MASTER'S THESIS 2022

# Concretizing CRISP-DM for Data-Driven Financial Decision Support Tools

Simon Grimheden, Joel Järlesäter


ISSN 1650-2884
LU-CS-EX: 2022-19

DEPARTMENT OF COMPUTER SCIENCE
LTH | LUND UNIVERSITY

EXAMENSARBETE
Datavetenskap

LU-CS-EX: 2022-19

# Concretizing CRISP-DM for Data-Driven Financial Decision Support Tools

## Konkretisering av CRISP-DM för datadrivna hjälpmedel för finansiell beslutsfattning

**Simon Grimheden, Joel Järlesäter**

# Concretizing CRISP-DM for Data-Driven Financial Decision Support Tools

Simon Grimheden
si4452gr-s@student.lu.se

Joel Järlesäter
jo3286ja-s@student.lu.se

April 25, 2022

Master's thesis work carried out at
the Department of Computer Science, Lund University.

Supervisors: Markus Borg, markus.borg@ri.se
Johan Crafoord-Larsen, crafoord@hetch.se

Examiner: Elizabeth Bjarnason, elizabeth.bjarnason@cs.lth.se

**Abstract**

To support development of applications utilizing Artificial Intelligence (AI) and/or Machine Learning (ML), so called data-driven applications, development process models such as CRISP-DM have been created. However, previous papers on the topic of CRISP-DM have concluded that the model lacks detailed method recommendations, hindering its use for developers without previous knowledge in the field. In this paper, we contribute to this research by creating a detailed CRISP-DM model, tailored for the domain of data-driven financial decision support tools by conducting a literature review and one case study consisting of two unit of analysis. To achieve this, we interviewed companies that would be potential stakeholders in the domain, as well as potential developers of an application within the domain. Our research found three main challenges when developing data-driven financial decision support tools, namely difficulty of defining main purpose, large number of interfaces, and uncertainty of data, and resulted in a detailed version of the CRISP-DM model. Our suggested concretization of CRISP-DM features a more holistic approach to evaluation, as well as concrete recommended activities for each phase of the original CRISP-DM model.

**Keywords**: CRISP-DM, Detailed CRISP-DM, Prototyping, Requirements Engineering, Data-driven applications, FX risk exposure

# Acknowledgements

We would like to thank our supervisor at Lund University, Markus Borg, for his unwavering support throughout this thesis project. Apart from thesis structuring and other formalities, Markus has provided us with insight in the academics of data-driven applications, requirements engineering, and much more, allowing us to better explore areas of interest.

We would also like to thank our examiner Dr. Elizabeth Bjarnason, without whom this thesis would be much more difficult to finalize. Dr. Bjarnason helped us iterate our research questions multiple times prior to project initiation in order to reach the best academical conclusions possible.

Last, but absolutely not least, we would like to thank all of the 12 interviewees whom have spent a lot of time participating in our research. This work was conducted in the context of the AIQ Meta-testbed, a project funded by Kompetensfonden at Campus Helsingborg, Lund University, Sweden.

# Contents

# Chapter 1
# Introduction

Artificial Intelligence (AI) and Machine Learning (ML) are topics which have been researched for many decades, but only in recent times have computing power become sufficient to make these technologies available to the general public [35]. In this thesis, we call applications utilizing AI and ML data-driven applications. Contrary to traditional information systems, data-driven applications are developed mainly on the basis of existing data, simulations, or recorded experience and not exclusively on the basis of rule-based knowledge [35], making them useful in assessing complex data and scenarios.

The majority of projects within the area of data-driven applications, run by less practically experienced developers, are usually done in an exploratory and unstructured way [57]. To support the development of these types of applications, development process models and frameworks such as Cross Industry Standard Process for Data Mining (CRISP-DM) have been created [35]. The CRISP-DM development process model is today the most frequently used [54]. However, detailed method recommendations for CRISP-DM are lacking [57, 59]. In this thesis, we will concretize CRISP-DM in the field of data-driven financial decision support applications by conducting a literature review and an improving case study [53]. The goal of this case study is to firstly, investigate the problem domain of financial decision support tools by conducting interviews with industry stakeholders, and secondly, detail CRISP-DM with regards to the challenges identified in the problem domain by interviewing potential developers and analyzing the activities they suggest for development in the domain.

**The Case Study** In this case study, will investigate two units of analysis, the first being challenges in development of an Foreign Exchange (FX) risk exposure application. FX risk exposure stems from companies and organizations working with multiple currencies as they may experience struggles to deal with wildly fluctuating exchange rates. Controllers, economists and business administrators constantly need to work on strategies to avoid major financial losses due to these fluctuations. Hence, currency rates and FX must be seen as a business risk in companies working with multiple currencies. Problems arise when businesses try to quantify their FX risk exposure. This is caused by complex relations between differ-

ent currencies, types of investments and strategies etc. Due to companies individual goals, strategies, risk aversion, as well as the vast number of accounts, invoices, investments, and receivables present in many larger companies, there currently does not exist a widely adopted tool for evaluating risk exposure. Thus, companies often approximate their exposure based on experience rather than data.

## 1.1 CRISP-DM

CRISP-DM is a process model describing the life cycle of a data mining project. The model provides a workflow based on 6 phases, with related activities to be completed during each phase. The models phases are to be utilized in a cycle, as an iterative approach is emphasized for successful implementation. Though the phases of the process cycle follow a distinct order, the actual order of execution varies based on the problem at hand, and repetition of steps might be necessary as the work progresses. The 6 phases of the CRISP-DM model are stated as follows [20]:

1. Business Understanding
2. Data Understanding
3. Data Preparation
4. Modeling
5. Evaluation
6. Deployment

**Business Understanding** The business understanding phase is the first step of the CRISP-DM model, where the purpose is to understand the requirements of the data model from a business perspective. When the purpose of the model is clear, the goal is then to transfer this problem description into a problem definition suited for data mining and make an initial plan for project completion [20].

**Data Understanding** The data understanding phase is where you start gathering available data necessary for the project, and familiarize yourself with it. This phase should also be utilized to analyze possible data quality problems, detect data subsets and theorize possible hidden information [20].

**Data Preparation** The purpose of the data preparation phase is to construct a dataset for use in the final model. This often requires activities such as selection, transformation, and cleaning of the raw dataset [20].

**Modeling** The modeling phase is where the actual ML-Modeling takes place. This involves activities such as selection of modeling technique and parameter calibration. As there often are multiple different modeling techniques that can be used for a single data mining problem, its important to consider the specific use case and choose a suitable algorithm. During this phase it can be necessary to go back to the data preparation phase, as some suitable techniques might require specific data types or attributes. During the modeling phase the initial plan for testing should also be constructed [20].

**Evaluation** The evaluation phase is where the built model is evaluated according to the initial problem formulated in the business understanding phase. It is important to ensure the quality of the model from both a data analysis perspective, as well as from a business perspective. If the model does not fulfill its purpose adequately, you should consider why it does not fulfill the goals, and then go back to a previous phase depending on the origin of the fault. When evaluation of the model is satisfactory, the creation phase of the model is complete [20].

**Deployment** The deployment phase is where the use of the model is decided. This involves how the model will be presented, such as on a website or in an application, or even as a report of simple key results. Depending on the project, this phase might not be part of the creation cycle, for example if a customer demands just the model and will perform their own implementation [20].

## 1.2 Academic Contribution

By answering the stated research questions in Section 1.3, we aim to increase the understanding of how a refined version of CRISP-DM can be used during prototyping and development of data-driven applications within a problem domain that has received little academic attention from this perspective. With our exploration of the problem domain we also aim to increase the understanding of the challenges developers face during development in the field of financial decision support tools.

## 1.3 Research Questions

Our aim is to answer the following questions:

**RQ1** What characterizes the problem domain for data-driven financial decision support applications?

**RQ2** How can CRISP-DM be applied when prototyping a data driven FX risk exposure application?

RQ2 will mainly be answered by focusing on the perspectives stated below.

RQ2:A How can CRISP-DM be detailed with methods from requirements engineering?

RQ2:B What would be the benefits and challenges of using the detailed CRISP-DM?

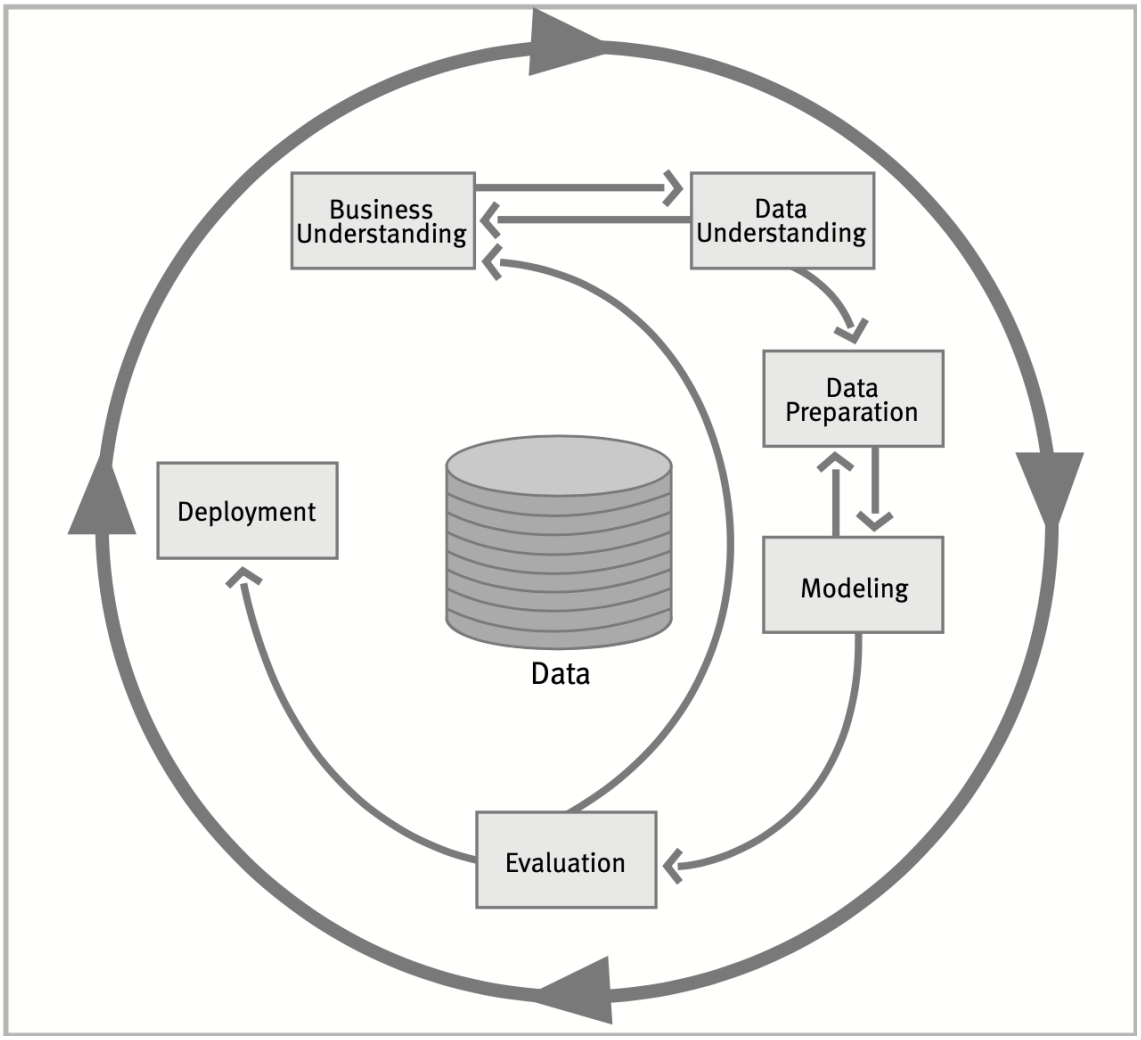



**Figure 1.1:** An overview of the CRISP-DM model

RQ2:C How do aspects related to data, e.g. quality, affect the detailed CRISP-DM?

RQ2:D How is the practice of prototyping affected when developing data-driven applications?

RQ2:E How is the practice of prototyping affected by the problem domain?

## 1.4 Related Work

In this section we will present the related works on the topic of this master thesis. It includes two other studies detailing CRISP-DM, one with a general purpose and one with a focus on quality management, one article on requirement engineering in data driven project, and two articles on common challenges when adopting data-driven applications in the financial industry. In general, we will make use of the articles regarding CRISP-DM and requirement engineering when answering RQ2, and the articles on data-driven applications when answering RQ1.

In 2021, Shailesh et al. [59] published an article in the journal *"Frontier in Artificial In-*

*telligence"*, stressing the need of extending the CRISP-DM process model. Their work resulted in what they call the Generalized Cross-Industry Standard Process for Data Science (GCRISP-DS). The new process model, or framework as they call it, is designed in order to allow dynamic interaction between different development phases. This to address data- and model-related issues for achieving robustness. The GCRISP-DS allows the users of the framework to iteratively move between all of its phases. It also extends the CRISP-DM phases to include multiple process methods as follows:

**Business Understanding -** Defining business objectives and project planning

**Data Understanding -** Data acquisition

**Data preparation -** Data processing, exploration, and descriptive analysis

**Modeling -** ML model implementation, feature engineering, feature selection, and ensemble learning.

**Evaluation -** Model accuracy, interpretability, transparency, and selection.

**Deployment -** No additions

More researchers have in recent time undertaken the challenge of concretizing the CRISP-DM process model. In 2018, Schräfer et al. presented in their report "Synthesizing CRISP-DM and Quality Management: A Data Mining Approach for Production Processes" the lack of detailed method recommendations for CRISP-DM [57]. They developed the Quality Management-CRISP-DM (QM-CRISP-DM) by concretizing CRISP-DM from a Quality Management perspective. The QM-CRISP-DM uses the six sigma framework's DMAIC phases (Define, Measure, Analyze, Improve and Control) as a basis. As the result, they proposed using a set of Quality Management tools for each phase in CRISP-DM process model.

Regarding requirements engineering for machine learning, Borg and Vogelsang [65] published an article on the topic in 2019. In the report they analyze data scientists' approach on elicitation, specification, and assurance of requirements and expectations based on four conducted interviews. As results to their work they present five crucial areas of requirements engineering of data-driven applications, namely quantitative targets, explainability, freedom of discrimination, legal and regulatory requirements and data requirements. With these areas of interest, they derive recommendations for aspects to take into consideration when conducting the requirements engineering tasks of elicitation, analysis, specification, and verification and validation in the development process of data-driven applications. These aspects will be utilized when making detailed recommendations for CRISP-DM in RQ2 of this master thesis.

In 2019, Dixon and Halperin [24] analyzed the difficulty of applying machine learning in the financial industry. In their research, they identified four major challenges. The first challenge was the "statistical illiteracy" of some people in the financial industry and their inability to co-integrate their discipline with financial time series analysis, financial modeling, and dynamic programming. Secondly, they point to the financial industry's disbelief that predictive signals may be drawn from non-stationary data or data sets which may not fulfil the criteria of traditional financial theory. Third, they present that models with relatively high accuracy tend to only deliver mediocre profits, and lastly they present some people in the

industry's incapability of accepting that machine learning can discover new and meaningful patterns.

In a conference paper published in June 2021, Nilsson Tengstrand et al. [49] analyzed the challenges of adopting The Scaled Agile Framework (SAFe), a framework for scaling agile methods in large organizations, in the banking industry. The resulting challenges were assigned to seven high level themes, namely Management and organization, Education and training, Culture and mindset, Requirements engineering, Quality assurance, Systems architecture, and Banking specifics. Some of the challenges we have found to be interesting in the context of this master thesis are "Difficult to create a shared vision and align the entire organization around common goals", "Difficult to break down requirements", and "Legacy systems are not easily adopted to agile ways of working".

## 1.5 Distribution of Work

We have worked in close tandem throughout most parts of this project. Both in regards of research and writing. In terms of writing, Joel Järlesäter had more responsibility regarding Chapter 3 *Research Method*, while Simon Grimheden contributed more to the writing of Chapter 7 *Discussion*.

Both Järlesäter and Grimheden participated when conducting all interviews and transcribing them, as well as when analyzing their content. Both authors were also involved in the process of deriving the research results and formulating the conclusions for each research question.

# Chapter 2
# Background

## 2.1 Requirement Engineering

Requirement engineering is a systematic approach of determining what a products shall achieve by elicitation, specification, prioritization and validation of system requirement. In general, requirements come from the users and other types the *stakeholders* of the system. However, requirement elicitation can be very difficult for many reasons. For example, stakeholders may experience difficulty expressing their needs, have conflicting demands and even demanding solutions which does not meet their real need [41].

### 2.1.1 Problem Domain

The problem domain is a term referring to all information describing the problems and constraints of a solution. This includes the goals that the stakeholders of the system wished to achieve, a description of the context within which the problem exist, and rules which defines the essential functions and constraints of the system.

### 2.1.2 Elicitation

Elicitation within the field of software engineering is the process of identifying and formulating the requirements of a system. Elicitation is an iterative process, without clear cut steps for completion. This stems from the fact that many stakeholders have problems expressing their true needs and requirements for a system, and often have not formulated its true purpose beforehand. Stakeholders also often have conflicting views, and one stakeholders requirements might not be possible to coexist with another's [41].

When eliciting requirements there are multiple different techniques that can be utilized. Advantages of elicitation techniques are very situational, and it is therefore important to

consider the situation at hand before deciding what technique to use. Common elicitation techniques include stakeholder analysis, interviews, and prototyping [41].

### 2.1.3 Prototyping

Prototyping is an agile requirement engineering practice and widely regarded as a core mean of expressing and exploring designs for interactive computer artifacts [13, 36]. In general, it is common practice to build prototypes in order to represent different states of an gradually advancing design, and to explore different options [36]. There exists a large number of different types of prototypes due to each type being suited for exploring certain aspects [62]. For example, creating a beverage container prototype utilizing recyclable materials most certainly involve the use and processing of physical materials in a physical environment, while creating a prototype for a new mobile applications may only require digital means. A prototype can be everything from a simple sketch to an incomplete version of the application, also known as a minimum viable product (MVP) [13].

Startups within the software industry develop new and innovative products and services under uncertainty and with restricted resources [51, 30]. One major success factor for startups is to validate the business ideas feasibility in the market early on [15]. A prototype is an early model, sample or release, which imitates one or more aspects or features of the final product. It can therefore be used to communicate, explore, and evaluate potential solutions [45]. Thus, allowing cost effective testing with real, potential, users [48].

According to requirement engineering practices, prototypes can be utilized for elicitation, testing and validation of requirements [18]. By utilizing prototypes two different levels of requirements may be obtained, namely product-level requirements and design-level requirements [41]. Experimenting with product-level requirements may bring an understanding whether required functionalities are feasible and useful while prototypes focusing on design-level requirements shall represent exactly how the user interface of a given functionality will be implemented.

**Prototyping Aspects** Current use of prototyping within agile requirements engineering is inefficient in obtaining feedback on the intended or correct *aspect(s)* [13]. In 2021, Bjarnason et al. presented the Prototyping Aspects Model (PAM) and concluded it may be used to support agile teams in reflecting on their prototyping practices. By utilizing PAM, teams would be able to make conscious choices regarding how to explore the solution space in an effective way, considering their goals and resources. In their report, the authors states practitioners should consider the following aspect:

**Purpose of Prototype -** The main object of this aspect is to answer the question *Why prototype?*. Prototyping can achieve multiple purposes, and the object of it usually varies throughout the life-cycle of the project. In their case study presented in the paper *A Model of Software Prototyping based on a Systematic Map*, Bjarnason et al. identified eight main purposes of prototyping, namely exploration, communication, incremental development, quality improvement, and validation & testing of business viability, market desirability, technical feasibility, and usability. Multiple purposes may though be satisfied simultaneously in a single prototype [13].

**Prototype Scope -** The main object of this aspect is to answer the question *What to prototype?*. The prototype scope, in this case, represent to what extent the prototype resembles

the final product in regards to the *breadth* and *depth* of the prototypes functionalities. The breadth of the prototype represents to what extent the prototypes functionalities covers the full product's functionalities while the depth represents how each and every functionality in the prototype covers the final functionalities utilities [13]. The scope of the prototype may also include other aspects, namely *visual appearance*, *interactivity*, and *data realism*. A prototype's visual appearance concerns aesthetic such as fronts, layout and elements in user interfaces. Interactivity is related to what extent a prototype imitates the final products behavior, and data realism to what extent the used data simulates normal and realistic use.

**Prototype Use** - The prototype use covers how it is used in order to achieve the purpose as well as in what environment it is presented and reviewed. Bjarnason et al. presents four main areas of prototype use, namely *stakeholder demonstration*, *scenario testing*, *free testing* and *internal use*, i.e., without any user presentation.

**Exploration strategy** - This aspect aims to answer the question *How to traverse the solution space over time?* [13]. Four strategies of traversing the solution space was initially stated by Tronvoll et al, and later renamed to increase clarity by Bjarnason et al. [62, 13], namely *point-based*, *parallel*, *optimization* and *flexible* exploration. The exploration strategy forms the basis of how resources are allocated, what instances to pursue and how uncertainties in the development are managed [13].

Point-based exploration refers to the idea of focusing on a single solution path, while parallel exploration refers to multiple solutions being explored simultaneously. In optimization exploration, solutions are judged by performance and only the most promising one is pursued. Lastly, flexible exploration refers to solution options being based on best-guesses, iterated, evaluated, and thereafter changed as required [13].

## 2.2 Foreign Exchange

The foreign exchange (FX) market is a global decentralized market of currency trading. This market sets the FX rates for every currency in the world. It handles all aspects of selling, buying and exchanging currencies at current or fixed prices. When considering the total trading volume, it is by far the world's largest market [52].

### 2.2.1 Foreign Exchange Risk

The variability of exchange rates introduces significant macroeconomic uncertainty affecting businesses operating in multinational open economies [22]. Due to the financial losses associated with failing to hedge such risks, currency exchange risk is one of the most researched risks facing companies all around the world [68]. Exchange rate fluctuations, particularly when combined with large time lags between the time of order and payment, greatly affect the cash flow and the business value of firms through their individual *exchange risk exposure* [22, 68]. To mitigate currency exchange risk, various *FX hedging strategies* may be used. While some firms desire full hedging to avoid any currency risk exposure, others implement hedging strategies in order to maximize excess returns rather than to simply reduce risks [21].

In order to asses individual corporations' currency risk exposure, multiple control variables have to be considered. According to Zhang [68] the following six should be taken into account:

**Foreign Sales Ratio** Liabilities denominated in foreign currencies are more likely in multinational enterprises. Hence, the likelihood of these firms being affected by foreign exchange fluctuations are high [40].

**Size** There exist different school of thought regarding how the size of companies affect their individual foreign risk exposure [68]. Gilson and Warner [31] argue that small-sized companies face a higher probability of financial distress compared to large-sized companies. Hence, they point out a positive correlation between company size and FX risk exposure. On the contrary, according to Nance er al. [47], there is a negative correlation between the size of an enterprise and its FX risk exposure since the larger the company size, the better it can manage risk exposure by hiring experts.

