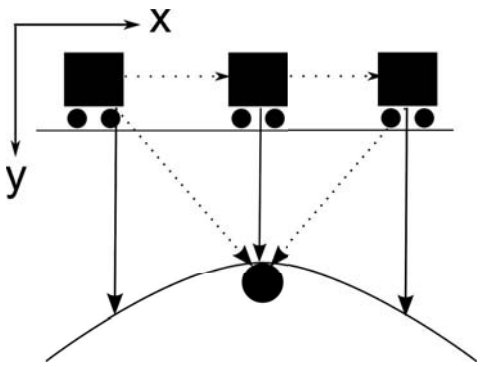

Figure 1.1: Schematic visualization of the data collection process.

et al. [2007]), or to assist archaeologist in finding ancient buildings (see, e. g., Daniels [2004] and Jol [2009]). For this thesis, I am defining information as humanly buried linear objects, e. g., pipes and cables, while all other object types are regarded as being clutter. Such clutter definition includes especially objects such as cans and land mines, as those show similar visual evidences compared to linear objects, though they are not of importance within the scope of my thesis. I am dealing with the *buried pipe localization problem* only.

Subsoil analysis by means of GPR data has become increasingly important in the last years, for its inexpensive capability to infer structures in subsoil. Having exact knowledge of the underground actually decreases maintenance costs on reconstruction works and accelerates the overall process for both, companies and engineers. The reason is simply stated by considering the usage of excavators, which may be used for a longer period of time before switching to manual soil digging in the very last moment for preventing damage to subsoil objects. Radar measurements are usually visualized as images, so-called radargrams, which contain geometric shapes of a certain kind to be identified. The most important geometric structure within the scope of my work to be identified is the hyperbola.

Figure 1.1 visualizes the data collection process. Therein, a measurement vehicle moves along the horizontal x direction, while at fixed intervals a radar wave is emitted in subsoil (y-axis). Such a radar wave propagates spherically in subsoil, while for two-dimensional visualization purposes, one can consider the propagation being a half-circle only. Whenever the radar wave hits an object, e. g., the pipe in the figure while the pathway of the wave is denoted by the dashed line, some part of it is reflected, while another proportion continues its propagation (not shown). The reflected energy (reflection intensities at certain 'pixel positions' within the radargram image) along with its characteristic pattern correlates with the type of the buried object. This results in higher reflection intensities for those objects which are dissimilar from their surrounding medium, e. g., a solid pipe compared to surrounding subsoil structures. The reflected energy is recorded and displayed at its corresponding wave travel time index (a depth index), denoted as the vertical solid line in the figure. Such a measurement at one position is called an *A-Scan*. Stacking multiple of these measurements for individual positions one after the other results in a radargram image, also being called a *B-Scan*. Three of those measurements are visualized in the figure. In this thesis, I will be using the term *lane* to refer to a single measurement drive, being the consecutive measurement of individual A-Scans at fixed intervals.

I will show throughout my thesis, and especially in chapter 4, that the interpretation of GPR data is a non-trivial task, also for the following reasons: (a) one needs to deal with previously unknown, unstructured subsoil structures and heterogeneous soil in general, which obscures the hyperbolic reflection patterns, and (b) one usually needs to apply site-dependent preprocessing techniques, to augment hyperbolic structures and to suppress unwanted artifacts. These obstacles render the easily phrased task of 'detecting hyperbolic reflection patterns in image data' a highly non-trivial task. I am aiming at assisting the hyperbola detection task by means of (semi-)automated detection models.

The data measurement process is usually performed by measuring multiple lanes in a parallel manner, resulting in multiple radargram images for a single *measurement site*. From an analysis of these radargram images, one is able to derive, e. g., supply maps depicting the locations of buried pipe structures, which may readily be used by on-site engineers when reconstruction works take place. My thesis presents reliable techniques to infer the location of a buried pipe on a radargram image alone, by presenting techniques for both, sparse (chapter 6) and dense (chapter 7) data analysis. Those techniques correspond to both, supervised and unsupervised Machine Learning techniques.

