents the values on 29 March 2007, at the end of the antarctic summer. Total ozone reached values even in excess of those found in the tropics

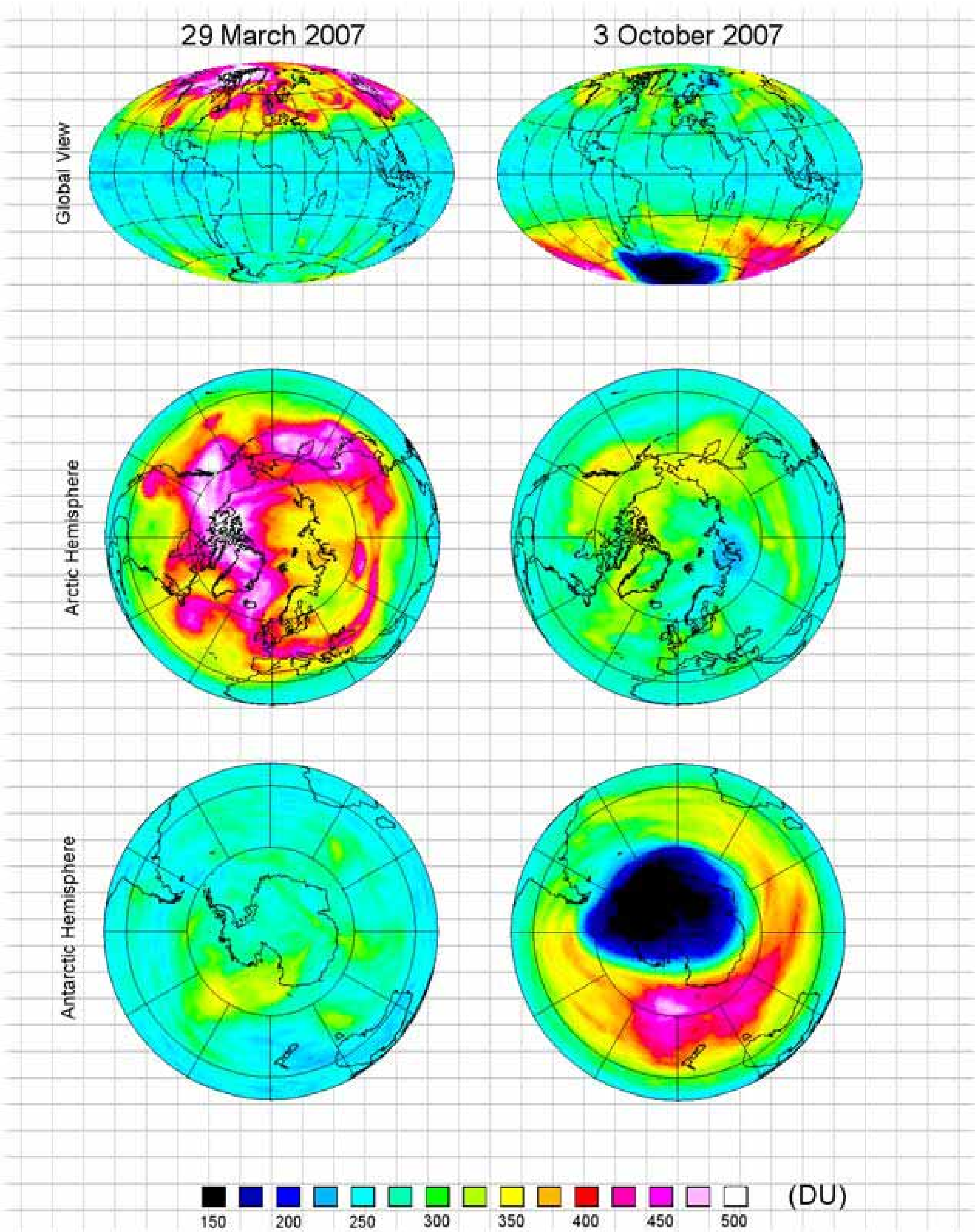


Fig. 12: Assimilated total ozone distributions on 29 March and 3 October 2007. (Credit: TEMIS)

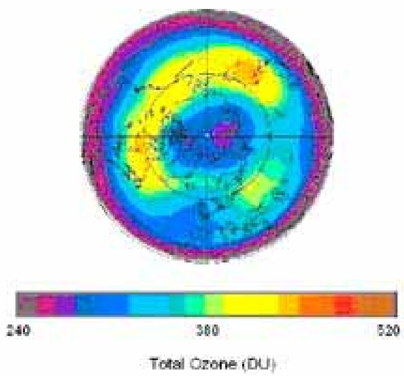


Fig. 14: Arctic total ozone distribution 1997, March average. This image shows the arctic ozone hole that developed at the end of winter 1996/97. In that period, the longest lasting polar vortex on record built-up over the Arctic and in March 1997, the average ozone column over the Arctic fell down to 354 DU, the lowest in the 20 years of observations. (Credit: NASA/GSFC)

thanks to the Brewer-Dobson circulation. The contrast to the right-hand panel, however, could not be more dramatic: here, at the end of the antarctic winter, the ozone destruction process confined within the stable polar vortex during the last few months destroyed some 50% of the ozone as compared to the summer values. Today, the antarctic ozone hole can reach a size exceeding the area of the United States of America.

It is important to understand why the situation is much more dramatic in Antarctica than in the Arctic. The large landmass constituting Antarctica is almost perfectly centred on the pole, allowing a strong circumpolar circulation of the jet stream building the polar vortex during the southern winter. This gives rise to the strongest polar vortex and the coldest temperatures within the vortex, allowing in turn the development of polar stratospheric clouds and consequently the largest ozone holes on Earth.

Local Ozone Data

While satellite views provide an excellent record of the Earth's ozone layer on a global scale, local data remain attractive for their (generally) higher temporal and spatial resolution. They also allow for comparison with series that were taken before the space age that were exclusively local data. In addition, such measurements play the role of "ground truth" data supporting the concurrent calibration of data from spacecraft.

In Switzerland, the Swiss Federal Office of Meteorology and Climatology (MeteoSwiss) is responsible for monitoring the status of the ozone layer over Switzerland. The data is supplied via their Internet site[13], from where the following data have been retrieved.

Total Ozone Over Arosa

It has already been mentioned that the Arosa time series represent the

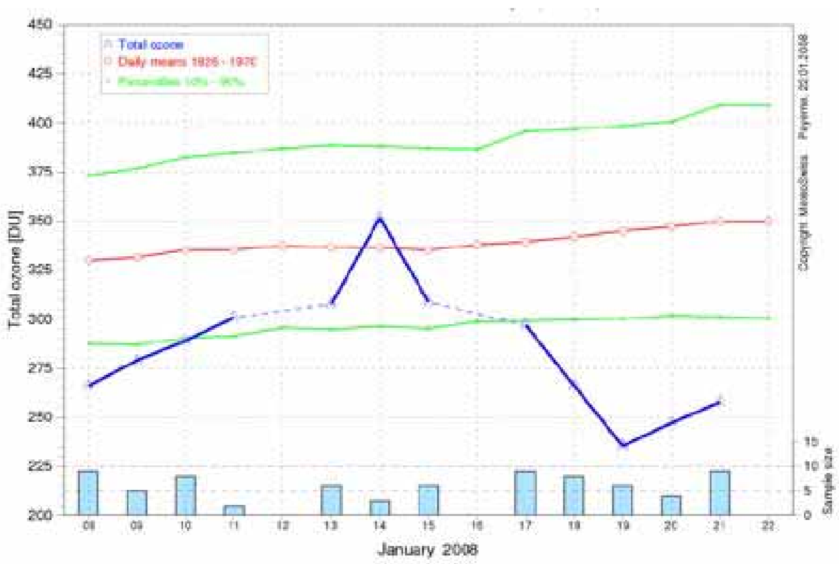


Fig. 15: Excerpt from a recent Arosa total ozone time series. This table shows the mean daily total ozone values over Arosa from 8–22 January 2008. The blue line segments represent actual daily mean values, while the red line shows mean values between 1926 and 1970, that is to say, before the ozone layer's degradation became measurable. The green lines mark the band around the mean values that contain 90% of the values measured from 1926-1970. The table shows a significant loss of ozone towards the end of the observation period that can be attributed to large-scale meteorological activities. As the total ozone measurement requires clear sky, some data from cloudy days are missing. Therefore, some parts in the actual blue curve were interpolated as indicated with dashed blue line segments. (Credit: Bundesamt für Meteorologie und Klimatologie MeteoSchweiz, 2007)

13 http://www.meteoschweiz.admin.ch/web/de/wetter/ozone_layer.htm

longest traceable total ozone record worldwide. This station continues to take daily measurements and to compare actual with former values **(Fig. 15)**.

As a result of large-scale meteorological activities and of the previously noted relationship between synoptic-scale weather system and total ozone, local ozone values can fluctuate considerably from day to day. This means that assessments of the status of the ozone layer from daily values are not feasible. To this end, long-term time series are required, and this is why the long Arosa record is of such a high scientific value.