**Growth** High company growth leads to increased opportunities of investment, which in turn leads to an increased need of external financing. Investors generally require stability and thereby reduced volatility of future cash flow and high credit ratings [68].

**Solvency** Firms with a high debt ratio, i.e., low solvency, have a greater probability of facing financial distress [33]. Thus, there exist a negative correlation between solvency and FX risk exposure.

**Liquidity** Corporations' liquidity represents FX hedging as it may be used to buffer FX rate movements. Thus, reducing the cost of financial distress. Consequently, a negative correlation between corporations' liquidity and FX risk exposure exists [10].

**Profitability** The higher the profitability of foreign sales is, the higher the FX risk exposure of the company is [22]. I.e., it exists a positive correlation between foreign sales profitability and FX risk exposure.

### 2.2.2 Foreign Exchange Hedging Instruments

Companies that are exposed to FX risk can utilize multiple different financial instruments to insure against unwanted exchange risk. The most common instruments used on the Swedish market include the following [42].

**Foreign Currency Swap** A foreign currency swap is an agreement between two parties to exchange interest payments and principal of a loan in one currency with interest payments and principal of a loan in a different currency. One use of foreign currency swaps is to hedge against balance sheet exposure of debt in foreign currencies. As a foreign currency swap involves exchange of both interest rate and principal, the instrument can sometimes be used to borrow money in a foreign currency at a more favorable rate than would otherwise be possible through a direct loan [11]. Foreign currency swaps constitute for over 50% of the Swedish FX derivatives market [42].

**Foreign Exchange Forward Contracts** A forward exchange contract is a contract between two parties to exchange an amount of currency at a specified date for a rate set at the time of contract agreement. Forward exchange contracts are the second most common instrument for exchange rate hedging [42].

**Foreign Currency Exchange Options** Foreign currency exchange options are a contract that give the holder the right to at a given date exchange a predetermined amount of currency at a set rate. A European exchange option consists of three factors decided at the time of contract agreement, the exchange rate (Strike), the time of expiry, and the amount of currency involved. There are two types of foreign currency exchange options: European and American. European exchange options only allow the buyer to exercise this option at the date of expiry, while American exchange options allow the buyer to exchange this option at any time before and including the time of expiry. Exchange options can be formulated as giving the holder either the right to buy (call) or sell (put) the specified amount of currency [11].

## 2.3 Data Quality

In this master thesis, we will use the European statistical system's (ESS) definition of data quality [27], used by e.g. the Bank of England to define quality of financial data [50]. ESS defines six data quality dimensions, namely (1) Relevance, (2) Accuracy, (3) Timeliness and Punctuality, (4) Accessibility and Clarity, (5) Comparability, and (6) Coherence. In this section, we will present ESS' definition of these terms.

**Relevance** - *"Relevance is the degree to which statistics meet current and potential users' needs. It refers to whether all statistics that are needed are produced and the extent to which concepts used (definitions, classifications etc.) reflects user needs."*. According to Bank of England, this can also refer to the degree of which data meets the users' needs [50].

**Accuracy** - *"Accuracy in the general statistical sense denotes the closeness of computations or estimates to the exact or true values."*.

**Timeliness and Punctuality** - *"Timeliness of information reflects the length of time between the availability of data and the event or phenomenon they describe.*

*Punctuality refers to the time lag between the release date of data and the target date when it should have been delivered, for instance, with reference to dates announced in some official release calendar, laid down by Regulations or previously agreed among partners."*

**Accessibility and Clarity** - *"Accessibility refers to the physical conditions in which users can obtain data: where to go, how to order, delivery time, clear pricing policy, convenient marketing conditions (copyright, etc.), availability of micro or macro data, various formats (paper, files, CD-ROM, Internet...), etc.*

*Clarity refers to the data's information environment whether data are accompanied with appropriate metadata, illustrations such as graphs and maps, whether information on their quality also available (including limitation in use...) and the extend to which additional assistance is provided by the NSI."*

**Comparability** - *"Comparability aims at measuring the impact of differences in applied statistical concepts and measurement tools/procedures when statistics are compared between geographical areas,*

*non- geographical domains, or over time. We can say it is the extent to which differences between statistics are attributed to differences between the true values of the statistical characteristic."*

**Coherence** - *"Coherence of statistics is therefore their adequacy to be reliably combined in different ways and for various uses. It is, however, generally easier to show cases of incoherence than to prove coherence.*

*When originating from different sources, and in particular from statistical surveys of different nature and/or frequencies, statistics may not be completely coherent in the sense that they may be based on different approaches, classifications and methodological standards. Conveying neighbouring results, they may also convey not completely coherent messages, the possible effects of which, users should be clearly informed of."*

# Chapter 3
# Contribution & Research method

The research method used in this master thesis consists of a literature review and one case study with two units of analysis. By conducting a literature review, we aimed to form the basis and gather the necessary theoretical knowledge needed to conduct the case study, which will be the main contributor to the result of this thesis. By studying the first unit of analysis, Unit of analysis A or Unit A, we will specify the problem domain of data-driven financial decision support tools. By studying the second unit of analysis, Unit of analysis B or Unit B, we will create a detailed CRISP-DM process model for the problem domain studied in Unit A. According to Runeson et al. [53], this case study is an improving case study of the CRISP-DM process model in the context of data-driven financial decision support tools. The case study's context and units of analysis are described in 3.1.1, and visually presented in Figure 3.1.

When studying the first unit of analysis, we interviewed companies, or so called potential customers, regarding development of a FX risk exposure application. The idea of the application is provided by a startup located at Hetch AB in Helsingborg, Sweden. The head of the startup is refered to as the "System owner" from now on. The aim of the application is to quantify the customer companies' FX risk exposure. Prior to the study of Unit A, we conducted a literature review of currency hedging and the current CRISP-DM practices to form a basis for the interviews. Unit A also consisted of a field study, conducted at one of the potential customers, with the aim of understanding the current practices and data sources better than through mere interviews – in line with recommendations to draw conclusions based on multiple sources of evidence in case study research [53].

An overview of the research method is visualized in Figure 3.2, with an overview of the process phases, activities and artefacts. With the findings obtained from Unit A we answer RQ1: *What characterizes the problem domain for data-driven financial decision support applications?*, which is presented in Chapter 7, Section 7.1. The description of the problem domain was used as the basis when designing the interview guide for Unit B. When studying Unit B, we interviewed developers, data scientists and IT consultants in order to give detailed method recommendations for similar projects in the domain. Using the findings from the second

unit of analysis we answer RQ2: *How can CRISP-DM be applied when prototyping a data-driven FX risk exposure application?*, presented in Chapter 7, Sections 7.2-7.6.

# 3.1 Overview

In this section we present the case study in its context, with the defined units of analysis in 3.1.1 *Context and Unit of Analysis*. We will also present an overview of this master thesis work flow in Section 3.1.2 *Work Flow Overview*.

## 3.1.1 Context and Units of Analysis

In general, a case study may target anything which is a "contemporary phenomenon in its real-life context" [67]. A case may be a development project, an individual, a group of people or employees, a process, etc. The project, individual, group of people etc., may also be a unit of analysis within a case [53]. In this master thesis, we have chosen to define the context of the case study as the domain, i.e. the domain of "Data-driven financial decision support tools". Within this context we conduct one case study, studying two units of analysis, focusing on prototyping within the domain as the case under study. The findings from the first unit of analysis, i.e. the problem domain, is later utilized when studying the second unit of analysis, Unit B.

Unit A consists of the potential customers of the FX risk exposure application described above. The goal of studying Unit A is to define the problem domain and thereby answer RQ1. This will be done by conducting the first two phases of the CRISP-DM process model, i.e. Business Understanding (BU) and Data Understanding (DU). As the problem domain will form the basis of the detailed CRISP-DM provided when answering RQ2, there did not exist any detailed CRISP-DM when we conducted BU and DU. Thus, we had to create an initial draft of method recommendations for these two phases prior to conducting them ourselves.

Unit B is the potential developers of the FX risk exposure application. The goal of studying this unit of analysis is to provide detailed method recommendations when using CRISP-DM in the problem domain defined in Unit A. An overview of the case study's context and unit of analysis is presented visually in Figure 3.1.

## 3.1.2 Work Flow Overview

In this section we will present the high level structure of this master thesis by explaining the layout of Figure 3.2. To ease the understanding of the figure, a map legend is provided in its upper right corner. As visualized in the figure, this master thesis consists of one case study, divided into six phases, and a total of 21 activities and artefacts. Each activity and artefact is labeled with a capital letter, e.g. "**A.** Literature Review". The workflow of this master thesis can be derived by following these activities and artefacts in alphabetical order, A to U. The work flow is also visualized in the figure with arrows connecting the individual activities and artefacts. Each phase, action and artefact is further described in Section 3.2 *The Case Study*.

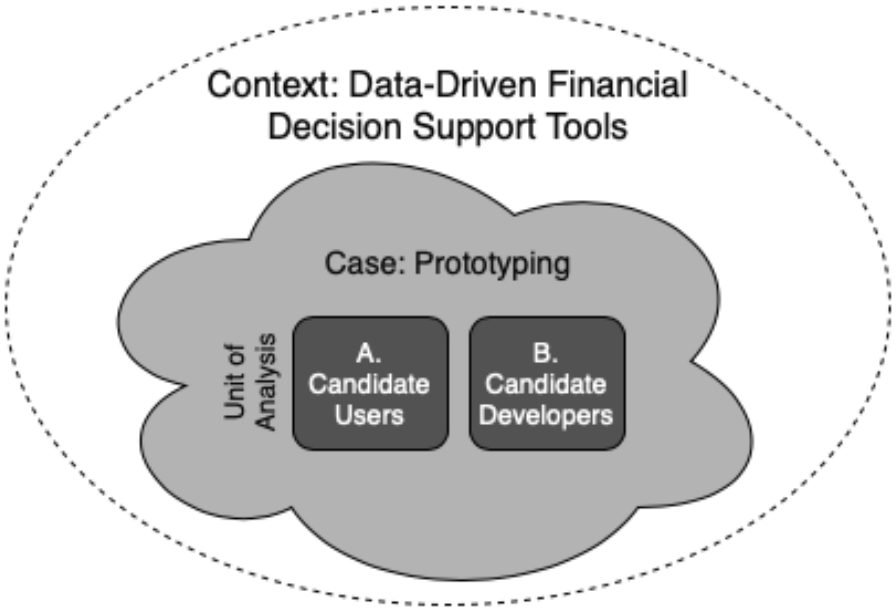



**Figure 3.1:** An overview of the case study's context and units of analysis.

## 3.2 The Case Study

In this master thesis project, we conducted a case study according to the methodology provided by Höst et al. [37]. In this section, we will present and describe the actions taken during the case study research when studying the two units of analysis, here referred to as "Unit A: Problem Domain" and "Unit B: CRISP-DM". We will divide the studies of each unit of analysis into three sections, namely (1) Planning, (2) Data Collection, and (3) Data analysis. Additionally, we will shortly describe the literature review conducted prior to defining related work and the initial draft of recommended methods for Business Understanding and Data Understanding.

In Unit A: Problem Domain we define the problem domain of data-driven financial decision support tools by interviewing representatives from five companies, see Table 3.2, active in different industries. A total of seven interviewees, with interviewees representing different corporate positions, were conducted as part of this unit of analysis. The selection of companies to participate in this case study is described in Section 4.2.1 and the selection of interviewees is described in Section 3.2.3.

When studying Unit B: CRISP-DM, we interviewed data scientists, IT consultants and developers in order to create a detailed CRISP-DM to suit the problem domain. A total of five interviewees were assigned to this unit of analysis. The participants are presented in Table 3.4 and their experience within the domain is briefly described in Chapter 5, Section 5.1.

Throughout this chapter we will refer to the activities, artefacts and phases presented in Figure 3.2. They will be referred to using brackets and the number or letter they have been assigned, e.g. Literature Review (Activity A), or Planning of Unit A (1). In short, phases are named numerically and activities and artefacts are named alphabetically. Note that the order of the following subsections do not strictly follow the chronological order of phases as they

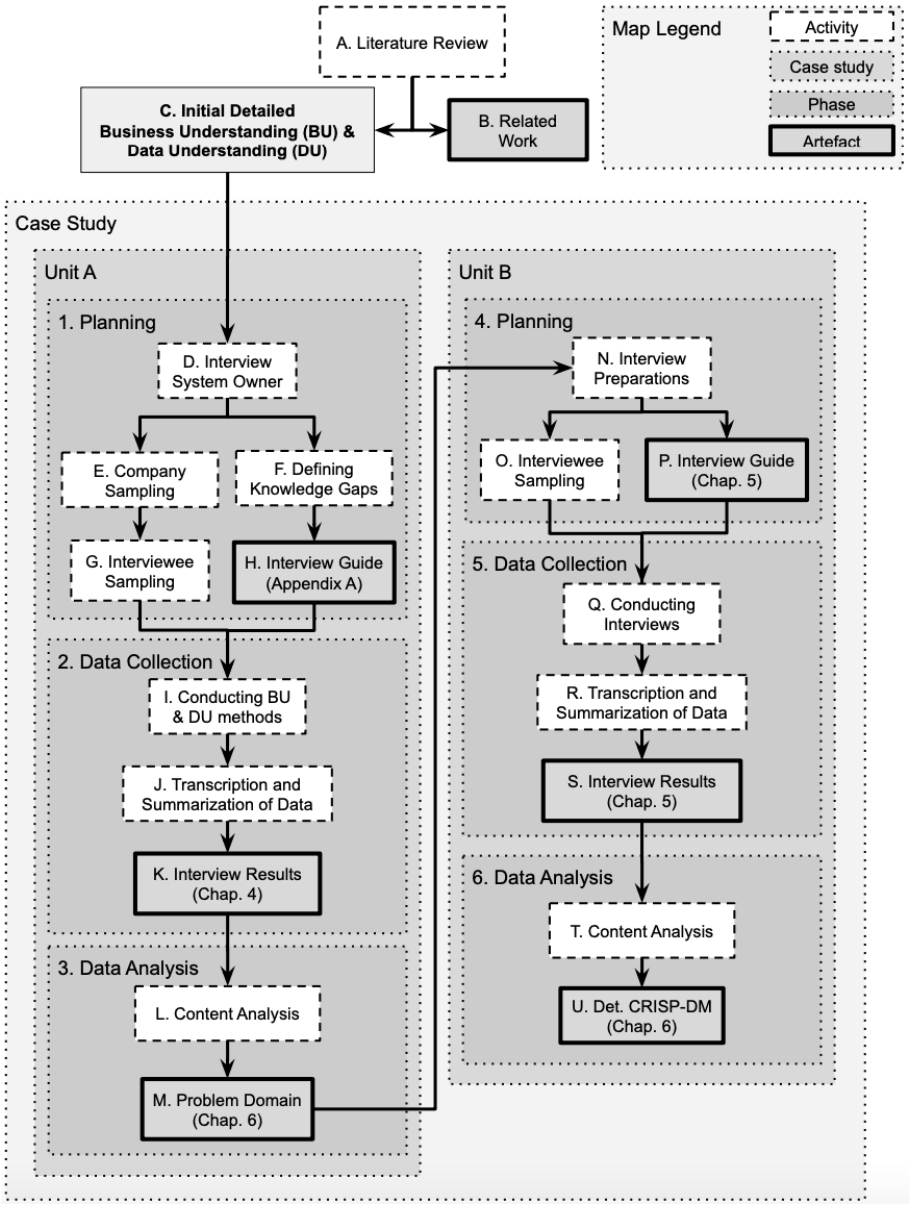



**Figure 3.2:** An overview of this master thesis' work flow. Is consist of one case study, divided into six phases, and a total of 21 activities and artefacts. Use the map legend in the figures upper right corner to ease the understanding of the work flow, and follow the "work flow"-arrows from the top down.

were conducted in the case study. We organize subsections as follows: Planning (Phases 1 and 4), Data Collection (Phases 2 and 5), and Data Analysis (Phases 3 and 6).

### 3.2.1 Literature Review

Starting off this master thesis, we conducted a literature review (Activity A) of related work, the FX market and requirements engineering practices. Following this literature review we defined related work (Artefact B) and created an initial draft of method recommendations

for the Business Understanding and Data Understanding phases for the CRISP-DM process model (Activity C). The idea of making this initial draft was to define the problem domain, and thereby answering RQ1, by conducting these phases when studying Unit A. Since the problem domain had to be defined in order to interview experts regarding method recommendations for CRISP-DM within this problem domain, we decided to create the initial draft ourselves.

In order to create this draft, or first version of the phases Business Understanding and Data Understanding, we conducted a literature search and review of current requirements engineering practices suitable for the business idea of the FX risk exposure application. This was done by searching for keywords related to CRISP-DM such as "Business Understanding, Data preparation, Data understanding, Requirement engineering, prototyping, data driven applications, data driven financial tools" on LUBsearch, relevant articles found serve as citations in this masters thesis, and can be found in the bibliography.

### 3.2.2 Planning

In this section, we present the actions taken during the planning phase of the studies of our two units of analysis.

**Unit A: Problem Domain** When planning the study of Unit A (1), we conducted an open interview with the system owner (Activity D) to obtain his ideas and thoughts of the system and its problem domain. For example, we discussed his experience in the FX market, his previous start-up with a similar business idea and the desired high-level functionalities of the financial decision support tool, hereafter referred to as “the system”.

With this initial information regarding the system, the problem domain and the FX market, we defined company criteria for the “potential customer” segment. By doing so, and only contacting and interviewing companies fulfilling these criteria, we aimed to minimize the risk of collecting misleading data and requirements from outside of the identified customer segment. The analysis of these criteria is described in Section 4.2.1 *Stakeholder Analysis*. To sample companies (Activity E) fulfilling the defined criteria, we used Google Search to find the largest export/import companies in Skåne (province in Sweden) and Sweden as a whole. Thereafter, we used *LUBsearch Databases A-Z Business Retriever* to obtain the individual corporations last three years’ financial statements to make sure they indeed had a large import/export ratio. Exactly what list of companies was used to find companies fulfilling the criteria will not be disclosed in this master thesis. This is in order to not threaten the anonymization of the participating companies. As the company sampling was finalized, we contacted employees within these corporations in order to obtain a list of interviewees for Unit A (Activity G).

The information obtained from the initial interview with the system owner was also used to define knowledge gaps regarding the knowledge domain (Activity F). These knowledge gaps were thereafter used to create an interview guide utilized when studying Unit A, see 3.1

**Unit B: CRISP-DM** Prior to planning the study of Unit B (4), we finalized the characteristics of the problem domain (Artefact M) by adopting content analysis, see 3.2.4 *Data Analysis*. With these characteristics in mind, we worked on finding companies working on projects with similar characteristics (Activity N). When the companies were identified,

we contacted personnel via mail whom we found to have relevant expertise to answer the questions we had in mind (Activity O).

Interview questions were also prepared (Artefact P) prior to the interviews in Unit B, see Table 3.3. The questions were derived from the problem domain defined when studying Unit A, the CRISP-DM process model, and the aspects mentioned in the research questions.

### 3.2.3 Data Collection

In this section, we will present the main source of information when studying both units of analysis, i.e. interviews, as well as how the data was collected and stored.

**Interviews** According to Lauesen [41], interviewing is an elicitation technique which is appropriate when exploring current work practices and potential challenges within a domain. By conducting interviews, you may also gain other types of information, e.g. regarding alternative solutions and and conflicting interests and demands. Many analysts consider interviewing to be the preferred elicitation technique due its versatile use, depending on what questions you ask. Nonetheless, while interviews are efficient when gathering information, other techniques may be needed to resolve identified conflicts and to validate the findings. For the reasons mentioned above, interviews have been the main technique for collecting data through out the whole case study, and the analysis of the data will be conducted using content anaylsis, see Section 3.2.4 *Data Analysis*.

In his book, Maciaszek [43] emphasises the importance of involving both customers, stakeholders and domain experts in the process of understanding the problem domain. When studying Unit A: Problem Domain, employees within the "potential customers" segment may be considered as part of several of these groups, depending on their corporate positions. The approach of interviewing different types of stakeholders and experts, and in line of the recommendations to draw conclusions based on multiple sources [53], we decided to not only interview one type of experts when studying Unit B, but to include data scientists, developers and IT consultants.

All of the interviews, except from the initial interview with the system owner, have been conducted using a semi-structured approach. The questions asked in each interview have been based on the purpose of answering the research questions as well as the corporate position of the interviewee. The two different interview guides are located in Table 3.1 and Table 3.3. The length of the interviews varied between 30 and 90 minutes. The majority of the interviews have been conducted through video meetings[1] while some were conducted via physical meetings or workshops. Some of the interviewees have been interviewed multiple times in order to verify correct interpretations during the content analysis or to fill knowledge gaps discovered later on in the thesis work.

**Unit A: Problem Domain** As the interviewees had been identified and contacted, and an interview guide had been created, we initialized the data collection phase (2) in the study of Unit A. The main source of information when conducting the recommended Business Understanding and Data Understanding methods (Activity I) was interviews conducted with different identified stakeholders. The interviewed stakeholders in this phase were all

[1]Note that this master thesis project was conducted during the Covid-19 pandemic

**Table 3.1:** A template of the interview questions asked in Unit A of this case study.

| Question | Interviewee Corp. Pos. |
|---|---|
| Q1 | Tell us about your company and your current company role |
| Q2 | How do you and your company currently work with FX hedging? |
| Q3 | What is the business goal of your current FX hedging? |
| Q4 | How do you decide on what FX hedging strategy to use? |
| Q5 | How do you currently quantify your FX risk exposure? |
| Q6 | What resulting data would you require from a FX risk quantification system? |
| Q7 | What data do you currently utilize in the decision making process? |
| Q8 | Are there any other interfaces/software that would need to be integrated? |

employed by potential customer companies, see 4.2.1. To support the data collection, interview questions were prepared prior to the interviews and revised iteratively. When studying Unit A, we used the interview guide found in Table 3.1.

All of the interviews were recorded and later transcribed (Activity J) in order to make searching for key takeaways easier and to enable content analysis. As all of the interviews were transcribed, we summarized the data and presented the results (Artefact K) in Chapter 4 *The Problem Domain*. Interviews regarding Unit A were conducted via online meeting tools. As part of validating our results from the interview in Unit A, we created a mock-up GUI, presented in section 4.2.6 *Requirement Analysis*, displaying the gathered requirements of the system. This activity was mere a method to ensure that the information gathered during the interviews was understood properly, and that the recommended process methods resulted in successfully conducted CRISP-DM phases.

To minimize the risk of missing important standpoints and perspectives, the selection of interviewees included at least one participant from each stakeholder category employed at a company defined as a "potential customer", see 4.2.1 *Stakeholder Analysis*. Depending on their corporate positions, they were either seen as a domain expert, data domain expert, or both. The assignment of these roles are described in Section 4.2.2 *Roles*. The interviewees which have participated in this unit of analysis are presented in Table 3.2. Due to multiple reasons, e.g. some employers being publicly listed companies and due to research ethics, all of them have been anonymized and assigned a letter ranging from A to G.