The analysis of GPR data, as an example setting for object detection in images, has been performed within the project 'Adaptive contactless Ground Pentrating Radar' - AcoGPR, lasting from October 2011 till September 2013. An overview of the problems and aspects tackled is given by Seyfried et al. [2012].

A short note on terminology

My thesis is approaching the GPR data analysis problem from the Machine Learning perspective. From the application setting, it was required to review and compare literature from this interdisciplinary domain. This results in difficulties when using, citing and embedding the terminology as used in related work.

As an example, LeCun et al. [1998] use the term 'energy' to denote a loss function (in Machine Learning terminology) to be minimized. However, from the 'radar perspective', the term 'energy' corresponds to a well-defined measure of the radiated electromagnetic radiation or of the signal received by the antenna, respectively. The radiated 'power' is directly connected to the electric and magnetic fields of the electromagnetic radiation, whereas 'energy' integrates 'power' over time. To describe the propagation of electromagnetic waves in the investigated structures (pavement, soil, underground supply lines, etc.), the natural measure of 'received signal strength' would be 'power'. In the context of pulse radar, however, it is more appropriate to use the 'energy' of the radar signal as measure of 'signal strength', because this more closely reflects the actual process of the sampling and analog-to-digital conversion of the signal.

In my thesis, I am not concerned with physical details of the radar system. For all aspects being specific to an underlying 'antenna system' which is used to measure the radar data, I am referring to Daniels [2004] and Jol [2009] who both have a much more comprehensive overview of the details of individual radar systems.

Instead, my thesis focuses on the analysis of GPR data from the Machine Learning point of view only: I will be treating the data as being a priori given. I am focusing on their interpretation and analysis by means of (semi-)automated Machine Learning techniques only. This has consequences in the precise usage of terminology when talking about radar-specific

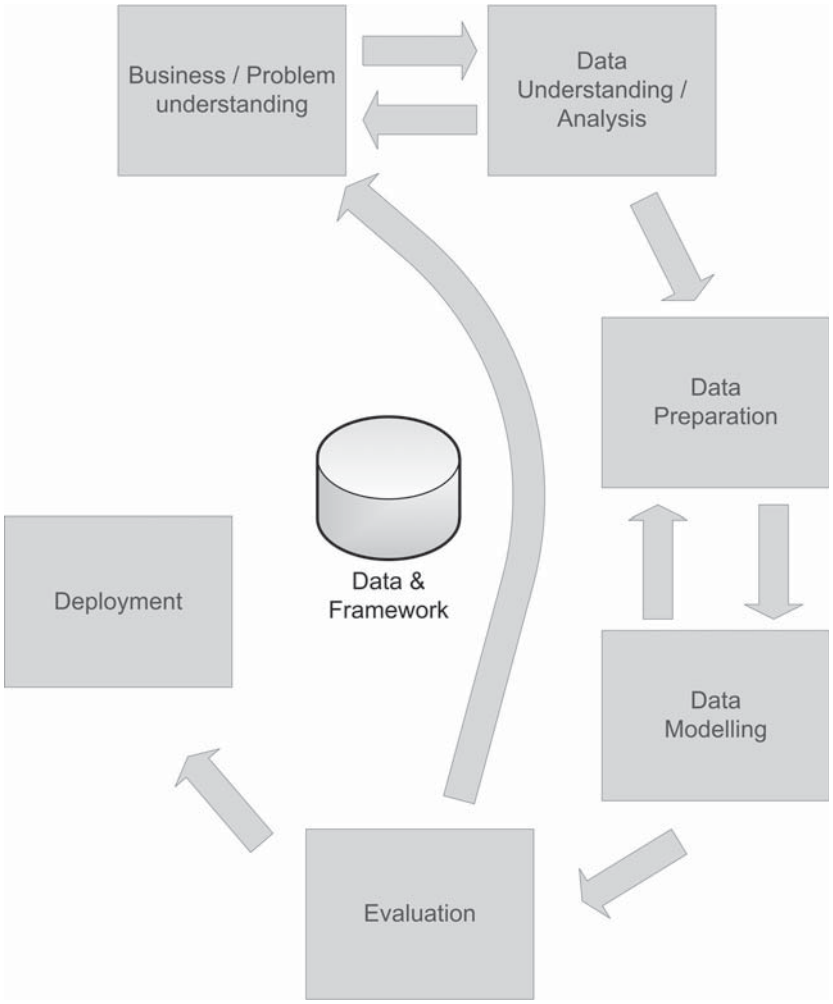

Figure 1.2: Adapted CRISP-DM for my analysis of CERD - Complex Engineering Raw Data.