Vertical Ozone Profiles

The total ozone values measure the sum of ozone gas from the surface up to the top of the atmosphere. In contrast, the distribution along the vertical axis shows where the ozone resides in the atmosphere. In Switzerland, this data is gathered by means of balloon probes that are launched regularly thrice per week from the MeteoSwiss aerological station in Payerne. As an example, **Fig. 16** shows the partial pressure[14] (in nbar) of ozone, a measure for the ozone particle density per volume, over height. The altitude distribution of ozone is represented for the different seasons of the year 2006. The seasonal maxima of ozone concentrations are situated somewhere between 20 and 25 km. Daily measurements, however, show major displacements of the maximum as well as significant changes over relative small altitude differences as was the case on 3 March 2006.

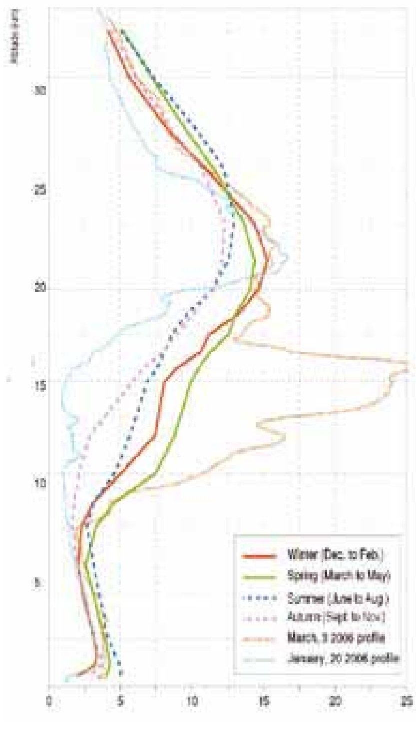


Fig. 16: The vertical ozone profile over Payerne, Switzerland. (Credit: Bundesamt für Meteorologie und Klimatologie MeteoSchweiz, 2007)

Outlook

We do not know the global stratospheric ozone concentrations in earlier times, specifically not before global industrialization with its ever-increasing discharge of CFC's. Nevertheless, we know for sure that the antarctic ozone hole will continue for decades as the relevant pollutants have lifetimes that can reach hundreds of years in the stratosphere. Still, many pieces of the ozone puzzle remain missing, and the science community is challenged by open questions such as how safe the CFC substitutes are, or whether other compounds exist that significantly deplete the ozone layer.

The Montreal Protocol provides a striking example of science at the service of humankind. By quickly piecing together the ozone puzzle, atmospheric researchers revealed the true danger of halocarbons even before the discovery of the Antarctic ozone hole, and urged the community to take decisive actions protecting the Earth's ozone layer. While this agreement represents a critical step toward saving the ozone layer, it has taught scientists and policy makers an invaluable lesson about addressing environmental problems. Beyond all differences that divide humanity there is one common vision to safeguard planet Earth as the home for our future generations.

14 The partial pressure is defined as the absolute pressure exerted by one component of a mixture of gases. It is thus a measure of the absolute number of molecules of the gas of interest. If related to the pressure of the entire mixture of gases, the partial pressure allows computing the relative fraction of the gas particles respective to the total air molecules.

The Author

Yasmine Calisesi, is both a Swiss and an Italian citizen; she studied physics at Geneva University and received her diploma in 1996. During her studies, her interest in atmospheric science was aroused by the reading of a book: "Gaia: a new look at life on Earth", by James Lovelock. Besides displaying a quite pessimistic view of Earth's future, Lovelock was also one of the first scientists to demonstrate the influence of mankind on its environment even at the most remote locations on Earth, forcing one's awareness of the proper responsibility in this issue, and triggering the willingness to contribute to a solution…

At the University of Bern, she studied the variations of stratospheric and mesospheric ozone by means of ground-based microwave instruments. She received her Ph.D. degree in Atmospheric Physics in 2000. Ozone was also the subject of her post-doctoral activities in the frame of a joint project of the University of Bern and MeteoSwiss, the Swiss national weather service.

When the International Space Science Institute decided to study its possible involvement in Earth Sciences in 2004, Yasmine Calisesi was asked to prepare the scientific and programmatic issues and to help assessing projects proposed by the international Earth Science community. She organized the workshop on Solar Variability and Planetary Climates in 2005 and co-edited the related subsequent volume i n ISSI's Space Science Series, volume 23. In 2006, she was awarded the Atmospheric Chemistry and Physics Award of the Swiss Academy of Sciences for her work on the continuous monitoring of ozone using microwave radiometry.

In 2007, Yasmine Calisesi received a call from the Swiss Federal Office for Energy to co-ordinate public energy research and education activities in Switzerland. She is now in charge of knowledge and technology transfer within the Swiss Federal Office of Energy.

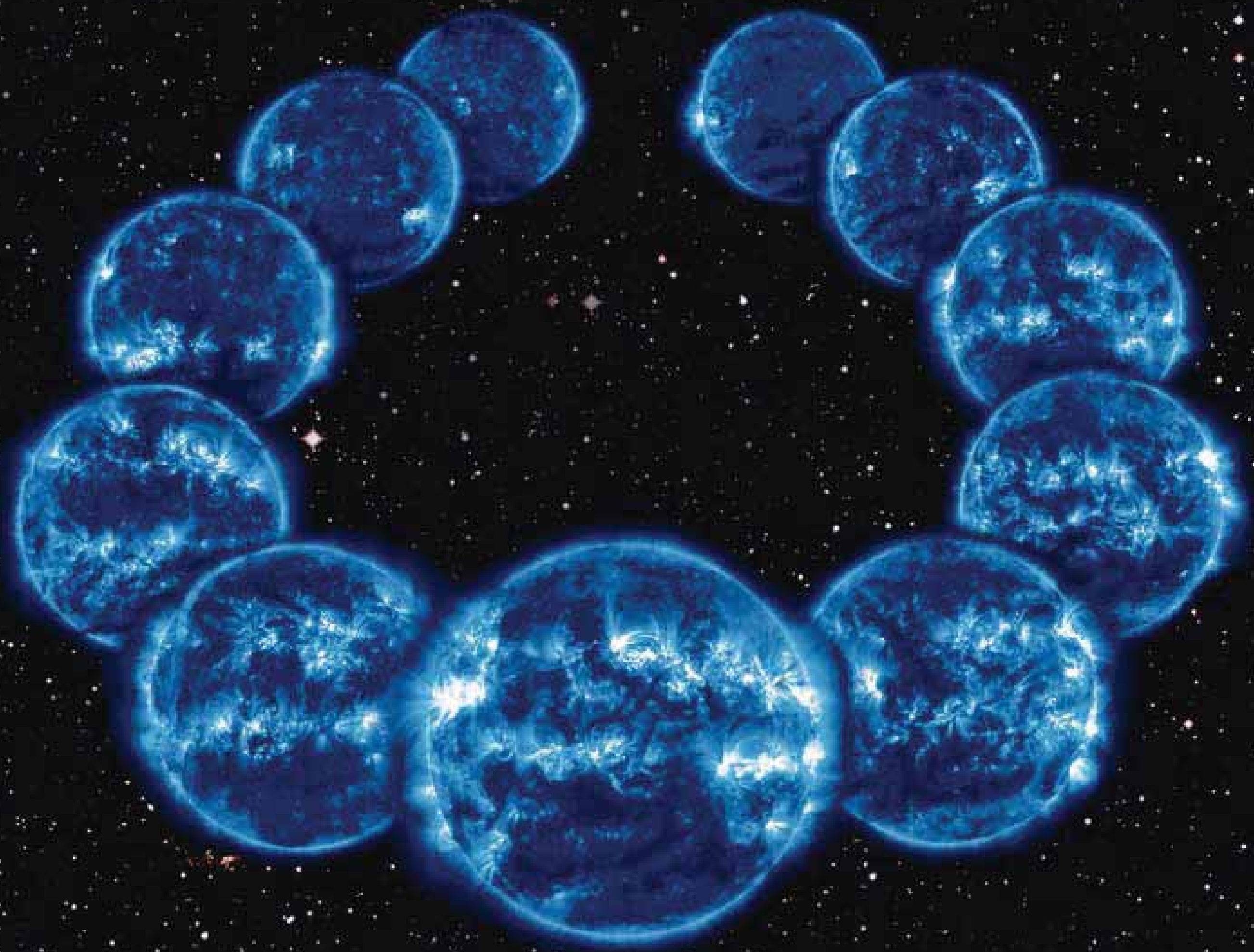
In her free time (whenever work permits!), she loves to run in the "Bremgartenwald", to go for long hikes (recommended: the Swiss National Park in Graubünden), or to explore Bern's surroundings by bicycle with her husband. Of course, she does not own a car, as this is far from necessary in Switzerland…

INTERNATIONAL SPACE SCIENCE INSTITUTE

SPATIUM

Published by the Association Pro ISSI

No. 22, August 2008



Solar Magnetism

Discovery and Investigation

Editorial

Nature has provided us with senses that are precisely tuned to our natural habitat. Our eyes carefully match the light spectrum of the Sun, but are sightless beyond. Our ears capture the sounds of our environment but are deaf for instance to the whistle of bats.