**Unit B: CRISP-DM** As the planning phase of the Unit B study (4) was finalized, we started the data collection to create the domain-specific detailed CRISP-DM (5). The data collected when studying Unit B was derived from interviews conducted with developers, data scientists and IT consultants (Activity Q) within the domain of the business case in Unit A, i.e. the FX risk exposure application. To support the data collection in this case study, interview questions were prepared prior to the conducted interviews, based on the problem domain of data-driven financial decision support tools, the CRISP-DM process model and the mentioned aspects of RQ2, see Section 1.3. The interview questions stated in Table 3.3 were used as a template during these interviews. Over the course of the case study, the interview questions were revised and improved to fill the knowledge gaps iteratively – in line with a flexible research approach [53]. As in Unit A, the interviews were recorded and thereafter transcribed to enable thematic analysis of the collected data (Activity R). As the interviews

**Table 3.2:** The interviewees which have participated in the requirement elicitation process. For each interviewee their employer (company), its last year's annual revenue, the participating interviewees' corporate positions (Corp. Pos.), and the company's main industry segment is presented.

| Interviewee | Company | Revenue (bn SEK) | Corp. Pos. | Industry |
|---|---|---|---|---|
| A | 1 | 2 | CFO | Clothing |
| B | 2 | 45 | Treasury Manager | Agriculture |
| C | 3 | 4 | Account Manager | Furnishing |
| D | 4 | 10 | Board Member | Manufacturing |
| E | 4 | 10 | Treasurer | Manufacturing |
| F | 4 | 10 | Treasurer | Manufacturing |
| G | 5 | 12 | Treasury Manager | Chemicals |

were transcribed, key takeaways form each interviewee were presented as results of this unit of analysis in Chapter 5 (Artefact S).

When conducting the interviews in order P1 to P5, we first asked the interviewees to present their experience of working with data-driven applications and CRISP-DM, within the problem domain, and to present similar information of interest. Secondly, we asked what actions they take in each of the CRISP-DM phases and if and how they would change their methodology when presented to the characteristics of the problem domain. Thereafter, we presented the actions taken when conducting the Business Understanding and Data Understanding phases in the Unit A study, and asked them whether they would like to add or remove any of them. Finally, we presented the approaches of the other interviewees to verify whether or not they approved the actions taken by the them.

As the different phases of the CRISP-DM process model focus on different topics and problems, we aimed to include three different types of experts to cover them all. Hence, Unit B consist of one IT consultant, three data scientists and one software developer with experience in development of data-driven applications. The hypothesis is that the IT consultant will have experience in understanding the problem domain and the customer values, the data scientists will have experience with data understanding and preparation, modelling and evaluation, and that the developer will bring important aspects to keep in mind then deploying the model.

### 3.2.4 Data Analysis

When analyzing the collected data (Activity L & T), we used the content analysis method [26]. The goal of content analysis is to create a descriptive presentation of qualitative data, i.e. data from interviews or other types of textual data. The idea is to portray the content of transcripts by identifying and highlighting common codes in the data collection. While some interpretation had to be done in order to form distinct codes in the transcripts, the interpretations were kept at a bare minimum. If interpretations were made, the interviewee was contacted in order to verify it.

In order to find codes in the transcripts, a condensed summary consisting of mere bullet points was written. In order to write the summarizing bullet points, we divided the inter-

**Table 3.3:** A template of the interview questions asked in this case study. Horizontal lines indicate separate sections of the interview. Approach refers to approach to developing data-driven solutions.

| Question | |
|---|---|
| Q1 | Tell us about yourself and your professional career |
| Q2 | What is your experience of working with CRISP-DM? |
| Q3 | How do you work with Business Understanding? |
| Q4 | How do you work with Data Understanding? |
| Q5 | How do you work with Data Preparation? |
| Q6 | How do you work with Modelling? |
| Q7 | How do you work with Evaluation? |
| Q8 | How do you work with Deployment? |
| Q9 | What is your view on the problems identified in the domain? |
| Q10 | How would your approach be modified to suit the problem domain? |
| Q11 | How would your approach be modified to suit prototyping? |
| Q12 | What is your view on the previous interviewees approaches? |
| Q13 | What is your view on the methodology used to explore the problem domain? |

**Table 3.4:** The interviewees which have participated in the detailed recommendation of CRISP-DM. For each interviewee their corporate position (Crop. Pos.) and main industry is presented.

| Interviewee | Gender | Interviewee Corp. Pos. | Industry |
|---|---|---|---|
| P1 | Male | Data Scientist | Business Communication |
| P2 | Male | Data Scientist | Financial Industry |
| P3 | Male | IT Consultant | Information Technology |
| P4 | Male | Developer | Information Technology |
| P5 | Female | Data Scientist | Machine Learning Consultancy |

views amongst the authors. As the bullet points were done, we together discussed each bullet point in order to asses whether or not it seemed like a code of interest. As all of the codes in each bullet point were identified, we aimed to combine similar wording from different interviews into one common set of codes based on content. A mix of individual codes and combined codes makes up the total data presented in Chapter 4 and 5.

A summary of the problem domain and the interviews from Unit B is presented in Chapters 4 and 5 respectively. The data included in these chapters are later analyzed and discussed with the identified codes in Chapter 7 *Discussion*.

# Chapter 4

# Results Unit A: Characterizing the Problem Domain (RQ1)

In this chapter, we will present the data collected when studying unit of analysis A, i.e. the potential customers of an FX risk exposure application. This chapter is derived from the summation of the data collected when conducting the initial business understanding and data understanding phases of the CRISP-DM process model, i.e. activity I and J in phase 2 of the case study (see Figure 3.2).

The goal of studying this unit of analysis is to gain an understanding of the problem domain and its specific challenges. The results presented in this chapter are derived from summation of the interview transcripts. Each Interviewee assigned to this unit of analysis is anonymized and named with a letter ranging from A to E, as showed in Table 3.2. The interviews in this case study have been conducted in a semi-structured way, utilizing the prepared interview questions in Table 3.1 as a template.

Prior to the activities reported in this chapter, we conducted activities C, D, E, and F. i.e. we planned this study of Unit A by creating an initial draft of business understanding and data understanding (Artefact C) in order to concretize what activities to conduct when defining the problem domain and its challenges, we interviewed the system owner (Activity D), sampled companies and interviewees for the study (Activity E), and created an interview guide (Activity F).

The main source of data when studying unit A is derived from interviews conducted together with the interviewees presented in Table 3.2. The interview questions are presented in Table 3.1. The interviews were divided into two separate sections. During the first one, i.e. “Business Understanding”, we discussed questions Q1 to Q5. Thereafter, during “Data Understanding” we discussed Q6 to Q8.

The resulting understanding of the problem domain and its specific challenges serves as the basis for the interviews with potential developers in Chapter 5, and the fundamental challenges on which our detailed CRISP is constructed, see Chapter 6.

## 4.1 Summary of the problem domain

By studying Unit A, the goal was to obtain domain-specific knowledge and to define the problem domain. The idea of defining the problem domain is to identify challenges within the domain that later on should be taken into account when creating a detailed CRISP-DM process model.

During the study of Unit A, three main challenges mentioned by multiple companies have been identified, namely *Large number of interfaces*, *Uncertainty of Data* and *Difficulty of defining main purpose*. These identified problems have been chosen as the characteristics of the problem domain since they have prevented the potential customers to implement similar systems before. The problems are further discussed in Chapter 7, Section 7.1 *The Problem Domain*.

## 4.2 Business Understanding

Business understanding is the first phase of the CRISP-DM process model. In this section we will present why we conduct the additional process steps and the results of conducting them. When conducting the activities assigned to business understanding, we analyze stakeholders and their demands, and gathered domain specific knowledge to clarify present procedures and problems when developing applications in the domain of data-driven financial decision support tools.

### 4.2.1 Stakeholder Analysis

Stakeholders are the people needed to ensure project success. The term includes people and organizations such as investors, daily users, business partners, authorities, and more [41]. Lack of stakeholder participation during requirements elicitation has been observed as the main reason of software failure since they are the main source of requirements [44]. Thus, it is crucial to identify all stakeholder and their interests in the system [41, 66].

Stakeholder analysis is a tool, set of tools or an approach used in the process of gathering information about actors, i.e. organizations or individuals, in order to understand their behavior, interrelations, interests and intentions [64]. To ensure maximum requirements coverage in regards to all stakeholders as well as their participation in the project, stakeholder analysis will be the first thing to be conducted in this case study.

In the stakeholder analysis we aim to answer the following questions: *Who are the stakeholders?*, *What are their interests in the system* and *Which stakeholders shall be involved in the elicitation process?*. Regarding the first question, there are currently four identified groups of stakeholders, namely (1) Potential Customers, (2) Daily Users, (3) The System Owner, and (4) Authorities.

**Potential Costumers** Potential customers we define as the companies which may be interested in purchasing the system under development. This is not to be confused with the individual employees within the corporations, as intended with the next group of stakeholders *Users*. To find potential customers, we had to define criteria for them to be interested in an FX risk exposure application. The identified criteria are stated below.

**Large revenue** For companies to be interested in an automated FX risk exposure application, they must handle a too large amount of data for the operations to be done manually in an efficient way. Hence, we will primarily focus on large enterprises.

*Case study criterion*: In the case study, we contacted companies reporting an annual revenue greater than one billion SEK.

**Large foreign trade** Since FX risk exposure only affects companies trading in foreign currencies, this must be seen as the primary criterion. Additionally, due to FX risk exposure being positively correlated to the foreign trade ratio, we will primarily focus on companies having a large foreign trade.

**Users** The users of the system are defined as the employees within the *potential customer* segment who will be using the system. In this case study, we decided to divide the users into two different "user types", namely daily and occasional users. The purpose of this differentiation was to ease the requirements analysis work later on. The idea was that it would better help us understand differences between application domain requirements and goal domain requirements, as out hypothesis was that daily users would be more inclined to specify application domain requirements. The identified stakeholders within each corporation are specified below.

**The Board of Directors** The Board of Directors is responsible for creating the corporate strategies and policies. The financial policies may, or in the case of the system's potential customers usually, include directives regarding if and how much the company should hedge its foreign trades. These decisions should according to all interviewees be based on the companies' FX risk exposure. Since the data provided by the system shall form the basis of the financial policies, the Board of Directors is a system stakeholder.

*User type*: Occasional user

**The Chief Financial Officer** The Chief Financial Officer (CFO) is responsible for all the organization's finance and accounting. One of the CFO's main responsibilities is making sure the corporation meets the strategies and goals set by the management and the directors. To be able to assess and evaluate the organization's hedging strategies and risk exposure, the CFO will need the data provided by the system, making them stakeholders of it.

*User type*: Occasional user

**The Treasury or corresponding department** According to the majority of the interviewees, the treasury, or a corresponding department, is responsible for the execution of the financial strategies set up by the Board of Directors and/or CFO. Treasurers will use the system more frequently than the other user stakeholders to continuously asses the organization's FX risk exposure and thereafter take appropriate actions.

*User type*: Daily user

To visualize the interrelationship of the user stakeholders, a generic organization chart displaying the common financial hierarchy of organization was created and presented in Figure 4.1. The identified stakeholders within each potential customer's organization are marked

in green. Additional corporate positions with direct relations to the identified stakeholders are also included in the figure. These are the CEO, vice presidents such as Chief Technology Officer (CTO) and Chief Operating Officer (COO), controllers and auditors.

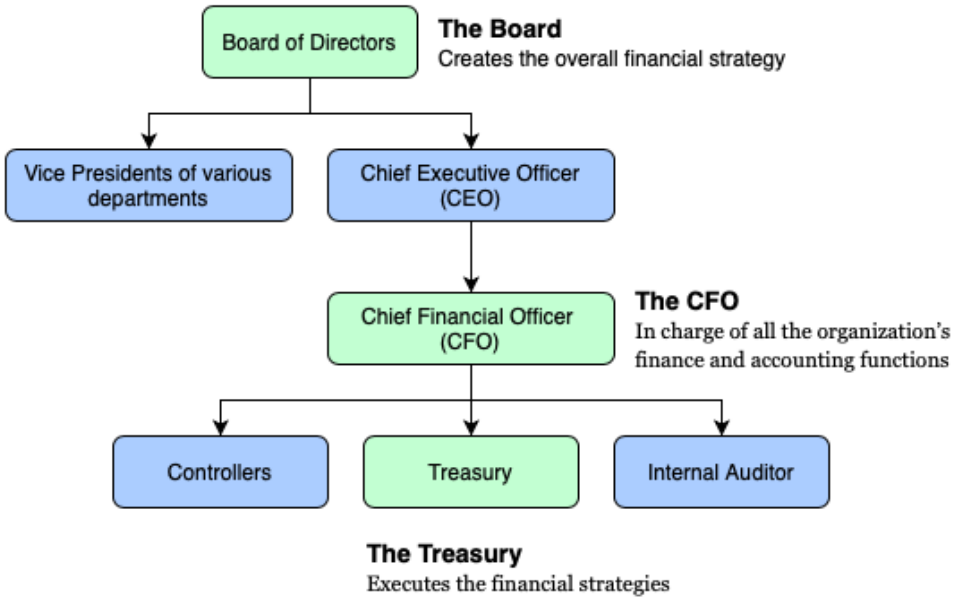



**Figure 4.1:** An simplified overview of the financial hierarchy of organizations. The identified stakeholders are colored in green.

**Data Scientist** As stated by Borg and Vogelsang [65], data scientists are one of the most important stakeholders when developing data-driven applications. A data scientist needs a skill set to help customers set reasonable targets, including statistics, domain understanding, and computer science.

**System Owner** Currently, the system or business idea owner is only one person. During a live demonstration of the constructed prototype's candidate GUI, see Section 4.2.6, the System Owner validated the prototype and gave additional feedback as to what should be included in an MVP. The System Owner stated that an additional issue in the domain is that of handling financial data. Handling of financial data for a publicly traded company is regulated by the Swedish financial supervisory authority *Finansinspektionen*. His expertise in the field therefore added an additional stakeholder to our initial analysis, Finansinspektionen.

**Authorities and Legal Experts** System stakeholders may also include authorities, such as inspectors, auditors and local government [41]. While these types of stakeholders do not have an active role in the requirements elicitation process, they may enforce regulations which have to be taken into account when developing the complete system. To identify these types of stakeholders, we conducted an interview with the System Owner, as he has domain specific knowledge from previous working experience. The identified stakeholders are stated below:

**Swedish Authority of Privacy and Protection** The Swedish Authority of Privacy and Protection is responsible for ensuring that Swedish organizations follow the European Union's General Data Protection Regulation (GDPR) regulations [3]. This stakeholder and the GDPR will be of interest since the system will store and use financial data that may be linked to individual persons.

**Swedish Financial Supervisory Authority** As handling of financial information concerning some organizations is regulated, the System Owner clarified during the interview that the Swedish Financial Supervisory Authority [4] must approve the data collection and storage. This is especially relevant if the corporations using the system are publicly listed.

**Legal Experts** Data-driven applications generally utilize a large amount of data which might have its use restricted by a third-party owner or other legal regulations. Thus, legal experts are of interest in such development projects. Borg and Vogelsang also state in their article [65] "We find it inevitable that requirements engineers working on ML systems must stay on top of legal requirements, and the data lineage must show that no illegal features have influenced the final data set used for training the ML models."

### 4.2.2 Requirements Engineering Roles in the Development Organization

In their article, Hesenius et al. [35] identifies four main requirements engineering roles in the development process of data-driven applications, namely Domain Experts, Data Domain Experts, Data Scientists and Software Engineers. Here, a role is not defined as a person but rather a specialization within the process. This means that a single team member may have multiple roles while a whole team, or set of people, may share a single role. Although the four roles mentioned by Hesenius et al. are relatively universal in the domain of data-driven applications, depending on the application itself additional roles might be needed. The main roles identified in our master thesis project, considered critical to the development of the envisioned system, are further specified below and summarized in Figure 4.2.

The Software Engineers are the ones in charge of the software development process and implementation of the system itself. The Software Engineers' tasks include integration of the data-driven module and implementation of necessary source code artifacts. The Software Engineers are also responsible for gathering and documenting model requirements for the data-driven module.

Data scientists are responsible for the data-driven solution component and the data foundation. Tasks may include identifying data relationships, visualizing data for Domain Experts and Domain Data Experts, choosing suitable ML algorithms and other tasks within the data exploration area.

The Domain Experts provide information regarding the application's business context. Generally, they are also the targeted users of the final system. Domain Experts have knowledge in use cases, the necessary business processes, regulations and more. In the project, it needs to be identified which Domain Experts should be included in the process.

Data Domain Experts possess knowledge in and access to data within the domain or specific company. While Domain Experts possess knowledge in business-level data inter-

pretation, the Data Domain Experts have a more technical view and approach. The Data Domain Experts' tasks include mapping business and domain information to corresponding data sources and providing necessary interfaces to the Software Engineers.

In this case, with the FX risk exposure application, we will not assign the roles to people in the project, mainly since we do not have a development team, nor will the application be developed during this master thesis. However, as Hesenius et al.'s research suggests, a development project of a data-driven application should have these roles assigned.

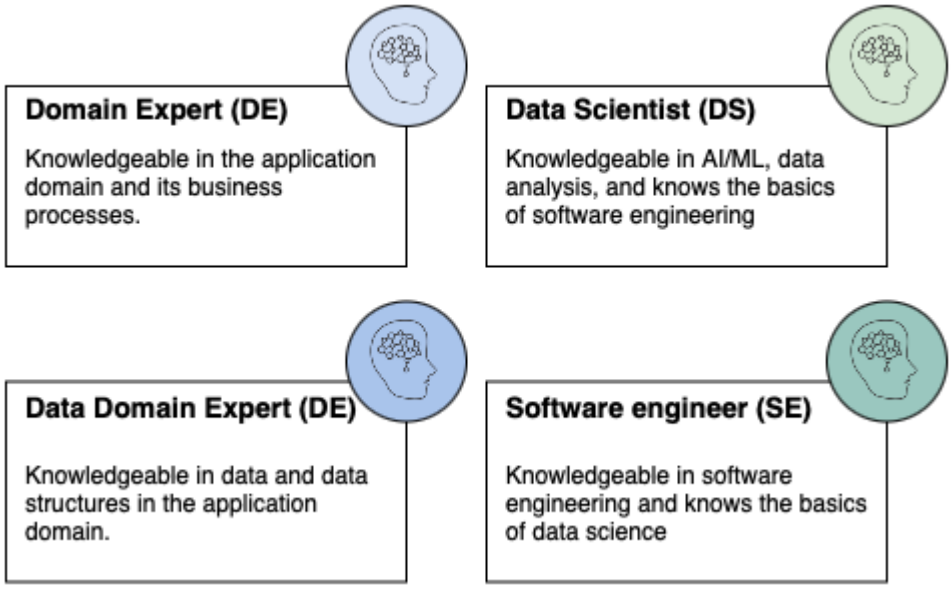



**Figure 4.2:** Skills and roles in the development process of data-driven applications according to Hesenius et al. Adapted from [35].

### 4.2.3 Purpose of Using the System

As part of the stakeholder analysis, we wanted to clarify the identified stakeholders' purpose of the system. This is to ensure the final prototype actually fulfills the stakeholders *true needs*, rather than what *they think they need*.

The purpose of quantifying FX risk exposure was unanimous for all of the interviewees, namely forming the basis of FX hedging strategies. Hence, the focus of the interviews became to figure out root causes of the problem, i.e. *"what are the purposes of the potential customer's FX hedging strategies"* and *"what data and assessments are needed to create such strategies?"*.

The purpose of FX hedging itself was somewhat divided. The common ground of FX hedging was to create *financial certainty*, allowing the corporations to plan for the future. The definition of financial certainty differed from company to company. While Company 1 hedged in order to stabilize their profit margins of certain product categories and cash flow from specific markets, Company 2 also hedged in order to fix prices of business acquisitions and other larger investments. In general, the interviewees agreed on that financial uncertainty is associated with volatile *liquidity* and thereby *profit margins*. Furthermore, while some companies (Company 2, 3, 4 & 5) performed FX hedging mere on the basis of creating financial stability, Company 1 also wanted to speculate in future currency rates to raise the

profitability.

### 4.2.4 Current FX Risk Procedures

The companies' current procedures of FX risk exposure calculations is of great importance since it will form the basis of the data-driven application to be developed. Here, the aim is to create an understanding of the interviewed companies' current procedures for FX risk quantification, to strengthen the understanding of the required functionality of the system.

**Time Interval** The time interval of FX risk exposure calculations differed from company to company. While Company 1 performed risk exposure calculations once every month, the frequency varied in Company 2, 3, 4, and 5. This is mainly due to Company 2, 3, 4 and 5 being company groups. Within these company groups, the data collection was mainly conducted in the individual company and the data was thereafter reported to the group treasury. Depending on the individual companies' annual sales, foreign trade ratio and more, they had a given time period ranging from one to four months in which they had to send their data to the group treasury.

**Data** The data needed to asses a company's FX risk exposure depends on what type of risk quantification they desire. While some (Company 2, 3 & 4) use the net revenue of each currency, others (Company 1 & 5) use other measurements such as correlated prices of raw materials and volatility.

In all of the cases, data from ERP, treasury, accounting systems as well as from banks will have to be collected. In the case of additional factors such as correlated asset prices, third-party data will have to be collected as well.

**Assessment** To asses what currencies to hedge, all companies 1-5 calculate the net revenue for each of them to use as a basis for the assessment. The larger the net revenue, the higher the risk. However, there exists no general approach of assessing the net revenues, nor the additional data such as correlated prices, since risk is subjective and based on an individual's or a company's risk aversion[1].

### 4.2.5 Current Challenges in FX Hedging

The information gathered during the interviews was consolidated and analyzed to discern key problems that the interviewees currently have with their FX risk exposure quantification. The following key problems were identified:

**Lack of common goal with hedging** The results presented in Section 4.2.3 and 4.2.4 indicated that the interviewed companies had differing reasons for their work with FX risk management. This indicates that there currently is no agreement on what would be the optimal method regarding FX hedging strategy and FX risk quantification.

[1] In economics and finance, risk aversion is the tendency of people to prefer outcomes with low uncertainty to those outcomes with high uncertainty, even if the average outcome of the latter is equal to or higher in monetary value than the more certain outcome. [38]

**Large amounts of interfaces** Interviewees stated large amounts of interfaces as one of the primary reasons for not incorporating assisting software when quantifying their FX risk exposure. According to company 2, 4 and 5, they are currently in the search of a similar system, but the cost of integrating ERP, treasury, accounting and banking systems is currently too high for the FX risk exposure systems to be valuable. Note that company 2, 4 and 5 are company groups and also the largest companies included in this research. Due to the fact that these enterprises are company groups, integration of subsidiaries' systems is also needed, resulting in greater costs of integration.

Interviewed companies with multiple subsidiaries (Company 2, 3, 4 & 5) reported that individual companies in the group use different service provides for ERP, accounting, treasury and banking system. Resulting in an exponential increase in the need of integration. Today they often use spreadsheets as an export format of data during internal communication to solve the problem. Company 2 also stated that this internal manual transfer of financial data sometimes can be a source of error and uncertainty.

**Uncertainty of data** Uncertainty of data was determined to be a factor complicating automation of FX risk quantification. As invoices and deliveries sometimes are entered incorrectly, all interviewed companies 1-5 reported that current financial data retrieved from their systems can be misleading without context. To prevent this, all the companies performed manual inspection of the data in some capacity before exporting to other information systems.