aspects of the overall analysis pipeline. As an example, subsection 4.2.2 on page 70 discusses strategies to preprocess radar data to enhance the visibility of subsoil structures: This is not only affected by the subsoil structure, but is also affected by the physical fact that the antenna system has an influence on the quality of the measurement data. To this end, I would like to thank my reviewers for outlining those parts in this thesis which were less precise. I reworked the corresponding sections for the final print.

1.2 Thesis Overview

I will present the outline of my thesis both, within the context of the CRISP-DM (Cross Industry Standard Process for Data Mining) process model for analysing data in general, as well as sequentially, by stating their content, year of development, and their prior publication status. Figure 1.2 shows the CRISP-DM process model for the context of this thesis, which I adapted and extended in order to suite the following presentation.

- **Business / Problem Unterstanding**

  This chapter (chapter 1) has informally introduced the GPR analysis problem, whereas chapter 2 presents a formal definition of the problem.

- **Data Understanding / Analysis**

  A thorough data analysis is conducted in chapter 4. Therein, I will give a brief overview of the phenomena and artifacts being present in my current data, which have influenced my decisions when building data analysis models in the second part of my thesis.

- **Data Preparation**

  Chapter 5 presents a semi-manual data labeling approach, which was required due to the characteristics of my data at hand. The individual, model-specific data preparation for learning and applying Machine Learning models and methods is discussed in the individual chapters 6, 7 and 8.

- **Data Modeling**

  I will be presenting data analysis models and methods in chapters 6, 7 and 8, whereas chapter 9 presents a follow-up technique, which can be applied, once object locations are found.

- **Evaluation**

  While individual aspects will be evaluated in the corresponding individual chapters, I will be performing a large-scale analysis in chapter 8.

- **Deployment**

  Individual software prototypes are deployed in individual use cases, which are described in section 3.8 on page 56. Chapter 9 presents an approach for a follow-up data analysis, which may be applied once object locations are known.

- **Data & Framework**

  Chapter 3 presents my conceptual framework along with its technical implementation, which has been used for realizing and implementing all methods presented in my thesis.

Let me note that the analysis and visual presentation, as shown in the individual chapters of my thesis, is primarily based on real world in-house data from a project partner. Without loss of generality, all my methods and techniques allow for an application in comparable application domains as well.

In sequential order, my chapters are developed and relate to each other as follows:

- **Chapter 2** is newly developed for my thesis and formalizes the data analysis problem.

- **Chapter 3** shows my efforts within the last seven years for developing a software framework suitable for running large scale Machine Learning experiments. The core component, the Bootstrap library, was released in autumn 2012 as Open Source software and available at Busche [2013].

- **Chapter 4** is newly developed for my thesis and presents a thorough data analysis for my data at hand.

- **Chapter 5** has been largely extended for presentation in this thesis, has been initially published in Busche et al. [2013a] and was presented at KDML 2013. The study was performed in four months in winter 2012 / 2013. It represents my approach for obtaining high-quality data labels for my data.

- **Chapter 6** has been extended for presentation in my thesis, was published in Busche et al. [2012a] and performed in winter 2011 / 2012. It presents analysis methods for hyperbola detection from sparse data.

- **Chapter 7** corresponds to the work as presented in Busche et al. [2013b]. It corresponds to efforts spent in spring / summer 2013 and presents an unsupervised approach for identifying hyperbolic reflection patterns in radargram images.

- **Chapter 8** presents a large-scale evaluation and is split up into two conceptual parts. The first part of evaluating the manual annotations as obtained in chapter 5 is accepted at KDML 2013 and published in Busche et al. [2013a]. The second part is submitted along with the achievements of chapter 7 and published as Busche et al. [2013b].