We can see the light of the Sun and enjoy its warmth, but it does not disclose its secrets to our senses. Technical means are required to overcome the limitations of our senses and to understand the daytime star. One of these secrets is the Sun's magnetic field that is responsible for such phenomena like the enigmatic spots on its surface that have been observed since the earliest times.

Baffled by the thought that a hot gaseous Sun should have a magnetic field, such ideas came up only very late. It was George Ellery Hale who suggested solar magnetism in 1908. The modern era of its exploration arrived around 1950, when new instruments provided detailed maps of the Sun's magnetic field. This led to the notion of a Sun which is not simply a leisurely shining disk, but rather a living entity with a huge range of fierce dynamic behaviours on all scales.

One of the major consequences of solar magnetism is the solar wind, a continuous high-speed stream of matter jettisoned by the Sun into its space environment. The concept of solar wind, as it is still understood today, was first proposed by Eugene Parker in 1958.

The year 2008 marks therefore not only the centenary of Hale's publication on solar magnetism but also the fiftieth anniversary of Parker's seminal paper on solar wind. This was reason enough for the International Space Science Institute to dedicate a workshop to solar magnetism in early 2008. The gathering was honoured by the presence of Eugene Parker who held a public lecture at the University of Bern on 23 January 2008.

The present issue of Spatium is devoted to solar magnetism and we are proud to present our readers herewith a written version of Professor Eugene Parker's fascinating lecture.

Hansjörg Schlaepfer
Brissago, August 2008

Impressum

SPATIUM
Published by the
Association Pro ISSI

INTERNATIONAL SPACE SCIENCE INSTITUTE

Association Pro ISSI
Hallerstrasse 6, CH-3012 Bern
Phone +41 (0)31 631 48 96
see
www.issibern.ch/pro-issi.html
for the whole Spatium series

President
Prof. Klaus Pretzl
University of Bern
Layout and Publisher
Dr. Hansjörg Schlaepfer
CH-8173 Neerach
Printing
Stämpfli Publikationen AG
CH-3001 Bern

Front Cover
Every eleven years, our Sun goes through a solar cycle. A complete solar cycle has been imaged by the sun-orbiting SOHO spacecraft in extreme ultraviolet light for each year of the last solar cycle, with images picked to illustrate the relative activity of the Sun.
(Credit: SOHO – EIT Consortium, ESA, NASA)

Solar Magnetism: Discovery and Investigation[1]

Eugene Parker, University of Chicago, USA

Introduction

"Were it not for magnetic fields, the Sun would be as uninteresting as most astronomers seem to think it is."

This statement is credited to R.W. Leighton[2] who, many years ago, pinpointed the Sun's important role in astrophysics when it comes to understanding a star's enigmatic behaviour. Thanks to its vantage location, our daytime star allows scientists to gather detailed data for which all other stars are much too far away. And these data reveal a degree of complexity so that even today many observations are still waiting for scientific interpretation. What makes the Sun really interesting is its magnetism that makes it an ever-changing and active star; without its magnetic fields, as Leighton pointed out, it would be an inactive ball of fire in the sky.

No wonder then that the Sun has been the object of fascinated observers for many centuries. Special attention has been paid to the sunspots that are dark stains waning and waxing randomly on the solar disk. It has been noted for a long time that these spots exhibit a periodicity of about eleven years with times of a larger number of sunspots alternating with times of fewer spots. Beyond this regularity, it came as a great surprise that from about 1650 to 1715 the Sun produced virtually no sunspots[3]. Obviously, the Sun seems to possess periodicities of even longer cycles, in any case much longer than a human's life span. Referred to as the Maunder Minimum, the period of unusually low solar activity in the 17th century coincided with an era of chilly temperatures on Earth. The Little Ice Age, as this period is called, was a time of lowered temperatures especially in the northern hemisphere: the summers were cool, the winters were long and cold, rivers froze, and sea ice was widespread. Its effects have been widely recorded by humans be it in artwork **(Fig. 1)**, documented in contemporary papers, but also by nature itself: they are preserved in tree rings and arctic ice. Although the causes of the Little Ice Age are still debated amongst scientists, it seems likely that solar activity in general plays a key role in determining Earth's climate: if the Sun changes its energy output, it affects the Earth's climate. And in a way yet to be understood, sunspots are related to the kind of output that matters for the Earth.

Fig. 1: Enjoying ice. Painted by the Dutch artist Barend Avercamp (1612–1679), this canvas shows how life was during the cold winter months in the Netherlands during the Little Ice Age, when the Sun produced extremely low numbers of sunspots. It is thought that this phenomenon is closely related to a period of chilly winters all over Europe.

---

1 The present text follows a lecture by Prof. Eugene Parker held at the University of Bern on 23 January 2008. The notes were taken by Dr. Hansjörg Schlaepfer and reviewed by Prof. André Balogh, ISSI.
2 Robert B. Leighton, 1919, Detroit – 1997, US-American physicist.
3 See Spatium 8: Sun and Climate.

In addition to sunspots, the solar corona was also observed for centuries at the times of total solar eclipses. Physics, however, was not yet ready to explain the common source of both phenomena: solar magnetism. The idea of solar magnetism began in the second half of the 19th century, most probably initiated by the resemblance of the solar corona to the pattern of iron filings around a bar magnet, **Fig. 2.**

In the following chapters we shall first describe some physical properties of the Sun and then report on the milestones in the history of the discovery and investigation of solar magnetism.

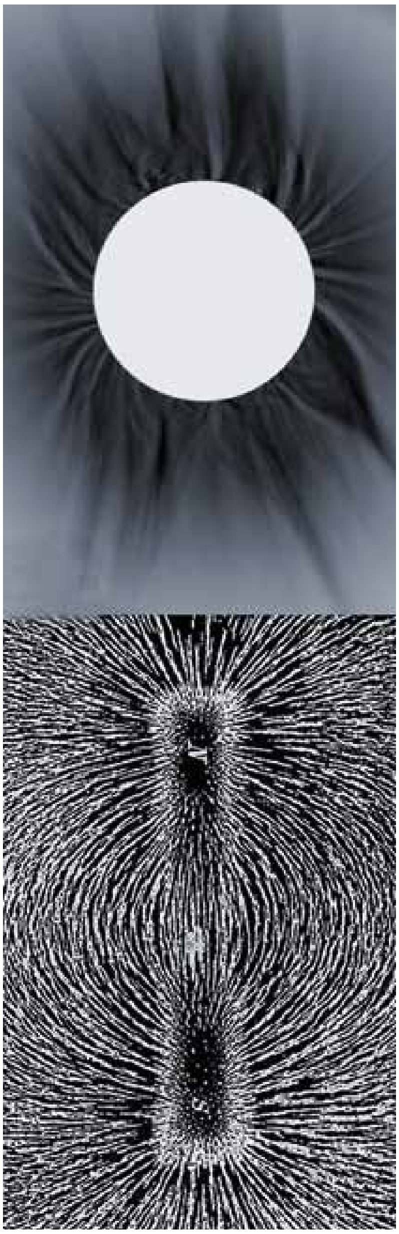


Fig. 2: The solar corona during eclipse shows some resemblance to the iron filings aligning along the magnetic field of a bar magnet.

The Sun, an Ordinary Star

The Sun is one of over 100 billion stars in the Milky Way Galaxy. It is about 25,000 light-years from the centre of the galaxy, and it revolves around the galactic centre once about every 250 million years. The Sun is a huge, glowing ball containing 99.8 % of the solar system's mass. During the process of its formation, some 4.6 billion years ago, pressure and temperature in its core reached the critical values needed to ignite nuclear fusion processes that generate enormous amounts of energy and cause the Sun to shine[4] **(see box Nuclear Fusion).** The Sun has enough fuel to stay active for another say 5 billion years. It is this nuclear fusion process that provides ultimately the energy required for life to develop on Earth. As a consequence of its high temperature, much of the Sun is made up of plasma, that is a gas of ionized atoms, mainly hydrogen. Carrying electric charges, such ionized atoms are susceptible to magnetic fields, and, as we shall see later, may themselves generate magnetic fields.

The Sun consists of a core that extends about one-fourth of the way to the surface, **Fig. 3.** Here, most of the nuclear fusion takes place at an estimated temperature of some

[4] See Spatium 2: Das neue Bild der Sonne.