**Lack of authority during creation of hedging strategy** All interviewed companies 1-5 reported that the policy for hedging and FX risk exposure is set by the board, and had not been revised in a long time. This gives the treasurer, CFO, and other related parties less incentive to actively work with improvement of their FX risk management strategy, as it gives them less room for their own ideas.

### 4.2.6 Requirements Analysis

The requirements analysis process is performed to transform the stakeholders' views of their desired services into a technical view of a required product that could deliver those services. In this process, one should aim to find and resolve conflicting requirements [8], and to form the basis of the upcoming phases in the CRISP-DM process model. For this case study, a minimum requirements analysis was conducted in order to form the basis of a GUI prototype to validate the findings from the study of Unit A. The GUI prototype is presented in Figure 4.3.

In Figure 4.3 a numerical FX risk quantity is represented over time in a diagram. The data used in the mock-up is randomly generated and has no association to any of the interviewed companies. In the case of this case study, the desired quantification was not unanimous and therefore we have avoided to explicitly name the quantification measurement in the diagram. However, since the data generally is uncertain and in many cases based on forecast and invoiced, we have represented the uncertainty of the future by creating three different expected outcomes, namely the green, orange and red curves.

In the upper left corner of the GUI prototype there are check boxes for different currencies to include in the diagram. This is based on the current procedures of companies

monitoring different currencies more or less. In the upper right corner there are two switch boxes named “Forecast” and “Show transactions”. The forecast switch box, currently activated, is there because not all companies wanted to quantify their future FX risk exposure, and thereby only the current, meaning that forecast should not be included. The exposure calculated according to the forecasts are to the right of the horizontal marking named ´´Today” and colored green, orange and red. Due to the uncertainty of forecasts, and thereby the results of the data-driven application, the uncertainty of the outcome is represented with this color schema. Depending on what type of exposure quantification will be used, red may either be the worst outcome, or the best. The switch box for transactions, currently not toggled, is a desired functionality by company 2, where the explicit due dates of all transactions is marked in the diagram according to their date and volume.

The lower part of the GUI contains two frames with numerical FX risk exposure quantifications. The one to the left presents the risk exposure of each currency in relation to the company’s main currency. In the case of the GUI, the assumption that all companies used Swedish Crowns (SEK) as their main currency was made. To the right, the total exposure of all currencies in relation to the main currency is presented.

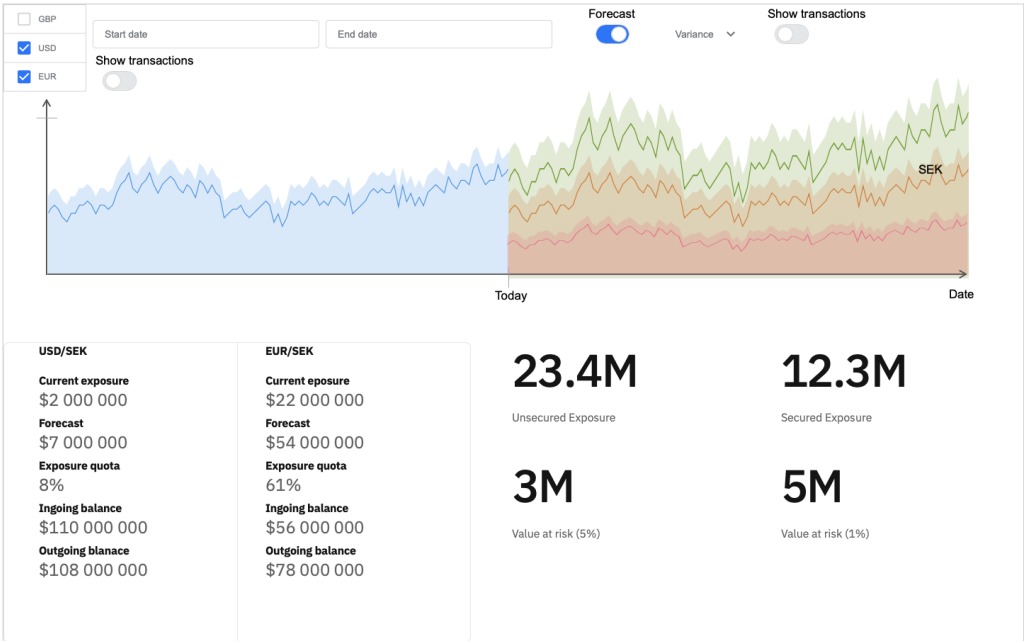



**Figure 4.3:** The created mockup used to validate the business understanding and the use of the system with the stakeholders.

Interviews were conducted to validate the requirements of the system by presenting and discussing the GUI. During this validation process, additional requirements were added to the GUI, namely the numerical exposures in the lower part of the GUI. The GUI presented in Figure 4.3 is the final revision which was approved by all of the companies.

## 4.3 Data Understanding

Data understanding is the process of examining available data and analysing its surface properties. This includes surface properties such as quantity and format of the data [20]. In data-driven applications the process of understanding the data is crucial, as data only is useful if it has potential to discover knowledge or reveal information [9].

### 4.3.1 Interfaces

To gather information about data availability, we decided to investigate the possible sources of data available in different organizations. Interfaces for data transfer of the identified sources were then examined.

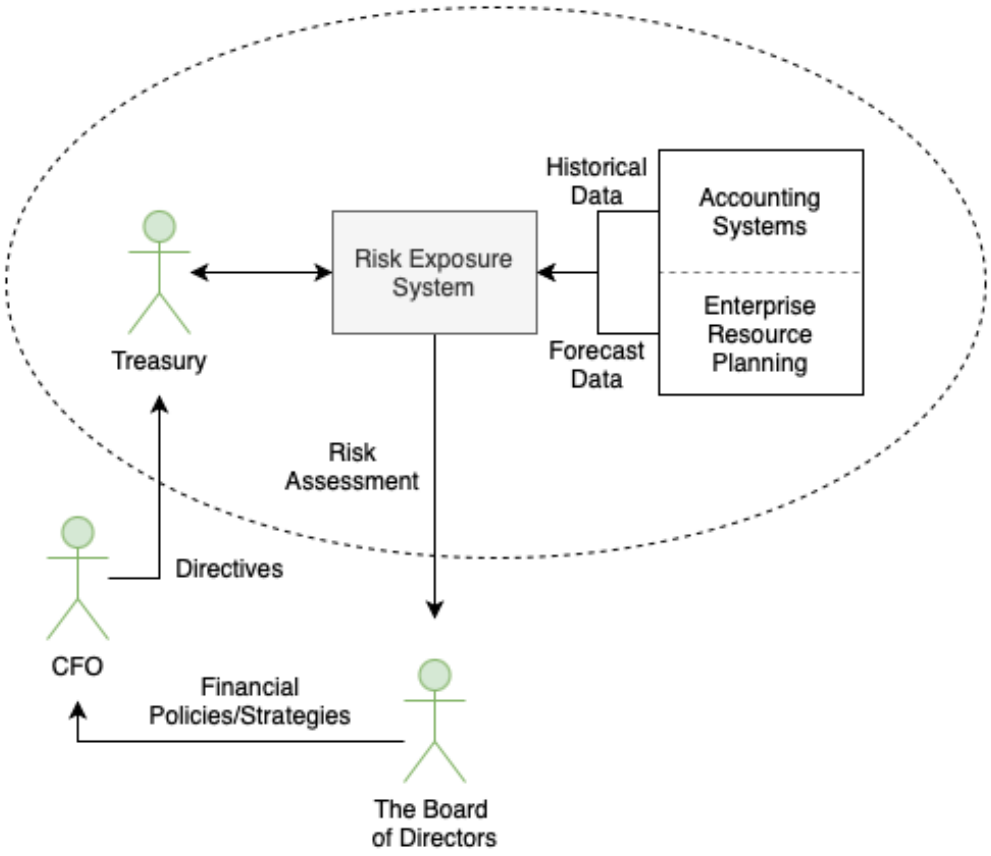



**Figure 4.4:** The system context in relation to the three identified stakeholders (marked in green) within each customer organization as well as to their internal data interfaces.

Companies 2-5 reported a lack of common interfaces for transmission of financial data. Company 2 stated that subsidiaries often use differing software for accounting and forecasting purposes, and therefore use spreadsheets as a medium for data transfer internally. The speadsheets are mostly transferred through email, which often causes problems related to version control and simple errors in data entry. The process of transferring data through

spreadsheets as email attachments is also time consuming compared to options such as a shared database or requests via API. As the organization of Company 2 consisted of multiple subsidiaries, they did not have directly access all operational data of their respective subsidiaries. They were therefore heavily reliant on reports constructed by the subsidiaries to gather information concerning their operations. They however stated that it would be possible to construct such an integration with their ERP systems, that would allow for direct access to operational data. Their banks did not provide interfaces for easy access to financial data due to security concerns, and they did not consider it possible to develop such an integration with their current banks. Their accounting systems did not provide data as easily retrievable as their ERP systems, but they considered it possible to develop such an integration.

To reduce the amount of time spent on aggregating financial data from subsidiaries, multiple interviewed companies suggested the use of a standard interface for data transfer. This interface would have to be compatible with common ERP (Enterprise resource planning) systems and accounting systems. Figure 4.4 presents a context diagram for the FX risk exposure applications based on this information.

### 4.3.2 Available Data

To calculate the net values, the system needs *historical records* of sales and purchases to calculate the historical FX risk exposure, and the *forecasts* of sales and purchases to calculate the current and future FX risk exposure. Furthermore, to calculate measurements associated with the volatility of each currency, historical and current *exchange rates* must also be gathered.

To further concretize the availability of data in the domain of financial decision support tools, the current information transfer process of Company 4 was studied as part of their interview.

#### ERP Systems

ERP systems are business suites used to manage organizations' day-to-day activities. Their main advantage is that they can combine multiple different processes such as accounting, supply chain operations, and warehousing using common databases. This reduces the occurrence of problems such as data duplication and double maintenance that may otherwise occur with the use of multiple different systems for these functions [7]. Due to the variation in what business functions organizations need, ERP systems used in large organizations are often customized to fit the organizations needs [29]. This means that organizations can customize their ERP systems to include functions for exportation of data in various formats, as all their data is stored in a common database.

Company 4 stated that their ERP systems as well as other common ERP systems were able to export data in various formats. Many ERP systems can therefore be considered able to export data in a format usable by the system this thesis concerns.

#### Accounting Systems

The interviewees differed regarding which accounting systems they used. Common accounting software used in large organizations include Fortknox [2] and Visma [5]. Both Visma and

Fortknox supports development of customized software integrations, which could provide a channel for data collection for the system [2] [5].

## System for Extraction of Financial Information

The use of an external source of data is necessary to collect real-time data of currency pair rates. As FX data and current prices of financial instruments are publicly available through financial institutions, it was determined that this data can be collected from an external source. An example of a free-to-use tool for gathering FX data is *fixer.io* [1]. FX data was deemed time-critical, as the prices of financial instruments and currencies need to be up-to-date for the system to be useful for decision making.

# Chapter 5
# Results Unit B: Detailing CRISP-DM (RQ2)

In this chapter, we will present the data collected when studying unit of analysis B (Unit B), i.e. when interviewing the data scientists, developers and IT consultants.

The interviews were divided into four separate sections. During the first section, i.e. "Ìntroduction", we discussed questions Q1 and Q2. Thereafter, during "General Approach" we discussed Q3 til Q8, during "Domain Specific Approach" we discussed Q9 and Q10, and during "Effects of Prototyping" we discussed Q11. Lastly, we validated the previous findings from Unit A and Unit B by discussing Q12 and Q13. This chapter is structured accordingly, Sections 5.1–5.4 summarize the interviews whereas Section 5.5 synthesizes the results.

**Introduction** Discussing the interviewee's professional and CRISP-DM related experiences (Q1-Q2).

**General Approach** Discussing the interviewee's approach to each of the CRISP-DM phases (Q3-Q8).

**Domain Specific Approach** Discussing if and how the approach would change with the identified characteristics of the problem domain (Q9-Q10).

**Effect of Prototyping** Discussing if and how the approach would change with the goal of prototyping in the domain (Q11).

**Validation of study** Discussing the previous interviewees statements to validate our previous findings. (Q12-Q13)

## 5.1 Introduction

In this section we will present the interviewees' professional and CRISP-DM related experiences. The data in this section is derived from the interviewees' answers to questions Q1 and Q2. See Table 5.1 for a summary of the information.

**Table 5.1:** Summary of the selection of interviewees. Abbreviations used in this table: Financial Technology (FT), Data Science (DS)

| Interviewee | Professional Experience | Education | CRISP-DM Experience |
|---|---|---|---|
| P1 | DS, FT | M.Sc. Engineering | No previous knowledge |
| P2 | AI/ML, FT | M.Sc. Engineering | No previous knowledge |
| P3 | It consultancy | M.Sc. Engineering | No previous knowledge |
| P4 | FT | M.Sc. Engineering | No previous knowledge |
| P5 | AI/ML consultancy | M.Sc. Engineering | No previous knowledge |

**P1** Interviewee P1 has a master's degree in engineering and professional experience working with data science and modelling of financial support tools in the tech industry. P1 had no previous knowledge of CRISP-DM, he however stated that his current workflow is very similar to that proposed by CRISP-DM and utilizes the same ideas.

**P2** Interviewee P2 has a master's degree in engineering and professional experience in working with data-driven applications within the healthcare and financial domains. Prior to being presented to this master thesis project, P2 had not heard about the CRISP-DM process model. However, P2 adds that he utilizes the same general workflow as provided by the model.

**P3** Interviewee P3 has a master's degree in engineering and professional experience working with project-based development on a consultancy basis. Prior to the interview, P3 had no knowledge of the CRISP-DM model. He, however, stated that it shares a similar workflow to that which he uses when planning projects. P3 states that in his current work, the data-related processes are designated to a team with the sole purpose of handling data-related issues, and he therefore has less experience working with data understanding and preparation than that of a data scientist.

**P4** P4 has a master's degree in engineering and some previous experience in developing data-driven applications, though he today mostly work with conventional software development. He has previously worked as full stack developer within the financial sector in Sweden and currently works as a developer in Denmark focusing on government technology. Prior to the interview, he had no knowledge of CRISP-DM, but said that the overall structure seemed reasonable when developing data-driven applications.

**P5** P5 has a master's degree in computer science and is the founder and CEO of an AI/ML consultancy. As a consultant, she has worked with development of data-driven applications within multiple industries, but mainly the financial. Prior to the interview, P5 had not heard about the CRISP-DM process model, but found the phases to be closely related to the one used in her firm.

## 5.2 General CRISP-DM Approach

In this section we will present key takeaways from the discussions regarding a general approach to each of the phases of CRISP-DM. These discussions are based on questions Q3 to Q8. For each question, we present a summary followed by a list of the identified codes. We describe the codes and provide a synthesis in Section 5.5.

### 5.2.1 Business Understanding

In this section we will present the key takeaways of the interviewees' approach when working with business understanding, i.e. the answers to question Q3.

**P1** P1 states that his need for development of data-driven models commonly is requested by other departments in his organization in need of a solution for a specific business problem. His department is then tasked with this process, which he states starts with a rigid business understanding. The first step of his business understanding is to discuss the case with colleagues within his own department. To fill any gaps in knowledge, he might then seek out domain experts from other departments of the organization that can provide more information about the business purpose and requirements of the model. Depending on the complexity of the business problem, he either gathers involved stakeholders for a rigid discussion of the expectations and requirements, or seeks out specific stakeholders for informal discussion with the aim of filling his gaps in knowledge.

**Identified Codes** *Stakeholder analysis, Define Purpose, Align Environment, Domain Specific Problems, Business Requirements Analysis, Consult Domain Experts*

**P2** According to P2, business understanding is one of the most important phases in the development of data-driven applications. Additionally, he says *"It's important to keep in mind there exist data scientists on various business levels. Some are more R&D-related data scientists, for whom this phase might not be as important. For me, generally working in start-ups, where the end goal it to provide the end-customers with something valuable, this is everything."*. However, he also adds that the definition of business understanding is highly dependant on the purpose of the model under development.

In general, P2 first tries to identify who has the knowledge and expertise needed to implement the model. Development projects of data-driven applications does in general not only consist of software engineers and data scientists, but also personnel with widely spread knowledge within the domain. He states that additional areas of interest might be finance, marketing, legal, healthcare, etc. Depending on if the project is ordered internally by a different department, or a customer, or if you come up with the idea yourself, it might be more or less difficult to identify all stakeholders. P2 adds, *"Usually I try solving this issue by conducting an interview with the person I assume has the broadest knowledge within the specific domain."*.

When trying to define the purpose of the model, P2 generally tries to define a single brief use case describing the functionality of the model and its value, e.g. *"This model will increase a corporation's financial security by predicting future cash flow"*. By doing so, you can use this short definition of the model to investigate which stakeholders obtain direct value from it

and sometimes even identify new stakeholders, he adds. However, creating such a use case might be hard since it often requires deep understanding of the business domain.

At the end of the business understanding phase, it should according to P2 be possible to define high-level business requirements. However, defining technical requirements at this stage it not really possible. Defining technical requirements in the domain of data-driven applications usually requires adequate data understanding. Validating high-level business requirements can be done in multiple ways, but creating some kind of high-level sketch, GUI mockup or diagram might be suitable.

**Identified Codes** *Stakeholder analysis, Define Purpose, Domain Specific Problems, Business Requirements Analysis, High Level Design Prototype*

**P3** Interviewee P3 states that in his experience, starting very broadly when understanding the business needs of his customers is a successful approach. He often begins by defining the top-level goals of the client's system via the use of workshops with relevant stakeholders. The goal of this being to identify the overarching system goals, which he then tries to break down into system requirements. These system requirements are then validated via a second workshop with the stakeholders, where the requirements are presented in the form of user stories with the goal of ensuring that all the required functionality is supported by the system.
**Identified Codes** *Stakeholder analysis, Define Purpose, Business Requirements Analysis, High Level Design Prototype, Workshop*

**P4** According to P4, business understanding means understanding the core and the value drivers of the business the system is meant to be used by or applied on. P4 adds, aspects such as how the business itself generates customer and monetary value and how the system affects that process is important to keep in mind. Which parties should be involved in the process of understanding the business is highly dependent on what system or model is under development. Due to his current profession with a focus on government technology, much of the business understanding has to be conducted prior to tender offers. In some cases, analysing the business prior to tenders might be difficult due to lack of early stakeholder participation.

P4 have also experienced the difficulty of truly understanding the business purpose. According to P4, the difficulty can usually be explained by the stakeholders' lacking technical expertise in the customer segment. Similar difficulties may also be due to stakeholders using old or obsolete systems as a point of reference when discussing the purpose of future models, and thereby failing to express the true needs.

**Identified Codes** *Stakeholder analysis, Define Purpose, Domain Specific Problems, Business Requirements Analysis, Legacy Systems*

**P5** In general, when developing models as a consultant, the business case of the model is already stated. This means according to P5 that *"[...] we generally can start working on the data understanding phase straight away."*. However, in order to understand if and when the model performs well enough, it is important to understand the business requirements and values of the model.

To ease the understanding of data exploration some sort of data dictionary[1] might be

[1] a dictionary mapping data to its attributes and description

needed. This can either be provided by the company itself or created by interviewing data domain experts.

**Identified Codes** *Define Purpose, Business Requirements Analysis, Data Dictionary*

### 5.2.2 Data Understanding

In this section we will present the key takeaways of the interviewees' approaches when working with data understanding, i.e. the answers to question Q4.

**P1** In his interview, P1 states that the data used in his work is often user-generated in some way. He therefore does not necessarily turn to a domain expert nor a data scientist to gather more information when analyzing data. Instead, he often turns to the source of the data generation, in his case the users of various systems. The motivation behind this is that it sometimes can be difficult to determine whether anomalies in the data is caused during its creation, or during transformation somewhere before it reaches him. Discussion with the users and seeing their workflow when generating the data can therefore be helpful to understand both the users' needs and their motivation behind what they do. In some cases, it can be that the cause of the error is that the user purposefully uses a system in a way not aligned with the software's purpose, due to it being impeding to their own work. In this case, the easiest fix might be to adjust the existing software so that its data output suits both the direct users of the software as well as future modelling using the data.

To facilitate the process of understanding data, P1 suggests using a top-down approach during data exploration, *"Start by analyzing data on a macro-level, and successively work your way down to a more granular level"*. The reason for this, he states, is that outliers often have less impact on larger data sets, and relations can be more obvious compared to when analyzing smaller data sets. Using a top-down approach therefore lets you understand the high-level relations and use that knowledge when exploring the data on a more granular level.

**Identified Codes** *Current Procedures, Data Exploration, Align Environment*

**P2** When acquainting himself with data to be used in the model, P2 generally does not use a specific method. Instead he uses more of an ad hoc exploratory analysis strategy, but this also depends on how the data is stored and structured. In P2's case, most of the data he has been working with has been stored in different SQL databases. He recommends other persons in similar situations to create Jupyter Notebooks and exploring the data by writing queries and plotting diagrams of various kinds. According to P2, this will hopefully bring an understanding of *"What kind of data we're working with, how much data we have and the characteristics of it"*. By plotting the data using different explanatory and descriptive data sets will hopefully also bring an understanding of how the data is correlated and what seems to be reasonable values.

According to P2, it is also highly important that you acquaint yourself with the statistical data prior to exploring it using plots. That involves studying values such as the means, variances, and distributions of the different data sets. For example, financial data is usually exponentially distributed which might be important to keep in mind. Data of such kind might be easier interpreted by, e.g. plotting it as logarithms or in ranges.

P2 also recommends that people participating in the Data Understanding phase make use of already existing packages for analysis of the data. He says *"There exists a great amount*

*of free packages people can use when acquainting themselves with the data sets. Python, for example, has Seaborn which I use a lot.".* Making use of similar packages will according to P2 drastically reduce the time needed to get an overview of the data you are working with.

Additionally, P2 emphasizes that it is hard to realize when to move on to the next phase and the importance of iteratively moving between business understanding and the data exploration. He says *"The arrows in CRISP-DM is actually quite misleading. In real life, you move between the phases as new knowledge gaps arise. However, all of the phases exist – it is just that they might be conducted in a different order.".* Finding alternative solutions or other variables which might add predictive value is not unusual. Since they might be of greater value for the stakeholders than the first idea, it is important to have continuous contact with the stakeholders regarding the progress and new insights.

**Identified Codes** *Data Exploration, Exploration Scripts, Data Requirements*

**P3** As P3 is not directly involved in the data extraction process, he does not have that much experience of concrete steps to use for data exploration and extraction. Instead, in his experience, data understanding is the process of ensuring that interfaces for data retrieval as well as extraction are available. P3 prefers to use a workshop-like approach, where he discusses data requirements and availability with relevant stakeholders, to collectively figure out an optimal way of designing the data structure of the system.