- **Chapter 9** has been published in Busche et al. [2012b] within the context of an image reconstruction framework and has been extended for presentation in this thesis. It corresponds to efforts spent in summer 2012 and presents an approach for estimating the exact curvature of reflection hyperbolas based on dense image data.

Chapter 10 concludes my thesis according to my claimed contributions in the next section and states five important aspects which directly allow for a continuation of my studies.

Please note that the manual annotations as obtained in chapter 5 have not been available when developing the contents of chapters 6, 7 and 9. Therefore, the evaluations presented therein only focus on simulated data, or show a qualitative evaluation on real world data.

1.3 Contributions

My thesis makes contributions to the current state of the art in almost all following chapters. Though I will be using the term ‘we’ for presenting the content in chapters 2 to 9, you should keep in mind that all aspects, insights and findings were - to the best of my knowledge - developed all by myself, unless otherwise explicitly noted or cited. Specifically, I will be stating individual contributions explicitly at the beginning of each chapter. From a broad perspective, my thesis contributes the following:

- **Creation of ground truth data for analysing GPR data**

  I will be gaining a high quality ground truth data for my current GPR data in chapter 5 and show their suitability in chapter 8.

- **Structured Analysis of intuitive GPR data analysis approaches**

  I will be investigating the hyperbola estimation problem in various aspects in both one-dimensional (chapter 6) and two-dimensional (chapters 7, 8 and 9) scenarios. The hyperbola estimation techniques are considered on both, sparse and dense data. I am investigating their robustness, e. g., to the data sampling process, when influenced by either or both, noise and jitter. All presented methods are based on a thorough data analysis as presented in chapter 4. Methods and techniques are presented from the perspective of the easily phrased task of finding hyperbolic reflection patterns in images. For approaching a solution, I am analysing intuitive approaches to this task.

- **Conceptual framework design for raw data analysis**

  I am presenting a flexible framework for conducting Machine Learning experiments in chapter 3. It allows for seamlessly considering many-to-many relationships between all, data, labels, and metadata.

1.4 Published Work

Some parts of my thesis are already published, e. g., in conference proceedings, as follows:

- Busche et al. [2009] present prototypes and end-user software, which have influenced the development of my conceptual framework in chapter 3.
- Busche et al. [2012a] present a sparse analysis approach for hyperbola detection and was revised for presentation in my thesis in chapter 6.
- Busche et al. [2012b] present a general framework for radargram image reconstruction and has been extended for my thesis in chapter 9.
- Busche [2013] is an online documentation of my Bootstrap library, as documented in section 3.6 on page 45.
- Busche et al. [2013b] compares both, the Hough Transform and the Kirchhoff Migration, and is presented in detail in chapter 7.
- Busche et al. [2013a] present my approach for obtaining an accurate radargram labeling. It is presented with advanced statistics and deeper general analysis in chapter 5.

Within the last five years of research in the lab, I furthermore coauthored the following papers which did not fit as a core contribution in my thesis. I need to note that from all those papers, synergies arose while working on individual topics of my thesis.

- Ruth Janning, *Andre Busche*, Tomas Horvath, Lars Schmidt-Thieme (2013): Buried Pipe Localization Using an Iterative Geometric Clustering on GPR Data, Artificial Intelligence Review, Springer.

- Ruth Janning, Tomas Horvath, *Andre Busche*, Lars Schmidt-Thieme (2012): Pipe Localization by Apex Detection, in Proceedings of the IET international conference on radar systems (Radar 2012), Glasgow, Scotland.
- Ruth Janning, Tomas Horvath, *Andre Busche*, Lars Schmidt-Thieme (2012): GamRec: a Clustering Method Using Geometrical Background Knowledge for GPR Data Preprocessing, in Artificial Intelligence Applications and Innovations (IFIP Advances in Information and Communication Technology 381), Springer, Heidelberg, Halkidiki, Greece, pp. 347–356.
- Artus Krohn-Grimberghe, *Andre Busche*, Alexandros Nanopoulos, Lars Schmidt-Thieme (2011): Active Learning for Technology Enhanced Learning, to appear in Proceedings of the European Conference on Technology Enhanced Learning (EC-TEL) 2011, LNCS, Springer.
- **Andre Busche**, Artus Krohn-Grimberghe, Lars Schmidt-Thieme (2010): Mining Music Playlogs for Next Song Recommendations, in Workshop Proceedings of Knowledge Discovery, Data Mining, Maschinelles Lernen 2010 (KDML 2010).
- Nguyen Thai-Nghe, *Andre Busche*, Lars Schmidt-Thieme (2009): Improving Academic Performance Prediction by Dealing with Class Imbalance, in Proceedings of the 9th IEEE International Conference on Intelligent Systems Design and Applications (ISDA 2009), IEEE Computer Society, pp. 878–883.