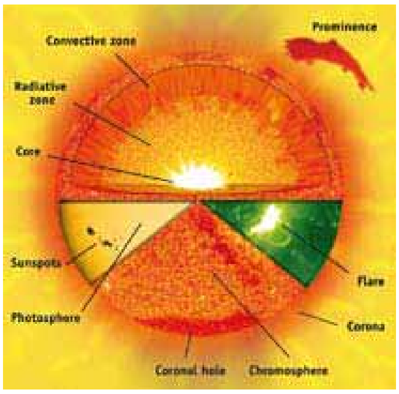


Fig. 3: Cut-away view of the Sun showing its different shells and external phenomena. (Credit: ESA/NASA)

15 million K. The core is surrounded by a shell, called the radiative zone, where the energy from the core travels through by radiation. Photons emerging from the core pass through layers of gas and are scattered by the dense particles of gas so often that an individual photon may take as much as a million years finally to reach the Sun's surface. It's a fascinating thought that the sunlight we see today may stem from the times of the dawn of mankind.

The outer limit of the radiative zone is about 70% of the way to the solar surface. Outside the radiative zone follows the convection zone that reaches to the Sun's surface. This zone consists of the "boiling" convection cells that are heated by the photons from the radiative zone below and bring up mass and magnetic fields to the surface of the Sun.

Just like the Earth, the Sun is surrounded by an atmosphere, although the Sun's atmosphere is different in most respects from that of the Earth. Its lowest layer is the photosphere (really the visible surface of the Sun) which is relatively cool, at 5,700 K, so that the gas is only partially ionized. The next zone up is the chromosphere, where temperature can reach some 10,000 K. The chromosphere is covered by the transition region, where the temperature rises very abruptly up to 500,000 K and more. This region is highly structured into loops and streams of ionized gas. As we will see later, these structures connect vertically down to the solar surface, and magnetic fields that emerge from inside the Sun shape them. The outermost layer of the Sun, the corona, is even hotter, generally in excess of one million K. It extends far into space and forms a large cavity called the heliosphere[5]. The Sun and all the planets are inside the heliosphere. Far beyond the orbit of Pluto the heliosphere joins the interstellar medium, the dust and gas that occupy the space between the stars. The flow of coronal gas into space is known as the solar wind. It makes the Sun lose close to 2 million tons of mass every second. At the distance of the Earth, the density of the solar wind amounts to about 10 to 100 particles per cubic centimetre.

The Sun is far from quietly shining. Rather, tremendous amounts of energy flow from its interior towards the surface and some of the energy causes violent events on scales that are many times the size of the Earth. Solar flares are hot violent explosions in the Sun's lower atmosphere where temperatures of up to 10 million K or more can be reached and electrons, protons and heavier ions are accelerated to near the speed of light. Other violent events are coronal mass ejections. These are ejections of large amounts of solar matter, up to 10^{12} kg, which equals a cube of rock of 1 km side length that is jettisoned out to space at speeds up to 2,700 km s^{-1}. These and further spectacular features of the Sun will be addressed later in more detail.

Nuclear Fusion is the process by which atoms join together to form other atoms with heavier nuclei. It is accompanied by the release or absorption of significant amounts of energy. The fusion of two light nuclei (lighter than iron or nickel) generally releases energy while the fusion of nuclei heavier than iron or nickel absorbs energy[6]. In the Sun, most of the energy is generated by the fusion of four hydrogen atomic nuclei, or protons, into a helium nucleus that consists of two protons and two neutrons. The difference in mass between four protons and the helium nucleus is converted into energy according to Einstein's $E = mc^2$ formula. The Sun converts mass into energy at the rate of 4.28 million tons per second thereby generating its unimaginable amount of energy.

---

[5] See Spatium 17: The Heliosphere: Empire of the Sun.
[6] See Spatium 13: Woher kommen Kohlenstoff, Eisen und Uran?

Discovery of Solar Magnetism

The Beginnings

The first recognition of magnetic fields in the Sun was in the second half of the 19th century, more or less contemporary with the growing realization that the activity at the Sun drives the terrestrial aurora through the emission of solar corpuscular radiation.

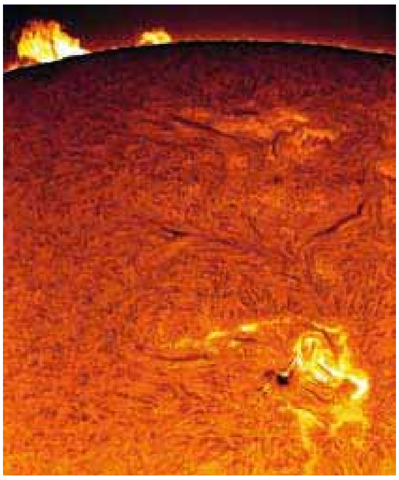


Fig. 4: The Sun's surface seen at H-alpha. Images taken at the H-alpha wavelength provide an insight into the processes on the Sun's surface, specifically the distribution of hydrogen. This table shows a very active sunspot region, larger than the Earth. Relatively cool regions appear dark while hot regions appear bright. On the left top, solar prominences are visible hovering above the Sun's surface. (Credit: Greg Piepol, sungazer.net)

In 1866, Sir Joseph Norman Lockyer[7] remarked that the coronal streamers, visible during an eclipse of the Sun, outlined a pattern resembling the magnetic field around a bar magnet. Lockyer, however, never published his thoughts and the concept of a general magnetic field for the Sun was proposed by others some years later. In 1912, solar magnetic fields of smaller scale were suggested by the strong filamentary appearance and dynamical behaviour of prominences.

In 1908, George Ellery Hale[8] published a seminal article in the Astrophysical Journal entitled "On the probable existence of magnetic field in sun-spots". This marks the first scientific recognition of magnetic fields in the Sun. His observations were made by his then newly developed spectroheliograph. This is basically a telescope which, however, allows the elimination of all but a very narrow part of the solar spectrum. Specifically, Hale used the H-alpha line (**see box H-alpha** and **Fig. 4**) from which he discovered that the filaments around sunspots sometimes show a strong spiral or vortex form.

Spurred by his groundbreaking discovery, Hale continued to observe the Sun and came to the conclusion that the Sun has an overall magnetic field of some 50 Gauss at the poles. (Note: the Earth's magnetic field has the strength of some 0.6 Gauss).

Subsequent observations, however, did not support this conclusion, suggesting that Hale was grappling with noise rather than a true signal. And there the observations rested until technological advances during the period of WW II provided a great leap forward in detector sensitivity.

Observing the Sunspots

Hale's idea that the sunspot is associated with – and is perhaps a consequence of – a rotary flow of gas motivated the British astronomer John Evershed[9] to look at sunspots near the limb to measure the Doppler shift arising from rotation about a vertical axis. He quickly discovered a radial outflow from the sunspot at the photospheric level, reaching a maximum of about 2 km/sec. Further, he found a radial inflow at higher levels, **see Fig. 5.**

Later, it was suggested that these flows near a sunspot are associated in some unspecified way with a vortex in the deep photosphere, believed at that time to be the origin of the sunspot magnetic field. The idea of a magnetic field associated with a vortex was fondly held, even if not demonstrable by theory. This prompted numerous scientists to propose theories, however, none of these has survived until today.

[7] Sir Joseph Norman Lockyer, 1836, Rugby, Warwickshire, UK – 1920, Salcombe Regis, UK, British scientist and astronomer.
[8] George Ellery Hale, 1868, Chicago – 1938, Pasadena, USA, US-American astronomer.
[9] John Evershed, 1864, Gomshall, Surrey, UK – 1956, Ewhurst, Surrey, UK, British astronomer.

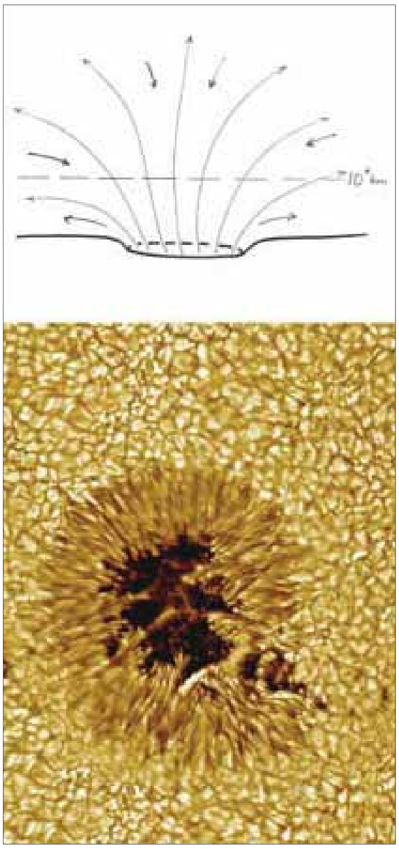


Fig. 5: Schematic diagram illustrating the outflow and the inflow of plasma above a sunspot (upper panel, illustration by Eugene Parker). The lower table shows a close-up picture of a sunspot that is slightly cooler and less luminous than the rest of the Sun. The Sun's complex magnetic fields create this cool region by inhibiting hot material from entering the spot. Sunspots can be larger than the Earth and typically last for only a few days. This high-resolution picture also shows clearly that the Sun's face is a bubbling sea of separate cells of hot gas. These cells are known as granules. A solar granule is about 1,000 kilometres across and lasts only about 10 minutes. (Credit: Vacuum Tower Telescope, NSO, NOAO).