**Identified Codes** *Data Exploration, Data Workshop, Model Requirement Analysis, Business Requirements Analysis*

**P4** When understanding the data, P4 usually first specifies the origin of the data. He says *"It's important to understand the origin of the data in order to clarify what process causes the data generation and thereafter how it should be interpreted and used".* Secondly, you should consider the data format prior to moving on to e.g. statistical data exploration. With statistical exploration, P4 means investigating statistics such as what values occur most often and how the data is distributed. This might be more intuitive when studying numerical values, but can also be done with other types of data such as text. P4 have some experience working with text models, and in such cases it might be of interest to study total word frequencies or bigrams[2].

P4 also recommends developing data exploration scripts for the data types most commonly used. Scripts can be used when conducting the statistical explorations presented above and much more. Creating scripts of such kinds may reduce the time of data exploration significantly. P4 adds, *"There exist a lot of free scripts, which very well can be used when exploring common data types".*

**Identified Codes** *Data Exploration, Identifying Interfaces, Data Relations & Format, Data Causality, Exploration Scripts*

**P5** When working with data understanding, P5 usually starts with exploration of a sub set of the data which will be used by the model in production. This will generally be the set of data used in a proof-of-concept (PoC). Firstly, studying the quantity and quality of the data and thereafter statistical values such a correlation, causality, etc.

**Identified Codes** *Data Exploration, Data Causality, Data Relations & Format, Proof-of-concept*

[2]A bigram is a pair of consecutive written units such as letters, syllables, or words.

### 5.2.3 Data Preparation

Data preparation is the process of cleaning and preparing data for use in the modeling phase. This includes activities such as removing duplicates, missing values, and possibly any outliers [6].

The purpose of data preparation is to ensure that available data is structured in a way that is useful for the purpose of the model. Depending on the source, raw data may contain noise, there may be values or attributes missing, and data originating from multiple sources may contain conflicting values. The various forms of errors that may need correcting are vast. It is therefore necessary to manually inspect and analyze what attributes and values are wanted and which are redundant or faulty. This process should preferably be conducted by a data scientist, as it requires a good understanding of the relationships between fields, attributes, and values. To successfully prepare data, it is vital to ensure that the previous steps of the CRISP-DM model have been adequately performed. A good business understanding is vital to be able to understand how the data may be manipulated without corrupting the data through errors such as wrongfully discarding values or drawing false relations between entries. Data understanding provides knowledge of the source of data, and its reason for inclusion in the model [55].

**P1** During discussion of problems that arise when working with multiple interfaces, interviewee P1 stated that an important factor to consider is standardization of interfaces. When developing you should strive to integrate as much of the transformation of data as possible into the data model, and therefore minimize the amount of transformation that is needed before the data enters the model pipeline. When asked about how this can be handled when you have multiple different interfaces producing similar data, interviewee P1 stated that there are two options. One way of handling it is to develop a standard interface that only allows input following a well-specified structure. This requires the data stemming from other systems to be transformed before being imported to the software database to follow the standard structure. A drawback of using this approach is that it shifts the responsibility for transformation of the data to its origin. It is therefore more difficult to ensure that the data is transformed correctly and that there is a stringent method of data preparation. This method can be useful when there are a large number of different formats for data representing the same values.

The second approach is to develop multiple different bridge-interfaces that serve to transform data from their original form to the form desired in the software database. This might increase the amount of work necessary to prepare data, but gives the data scientist working with data preparation more control over how the data is transformed before it enters the software database. This approach can be more time consuming, as it can require development of multiple interfaces for importing data.

**Identified Codes** *Identifying Interfaces, Data Warehouse*

**P2** P2 points out that it is important to keep in mind that the data you initially retrieve is often of bad quality. It might be due to bugs in the application posting the data, because of human errors of any sort, or something else. Therefore, sanitation of data is of great importance. With adequate domain knowledge, gained from the business understanding phase, it should not be too hard to identify incorrect outliers when for example plotting the data. P2 also emphasises the importance of iteratively moving between the data understanding and

preparation phases, i.e. repeating the process of retrieving, exploring and sanitising the data.

P2 adds that in a start-up, validating all interfaces, data sources, etc. might not be needed since everything needed for the model usually is easily accessible as well as stored and structured in the same way. However in larger corporations, especially company groups, validating the desired flow of data might be needed to make sure everything is taken into account when moving on to the data preparation phase.

**Identified Codes** *Data Sanitization, Data validation*

**P3** P3 states that he does not have that much experience working with actual data preparation, as they have a team designated for the task. He does however state the importance of ensuring that the data understanding is adequately performed and that the desired data structure and format is verified with relevant stakeholders as soon as possible. He states that in larger projects, extraction of data can often be a demanding task as it is dependent on external interfaces for data retrieval.

**Identified Codes** *Identifying interfaces*

**P4** According to P4, the data preparation needed may vary depending on the findings from the data understanding phase. But in most cases you have identified some problems with the data which you will have to solve. In the case of smaller data sets, it might be possible to sanitize the data manually, but in most cases automated sanitization is needed. He adds *"[...] especially in the case of machine learning, the idea is to scale the product using a continuous flow of new data. It is usually not possible to have employees manually sanitizing the data when the model is deployed."*, emphasising the importance of automated sanitization scripts. Automated sanitization scripts may according to P4 include functionalities such as reformatting data and data filtering.

In the case of data-driven applications, some sort of labeling or encoding of the data is also needed. P4 says *"In some cases, it is possible to automate this process as well but in many cases human interaction is needed in order to at least verify the labeling."*.

The last to consider is the final format of the data, P4 says. The model will have some sort of input, and it is essential that the data is formatted and structured in the correct way prior to moving on to the modelling phase.

**Identified Codes** *Data Sanitization, Data Categorization and Encoding, Sanitization Scripts*

**P5** P5 emphasises the importance of an iterative workflow between data understanding and exploration, and data preparation. When preparing the data you might find the specific data set to be too small for the purpose of the model, or of too bad quality. This means that you will have to reverse the process and find alternative solution paths. Additionally, data exploration, which is the most important part of data understanding, is not possible without first collecting the needed data. Data preparation also includes actions such as categorization, encoding and sanitization of the data. The amount of work needed for each of these actions is highly dependant on the data and what modelling technique will be used.

During data preparations, it is also important to investigate the sensitivity of the data. Especially within the domain of financial technology, the data is retrieved from a large amount of data warehouses. P5 adds *"To study the sensitivity of data sources and data sets is important. It might be investigating whether it will be difficult to obtain a real-time data flow in the future."*. P5

also stresses the fact that in many cases within the financial industry, data owners are not willing to mediate data due to regulations such as GDPR, or other factors.

Depending on what type of data is used and the origin of the data, it might be feasible to create a final data warehouse when finalizing the data preparation phase. However, in some cases due to the data owner issues stated above, it might not be possible to collect and store all the data in-house.

**Identified Codes** *Data Exploration, Data Sanitization, Data Categorization and Encoding, Data Warehouse, Data Quality, Data Relations & Format*

### 5.2.4 Modelling

In this section we will present the key takeaways of the interviewees' approaches when working with modelling, i.e. the answers to question Q6.

**P1** P1 states that in his experience, selection of modelling techniques can often involve multiple stakeholders. The reason for this being that the data team often has a good understanding of what techniques can lead to what results, and the business team that requested the model has a good understanding of their requirements. When the previous phases of the process model have been completed, there can however still be some decisions left to make. He states that his team sometimes can gain valuable understanding during the data collection and understanding process that can lead to them making suggestions for other models than were initially required, if they believe that it might be of more value to the team.

**Identified Codes** *Selecting Modelling Technique, Multiple Models*

**P2** According to P2, many of the problems in the modelling phase are created due to a lack of understanding of the causality of the data. By not including causality, the model will often only work when using the exact same data as during the learning phase. P2 adds *"It's important to understand the data generation process, and what causal relationship have caused the data."*. The causal relationship can either be included in the model, or be used merely to evaluate and better understand the model's output.

P2 points out that the process of identifying causal relationships can sometimes be relatively easy when studying mathematical or physical events. However, when building models within the domain of healthcare or financial services, it may be very difficult. He says *"Models often fail, and even so when taking causality into account. However, if you do not take causality into account, it is almost certain it will fail at some point of time."*.

**Identified Codes** *Data Causality, Selecting Modelling Technique*

**P3** According to P3, one of the most difficult parts of modelling is understanding when the model is good enough for use. To handle this, he tries to set clear target goals before beginning to work. He states that an important factor to consider is that there always exist ways of improving the model. However, you need to realize when you have done a good enough job and could rather be spending your time on tasks with greater return regarding the end users satisfaction with the system.

**Identified Codes** *Define target metrics, 80/20*

**P4** According to P4, the first thing to do when starting the modelling phase is to research what type of modelling techniques have been utilized successfully in the same domain with similar data. By doing so, you may drastically reduce the time spent in the modelling phase. P4 adds *"It might be useful to investigate the effectiveness of multiple modelling techniques, but by investigating similar projects in the domain you might at least get some idea of what should work and rule out some others."*.

**Identified Codes** *Selecting Modelling Technique, Multiple Models, Previous Models*

**P5** According to P5, it might be needed to further specify correlations and causality of the data when working on the model. However, most of these investigations should be conducted prior to the modelling phase.

Prior to starting the model building, P5 suggests dividing the total data set into three sub sets, namely (1) Training data set (2) Evaluation data set, and (3) Testing data set. The idea of splitting the data into three sets is to train and evaluate the models on the same basis utilizing data sets 1 and 2, and lastly conducting a final testing of the best performing model utilizing the third data set. We note that this approach to splitting the data set is in line with established best practices in data science.

During the modelling phase, it should be decided what top level features should be implemented during this iteration. As the features are decided, it is time to investigate what algorithms and modelling techniques could be used in order to create them. A good practice is to build and later evaluate multiple algorithms and modelling techniques in order to asses which are the best suited in this case.

P5 adds *"The work when developing data-driven applications is truly iterative. However, it is at this phase the work turns more straight forward. Here, you have to decide what features to include and what models to build and test."*.

**Identified Codes** *Data Causality, Selecting modelling technique, Data Subsets, Multiple Models*

### 5.2.5 Evaluation

In this section we will present the key takeaways of the interviewees' approaches when working with evaluation, i.e. the answers to question Q7.

**P1** Interviewee P1 stated that one shortcoming of the original CRISP-DM model is that it emphasizes validation as a step to be conducted after the modeling phase. The interviewee on the contrary pointed to the need for continuous validation through all steps of the process model. The interviewee also stated that an important aspect to consider when validating the system model is that feedback can originate from multiple different instances. One example of this is that the use case of the development can make feedback from one stakeholder more important than feedback from another. It is therefore important to consider this when evaluating the system model. The interviewee therefore suggests that evaluation can be a part of all steps of the CRISP-DM model, instead of a separated entity.

The interviewee also states that evaluation can have multiple different goals depending on the context. It is therefore important to clearly state *what* the goal of the evaluation is before it is conducted. This is also an aspect that can be improved by conducting validation after each step in the CRISP-DM model, e.g. evaluating the business understanding after it is conducted to ensure that it was adequately performed.

**Identified Codes** *Continuous validation, Validation & Verification, Model Metrics*

**P2** According to interviewee P2, the evaluation phase of CRISP-DM can be divided into two different types of evaluation, i.e. evaluation of the business requirements and evaluation of the model itself. This however, differs from the standardized view on evaluation in the CRISP-DM model, where evaluation refers to mere evaluation of the model and process themselves [19, 46].

Regarding evaluation of business requirements, P2 says *"It's important to conduct validation in order to conclude whether the model is good enough to create the business values which initially started the project"*. Thus, it is important to closely relate all the evaluation activities to the domain knowledge previously obtained during the business understanding phase, and assuring that knowledge is correct. The iterative movement between interviewing stakeholders and e.g. understanding the data and building the model is usually due to the assurance of the domain knowledge.

With evaluation of the model, P2 refers to model metrics such as accuracy and ROC curves[3]. Some of these metrics should be based on the business requirements and be the final decision whether the model performs well enough to be deployed, and some may be mere technical metrics such as response time. However, creating business related metrics without previous validations and assurance of business requirements may lead to misleading evaluations of the model.

P2 also stresses the difficulty of determining when the model performs well enough and how it will perform in the future. He says *"It is hard to know how well the model will perform in real-time environments. [...] Especially since the characteristics of data changes over time, and this is especially true for the financial market."*. Thus, before moving on to deployment, he recommends determining which model metrics and thresholds should be monitored when the model is in operation. The thresholds can be used to decide if and when it should be taken out of use and they should therefore be closely related to the business requirements.

**Identified Codes** *Model metrics, Future Acceptance Criteria, Validation & Verification, Continuous Validation, Monitor Acceptance Criteria*

**P3** Interviewee P3 states that evaluation of the development process is something that he often does along with the other phases of the CRISP-DM model. In his experience, it is often useful to validate the findings of each step with stakeholders as soon as possible, as it both gives the stakeholders more insights into the development process, as well as minimizing the risk of development being headed in the wrong direction.

**Identified Codes** *Continuous Validation*

**P4** Regarding evaluation, P4 states that *"The best type of evaluation is some sort of model metrics"* and adds that such metrics should be closely related to the business problems defined in the business understanding phase. However, when testing the model it is important to keep in mind what data is used. For example, is it a fraction of the data sets which will be used in production, is it the complete data set or just a data set generated for the purpose of testing. P4 states that any data set may be used while evaluating the model as long as it

[3]A ROC curve is a plot which illustrates the diagnostic ability of a binary classifier dependant on varied discrimination. The ROC curve is created by plotting the true positive rate against the false positive rate at various threshold settings.

is representative of the data used in production. By using a non-representative data set, the model metrics will be misleading and fail to reach the needs of the stakeholders.

Other performance metrics may also be of interest in order to assure that the business requirements are met. For example, some larger models may not be feasible when deployed. According to P4, such limits may be due to the amount of data which needs to be processed in a specific time frame. In order to create correct performance metrics, it is therefore important to understand the current procedures of the business which will use the model.

Before moving on to the deployment of the model, P4 stresses the importance of creating *"[...] good enough criteria for the model"*. P4 means that it is difficult to create a fully representative data set when studying the model metrics. Hence, the true testing of the product will happen as it is deployed. Thus, creating some sort of acceptance criteria for the model in production is good practice.

**Identified Codes** *Model metrics, Future Acceptance Criteria, Representative Data, 80/20, Verification & Validation*

**P5** According to P5, evaluation shall be split into evaluation and testing. The difference is that all models are evaluated using the evaluation data set, and the best performing model is thereafter tested using the test data set. When testing the best performing model, it is important to assure that the test data set fairly represents the total data population in production.

Regarding both evaluation and testing, P5 recommends defining model metrics closely related to the value creating process of the model. What metrics are best suited in the case of a specific model may depend on the data, modelling technique, and much more.

**Identified Codes** *Model metrics, Validation & Verification, Data subsets, Representative Data*

### 5.2.6 Deployment

Deployment refers to the process of constructing a strategy and plan for the deployment of the final product [20]. The deployment strategy should include a plan for monitoring and maintenance of the data and the model. A poorly developed plan for deployment can lead to issues with periods of incorrect usage of data, as the product user might not fully understand its use. A final report of the development process can be included in the deployment process, as it provides useful insights in the system model for its users.

When prototyping, a plan for deployment might not always be necessary. It can however be useful to prototype with deployment in mind, to facilitate reuse of the prototype in an eventual final product. Making a report of a prototyping iteration can also be useful, as it can serve as documentation in future work. A final report of a prototyping iteration does not however need to be as comprehensive as a final report of a data mining process due to the often short time period being spent on a prototyping iteration.

**P1** As the models developed by P1 often are created by request of other departments within the company, metrics for use after deployment are sometimes not of the utmost importance. Instead, they often rely on user feedback to complement their own evaluation of the model.

P1 also states the importance of understanding that deployment rarely is the last step of the development of data-driven applications. In his work, he often deploys a model once he

feels that it has reached a "good-enough" state, and often returns to it in the future to both make improvements and to often reuse parts of the model in new applications.

**Identified Codes** *Plan for Reassessment of Model, User feedback*

**P2** In Section 5.2.5 Evaluation, P2 mentions that the characteristics of data changes over time. Thus, it is important to evaluate the model, using model metrics and decided thresholds, when it has been deployed. These thresholds can be used to determine whether the model should be modified or taken out of use. Creating automatic warnings which can be sent to the responsible parties may be useful when monitoring such thresholds, he adds.

**Identified Codes** *Plan for Reassessment of Model, Monitor Data Characteristics, Monitor Acceptance Criteria*

**P3** When deploying a model, P3 states that it is important to not consider it the end of development. In his experience, a plan for follow-up of the system is very valuable, as it ensures that responsibility is assigned to a stakeholder to ensure the future use of the system. P3 also stresses the importance of considering the context in which the system will reside, and have a plan for issues such as "if an interface is to be upgraded or changed, what impact will that have on the system model?".

**Identified Codes** *Plan for Reassessment of Model, System Context*

**P4** P4 does not have much experience in deploying data-driven applications. However, he recommends to keep in mind the same things as when deploying any software. He finds that there are no major differences in deployment of various types software. Things such as certificates, servers and of course a working versions of the software should be in place prior to moving on to the deployment phase.

In the case of data-driven applications, P4 says *"Someone should also be in charge of monitoring the set up good-enough criteria."*. Failure to meet the defined needs of the customer while in production might not be as easily identified as with conventional software.

**Identified Codes** *Monitor Acceptance Criteria*

**P5** P5 stresses the fact of introducing a Service Level Agreement (SLA) prior to delivery and deployment of the data-driven application. An SLA is a contract between the service provider and the service user. Particular aspects of the service, e.g. quality, availability and responsibilities, should be included in the SLA.

As the model is deployed, it is important to monitor its performance. The characteristics of the data change over time, especially when working with financial data and for example text data. In the case of text data, new words are created and the syntax of everyday sentences might be changed. P5 adds *"The most important thing to remember is that a model is never fully finalized. Relearning is in almost all of the cases a must."*.

According to P5, it is also good practice to evaluate the collection and storage of real-time data after the model has been deployed.

**Identified Codes** *Plan for Reassessment of Model, Monitor Data Characteristics, Monitor Acceptance Criteria*

## 5.3 Domain Specific Challenges

In this section we will present key takeaways from the discussions regarding how the interviewees would deal with the domain specific challenges identified in Section 4. These discussions are based on question Q10. The challenges discussed were (1) Large number of interfaces, (2) Uncertainty of data, and (3) Difficulty of defining main purpose.

**P1** *Large amount of interfaces*: When dealing with multiple interfaces, the importance of understanding the data increases. P1 states that he usually tries to be as flexible as possible with regards to compatibility with interfaces required in the system context. His suggested solution is therefore to ensure that the software can handle input in whatever format the interface requires. This ensures that the data is not transformed in any way unknown to him before entry into the data lake[4], which otherwise might possibly be a source of error that can be very difficult for him to identify.

*Uncertainty of data*: To deal with the issue of data being unreliable as suggested by the Unit A characterization of the problem domain, P1 suggests the creation of a data lake and creation of a standard structure for data input. He also mentions an increased importance of ensuring that all stakeholders are aligned as suggested by him earlier. The reason for this being that data science and modelling in itself cannot provide a solution for data being unreliable, and it is instead a problem best solved by discussion with domain experts to come up with a strategy for how to deal with the issue.

*Difficulty of defining main purpose*: When discussing the problem of multiple definitions of risk, and different stakeholders wanting different ways of quantifying it, interviewee P1 does not think the right approach is to offer multiple different ways of quantifying the data. The reason for this being that it can be confusing for users of the system since the same data then can give multiple different outputs. Instead, he stresses the importance of aligning the stakeholders during the initial business understanding. He states that there being multiple "correct" outputs can lower the credibility of the model.

**P2** *Large amount of interfaces*: When working with multiple data sources, P2 usually starts by creating a so called "data lake". The data stored in a data lake must not have a defined purpose in the future model. Additionally, in the data lake all data is stored in its raw format. Creating a data lake is always followed by the process of data quality assurance, i.e. data sanitization. The end goal when working with multiple interfaces will always be to create a single data warehouse [5], with the needed data structure and format.

*Uncertainty of data*: According to P2, it is hard to define a general approach for dealing with uncertainty of data. In the case of FX risk exposure applications, where the uncertainty of data originates from direct human interaction with the data, forecasts, etc., P2 says that the best approach to ensuring data quality is do conduct a meticulous data sanitization.

P2 also advocates the idea of aligning the system environment in order to ease the process of dealing with uncertainty of data and data sanitization. By aligning the enterprise's view on different data sets, it would be easier to create "logical rules" to identify incorrect data and thereby creating automated sanitization scripts.

[4] A site for collection of data in its raw format

[5] A site for storage of structured data

*Difficulty of defining main purpose*: P2 had no direct idea of how to solve the problems related to the difficulty of defining a main purpose.

**P3** *Large amount of interfaces*: P3 has some experience working with large amounts of interfaces. He says *"In general, banks, ERP systems and other types of larger system providers usually offer some kind of sandbox environment in which you can explore the data collection process."*. Thus, he recommends to get familiar with the individual data responses and formats prior to construction of any in-house storage.

Gathering all needed data in its raw format, in a common data lake, prior to sanitization and reformatting is a good practice. By doing so, you avoid losing important metadata. Exploring the data in its raw format is also a good practice, since it brings a better understanding of the origin and what process generates the data.

*Uncertainty of data*: If uncertainty of data exists, you should according to P3 try to identify what causes the uncertainty and ideally solve it. However, that might not always be possible. For example, if the uncertainty of the data is related to forecasts and predictions of future sales, as the case with an FX risk exposure application, the decision whether or not to include such data in the model should be taken by the stakeholders. When working in the domain of financial technology, uncertainty of data is the root cause of many problems. Nobody knows how the future will play out, and its up to the stakeholders if the model should take such predictions into account or not.

*Difficulty of defining main purpose*: When asked about the problem of there being multiple conflicting business goals identified during the business understanding phase, P3 suggests the need for compromises that sometimes arise during development with multiple different stakeholders. When facing similar problems in his work, he tries to make all stakeholders agree on a common goal that can be used during development.

**P4** *Large amount of interfaces*: P4 has experience working with a large amount of interfaces, and especially in the domain of financial tools. He says *"At my previous employer we used Apache Kafka for this purpose. It is a system built to help developers solve problems related to a large amount of interfaces."*. He recommends people working with financial services to utilize software solutions like Apache Kafka[6].

*Uncertainty of data*: Regarding uncertainty of data, P4 emphasises the importance of understanding the origin of the data. By doing so he says, the process of understanding if and how it may be used will be easier. Adequate data sanitization is also needed, and as he pointed out in Section 5.2.3, automated data sanitization is preferred.

*Difficulty of defining main purpose*: According to P4, when trying to define complex problems to be solved by software a good idea is to partition the problem into sub-problems and discuss how they might be solved individually. P4 adds, if one still finds it hard to solve the sub-problem it might be suitable to investigate alternative solutions to a data-driven application. He says *"It is essential to be able to validate the result of a data-driven application. Without understanding the problem, you cannot decide whether the model functions as desired or not."*.