CHAPTER 2

Problem Definition

This chapter formalizes the GPR data analysis problem, which was introduced in the last chapter by specializing it as a subtask of the more general radar data analysis problem. The formalization presented here is applicable for any analysis of two-dimensional GPR data, e. g., for land mine detection, or leakage detection, though being specialized in section 2.5 for the actual GPR subtask tackled in this thesis: the pipe recognition problem.

The outline of this chapter is visualized in Figure 2.1 and comprises three states, along with four transitions to be discussed in the following sections:

- The *Real World Situation* corresponds to the final goal of the GPR data analysis process, as discussed in this thesis. The real world situation is derived by interpreting the output of Machine Learning techniques to be developed.

- The *Measured Situation* is gained by performing radar measurements on targets present in the real world. Examples include raw radar data measurements, e. g., from air traffic control or - for our case - Ground Penetrating Radar data.

- The *Machine Learning Interpretation* corresponds to the data interpretation of the measured situation and is dependent on a certain application scenario, e. g., the deduction of airplane locations out of air traffic control data, or the location of buried objects out of GPR data. For a robust and reliable interpretation in varying situations, it probably has to be complemented by probabilistic models, capable of establishing a hypothesis based on prior knowledge (a mapping from input features to labels).

This chapter first discusses the transition from real world situations to measured situations in section 2.1. The requirements for applying Machine Learning techniques to any problem are discussed in section 2.2. Sections 2.3 and 2.4 subsequently formalize the transition from

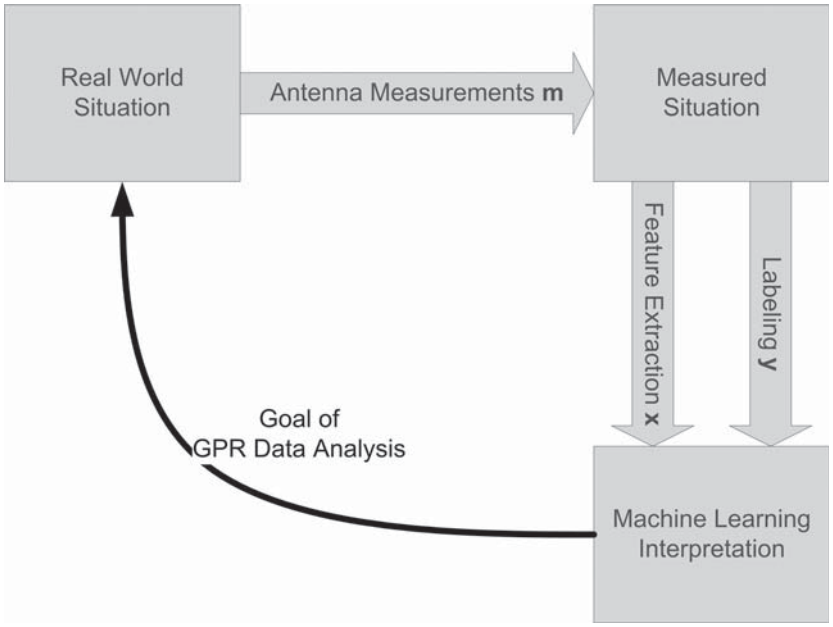

Figure 2.1: Visualization of the abstract (Ground Penetrating) Radar data analysis process: based on a real world situation, an antenna is used to obtain a measured situation. Such a situation is converted into a suitable format / interpretation for applying Machine Learning techniques which are able to output a hypothesis on the actual real world situation.

the continuously measured situation to a suitable discrete Machine Learning interpretation. The actual problem tackled in this thesis is defined in subsection 2.5.