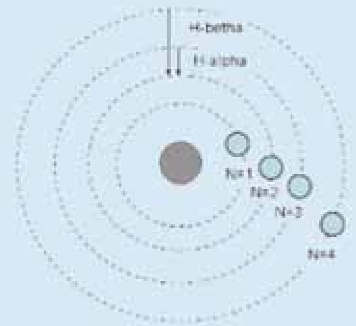
Struggling for Theoretical Understanding

Quite apart from the theoretical problem of the origin of the magnetic fields of the Sun, there was the stark fact of their existence, with wide-ranging implications. Hale's idea of a 50 Gauss polar magnetic field stimulated the construction of diverse theories over the next several decades. For instance, such a strong polar field represents a substantial investment in energy. A rough estimate delivers an energy content of the Sun's magnetic field at that strength of an equivalent of about 50 sec of the solar energy output. Not impossible, but impressive nonetheless!

Now, the magnetic field stipulated by Hale extrapolates out to the orbit of Earth to be 2.5×10^{-6} Gauss. Alfvén[10] seized upon this dipole field to argue that cosmic rays are a local solar phenomenon, created by the Sun and trapped within the enormous dipole magnetic field of the Sun. Alfvén's concept was in opposition to the view espoused by Fermi[11], that the cosmic rays originate in interstellar space and circulate freely along the interstellar magnetic field lines.

Alfvén's concept of localized cosmic rays played a crucial role in his idea that half of the stars in the Galaxy are composed of antimatter, distributed randomly among the stars of ordinary matter. For it was apparent from exposures of photographic emulsions above the terrestrial atmosphere that cosmic ray particles are made up of protons, with smaller numbers of heavier nuclei, and with no more antiprotons than would be expected from collisions of cosmic ray particles

H-alpha is the name of a specific emission line of hydrogen at 656.281 nm in the red part of the visible spectrum.



In the Bohr model of the atom, electrons exist exclusively in quantized energy levels surrounding the atom's nucleus. These energy levels are described by the principal quantum numbers n = 1, 2, 3, etc. Electrons may only exist in these states, and may only transit between these states. The set of transitions from $n \geq 3$ to n = 2 are called the Balmer series, specifically

- the transition from energy level 3 to energy level 2 is called Balmer-alpha or H-alpha
- from n = 4 to n = 2 is called H-beta, etc.

As this light is generated by ionized hydrogen, it allows tracing hydrogen in astronomical objects.

10 Hannes Olof Gösta Alfvén, 1908, Norrköping, Sweden – 1995, Djursholm, Sweden, Swedish plasma physicist and Nobel Prize laureate 1970.

11 Enrico Fermi, 1901, Rome – 1954, Chicago, Italian physicist, Nobel Prize laureate 1938.

with the nuclei of the ambient interstellar gas. On the other hand, Fermi's galactic concept of cosmic rays applied to Alfvén's matter-antimatter stars would predict that half of the cosmic rays are antiprotons, contrary to observed facts. Solar local confinement of cosmic rays was essential for the matter-antimatter universe concept.

The theoretical problem of solar confinement of cosmic rays was that a solar dipole magnetic field of 50 Gauss could not possibly confine, even temporarily, a proton above some 4 GeV, whereas the cosmic ray energy spectrum is observed to extend unbroken to much higher energies[12]. So cosmic rays are galactic in their distribution and provide proof that there are few, if any, antimatter stars in the Galaxy. We see that Hale's idea of a 50 Gauss solar field set in motion an interesting train of thought.

There were other important consequences of Hale's 50 Gauss solar magnetic field. It was recognized that the Sun sometimes emits bursts of "solar corpuscular radiation", presumed to consist of equal numbers of electrons and protons, at speeds of the order of 1,000 km/sec. This plasma impacts the Earth's magnetic field a day or two after emission from a big flare on the Sun, producing a geomagnetic storm that is a temporary disturbance of the Earth's magnetosphere, see **Fig. 6.** Magnetic storms usually last 24 to 48 hours, but some may last for many days. In 1989, such an electromagnetic storm disrupted power throughout most of Quebec, and it caused auroras as far south as Texas.

So one may ask, if Hale had been correct with his estimates of 50 Gauss polar magnetic fields of the Sun, how is it that a burst of solar corpuscular radiation ejected by a solar eruption passes freely out through space at that speed? The difficulty is obvious today, but a hundred years ago the solar corona was a mysterious entity. It was not until the pioneering work of Grotrian, Edlen, and Lyot in the 1940s that the corona was convincingly understood to be a million degree plasma with a density of about 10^8 ions/cm^3. Even then the solar corpuscular radiation was considered as being emitted from active regions in the form of intense beams. It was not until much later that the expansion of the corona was demonstrated and understood, producing the supersonic solar wind.

The solar wind was recognized by Eugene Parker in 1958 as the conventional solar corpuscular radiation phenomenon, **Fig. 7.** It is amusing to note, then, that Hale's 50 Gauss dipole magnetic field at the Sun would overpower the tenuous corona, holding it in a tight embrace with a magnetic pressure some 300 times the pressure of the coronal plasma.

The Hard Vacuum Concept

Now, prior to about 1950, space was generally regarded as a hard vacuum, devoid of free electrons and ions except for the occasional fast particles from solar eruptions passing by Earth and conveniently dis-

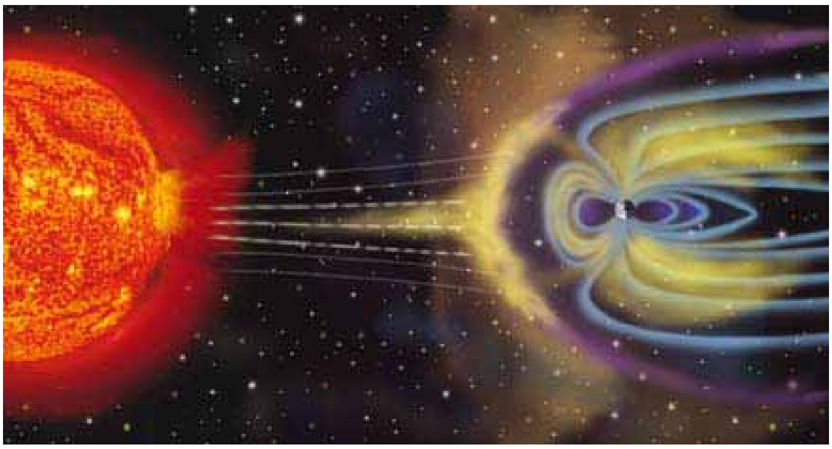


Fig. 6: Artist's view of a solar burst interacting with the Earth's magnetic field. (Credit: NASA/GSFC)

[12] See Spatium 11: Cosmic Rays.

DYNAMICS OF THE INTERPLANETARY GAS AND MAGNETIC FIELDS*

E. N. Parker

Enrico Fermi Institute for Nuclear Studies, University of Chicago

Received January 2, 1958

ABSTRACT

We consider the dynamical consequences of Biermann's suggestion that gas is often streaming outward in all directions from the sun with velocities of the order of 500–1500 km/sec. These velocities of 500 km/sec and more and the interplanetary densities of 500 ions/cm^3 (10^{14} gm/sec mass loss from the sun) follow from the hydrodynamic equations for a 3×10^6 °K solar corona. It is suggested that the outward-streaming gas draws out the lines of force of the solar magnetic fields so that near the sun the field is very nearly in a radial direction. Plasma instabilities are expected to result in the thick shell of disordered field (10^{-5} gauss) inclosing the inner solar system, whose presence has already been inferred from cosmic-ray observations.

Fig. 7: Facsimile of Eugene Parker's Article in Astrophysical Journal, volume 128, p. 664 (1958).

appearing into the "infinite" void outside the solar system. The solar particles somehow "cleaned up" after themselves, leaving no stray remnants to provide electrical conductivity in interplanetary space. An extreme consequence of this concept was put forth by Lord Kelvin at the end of the 19th century. By that time, science was aware of the fluctuations of the geomagnetic field that recur with the 27 day period of rotation of the equatorial regions of the Sun, and of the stronger fluctuations that often follow a day or two after a large flare on the Sun. A Sun – Earth transit time of a day or two implies a speed of the order of 1,000 km/sec, obviously particles of some kind. However, Kelvin had the idea that a magnetic variation here at Earth could only represent a variation of the general dipolar magnetic field of the Sun. Kelvin confidently announced that no possible magnetic field activity at the Sun could provide the observed fluctuations of the geomagnetic field, and, hence, solar activity could have nothing to do with the geomagnetic fluctuations. Spoken with the authority of Kelvin, there was no further pursuit of the geomagnetic fluctuations until Sydney Chapman[13] entered the field in 1918. He pursued the effects of the solar particle emission and the impact of those particles against the magnetic field of Earth. Together with V. C. A Ferraro[14] he showed how the impact creates the initial compressive phase of a geomagnetic storm.