**P5** *Large amount of interfaces*: With P5's experience in working with data-driven applications in the domain of financial services, she agrees on the issues related to the large amount

[6]https://kafka.apache.org/

of interfaces. She also adds *"Many financial service providers are aware of these issues and are currently working on solutions. They're creating data catalogs, data dictionaries, etc. However, it will probably take a long time until these issues are fully solved."*. Prior to initializing the work of integrating multiple interfaces, P5 recommends gathering all available documentation provided by the data owner.

Unfortunately, there exists no shortcut when working with multiple interfaces. She adds *"If the data is located in an old mainframe computer, the only way of obtaining the needed data is to gather it somehow."*.

*Uncertainty of data*: With adequate data sanitization, bad quality data should be removed from all of the data sets used when building and evaluating the model. It might also be preferable to discuss how similar issues should be handled in the real-time data flow when deployed.

*Difficulty of defining main purpose*: In general, P5 does not work closely with defining the business cases and requirements. As mentioned in Section 5.2.1, in her context, the business case of the model is often provided by the customer prior to the project start. However, partitioning problems into sub-problems can usually be a good practice when working with complex tasks.

## 5.4 Effects of Prototyping

In this section we will present the key takeaways regarding how the interviewees would adapt their process when prototyping models.

**P1** P1 states that in his experience, his workflow is not really impacted by whether he is prototyping a model or not. He states that the process of developing data-driven models is already agile, and therefore is very similar to that of prototyping whether the purpose is to develop a prototype or not.

**P2** P2 says that depending on what kind of prototype is meant to be developed, the approach in CRISP-DM might be altered somewhat. For example, if the purpose of a prototype is to investigate whether a model idea would create customer value or not, it might be good enough to create a simple mockup. However, when prototyping data-driven applications and models the most common type of prototype is a PoC, meaning that all of the phases have to be conducted.

When asked why the approach does not differ when prototyping, P2 emphasises that the nature of data science is very similar to prototyping. Development of data-driven applications is about understanding a business problem, investigating whether there exist data to support some kind of solution and how that solution may be implemented. He adds, *"The nature of data science is exploratory, which in many cases is the main purpose as when prototyping a product"*.

Depending on the complexity of the system and the time frame, it might be suitable to prototype a part of the system addressing a sub-problem instead of the full system at once.

**P3** P3 states that he does not have that much experience working with prototyping specifically, and that it instead is something they use as a tool for validation during software devel-

opment. His opinion is that it should not affect his development process much, as he believes that development of a prototype can be seen as the first iterations of an agile workflow, where an MVP often is created during the early phases.

**P4** According to P4, the main goal of prototyping in the domain of data-driven applications is usually to check the feasibility of a solution or to explore alternatives. The least wanted outcome when prototyping is to realise the impracticability of a solution you have invested a great amount of time into.

**P5** According to P5, when prototyping data-driven applications, the approach is quite standardized in terms of three phases. In first phase, a PoC has to be built. In terms of data use and functionality, the PoC is quite limited. Usually when developing a PoC, a small sub set of the total data population can be used. For example, when building a PoC-model it might be preferable to apply the model on a single product instead of the whole assortment of products. In general, the approach specified in the CRISP-DM process will not be modified to suite development of a prototype. However, the last phase, i.e. deployment, will only be conducted if the prototype is found to be good enough to be used by the customer. If the PoC is successful and sufficiently predictive signals can be identified, then the next phase will be to develop an MVP.

When developing an MVP, it is important to include a larger set of data compared to the data used in the PoC-model. Usually, during this phase of prototyping the first usable GUI is developed accompanied by adequate back-end functionalities. As the name states, an MVP could be used by the customer in order to obtain feedback and to validate the product as a whole.

The final phase will include monitoring the model's performance in production, adding desired functionalities and expanding the data sets.

## 5.5 Synthesis

This section describes the results from the content analysis. First, we describe all 40 identified codes identified in the interviews. Second, Table 5.2 maps interviewees to the codes. Finally, we discuss patterns in where the codes were identified, e.g. contextual factors that might have an impact.

**Stakeholder analysis** Investigating and mapping stakeholders of the system.

**Define Purpose** Defining and stating the purpose of the system.

**Align Environment** Ensuring that all parties are in agreement over the meaning of terms and data relevant to the project.

**Domain Specific Problems** Discussing and stating domain specific challenges posed by the problem domain in which the system resides.

**Business Requirements Analysis** Summarizing and prioritizing the stakeholders' expectations on the system.

**Consult Domain Experts** Consulting domain experts to gain deeper knowledge of the domain in which the system resides.

**High Level Design Prototype** A prototype of the system showcasing the desired use and functionality of the system, e.g. design sketch, mockups, use cases etc..

**Legacy Systems** A system considered outdated in relation to the system model.

**Data Dictionary** Creation of a dictionary mapping data to its attributes and description.

**Current Procedures** Investigating and specifying current procedures used to fulfill the business purpose of the system.

**Data Exploration** The process of exploring the data set to gain a deeper understanding of its characteristics.

**Exploration Scripts** Scripts used to automate parts of the exploration process.

**Data Requirements** Condensing the business requirements into data requirements.

**Data Workshop** A workshop with the specific purpose of investigating available data.

**Identifying Interfaces** Identifying available and necessary interfaces that the system will interact with.

**Data Relations & Format** Investigating the relations among data as well as its format.

**Data Causality** Investigating causality relevant to the data set.

**Proof-of-concept** A low level design prototype used to showcase the feasibility of the approach.

**Data Warehouse** Creation of a structured database that contains all data required by the system

**Data Sanitization** Cleaning and formatting of an unstructured data set.

**Data Validation** The process of validating data.

**Data Categorization & Encoding** The process of labeling data.

**Sanitization Scripts** Scripts used to automate parts of the sanitization process.

**Data Quality** Investigating data with regards to quality requirements, such as the ESS standard.

**Selecting Modelling Technique** The process of selecting modelling techniques to use in the modelling phase.

**Multiple Models** Creation of multiple models rather than a single one.

**Define Target Metrics** Stating the threshhold values that the model's evaluation metrics need to reach.

**80/20** A theory stating that 80% of the work requires 20% of the total effort, and the remaining 20% of the work requires 80% of the total effort.

**Previous Models** The usage of experience gained in previously developed models.

**Data Subsets** Dividing the data set into smaller subsets.

**Continuous Validation** To conduct validation in smaller sub steps rather than validating as a separate phase.

**Validation & Verification** The difference between evaluating the usability of the system (Validation) and the functionality of the system (Verification).

**Model Metrics** The metrics used to measure the success of the model.

**Future Acceptance Criteria** Treshhold values used to ensure that the model does not degrade to the point where it is no longer useful.

**Monitor Acceptance Criteria** The process of evaluating the acceptance criteria according to future events that might impact the models usage.

**Representative Data** The process of ensuring that a collected data sample is representative of the entire population.

**Plan For Reassessment of Model** Creation of a plan for when the model needs to be reassessed and possibly retrained.

**User Feedback** Usage of user feedback to evaluate the model.

**Monitor Data Characteristics** The process of monitoring the data characteristics and evaluate events that might impact its characteristics.

**System Context** The contexts in which the system resides, e.g. depicted as a context diagram.

Table 5.2 illustrates the mapping between interviewees and the codes identified in Section 5.

During the validation of previous interviewees' suggestions, our interviewees had no input that would discredit the previous suggestions. This suggests that the identified codes are all relevant in the context of prototyping data-driven financial decision support tools. This also suggests that the identified codes are not contradictory to one another, meaning that they can coexist in a development methodology. Our suggested concretization of CRISP-DM has therefore included as many of the identified codes as possible.

The spread of codes identified among interviewees might also suggest that there exist multiple ways of achieving the same end goal in regards to prototyping data-driven financial decision support tools. Another reason for this might be that it is difficult for the interviewees to generalize their methodology, as it might be very dependent on the specific context. Data exploration for example might require very different sub-activities depending on the data at hand and the business context.

**Table 5.2:** Mapping interviewees answers to identified codes. An X in the table represents that the interviewee has discussed the subject.

| | P1 | P2 | P3 | P4 | P5 |
|---|---|---|---|---|---|
| Stakeholder analysis | X | X | X | X | |
| Define Purpose | X | X | X | X | X |
| Domain Specific Problems | X | X | | X | |
| Business Requirements Analysis | X | X | X | X | X |
| Align Environment | X | | | | |
| Consult Domain Experts | X | | | | |
| High Level Design Prototype | | X | X | | |
| Legacy Systems | | | | X | |
| Data Dictionary | | | | | X |
| Current Procedures | X | | | | |
| Data Exploration | X | X | X | X | X |
| Exploration Scripts | | X | | X | |
| Data Requirements | | X | | | |
| Data Workshop | | | X | | |
| Identifying Interfaces | X | | X | X | |
| Data Relations & Format | | | | X | X |
| Data Causality | | X | | X | X |
| Proof-of-concept | | | | | X |
| Data Warehouse | X | | | | X |
| Data Sanitization | | X | | X | X |
| Data Validation | | X | | | |
| Data Categorization & Encoding | | | | X | X |
| Sanitization Scripts | | | | X | |
| Data Quality | | | | | X |
| Selecting Modelling Technique | X | X | | X | X |
| Multiple Models | X | | | X | X |
| Define Target Metrics | | | X | | |
| 80/20 | | | X | X | |
| Previous Models | | | | X | |
| Data Subsets | | | | | X |
| Continuous Validation | X | X | X | | |
| Validation & Verification | X | X | | X | X |
| Model Metrics | X | X | | X | X |
| Future Acceptance Criteria | | | | X | |
| Monitor Acceptance Criteria | | X | | X | X |
| Representative Data | | | | X | X |
| Plan for Reassessment of Model | X | X | X | | X |
| User Feedback | X | | | | |
| Monitor Data Characteristics | | X | | | X |
| System Context | | | X | | |

# Chapter 6

# Detailed CRISP-DM (RQ2)

In this section we answer RQ2: *"How can CRISP-DM be applied when prototyping a data driven FX risk exposure application?"*. This was done by creating a detailed CRISP-DM process model for prototyping in the domain of data-driven financial decision support tools, based on the findings in Chapter 5. An overview of the detailed CRISP-DM is presented in Figure 6.1. An obvious difference compared to the original CRISP-DM presented in Figure 1.1 is that we depict the six main steps without any arrows – instead we emphasize that development can move between the steps in any order during prototyping.

To summarize RQ2, we have created a detailed CRISP-DM with 32 recommended process methods. It is important, as mentioned by almost all interviewees, to stress the fact that development of data-driven applications and data science in itself is characterized by an agile workflow. The detailed CRISP-DM process phases do not have to be conducted in a specific order, which is why we have chosen not to draw workflow arrows in Figure 6.1. In relation to previous work on creating detailed CRISP-DM [59, 57], the goals of each phase seems to be the same. However, recommended process methods differ to some extent due to the goal of detailing the CRISP-DM process model. For example in the article published by Schäfer et al., the process model was detailed on the basis of including quality assurance practices. Thus, the "Quality Management"-CRISP-DM includes several activities for assessing the need of tools, documentation, and testing. In general, the detailed CRISP-DM provided in this thesis has more in common with the Generalized-CRISP-DM published by Shailesh et al. [59] in 2021. However, the GCRISP-DM does not include recommended process methods for the Business and Data Understanding phases, which we have provided with a focus on requirements engineering.

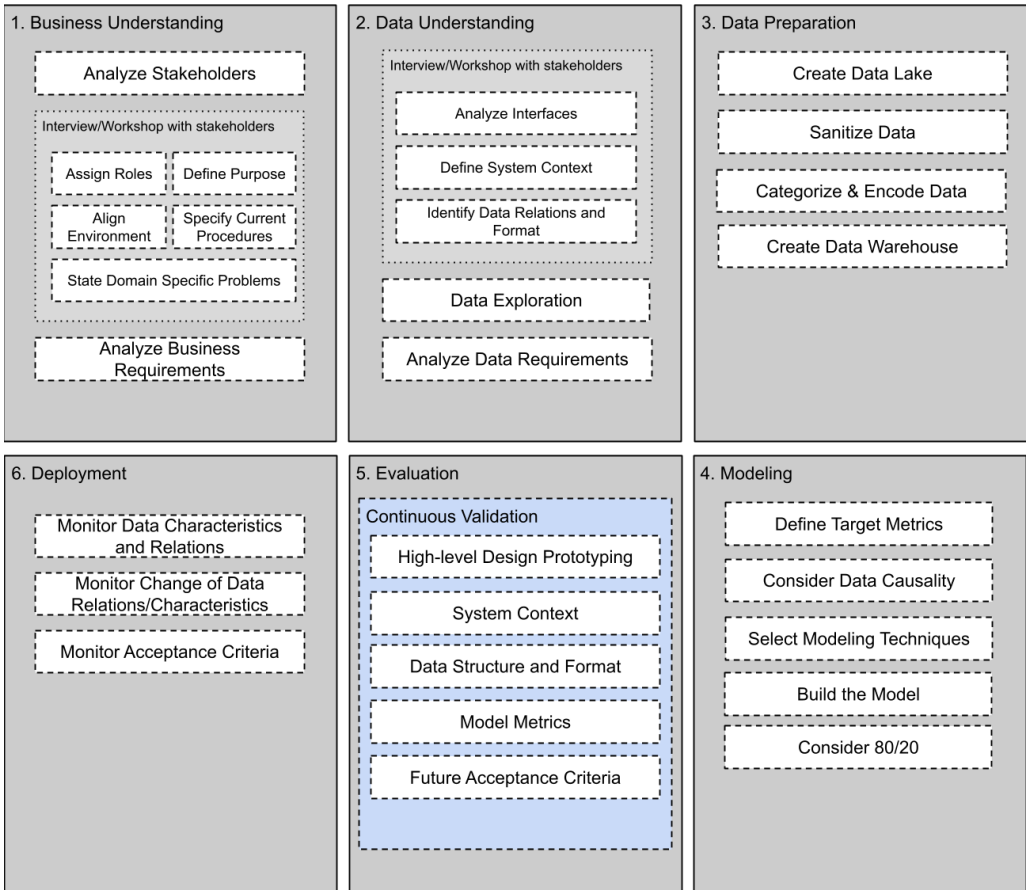



**Figure 6.1:** An overview of our detailed CRISP-DM model. White boxes refer to activities further described below, I.e. our main contribution to the detailing of CRISP-DM. The light blue box refers to the concept of continuous validation further described in Section 6.5.

## 6.1 Business Understanding

The process of identifying stakeholders of the system can often be difficult, as not all stakeholders necessarily have an equal interest in the system. Authorities were identified as a stakeholder of the system, however, they have no interest in the actual design of the system as long as it meets their legal requirements. Comparing authorities to for example the system owner on an equal basis may therefore be considered misleading from the perspective of prototyping. Assuming that the purpose of the system is to remain legal, authorities may therefore always be considered a stakeholder in all systems as the alternative is a software which is in some way illegal.

To ensure that we had accurately identified the stakeholders of the system, we continually validated the findings during interviews with the system owner as well as the companies from the case study. This was successful as their lack of additions to the stakeholder analysis indicated that all major stakeholders were already identified, i.e. we observed a form of qualitative saturation [56]. Stakeholders that may have been missed can therefore be assumed to have such a low stake in the system that they may as well be disregarded.

In total, we propose seven additional substeps in the Business Understanding phase of detailed CRISP-DM.

**Analyze Stakeholders** The findings from our stakeholder analysis suggest the problem domain is shaped by the different needs of different users of the platform. The treasury is often tasked with oversight of the more operative parts of the organization's business, while the board, as well as the CFO, are mainly concerned with the results of the operational work as they are reflected in the company's financial reports and statements. In the domain of FX risk exposure quantification this was reflected in that the treasuries were more interested in features that can break down their FX risk exposure, and identify the root causes of their exposure. The board members and CFOs were however more interested in features that facilitate summation of the companies' total exposures. As sharing of financial data in companies can be sensitive in some organizations, it can therefore be advisable to design systems with different access levels depending on the user's position within the organization, in line with common cybersecurity practice [17]. This enables the use of a shared database for all data no matter the organization's policy regarding sharing of financial information.

**Assign Roles** When developing data-driven applications certain expertise can be assigned to roles according to Figure 4.2. This allows associated stakeholders to focus on one aspect of the development methodology instead of requiring that all stakeholders are equally invested in all different parts of the development.

Our suggested roleset that should be assigned at the start of the project are as proposed by Hesenius et al. [35]:

- Domain Expert
- Data Scientist
- Data Domain Expert
- Software Engineer

Defining roles and delegating responsibilities for the different parts of the model ensures that the steps of the detailed CRISP-DM model are conducted in a way that is both feasible from a data perspective, as well as fulfils the needs posed by the problem domain at hand.

**Define Purpose** We suggest the activity of defining the business purpose of the system model when developing using our detailed CRISP-DM model. As suggested by all interviewees (P1 to P5), this is incredibly useful as it ensures that the business goal is clearly stated, and it can give valuable insights into the business needs that will form the basis of the model requirements and model metrics.

**Align Environment** Alignment was not anything we identified as a particular challenge during analysis of Unit A, but instead something which was suggested to be added to the Business Understanding phase by P1 in Unit B. Here we define alignment as *the activity of aligning the company's view of individual data categories, and other qualitative and quantitative metrics regarding the system.* The purpose of alignment is to give the parties participating in

the elicitation process a common ground to discuss system requirements, in which they all understand what specific terms mean, and to make it easier for the data scientists to understand, categorize and clean the data.

**State Domain Specific Problems** A useful activity during the Business Understanding phase can be to identify and state the domain specific challenges that will need to be managed during the development project. The reason why we identified this as a separate activity is that the domain specific challenges can seem minor from an outside perspective, while having a great impact on the resulting development process according to the interviewees. Stating the domain specific challenges also serves to validate an accurate business understanding, as the challenges will be heavily related to the specific business domain.

**Specify Current Procedures** One way of eliciting aspects of the required functionality of the software solution is by investigating the current procedures for fulfilling the business function the software seeks to replace. This is an effective way of both gaining knowledge of the specific use case of the system, as well as the target metrics required for the model to be viable. The current procedures can then serve as a comparison when evaluating the model from the perspectives of both costs and accuracy. As all interviewed data scientists stated, it can sometimes be difficult to analyze the viability of a model without some concrete process that can serve as a comparison. Our suggestion is therefore that when possible, current procedures for fulfilling the business functionality should be investigated and stated as well as quantified regarding viability.

Depending on the specific use of the model, there might not always exist current procedures that mirror the use case of the model. If so, an alternative is to aggregate multiple different business functions that collectively reflect the functionality fulfilled by the model. This method can be more time consuming, as it might be difficult to quantify the functionality of a collection of different business functions. Our suggestion is however to whenever possible consider alternate business functions that fulfill the same purpose as that of the system, as it would provide valuable insights into the alternatives of the system and can therefore function as complementary frames of reference.

**Analyze Business Requirements** As the final substep in the Business Understanding phase, we propose business requirements analysis. The aim of this activity is to concretize the requirements of the business domain into target requirements for use in the modeling phase. This activity should involve appropriate stakeholders and business domain experts, to also serve as validation of the business understanding phase.

## 6.2 Data Understanding

As stated in Section 4.3.2 Available data, the interfaces available for data extraction are very dependent on the specific information systems in use at an individual organization. It is therefore very difficult to generalize the findings into a software solution suitable for all potential customers. We therefore had to consult the system owner as well as generalize based on available information.

We propose five new substeps in the Data Understanding phase of the detailed CRISP-DM model.

**Analyze Data Requirements** When developing data-driven applications, a crucial step is to identify the data necessary for the system. An adequately performed business understanding should provide clear business problems that can be transformed into software requirements. These requirements should then be analyzed from the perspective of the data required to fulfill the software requirements. The aim of the analysis is to clearly state the data requirements necessary for successful modeling according to the business goals stated in the Business Understanding phase. The data requirements analysis should include data requirements, data format requirements, and data quality requirements if applicable.

As stated by some of the interviewees, *"[...] the data is what the data is, and that it is up to the data scientist to make the most out of it"*. This view on data is also observed by Borg and Vogelsang [65]. In their article, they discussed regarding the use of the data requirement standards ISO 25012 and 25010, but found data scientist to not be fond of them. They add that one data scientist said "*You could try, but it won't help*", suggesting that defining concrete data requirements may not be the ideal path to take in the domain of data-driven application.

However, as mention in Section 2.3, the Bank of England uses the ESS definition of data quality, suggesting that some data characteristics are desired in the domain of financial services. Hence, we suggest taking these characteristics into account when conducting the Data Understanding phase.

More regarding the ESS characteristics and how the may be taken into account throughout the whole detailed CRISP-DM process will be discussed in Section 7.4 *Data Quality*.

**Define System Context** The system context is the context in which the system resides. This environment should be specified to ensure an accurate understanding of the interfaces interacting with the system model. This system context should include possible interfaces that the system will interact with, and state their nature and defining characteristics.

**Analyze Interfaces** As suggested in our results, analyzing available interfaces for data extraction as well as their defining characteristics is a necessary step in all data-driven software. Different interfaces commonly provide different ways of data extraction. The system context, as described above, can be used in order to ensure that all interfaces are taken into account during this process.

**Identify Data Relations and Format** As P1 stated, one of the main tasks of the data understanding phase is to identify the origin, relations and formats of data. The purpose of this task is to gain an understanding of how changes in one sort of data can affect others, i.e. causal relations as stressed by interviewees P2, P4 and P5.

**Perform Data Exploration** As interviewee P2 stated, there is no universal way of analyzing available data. Data exploration is therefore necessary to gain an understanding of available data and its properties. A suggested way of exploring available data is by constructing multiple data plots with differing inputs, as well as calculating some standard metrics such as variance, mean, and median. The interviewed data scientists stated that plotting the

data in multiple different ways and comparing its key metrics to their expectations based on their preunderstanding of the data is an effective approach to data exploration.

## 6.3 Data Preparation

As large numbers of interfaces and the need of extensive integrations have been suggested as a major reason why corporations lack data-driven financial software today according to the interviewed companies, data preparations must be seen as a phase of great importance in this domain according to P2, P4 and P5. Data will have to be extracted from various ERP systems, banks, treasury systems and other financial services in order to asses a corporation's financial situation.

During our interviews in Unit B, we focused on the issue of the large number of interfaces when discussing question Q4 and Q10. It seems as the majority of all the interviewees had experience working with this issue, and IT Consultant P2 even mentioned that they have individual teams in each project working solely on these types of integrations.

In this section of the detailed CRISP-DM, we will present the key actions which the interviewees found to be essential when handling multiple data storages and interfaces in the development of data-driven applications. The key actions, or data phases, are visually represented in Figure 6.2. We propose four novel substeps in the Data Preparation phase of the detailed CRISP-DM.