The contribution of this chapter is the following:

- we define a formalism for the analysis of GPR data for its later use throughout this thesis.

2.1 Radar Data Analysis

The most general definition of Engineering Data Analysis in our sense is the detection of certain aspects within Complex Engineering Data which is obtained while measuring certain real world objects. While Complex Engineering Data in general may correspond to nearly all kinds of sensor data measured, e. g., over time and / or space, in various application domains, our special focus in this thesis will be the analysis of radar data sampled over time at different locations, though further example domains will be presented to show the generality of some of our proposed techniques and methods.

The general process of obtaining radar measurements is to emit a radar wave using an 'emitting antenna' and to record the load (voltage) being present at a 'receiving antenna' by sampling over time. Example domains for the application of radar technology include air traffic control and, as tackled in this thesis, subsoil imaging. One may instantly visualize such a measurement as a time-vs-intensity plot, or after performing a Fourier Transform, as an

amplitude-vs-frequency diagram. Alternatively, one may also directly sample the amplitude in the frequency domain at certain measurement locations. Our concrete problem at hand tackled in this thesis is the analysis of GPR. We refer to the term 'Ground Penetrating Radar' as being the application of radar technology for analysing subsoil structures. As radar waves in general propagate in the whole three-dimensional space, we will be using the term GPR to denote those radar technologies which are constrained to the task, e. g., by adding an isolation to the overall system to reduce the propagation of the wave to free air.

The quality of GPR measurements in general is influenced by many factors, including (a) the kind and design of the antennas used, (b) their operating frequencies and emitting energy, (c) their relative spatial location, (d) the first ground layer, e. g., concrete or asphalt, (e) subsoil types, e. g., sand, loam and water, (f) relative differences in physical properties, e. g., between (d) and (e), as well as (g) the size of the buried object.

This thesis is not concerned with details about physical aspects influencing radar measurements in general, but with their semi-automated analysis. Anyhow, some physical aspects need to be shown to outline and augment difficulties, and to serve as a motivation on why certain problems have been tackled in such a way. To this (physical) end, we will merely focus on intuitive explanations of effects rather than on their exact formalization. As such, we will consider the measured radar data 'as is', without any possibility to improve their quality by means of, e. g., exchanging physical components while data measurements take place.

The general aim of radar data analysis is the inference of real world objects (e. g., their location, shape, etc.) based on the measured radar data. Depending on the antenna used, each single radar measurement (in the frequency domain) is usually present in the domain of complex values $\mathbb{C}$. According to Daniels [2004], p. 251, the resultant radar measurements are a combination of each individual component which is used when building the overall antenna system. As we are aiming at analysing the radar measurement data by means of Machine Learning methods, we will not further elaborate on individual components as well as their individual influences to the measurement process. We rather refer the interested reader to Daniels [2004] who has a much more comprehensive overview of antennas as we are able to provide here.

If we denote the set of (arbitrary) individual real world objects by $\mathcal{O}$, any $o \in \mathcal{O}$ corresponds to a single real world object. Examples for $o \in \mathcal{O}$ include, e. g., a single pipe of certain length, direction and location on earth, or a land mine of certain (relative) depth, diameter and orientation. Now, let the power set of $\mathcal{O}$, $P(\mathcal{O})$, denote any combination of real world objects, whereas an $o \in P(\mathcal{O})$ denotes a certain, fixed combination of real world objects. We may define m to be an antenna which may be used to measure an arbitrary structure $o \in P(\mathcal{O})$ as follows:

Definition 2.1 (Antenna)**.** *Let m denote an arbitrary antenna function alike those defined in Daniels [2004], chapter 2 and section 7.2 therein. The antenna then maps any real world object structure $o \in P(\mathcal{O})$ into a partial observation of a multidimensional measurement space $\mathcal{M}^n$, for $n \gg 0$. Each spatial location $v \in \mathcal{M}^n$ is mapped to a single, complex observation $c \in \mathbb{C}$ as follows:*

$$m : P(\mathcal{O}) \rightarrow Maps(\mathcal{M}^n \rightarrow \mathbb{C}) \tag{2.1}$$