Finally note that simultaneously with the "hard vacuum" concept of space there was the thriving contrary notion that the faint sky brightness known as the zodiacal light represented scattering of sunlight from free electrons. The observed intensity of the zodiacal light indicated about 500 electrons/cm^3 at the orbit of Earth. Not a vacuum hard enough to allow electrostatic fields in interplanetary space. Yet the idea of interplanetary electrostatic potential differences of 10^9 Volts or more was introduced in the middle of the 20th century to explain the recently discovered variations in the cosmic ray intensity. What produced the electrostatic fields was never specified. Such are the scientific dilemmas sometimes forced upon scientists when interpreting their observations. Today it is appreciated that the zodiacal light represents the scattering of sunlight by dust grains, of cometary and meteoritic origin, orbiting the Sun.

13 Sidney Chapman, 1888, Eccles, England – 1970, Boulder, Co., USA, British physicist and mathematician.
14 Vincenzo Consolato Antonio Ferraro, 1907–1974, Italian/British applied mathematician.

The Modern Era of Solar Magnetism Investigation

The modern study of solar magnetic fields is best defined as beginning with the invention of the magnetograph **(see box The Magnetograph)** in the middle of the 20th century. That was the step that sparked the research of solar magnetism and expanded into a new world of exploration as observations turned up the immense complexity of the active magnetic fields on the Sun, ranging from the dynamical small-scale fibril structure of the photospheric "magnetic carpet", to the active fine structure of the "quiescent" prominences, to the universal magnetic flaring phenomenon, to the coronal X-ray emitting regions, to the coronal mass ejections, etc.

If one thing is clear, it is that magnetic fields are the driving factor in solar activity. Without magnetic fields the Sun would be a serene, gently convecting, self-gravitating gaseous sphere: the classical concept of a star. But the magnetic fields change all that, driving many mysterious dynamical activities that are still only partly understood. The essential point is that observations lead the way into the mysterious world of solar and stellar magnetic activity, with theory working to understand the dynamical phenomena discovered by the observations. The effort is aimed at developing an understanding of the physics of solar activity, and stellar activity in general. The overall subject of solar magnetism - synonymous with solar activity - has become so extensive that we can do little more than summarize some of the major phenomena below.

The Magnetograph is a scientific instrument that allows remote sensing of magnetic fields. It exploits the Zeeman effect as outlined below:

In the box H-alpha on page 7 we noted that the electrons orbiting an atomic nucleus can occupy only a specific set of quantized levels. Furthermore, electrons can transit between these quantized levels. Such a transition gives rise to a loss or gain of energy. As stated in that box, the H-alpha line specifically corresponds to the transition of the electron of a hydrogen atom from level 3 to level 2 whereby a photon of wavelength 656.281 nm is emitted. This spectral line is produced by the hydrogen atom exclusively: it constitutes therefore a fingerprint of the hydrogen atom. As the laws of physics are universal, a hydrogen atom emits this very line irrespective of whether it is in an Earth-bound laboratory or billions of light-years away. This and other spectral lines may be measured in the lab and used to trace the respective atoms throughout the universe.

Now, light waves are capable of carrying much more information. It was in 1886 that Pieter Zeeman[15] observed the spectrum of the light emitted by a sodium flame. More specifically, he placed the sodium flame between the poles of a strong magnet. Of course, he expected to find the typical emission lines of sodium, which he did of course, but, remarkably, these lines were splitted into several individual lines, a fact he correctly attributed to the effect of the magnet. From more laboratory experiments he found that the amount of splitting is dependent on the magnitude of the magnetic field strength at the location of the light emitting atoms. Light waves, therefore, can be used to trace the magnetic (and also the electric) fields at the location where they were generated.

The splitting of a spectral line into several components in the presence of a magnetic field is called the Zeeman effect. It is this physical process that allows a magnetograph to display distant magnetic field strengths and it is the magnetograph which set the stage for the next quantum leap in the study of the Sun.

[15] Pieter Zeeman, 1865, Zonnemaire, the Netherlands – 1943, Amsterdam, Dutch physicist, Nobel Prize laureate 1902.

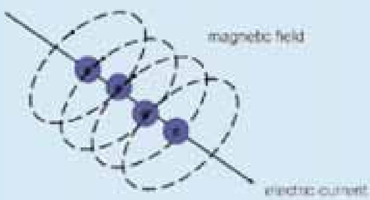
The Major Phenomena

The Origins of Solar Magnetism

Applying Ampère's law **(see box Ampère's Law)**, one concludes that at the root of the Sun's global magnetic field there must be strong electric currents. This is confirmed by the following reasoning: as the Sun is a huge ball of gaseous plasma it allows the gases at different latitudes rotating at different rotation speeds: while at the equator one full revolution is completed within about 25 days, near the poles the period is about 35 days. Between the different shells of the Sun rotating at different rates, a shear is created, which in turn creates an electric current deep within the Sun. It is this current that produces the Sun's global magnetic dipole field. Such a system is called magnetohydrodynamic dynamo. Yet, the detailed mechanism of the solar dynamo is not known and is the subject of current research. Theoretically, depending on the structure of the flow, the dynamo may be self-exciting and stable, self-exciting and chaotic, or decaying. The Sun's dynamo is self-exciting and chaotic: the direction of the field reverses about every eleven years, causing the sunspot cycle and magnetic field lines rise to the surface of the Sun and manifest themselves as sunspots on the surface.

Ampère's Law

Some 200 years ago, André-Marie Ampère[16] discovered the basic laws of magnetism. He found that an electric current creates a magnetic field around it as outlined below. The strength of the magnetic field is proportional to the electric current, while the direction of the magnetic field is orthogonal to the direction of the electric current.



For comparison, the Earth also possesses a magnetic field. It is generated by the flow of conducting fluid in the core across a pre-existing magnetic field, which generates electric currents that in turn reinforce the magnetic field. In contrast to the Sun, however, the reversal of the Earth's magnetic polarity is chaotic with periods between tens of thousands and many millions of years. Just like in the case of the Sun, there is no conclusive theory as to why the Earth reverses its magnetic field.

Returning to the Sun, the net effect of these processes is a dynamo that creates bands of east - west fields that migrate from middle latitudes to the equator over a period of some years, duplicating the 11-year magnetic cycle that one is

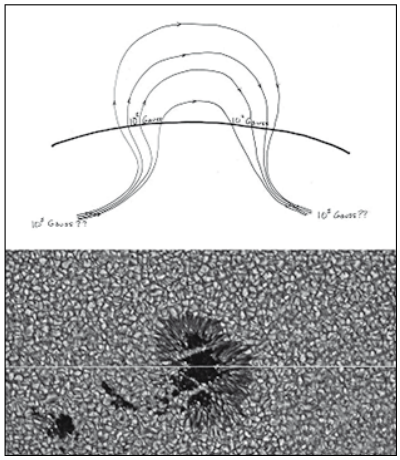


Fig. 8: The secrets of a sunspot: The lower panel shows a sunspot on the Sun's surface. Hot and cold regions are shown light and dark, respectively. The upper panel shows a theoretical interpretation of the magnetic fields producing the sunspot. (Credits: Eugene Parker, upper panel; Vacuum Tower Telescope, NSO, NOAO, lower panel)

Fig. 9: This image shows the looping structures of the magnetic field above the Sun's surface. It was acquired in the ultraviolet light from Fe IX/X (hot gas of iron atoms) by NASA's TRACE spacecraft. (Credit: NASA)

[16] André-Marie Ampère, 1775, Lyon, France – 1836, Marseille, French physicist, main discoverer of magnetism.

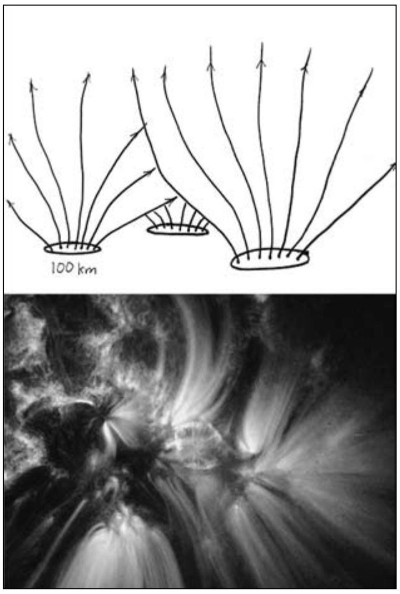


Fig. 10: Microfibrils: The Sun's surface displays a fine structure called microfibrils. The upper table shows a theoretical interpretation of the formation of such fibrils by Eugene Parker. Microfibrils consist of hot gas at about one million K, which emits extreme ultraviolet light as shown in the lower table. It occurs in large patches some 100 km in extent, and appears between 500 and 1,000 km above the Sun's visible surface. These microfibrils are typically persisting for tens of hours. (Credit: NASA)