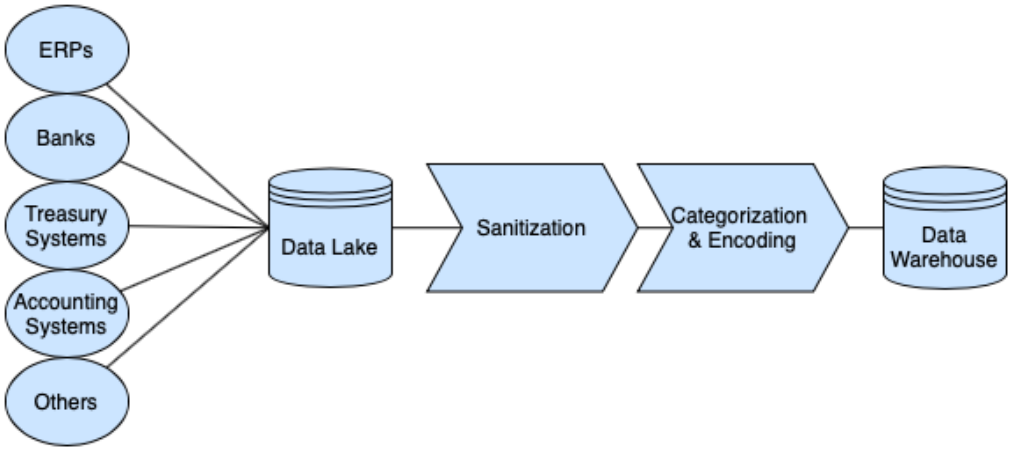



**Figure 6.2:** The process of data preparation for financial decision support tools. Each of the data stages or processes are described in an individual paragraph below.

**Create Data Lake** As mentioned by P2, data lakes are an increasingly popular concept in data science to store large amounts of data in its raw format. Contrary to a data warehouse, the reason of storing data must not be defined in a data lake. According to data scientists P1, P2 and P5, creating a data lake is a good way of exploring and understanding the data which is later to be used in the model.

For the reason mentioned above, creating a data lake with all the data to be used in the model should be the first step when conducting data preparation. This allows the data

scientists to explore the data by for example plotting it in various formats, which according to P2 is the best way of truly understanding it in the way needed for the upcoming sanitization.

**Sanitize Data** Data sanitization is the process of irreversibly removing or destroying stored data. It is important to use the proper technique to ensure that all data is purged. Data lineage[1] should be considered when sanitizing data, to minimize the risk of error during future use of data [39]. Data may be sanitized due to multiple reasons. It might for example be that the data is obsolete, redundant, trivial or incorrect.

During the interviews, interviewee P5 pointed out there are data sanitization approaches such as heuristic and meta-heuristic data-sanitization which are useful in order to, for example, remove sensitive data. However, there are no current algorithms or concrete approaches to solve the identified problem of *uncertainty of data*, mentioned in Unit A. According to P1, P4, P5, the best way of solving such issues in an adequate way is to acquaint yourself with the data by plotting it in various ways. By doing so, with the business understanding obtained in the first phase, you should be able to identify incorrect outliers in the data.

Interviewee P1 also suggests exploring and sanitizing the data using a top-down approach. This means acquainting yourself with the data in a larger perspective and gradually focusing on smaller and smaller subsets. Doing so supports the task of finding incorrect data.

**Categorize and Encode Data** As the data is sanitized, it will thereafter be categorized and encoded according to its use, origin, or in what categories are useful for the particular application under development. Encoding is the process of assigning each category a numerical value for the model to recognize its origin. In this stage, the alignment made regarding the system's environment and the domain might be useful in order to obtain the needed data categories. This substep is a fundamental activity for any subsequent supervised machine learning.

**Create Data Warehouse** The data warehouse is the storage for the finalized data to be used in the model. The data stored in the data warehouse has a defined purpose, is encoded, structured in a relevant way and ready to be queried and used in the model. Usage of a data warehouse ensures that data scientists have readily access to all relevant data needed for successful modeling. The data warehouse also serves as a bridge between the data lake and usage of data in the system model. This minimizes the risk of incorrect or inaccurate data entering the model due to the previous processing steps taken to sanitize, categorize and encode data.

## 6.4 Modeling

In this section we propose five substeps in the Modeling phase of detailed CRISP-DM.

**Consider Data Causality** The interviewed data scientist P2 suggested that causality is a very important factor to consider when modeling. Thorough analysis of data relations and consideration of the causality between them is vital to ensure that the model does not

[1]The chain of transformation that the data has followed during its lifetime

give false outputs based on relations that are not accurate representations of reality on which they are based. This is in line with the literature on causal modeling in the financial domain [32].

P2 also stated that the data in itself does not provide any information of the internal causal relationships. Analysis of causality is therefore heavily related to the business understanding and a deep understanding of the business problem the model seeks to solve. Analysis of causality can therefore with merit involve domain experts to ensure that correct conclusions are drawn.

A possible pitfall of statistical modeling emphasized by the interviewed data scientists is that correlation of data is presented as causality between data. Correlation among data can be proven using statistical methodology, causality on the other hand can not – unless leaving frequentist statistics to instead perform Bayesian causal analysis [34], which we consider out of the scope of this thesis.

**Define Target Metrics** During the modeling phase the data scientists suggested that target metrics for determining the viability of the model needs to be stated. When deciding target metrics, current procedures for accomplishing the task can be considered and used as comparison, if available. If no current procedures can be used for comparison, it can sometimes be difficult to decide target metrics as there are no universal values for determining the success of a model apart from comparison with existing solutions. Depending on the application, a model with an accuracy better than a random walk (making choices by random chance) can be extremely valuable, while other applications may require a close to perfect accuracy to be considered successful.

The target metrics should therefore be closely related to the overarching business goals and decided in collaboration with the domain experts as well as the stakeholders of the system. A plan for reassessment of the target metrics and viability of the model should also be constructed. This is important, as the model might behave differently in the future due to changing data and changing relations between it – a phenomenon known as *drift* in machine learning [63].

**Select Modeling Techniques** The interviewed data scientists suggested selection of modeling techniques to be one of the most crucial steps in the modeling phase of the development process. When selecting modeling techniques to use, deciding factors include the nature as well as properties of the input data. Successful business understanding, data understanding and data preparation should however facilitate the process of selection as it provides fundamental knowledge of the requirements of the modeling technique.

**Build the Model** The building of the model is very dependent on the specific situation at hand. However, as suggested by P5, it can be useful to build multiple models. This allows for selection of the best performing model for use in the final system. We therefore suggest building multiple models when developing, as it can be difficult to predict the performance of a specific model beforehand. The activity of selecting the best performing model among several candidates is referred to as model selection in data science.

**Consider 80/20** As suggested by interviewee P1 as well as supported by P2 and P3, consideration for diminishing returns when modeling is a valuable thought. The 80/20 rule

can be used as a mental guideline to consider when the model might be good enough for use, and further work spent on it might be time better utilized on other tasks. As the interviewees also stated, it can be difficult to approximate when the model is good enough, but as they stress, the 80/20 rule is useful in that it can serve as a guideline for consideration when finalizing the modeling phase. The 80/20 rule is commonly used in the finance industry, where it is known as the Pareto principle [25].

## 6.5 Evaluation

As mentioned by P2, the evaluations performed throughout the CRISP-DM process model can be divided into either validation of business requirements or verification of technical requirements. This differs from the original definition of evaluation in the CRISP-DM process model, in which evaluation only refers to the evaluation of the model itself. The most important concept mentioned by the interviewees regarding validation and verification is that it is a continuous process, which is visualized in Figure 6.3. The need of validation in the development process of data-driven applications is also stressed by Borg and Vogelsang [65] in their article about requirement engineering of data-driven applications.

There exists a lot of different validation and verification methods, and in this section we will present the ones found to be suitable to validate and verify the recommended process substeps in this detailed CRISP-DM. In total, we present six additional substeps in this phase.

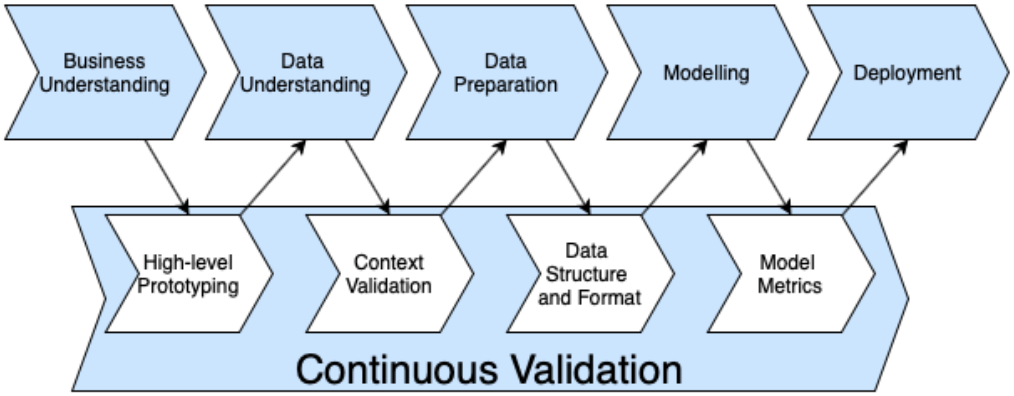



**Figure 6.3:** An overview of our suggested methodology for continuous validation.

**Continuous Validation** The idea of continuous validation is to minimize effort spent on tasks which will not provide actual business value after deployment of the system model [60]. If validation is conducted at distinct intervals, time spent developing between the validation steps might be obsolete as they rely on a previous understanding of the business or data requirements.

The concept of continuous validation was also supported by the interviewees. This strengthens our notion of it being a successful way to further extend the detailed CRISP-DM model. The concept of continuous validation is in line with general best practices in software engineering, i.e. continuous engineering [28].

**High-level Design Prototyping** Our method of validating the business understanding phase by constructing a high-level prototype, a GUI mockup, was considered very successful. As there is a very differing level of technical knowledge between the different stakeholders, utilizing a low-level prototype showcasing the *use* of the system is valuable. A more feature-complete prototype can then be used to complement the GUI mockup, to showcase the logic and more technical aspects of the system. This ensures that the system fulfills both the requirements regarding usability, as well as the technical functionality.

Presenting GUI mockups are useful when validating business requirements without going into detailed technical requirements. This kind of validation is also approved by all of the interviewees in the early phases of the development process. Thus, we recommend to create one or more high-level prototypes directly after the Business Understanding phase in order to validate results prior to moving on to the Data Understanding phase.

**System Context** As stated in Section 4, the large amount of interfaces and integration needed in the financial domain is a common issue. Context diagrams are useful to validate the findings regarding interfaces of the system under development. Thus, we recommend creating a context diagram directly after the Data Understanding phase in order to make sure all interfaces are taken into account prior to moving on to the Data Preparation phase.

**Data Structure and Format** The end goal of Data Preparation is to create a data warehouse with all the data needed to build the model. To make sure the data is structured and formatted in the desired way prior to the modeling phase, some kind of data structure and format validation is recommended. This can for example be done by drawing an E/R-model (Entity relationship diagram).

**Model Metrics** Model metrics refer to metric evaluation of the performance of the model itself. Depending on the characteristics of the model, various metrics may be utilized to verify its performance.

In case of binary classifiers, it might be suited to describe the models performance by plotting a ROC curve, or by creating an accuracy measurement. However, in the case of more "fuzzy logic"-like results it might be harder to define evaluation metrics.

In the case of Time Series Forecasting, which is often used in the domain, commonly used model metric are Root Mean Square Error (RMSE) and Mean Absolute Percentage Error (MAPE) [23].

**Future Acceptance Criteria** To ensure the future viability of the model, we propose that future acceptance criteria should be stated. As the specific criteria suitable for evaluating the model might differ depending on the specific model, it is important to consider the metrics used to evaluate the model when deciding. One key metric suitable for use can be the model lift, which describes how much better the model predicts values compared to random chance while considering the populations[2] distribution [65]. As suggested by P4, other metrics not directly related to the prediction outcome of the model can also be use-

[2]The entire data set

ful, such as performance metrics. Useful performance metrics can include the time frame required for predictions, as well as the amount of data necessary.

## 6.6 Deployment

As the process of deployment refers to the process of ensuring the future use of the data mining application in the initial CRISP-DM model, our way of adapting the deployment process is a significant alteration compared to the initial process model. A point can be made that the deployment stage may be skipped during development of prototypes using CRISP-DM as the steps from our extended framework may be integrated in the modeling step of the detailed CRISP-DM process model. We however feel that it is best to keep the phase, as removal of the phase may downplay its importance during development. As the main goal of any prototype likely is to in the future actually utilize either the information gained during prototyping, or the actual prototype, to facilitate development of a final product meeting the requirements posed by the domain as well as the stakeholders.

The process of prototyping does in itself not provide any value to the deployment phase other than in a less time consuming way explore the use of the imagined final product. We therefore believe that the future use of a prototype always needs to be considered when prototyping to minimize extra workload in the future, i.e., wasted effort. The research literature clearly shows that a substantial part of technical debt in data-driven applications reside in the operations of deployed models [58].

In the remainder of this section we present three additional substeps in our detailed CRISP-DM.

**Monitor Data Characteristics and Relations** As the interviewed data scientists stated, relations and characteristics among data are not necessarily constant during the entire lifetime of the models. It is therefore important to consider and monitor the input data to identify any changing relationships. This can be difficult to do from just analysing input data, and it therefore requires a good business understanding. Changes in relations among data can therefore be identified by analysing the overarching business functions that serve as the data origin. For example, if an interface intertwined with the system model was to change, the resulting data might change format if this is not considered. Automated data validation tools such as Great Expectations[3] can play an important role.

**Monitor Acceptance Criteria** According to the interviewed data scientists, a data-driven model's performance is heavily reliant on the input data. It is therefore important to have a plan for monitoring the decided acceptance criteria to ensure that the model is accurate enough for use. This plan should include concrete steps for how the model will be supported and maintained to ensure its future use. The plan should preferably include assigned stakeholders responsible for ensuring its execution if required.

**Plan for Reassessment of Model** As suggested by P1, P2, P3 and P5, creation of a plan for reassessment of the model is a very valuable activity to ensure the future use of

[3] https://greatexpectations.io/

the model. As the model's performance may be affected by factors such as drift and changes of relations among input data, we suggest this activity to be conducted during development using our detailed CRISP-DM model.

# Chapter 7
# Discussion

In this master thesis project we have investigated the problem domain of data-driven financial decision support tools (RQ1), as well as concretized CRISP-DM with regards to prototyping within the domain (RQ2).

## 7.1 RQ1: The Problem Domain

To summarize RQ1, we have identified three main challenges posed by the problem domain of financial decision support tools. These are *Difficulty of defining main purpose*, *Large number of interfaces*, and *Uncertainty of data*. The two first challenges can also be found in the article written by Nilsson et al. [49] in their challenges A3, D2, F1, and F2. The challenges presented by Dixon and Halperin [24] are more focused on the individuals working in the financial industry rather than specific technical or organizational challenges. Thus, we did not identify any direct connections with their challenges in relation to the ones we identified.

This section will answer RQ1: "*What characterizes the problem domain for data-driven financial decision support applications?*" by discussing the identified characterizing challenges posed by the problem domain.

**Difficulty of Defining Main Purpose** One of the most defining challenges posed by the problem domain is that of how FX risk is to be defined. This is not something stressed by the interviewed companies, but rather a conclusion from the analysis of the interview transcripts. The interviewed companies all had very differing views on where and how FX risk exposure arises. This can be problematic during development of software in the problem domain as it leaves two "philosophical" choices when designing the solution. The first option is to give the users of the system multiple ways of choosing how risk is calculated. This would ideally let each user select how they want to define risk and then use the software normally with all values being adjusted according to their preferred risk calculation.

The downside of using this approach is that it can be confusing for the system users as information gathered from the system might differ based on the choice of calculation method. Two users in the same company might therefore get contrasting views of the company's risk exposure, even though they are using the same software and the same data.

The other approach is to develop the software using a pre-determined way of calculating risk. The problem with this approach is that the system might not be viewed as useful by companies that currently calculate their FX risk with a different approach. This would however ensure that information retrieved from the system always is dependent on nothing more than the underlying data, thus mitigating the risk of confusion stemming from the software.

This challenge is also identified by Nilsson et al. [49], with their challenges named A3: *´´Difficult to create a shared vision and align the entire organization around common goals"* and D2: *Difficult to break down requirements.*

**Large Number of Interfaces** As stated in Section 4, the problem domain is characterized by the use of multiple different information systems for storage of data (Company 1, 2, 3, 4 & 5). Development of software in the problem domain therefore relies on support for multiple different interfaces for data retrieval. This might not be a unique factor characterizing only the problem domain of financial decision support tools, but it poses a clear challenge for the development. Both multiple interviewed companies (Company 2, 4, 5) and P5 also stated that some interfaces such as accounting systems and banks can be notoriously difficult to extract information from, and we therefore consider this a considerable challenge in the problem domain. This finding is in line with previously reported challenges in the banking domain [61].

The difficulty of working agile, as the nature of prototyping and CRISP-DM is also identified by Nilsson et al. [49]. In their report they present two challenges associated with legacy system, namely F1: *"Legacy systems are not easily adopted to agile ways of working"* and F2: *"Complexity and interdependencies between legacy systems are hard to deal with"*

**Uncertainty of Data** Another defining problem of the domain is that some data is not absolute (Company 1, 3 & 5), in the sense that it can be subject to change due to outside factors at a future time. The financial industry commonly uses forecasts and accounted values which due to their nature contain considerable uncertainty. A receivable[1] for example, does not guarantee actual payment of its full amount, as a customer might not ultimately pay for the goods. Due to this, software solutions used in the domain need to be able to adjust accordingly.

[1] A claim for payment held by the business for goods/services ordered but not yet paid for by a customer

## 7.2 RQ2A: Requirement Engineering in CRISP-DM

To summarize RQ2A, traditional requirements engineering practices fit well into the Business Understanding step of the CRISP-DM process model. However, practices can also be added to all of the other phases as well. By redefining evaluations to include both validation and verification, instead of mere evaluation of the model itself, requirements engineering becomes a continuous process throughout the whole process model, i.e. integrated requirements engineering [12]. Due to changes in data characteristics over time, so called drift, requirements engineering also becomes an important process when monitoring future acceptance criteria as the model is deployed.

In this section we will answer RQ2A *How can CRISP-DM be detailed with methods from RE?*. This will be done by discussing the requirements engineering activities presented in the detailed CRISP-DM in Section 6 above.

The detailed CRISP-DM has been concretized on the basis of prototyping an FX risk exposure application, which is a requirements engineering activity per se. More on how the practice of prototyping is affected when prototyping a data-driven application is discussed in Section 7.5.

In the detailed CRISP-DM, traditional requirements engineering activities are generally assigned to the Business Understanding phase. These include activities such as stakeholder analysis, elicitation, and business requirements analysis. However, some activities have been added in order to fit prototyping of data-driven applications and the problem domain. For our detailed CRISP-DM, customized for prototyping of data-driven applications in the financial domain, we have added assignment of roles based on the needed requirement engineering roles identified by Hesenius et al. [35]. This in order to assure the project participants' needed expertise for it to succeed. To suit the problem domain, with the difficulty of defining a common purpose of a data-driven components, we have added alignment prior to the elicitation of purpose. This in order to give the participating stakeholders in the elicitation process common definitions of terms used when discussing the system and the potential customers' current practices.

Regarding evaluation, we have redefined it to suit the practice of requirements engineering and the participants in Unit B's recommended workflow. Instead of evaluation referring to only the model itself, it might also be seen as two different types of evaluation, namely verification and validation. With this definition, verification would refer to the traditional definition of evaluation in the CRISP-DM process model, i.e. evaluation of the model, while validation would be a continuous process throughout the whole project, as discussed in Section 6.5. In Section 6.5 we also present four validation and verification activities tailored for the problem domain discussed in Section 7.1.

Data requirements are also part of traditional requirements engineering practices [41]. However, this was not something any of the interviewees worked with. As mentioned by one of the interviewees *"The data is what the data is, and it's up to the data scientist to make the most out of it"*, which is also a view on data requirements identified by Borg and Vogelsang in their study on requirements engineering for machine learning [65]. More on how to assure data quality when developing data-driven applications in this domain is discussed in Section 7.4.

Traditional requirements engineering activities are also needed during the modeling step

and even as the product is deployed. In these steps, evaluation metrics and future acceptance criteria of the system have to be created, which should be closely related to the identified business goals. Creating good metrics and acceptance criteria might be difficult, but adequate requirements elicitation and analysis should ease the process according to the interviewees in Unit B. As mentioned by the interviewees, characteristics of data changes over time, hence creating a need of monitoring the model's performance when receiving real-time data during operations. Monitoring the model's performance and reevaluating the acceptance criteria become a important parts of the continuous validation process.

## 7.3 RQ2B: Benefits and Challenges

To summarize RQ2B, the CRISP-DM model's strength in that it is very general, can also be considered its weakness as it will inevitably need refinement based on the specific project at hand. We therefore consider the original model a useful framework for determining a general approach for development of data-driven software. Our detailed CRISP-DM, showcases how the original process model can be customized for a specific application context.

This section will answer RQ2B: *What would be the benefits and challenges of using the detailed CRISP-DM?* based on an assessment of the original CRISP-DM and our proposed extensions provided in Section 6.

One strength of the CRISP-DM model is that it is very general and can as some interviewees stated be related to "common sense" during development of data-driven software. Part of CRISP-DM's value therefore stems from its ease of implementation in an already existing workflow. As P1, P2 and P5 stated, CRISP-DM very closely represented their existing workflow even though they had no knowledge of the process model beforehand. They therefore considered CRISP-DM useful, as it clearly defined processes they were already familiar with in a way that gave further inspiration for improvements to their respective development methodologies.

On the contrary, multiple interviewees also viewed this as a weakness of the model, as it does not provide concrete steps to ensure successful completion of the different phases. They therefore considered it to be lacking substance if it was to be used in a real working scenario.

This is also related to CRISP-DM's emphasis on a highly generalized workflow, and that the phases may need to be iterated in different order depending on the concrete application. Some interviewees considered this positive as many real scenarios provide unique challenges, requiring a somewhat customized workflow. Others considered this to be detrimental to the value of CRISP-DM, as it further generalizes the process model in a way so that it may be difficult to actually use further than as a high-level conceptual model.

## 7.4 RQ2C: Data Quality

To summarize RQ2C, multiple method recommendations have been added to the detailed CRISP-DM in order to ensure data quality. The relevance of the data is ensured by iterative work between business understanding, data understanding and data preparation. This work

consists of data identification, data collection and data sanitization. The accuracy of the data, and thereby the model, is assessed using model metrics. To efficiently compare modeling techniques, it is important to create three subsets of the data, namely a (1) Training set, a (2) Validation set, and a (3) Testing set. The timeliness of the data vary depending on the purpose of the model. If the model is evaluating historical risk exposures, high quality data can be ensured, while utilization of forecasts adds uncertainty that contributes to lower data quality. The accessibility of data in the financial domain is currently bad, but at the time of writing also improving. Unfortunately, there exists no shortcut to collecting financial data and therefore no method recommendations have been added in order to solve this problem. Comparability of data is in general possible in the domain of financial services. However, it is important to analyze changes in the characteristics of the data and the performance of the model. In order to do so, creation of future acceptance criteria and monitoring have been added as recommended methods in the detailed CRISP-DM. No methods have been added to analyze or improve coherence of data in the detailed CRISP-DM.

In this section we answer research question RQ2C: *How do aspects related to data, e.g. quality, affect the detailed CRISP-DM?*. The answers are based on the problems discussed in Section 7.1, and answered with the information collected from the interviews conducted in Unit B: CRISP-DM. Furthermore, we complement our discussion with findings from previous academic publications on the topic.