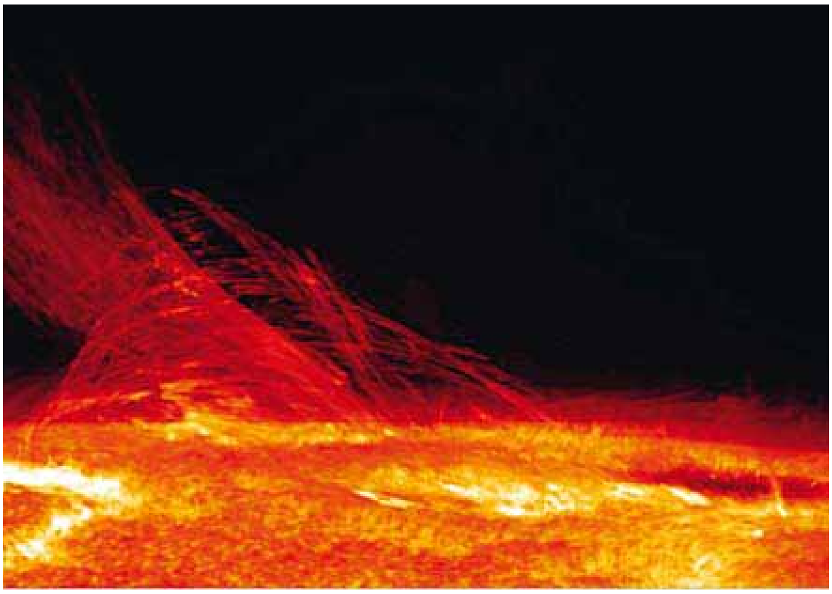


Fig. 11: The plasma of the Sun captured by the Hinode spacecraft's Solar Optical Telescope. Hinode (Japanese for sunrise) is a project to study the Sun, led by the Japanese Aerospace Exploration Agency (JAXA) in collaboration with international partners. It aims at exploring the magnetic fields of the Sun, and improving our understanding of the mechanisms that power the solar atmosphere and drive solar eruptions. This image reveals the filamentary nature of the plasma connecting regions of different magnetic polarity. (Credit: JAXA, NASA)

seeing at the surface. Unfortunately, we cannot see the magnetic fields below the surface to establish in what convective flows and at what depths the dynamo is operating. So there are a dozen or more configurations that can be constructed with enough free parameters to fit the surface observations. The theoretical dynamo is further challenged by such phenomena as the Maunder Minimum, when there were no sunspots on the Sun and the magnetic activity of the Sun appeared to be switched off.

The theory of the magnetohydrodynamic dynamo effect has become an extensive topic in theoretical physics establishing that almost any sufficiently vigorous flow of conducting fluid lacking simple symmetry has the ability to increase the magnetic flux, i.e. amplify or regenerate magnetic fields. Unfortunately the hydrodynamics of the fluid motion is a more formidable problem, and has not yet been solved for the convective zone and the non-uniform rotation of the Sun.

Sunspots

Sunspots are part of the bipolar magnetic regions that form with the upward bulging of an east – west magnetic field lying somewhere near the bottom of the convective zone, at a depth of about $\frac{2}{7}$ of the solar radius, see **Fig. 8 and 9.** The form of the upward bulge suggests the appellation Ω-loop. Sunspots form in the regions where the Ω-loop extends through the visible surface of the Sun. It is, however, not understood why the laws of physics compel the formation of sunspots in these surface magnetic regions.

Now it should be noted that the buoyant rise of the Ω-loops through the surface of the Sun as outlined in **Fig. 8** poses a serious theoretical challenge that has not yet been resolved.

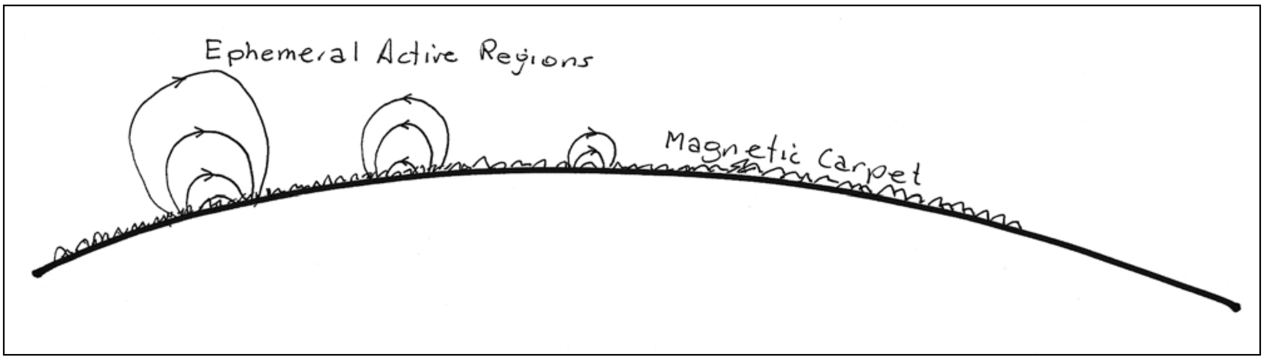


Fig. 12: The fine structure of the Sun: The smaller bipolar magnetic regions on the Sun's surface are called ephemeral active regions while the even smaller structures are termed magnetic carpet. (Credit: Eugene Parker)

Fibril Texture

It came as a substantial surprise about 25 years ago to discover that the photospheric magnetic fields exhibit a microfibril structure rather than the expected continuum indicated by the limited resolution of the magnetographs. Some clever scientific detective work showed that the individual fibrils have magnetic fields of 1–2 kilogauss with diameters of only 100 km, well below the resolution of the magnetographs. The observed mean fields of 5–10 Gauss in quiet regions and 100 Gauss in active regions represent the spacing of the fibrils. Why the magnetic field is in this enhanced magnetic energy state (compared to being spread out smoothly) is not clear. The individual fibrils are observed as bright spots against the normal photosphere **(Fig. 10)**. The waxing and waning of fibril numbers as the active regions vary over the 11-year magnetic cycle is responsible for most of the solar luminosity variations of one part in 10^3 in step with the magnetic activity.

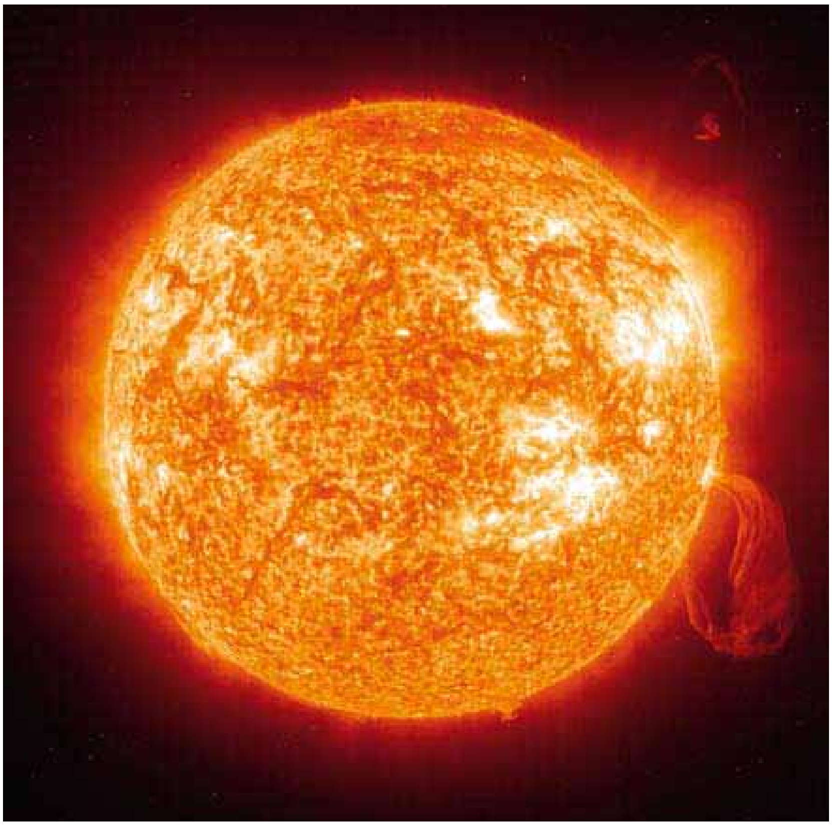


Fig. 13: The ESA/NASA SOHO spacecraft captured this image of a magnificent prominence above the Sun's limb. Seen at the lower right, streams of relatively cool dense plasma are lofted along looping magnetic field lines extending outward about 30 times the diameter of Earth. Far above the limb at the upper right, a disconnected ghostly arc surrounds a dark cavity with bright central emission. These features are telltale signs of a coronal mass ejection. (Credit: SOHO-EIT Consortium, ESA, NASA)

The bipolar magnetic regions created by the emergence of Ω-loops have maximum longitudinal (east - west) dimensions of the order of 300,000 km, i.e. a significant fraction of the solar radius, and they appear with increasing frequency on smaller scales down to a few times 10 km. Below about 20,000 km they are called "ephemeral active regions", see **Fig. 12.** These are distributed over all latitudes and with the east - west alignment fading away toward the smaller end of the scale. Ephemeral active regions do not develop sunspots, and the smaller ones vary but little with the 11-year magnetic activity cycle. It has been suggested that their magnetic fields have an origin different from the main east - west magnetic field deep in the convective zone.

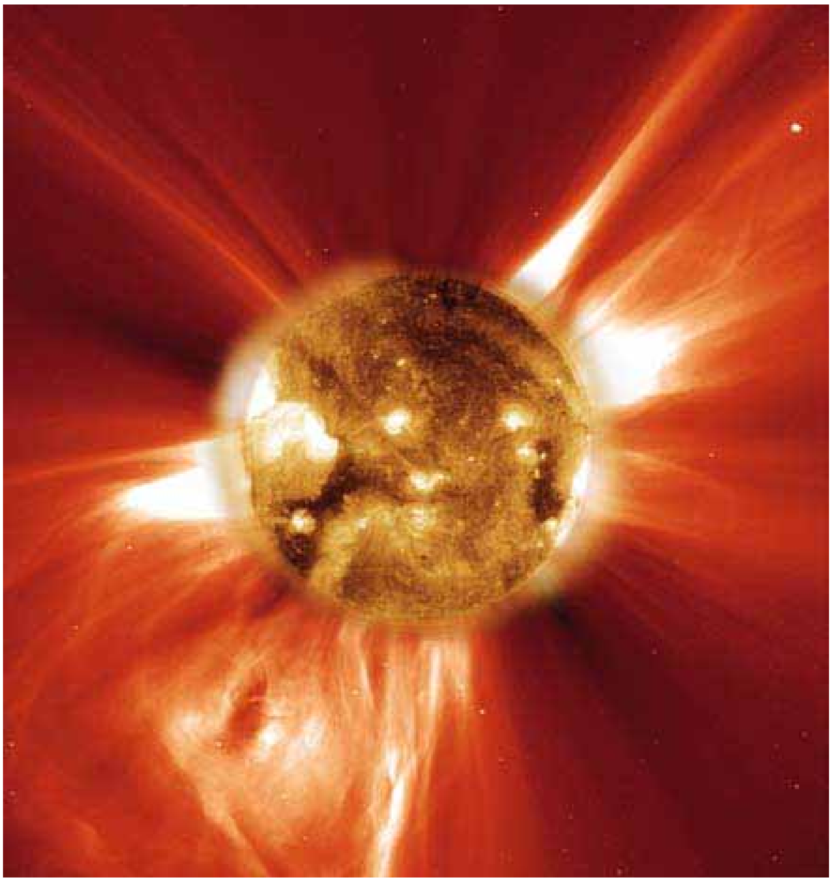


Fig. 14: Coronal Mass Ejections: This image shows many erupting filaments lifting off the active solar surface and blasting enormous bubbles of magnetic plasma into space. Direct light from the Sun is blocked in the inner part of the above image, and replaced by a simultaneous image of the Sun in ultraviolet light. The field of view extends over two million kilometres from the solar surface. Near the minimum of the solar activity cycle CME's occur about once a week, but near solar maximum rates of two or more per day are typical. (Credit: SOHO Consortium, ESA, NASA)

The Living Sun

Advanced instrumentation such as on ESA's SOHO spacecraft, or NASA's Advanced Composition Explorer (ACE) and Transition Region and Coronal Explorer (TRACE) spacecraft have provided evidence of a photosphere peppered by the small magnetic fibrils with continual dissipative activity as the squirming fibrils reconnect among their neighbours. These small-scale fields are referred to as the "magnetic carpet", with which the photosphere is very much alive.

Now the Ω-shape of the magnetic fields of the bipolar active regions traps coronal plasma, preventing it from expanding away from the Sun to join with the solar wind from the regions of weak open magnetic fields. The enclosed plasma is heated to temperatures of 1–10 million K with the density rising above the normal corona (10^8 ions/cm^3) to somewhere in the vicinity of 10^{10} ions/cm^3. At this enhanced density, the X-ray emission becomes significant, providing the quiet time X-ray luminosity of the Sun. The occasional large flare of course produces a very intense burst. The X-ray emitting temperatures in the enclosing Ω-shaped fields appear to be the result of many tiny (micro, nano, and pico) flares in the interlaced field lines of the Ω-fields. The interlacing is carried on continuously by the swirling gases at both foot points of the Ω-fields. This is evidently the reason why stars with convective zones like the Sun are all X-ray emitters.

Coronal Mass Ejections

A particularly interesting facet of solar magnetic activity is the spectacular coronal mass ejection (CME, **see Figures 14 and 15).** Such an event is created when the convection below the surface of the Sun increasingly deforms the Ω-loops. The internal deformation and twisting of the field accumulates in the bipolar Ω-loop to the point where the increasing internal energy produces instability, exploding outward while rapid magnetic reconnection cuts the loop free from the photospheric foot points to allow escape from the Sun. The coronal mass ejections escape with speeds up to 2,000 km/sec. It is the impact of coronal mass ejections that are directed towards the Earth against the geomagnetic field that produces the terrestrial magnetic storm phenomenon.

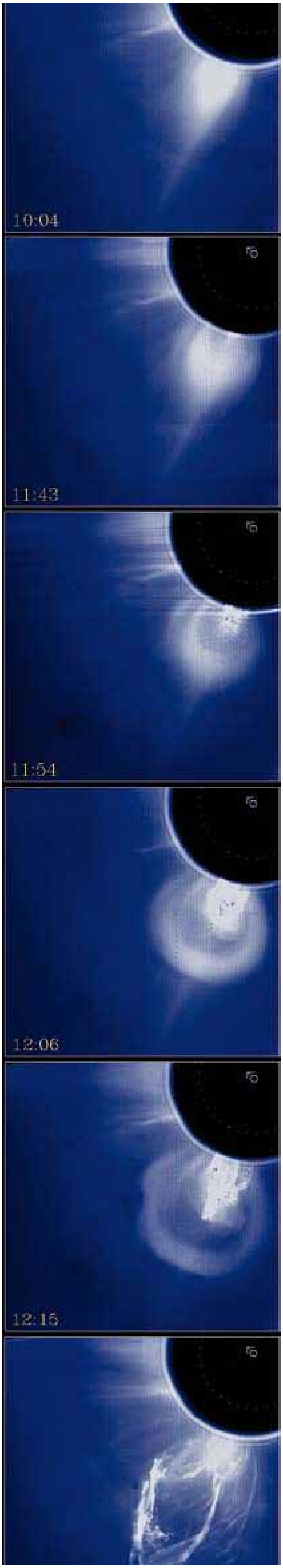


Fig. 15: The dynamics of a coronal mass ejection and prominence eruption observed in white light from the SMM (Solar Maximum Mission) spacecraft. The time of each panel increases from the top to the bottom. The dashed inner circle in each panel is the solar radius, the occulting radius is at 1.6 solar radii. (Credit NASA/GSFC)

Outlook

Never believe an observation for which you do not have a theory. This dictum by Sir Arthur Stanley Eddington[17] is the paradigm for all those active in the field of solar magnetism, where observations have always led the way at every step, with theory running along behind in an ongoing attempt to catch up.

Future generations of scientists will be challenged to recognize the clues for an advanced understanding of our daytime star. New means of observations based on refined theoretical concepts and advanced instruments with much greater spatial and temporal resolving power may lead the way to further understanding. In the meantime it is reassuring to know that the Sun will continue to shine as it mercifully did since 4.6 billion years.

---

17 Sir Arthur Stanley Eddington, 1882, Kendal, UK – 1944, Cambridge, UK, British astrophysicist.

SPATIUM

The Author

The general curiosity of young Eugene Parker about why things do what they do captured his interest as a young child. By the time he was in high school, he found physics class so interesting that it sealed his career path: astrophysics is simply applying the laws of physics to large-scale phenomena that cannot unfold inside a laboratory. Later, Eugene Parker visited Michigan State University and California Institute of Technology. Then, he spent four years at the University of Utah and since 1955 has been at the University of Chicago, where he has held various positions in the physics department, the astronomy and astrophysics department, and the Enrico Fermi Institute.

In the mid-1950s the British mathematician Sydney Chapman calculated the properties of the gas in the Sun's corona that must extend way out into space. At about the same time, the German scientist Ludwig Biermann observed that a comet's tail always points away from the Sun prompting him to postulate that this happens because the Sun emits a steady stream of particles that push the comet tail away. Parker realized that the heat flowing from the Sun in Chapman's model and the comet tail blowing away from the Sun in Biermann's theory had to be the result of the same phenomenon and that Chapman was right near the Sun and Biermann was right far from the Sun.

Initially the opposition to Parker's theory on the solar wind was strong and the paper he submitted to the Astrophysical Journal in 1958 was rejected by two reviewers. It was in the 1960s only that his theory was fully confirmed through direct satellite observations of the solar wind. This made Eugene Parker the leading authority on the solar wind and the effects of magnetic fields in the heliosphere. His work has greatly increased understanding of the solar corona, the solar wind, the magnetic fields of both Earth and Sun, and their complex electromagnetic interactions. He has seen spacecraft go out into the heliosphere and confirm the theoretical models he had developed before the space age, when observations of comet tails provided nearly all the data available.