To answer RQ2C, we will make use of the European Statistical System's (ESS) definition of data quality [27], used by e.g. the Bank of England to define quality of financial data [50]. ESS defines six data quality dimensions to asses the quality of data, namely (1) Relevance, (2) Accuracy, (3) Timeliness and Punctuality, (4) Accessibility and Clarity, (5) Comparability, and (6) Coherence [27]. The definitions of the data quality dimension are presented in Chapter 2 *Background and Related works*, Section 2.3.

### 7.4.1 Relevance

Relevance refers to the degree of which statistics meets current and potential users' needs. As mentioned in Section 2.3, relevance may also be extended beyond the scope of statistics to encompass also data, i.e. to refer to the degree of which data meets the users' needs [50]. The relevance of the data is ensured during the business understanding, data understanding and data preparation phases as needed data is identified, collected and sanitized.

### 7.4.2 Accuracy

The accuracy refers to the closeness of estimates and computations to the exact or true values. In development of data-driven applications this is tested and validated in the evaluation phase of the detailed CRISP-DM process model. As mentioned by multiple interviewees, it is important to define evaluation metrics closely related to the business requirements set up during the business understanding phase in order to assess the model adequately. As mentioned by interviewee P5, in line with established data science practices, it is also important to create three subsets of data used during development and evaluation of the models, namely a training data set, a validation data set, and a testing data set, in order to be able to compare different modeling technique appropriately.

Due to changing characteristics of data, it is also important to assess the model's performance and accuracy after deployment. This quality assurance is taken into account in our detailed CRISP-DM model as defining future acceptance criteria and monitoring such criteria are defined as corresponding substeps in the evaluation and deployment phases.

### 7.4.3 Timeliness & Punctuality

The timeliness of the data refers to the time between the availability of the data and the phenomenon they describe. Depending on what type of data will be used in the learning and utilization of the model, the timeliness may vary. For example in the case of the FX risk exposure application, if long-term forecasts are used in the assessment of a corporation's financial condition, the data will have an inferior timeliness and thereby a low quality. This in turn will probably result in lower accuracy in the model's predictions. However, when assessing historical rather than future FX risk exposure, the exact time of transactions and currency fluctuations may be utilized. Thus resulting in higher quality data. In terms of third party data the timeliness might be time-critical, as with the case of FX rates. However, what level of timeliness is needed is dependent of the users' time interval of FX exposure assessment.

To conclude, the timeliness of the data in the financial domain will be highly dependant on what type of financial records will be used in the model. Forecasts, and especially long-term forecasts, will result in inferior timeliness and hence low quality data. When assessing previous risk exposures, the exact time of transactions and other data points may be utilized and thereby resulting in good timelines and high quality data.

### 7.4.4 Accessibility & Clarity

Accessibility refers to the conditions in which users can obtain data. As motioned by e.g. interviewee P5 who has experience in working within the financial domain, the accessibility of financial data is low. Sometimes it may take up to months to collect the data needed in order to prototype a proof-of-concept or to develop a model. However, P5 also added that there exists no alternative to obtaining data, it just has to be done. Thus, no method recommendation has been added to the detailed CRISP-DM in order to solve this data quality assurance.

As mentioned when studying Unit B, financial institutions and companies are aware of the problems associated with the bad accessibility and working on improving it.

### 7.4.5 Comparability

Comparability refers to which degree data may be compared over time and/or across domains. In the case of the FX risk exposure application, data may in general be compared over time such as the data regarding sales forecasts, financial transactions, and FX rates. However, as mentioned by multiple interviewees it is important to analyze the change of the characteristics of the data over time. This is due to the changes affecting the model's accuracy and may thereby be a reason for relearning the model. If and when the characteristics of the data

have changed significantly, the model will probably not perform according to the "future acceptance criteria" set up prior to deployment, as mentioned in the detailed CRISP-DM. It is therefore of great importance to introduce and monitor such criteria.

### 7.4.6 Coherence

Coherence relates to the degree to which data derived from different sources or methods but concerning the same phenomenon are consistent with each other. No method recommendation has been added to the detailed CRISP-DM in order to asses this dimension as we did not find it to be relevant in this particular case.

## 7.5 RQ2D: Prototyping Practice for Data-Driven Applications

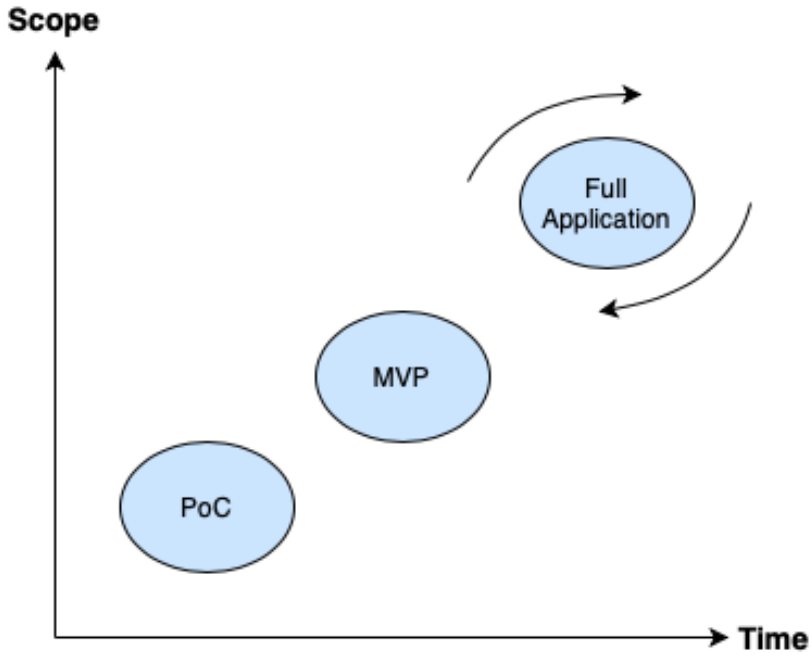



**Figure 7.1:** The process of prototyping a proof- of-concept, a minimum viable product, and finally a full application is a process of one gradually evolving product. The gradually evolving product is represented by bubbles using two dimensions, namely (1) Time in development and (2) Scope of the prototype in term of breadth and depth of the prototype's functionalities.

To summarize RQ2D, our research suggests that the general approach of prototyping data-driven application includes three distinct phases, namely (1) prototyping a PoC, (2)

prototyping an MVP, and (3) iteratively expanding the MVP until it is a full application (preferably guided by the 80/20 rule). The PAM has been applied when analyzing the prototyping practice in the domain of data-driven applications. Regarding prototype scope, our findings suggest that prototyping in the data-driven domain is characterized by a gradually expanding width of the functionalities while continuously keeping a great depth. It also suggest that the prototyping practice is characterized by a combination of a parallel and optimization exploratory strategy.

When studying Unit B, the interviewees were asked if and how their approach utilizing the detailed CRISP-DM would change when prototyping instead of developing a data-driven application. In this section, we will present and discuss the findings related to prototyping of data-driven applications, and thereby answer RQ2D: *How is the practice of prototyping affected when developing data-driven applications?*. This will be done by relating the answers of the interviewees to the Prototyping Aspects Model (PAM) [13].

### 7.5.1 The General CRISP-DM Approach

When discussing prototyping of data-driven applications, most interviewees stated that their approach would not change compared to development of a data-driven application. In general, the first thing data scientists need to conclude when developing and prototyping data-driven applications is whether or not there exist enough predictive signals to make development of the application possible. This is generally done by prototyping a Proof of Concept (PoC). The interviewees in Unit B suggest that the general approach of prototyping a data-driven application is to first develop a PoC, thereafter a minimum viable product (MVP) adding a GUI and back-end functionalities and thereafter the full product.

The reason why the method recommendations in the detailed CRISP-DM do not change when prototyping, is due to the need of all of the phases when prototyping an MVP. Prototyping a PoC also needs all of the phases in the detailed CRISP-DM process model, except from deployment, to be conducted. However, when taking the aspects of PAM into consideration, we can analyze the practice further than only through the actions taken in the individual phases.

The general approach of prototyping data-driven applications is visualized in Figure 7.1 using the maturity of the project as the x-axis and the prototype scope as the y-axis. The prototyping scope is discussed in Section 7.5.4, the exploration strategies used when moving between the prototypes is discussed in Section 7.5.5, and the 80/20 rule symbolized by arrows surround the “Full Product” bubble is discussed in Section 7.5.6.

### 7.5.2 Prototype Purpose

Our research suggest that there exist two general purpose of prototyping data-driven applications, namely (1) to validate the implementability of an application, and (2) to find the most effective modeling technique and thereby the “golden model” to base the final product on.

Let us divide the prototyping practice of data-driven applications into two phases, namely development of a PoC and thereafter a MVP, as mentioned in Section 7.5.1. By doing so, we may define the purpose of prototyping better, as our research suggest that it varies depending

on the type of prototype. When developing a PoC, the main purpose seems to be to obtain a relatively quick understanding of whether the final product is implementable and if there exist data describing the phenomenon well enough to create a model.

However, as the team finds the implementability and data to be good enough, there is no further need of focusing primarily on these aspects. Instead, the focus shifts to finding the "best candidate" modeling technique with can be used in the final product, when developing a MVP.

### 7.5.3 Prototype Use

The use of prototypes in the case of data-driven application is closely related to it's purpose, namely testing implementability and data, and evaluating modeling techniques. As mentioned my many interviewees, early stage models such are those included in the prototypes are frequently evaluated using evaluation metric. Thus, our research suggest that prototypes main use, in the case of data-driven applications, is to be evaluated in order to investigate the stated areas of interest mentioned in *Prototype Purpose*, Section 7.5.1.

### 7.5.4 Prototype Scope

On of the aspects in PAM is the scope of the prototype, i.e. the breadth and depth of the prototype's functionalities, as well as the prototypes visualization, interactivity and data realization. As stated in the previous sections, the common practice of prototyping a data-driven application is to first prototype a PoC, thereafter a MVP, and finally a full product.

As stated by e.g. P5, the PoC can be a very limited application in term of functionalities. Additionally, the sets of data used when developing a PoC may also be limited compared to the real-time data sets meant to be used in production. However, in order to asses the application's feasibility by evaluating its performance of a single functionality, the functionality itself has to be developed fully. As the project precedes and the prototyping shift from a PoC to a MVP, the use of data-driven functionalities, GUI & back-end utilities, etc., gradually expand.

Due to the reasons stated above, our research suggest that a preferred prototyping practice of data-driven application is to gradually expand the width of the prototyping scope while keeping a continuous great depth. The prototypes visualization and interactivity seems to be rather low prioritized aspects in the early stages of development process by the interviewees, and something which is added if the PoC proves the application's feasibility. However, data realization is stressed by many interviewees as a critical aspect of prototyping data-driven application. Without the use of realistic data, it is difficult if not impossible to determine the accuracy, feasibility and implementability of the application, which tends to be the main purpose of prototyping data-driven applications according to our research.

### 7.5.5 Exploration Strategy

Another aspect in PAM is the exploration strategy, i.e. the aspect related to the idea of how to traverse the solution space of the application. In their article, Bjarnarson et al. present

four exploration strategies used when prototyping, namely *point-based*, *parallel*, *optimization* and *flexible* exploration. These exploration strategies are described in Section 2.1.3.

In general, in the early stages in the development of data-driven applications, multiple models are developed and evaluated parallel, suggesting that a parallel exploration strategy is used. This practice is used until the development of the final product for which the best performing modeling technique has to be decided, suggesting a optimization strategy is used.

Using the above stated reasons, our research suggests that a combination of parallel and optimization strategy is generally used when prototyping data-driven applications. As the prototyping proceeds, the worst performing models may be removed, but it is only when developing the final product a best performing model has to be decided.

### 7.5.6 The 80/20 rule

The 80/20 rule, as interviewee P1 stated, is a rule based on the idea that 80% of a model's value is accomplished in 20% of total effort spent. On the contrary, the remaining 20% of the model's value is accomplished in the remaining 80% effort.

As interviewee P1 stated, one important aspect to consider when prototyping data-driven applications is that effort put in optimizing the model has a diminishing return. Using the 80/20 rule when considering whether to move onto the next step is therefore our suggestion when working with prototyping data-driven applications. This is further complemented with the suggestion of using an agile workflow and completing the steps of the detailed CRISP-DM model in iterations. The use of the 80/20 rule is further strengthened by the nature of data-driven applications, as multiple interviewees stated that the model will never be fully developed and finalized due to the need of evaluating the real-time data and retraining of the model.

A difficulty with the 80/20 rule, originating from the same issue that provides its value, is that it is difficult to estimate the maximum potential of the model. It can therefore be difficult to estimate when the 80% threshold has been reached. For that reason, the 80/20 rule should function more as a guiding thought to be considered when modeling, implying that progress to the next step of the model should be considered once the improvement in results give considerably diminishing returns. The 80/20 rule is represented in Figure 7.1 using circulating arrows surrounding the "Full Products" bubble.

## 7.6 RQ2E: The Problem Domain's Effects on Prototyping

To summarize RQ2E, we find that the problem domain only has a minor impact on the practice of prototyping. All three identified challenges affect the process of prototyping negatively, however both uncertainty of data and difficulty of defining main purpose are challenges in which a low-cost prototype is better suited than a full-scale system development, as it allows for failure and adjustments at a lower cost. This appears to be a wise option, as data science projects are repeatedly reported as intrinsically agile [16].

In this section we will answer RQ2E: *How is the practice of prototyping affected by the problem domain?* based on the challenges in the problem domain identified in Section 7.1. The three identified challenges in the problem domain are (1) The large number of interfaces, (2) Uncertainty of data, and (3) Difficulty of defining main purpose.

### 7.6.1 Large number of interfaces

The large number of interfaces will affect the possibility of time efficiently prototyping a data-driven application within the domain of financial decision support tools. In general, it is preferred to utilize a sub set of data adequately representing the real-time data used in the system during learning, evaluation and testing of a model, as stated by interviewee P5.

### 7.6.2 Uncertainty of Data

One factor identified regarding the problem domain is that of the need for uncertainties to be clearly communicated to the end user. In the domain it is very common with values to be associated with some type of risk, but in some applications it is crucial that values are absolutely correct. An application utilizing machine learning such as regression to extrapolate information therefore needs to be able to provide the user with information regarding what data is extrapolated and might be associated with risks stemming from the accuracy of the software model.

As interviewees stated, many decisions however end in informed decisions *without* perfect information, and the risks associated with software models can therefore probably be mitigated as long as the user understands its limitations.

### 7.6.3 Difficulty of Defining Main Purpose

To handle the challenge of defining the main purpose of the software, as suggested by multiple interviewees, we stress that aligning the stakeholders is a crucial step. A successful alignment should ensure that the software fulfills the needs of the stakeholders even though they might have had differing views of the needs of the system beforehand.

In the scope of prototyping, this challenge might also be easier to overcome in that a prototype can serve as a tool for ensuring alignment between the stakeholders. As a prototype requires less effort for implementation, the prototype can be used to evaluate the alignment of the stakeholders before development of the final product. The use of prototypes for alignment purposes has been reported as an industry practice also in previous research [14].

## 7.7 Thesis Work Validation

This section discusses identified threats to the validity of this thesis. We focus our discussion on external validity and reliability [53], i.e. the generalizability of our conclusions and to what extent other researchers would reach the same conclusions if replicating our case study.

### 7.7.1 Interviews

Before conducting the initial interviews we constructed an interview guide with closed questions. This guide was used during the first interviews, but soon showed to be more constraining then useful for extracting the most amount of information from the interviewees. An indicator of this was that many interviewees stated new important information during informal conversation after the interview was over, and we therefore decided to evolve the guide into more open questions using a semi-structured approach (see Table 3.1 and 3.3). The benefit of using a semi-structured approach was that we could better adapt the interview based on the corporate position, knowledge, and experience of the interviewee. As some interviewees had more knowledge of their respective organization relative to others, this approach generally extracted more useful information from each interviewee.

A disadvantage of using the less-structured approach was that it makes it more difficult to systematically compare the answers of the different interviewees. More analytical effort is therefore required to weigh and compare the interviewees' answers to draw conclusions and it is possible that another set of researchers would have interpreted the results differently. This is also a possible threat to the validity of the stakeholder analysis, as our extrapolation may not be an absolutely accurate depiction of individual interviewees' true needs. Another possible source of error is that the interviewees' organizations may differ compared to the typical organizational structure depicted in Figure 4.1. As the interviewees all represented different companies, it is possible that the the interest of an interviewee with one corporate position may not correspond to the interest of an interviewee with the same position at a different company. For example the CFO of a smaller company might be more involved in execution of their financial strategies than in a typical organization, a responsibility typically associated with treasury or corresponding departments, compared to a larger company with more personnel in the financial departments.

To account for these threats, our approach when summarizing and drawing conclusions from the interviews has been inclusive to preserve as much information as possible. Furthermore, we highlight that a future larger study covering additional organizations, as well ass more roles per organization, would support the generalizability of our findings.

In the scope of prototyping, it is however always difficult to weigh the true needs of different stakeholders against each other. We therefore consider our method of conducting and analyzing the interviews to be good enough for the purpose of the thesis.

### 7.7.2 Detailed CRISP-DM

One possible threat to the external validity of the suggested concretization of CRISP-DM (RQ2), is that it might be over-fitted to the needs in problem domain (RQ1). Generalization of the model and use in other domains might therefore encounter problems unaccounted for in the concretization. To minimize this risk, the interview guide was constructed so that the problem domain of financial decision support tools was not mentioned until after the CRISP-DM model had been discussed thoroughly. The process of recruiting interviewees however required interaction with the interviewees in which the problem domain had to be stated, and they might therefore be biased towards activities suitable for the problem domain.

The concrete activities suggested in our model are also created to serve the problem domain of financial decision support tools, and the stakeholders identified in the context. As

the concrete stakeholders in every system have differing needs, it might be possible that other activities might be better suited when developing in different problem domains. Still, we believe that our detailed CRISP-DM can inspire development of data-driven decision support beyond the financial domain but future studies will have to validate this hypothesis.

## 7.8 Thesis Work Validation

This section discusses identified threats to the validity of this thesis. We focus our discussion on external validity and reliability [53], i.e. the generalizability of our conclusions and to what extent other researchers would reach the same conclusions if replicating our case study.

### 7.8.1 Interviews

Before conducting the initial interviews we constructed an interview guide with closed questions. This guide was used during the first interviews, but soon showed to be more constraining then useful for extracting the most amount of information from the interviewees. An indicator of this was that many interviewees stated new important information during informal conversation after the interview was over, and we therefore decided to evolve the guide into more open questions using a semi-structured approach (see Table 3.1 and 3.3). The benefit of using a semi-structured approach was that we could better adapt the interview based on the corporate position, knowledge, and experience of the interviewee. As some interviewees had more knowledge of their respective organization relative to others, this approach generally extracted more useful information from each interviewee.

A disadvantage of using the less-structured approach was that it makes it more difficult to systematically compare the answers of the different interviewees. More analytical effort is therefore required to weigh and compare the interviewees' answers to draw conclusions and it is possible that another set of researchers would have interpreted the results differently. This is also a possible threat to the validity of the stakeholder analysis, as our extrapolation may not be an absolutely accurate depiction of individual interviewees' true needs. Another possible source of error is that the interviewees' organizations may differ compared to the typical organizational structure depicted in Figure 4.1. As the interviewees all represented different companies, it is possible that the the interest of an interviewee with one corporate position may not correspond to the interest of an interviewee with the same position at a different company. For example the CFO of a smaller company might be more involved in execution of their financial strategies than in a typical organization, a responsibility typically associated with treasury or corresponding departments, compared to a larger company with more personnel in the financial departments.

To account for these threats, our approach when summarizing and drawing conclusions from the interviews has been inclusive to preserve as much information as possible. Furthermore, we highlight that a future larger study covering additional organizations, as well ass more roles per organization, would support the generalizability of our findings.

In the scope of prototyping, it is however always difficult to weigh the true needs of different stakeholders against each other. We therefore consider our method of conducting and analyzing the interviews to be good enough for the purpose of the thesis.

### 7.8.2 Detailed CRISP-DM

One possible threat to the external validity of the suggested concretization of CRISP-DM (RQ2), is that it might be over-fitted to the needs in problem domain (RQ1). Generalization of the model and use in other domains might therefore encounter problems unaccounted for in the concretization. To minimize this risk, the interview guide was constructed so that the problem domain of financial decision support tools was not mentioned until after the CRISP-DM model had been discussed thoroughly. The process of recruiting interviewees however required interaction with the interviewees in which the problem domain had to be stated, and they might therefore be biased towards activities suitable for the problem domain.

The concrete activities suggested in our model are also created to serve the problem domain of financial decision support tools, and the stakeholders identified in the context. As the concrete stakeholders in every system have differing needs, it might be possible that other activities might be better suited when developing in different problem domains. Still, we believe that our detailed CRISP-DM can inspire development of data-driven decision support beyond the financial domain but future studies will have to validate this hypothesis.

# Chapter 8
# Conclusion

We have studied the domain of data-driven financial decision support tools, how the CRISP-DM process model can be used when prototyping in this domain, and how the practice of prototyping is affected by it. In this chapter, we briefly present the context of our study, the research, the contributions we set out to achieve, the research method and our conclusions.

Artificial Intelligence (AI) and Machine Learning (ML) have been researched topics for many decades, but only in recent times have computing power become sufficient to make these technologies available to the general public. To support development of applications utilizing AI and/or ML, so called data-driven applications, development process models such as CRISP-DM have been created. However, previous papers on the topic of CRISP-DM have concluded that the model lack detailed method recommendations for its respective phases. In this project, we aimed to contribute to this research by creating a detailed CRISP-DM for the domain of data-driven financial decision support tools and study how the approach in the model would be modified in order to support prototyping in the domain.

In order to create a detailed CRISP-DM for the specified domain, we first had to research what characteristics it has. This study was covered by studying RQ1: *What characterizes the problem domain for data-driven financial decision support applications?*. The detailed CRISP-DM for the specified domain was developed when studying RQ2 *How can CRISP-DM be applied when prototyping a data-driven FX risk exposure application?*, taking the aspects in the sub questions RQ2:A to RQ2:E into account.

The research method used in this master thesis is an improving case study, following the guidelines provided by Runeson et al. [53]. The case study included two units of analysis called Unit A and Unit B. Unit A was studied in order to obtain the characteristics of the problem domain and involved identified potential customers for an FX risk exposure application. Unit B was studied in order to create the detailed CRISP-DM and consisted of the potential developers of an application in the domain.

By studying Unit A, three main issues when developing data-driven financial decision support tools were identified. These were (1) a large amount of interfaces, (2) uncertainty of data, and (3) difficulty in defining the main purpose of the data-driven modules included in

the applications. These identified main challenges were presented during discussions with the interviewees of Unit B to explore how the detailed CRISP-DM model can be adapted to suit the problem domain of data-driven financial decision support tools.

By studying Unit B, we created a detailed CRISP-DM process model for the purpose of prototyping in the domain identified when studying Unit A. The detailed model is presented in Chapter 7, Section 7.2. This model is based on the insights gathered during our study of Unit B. The model consists of the same phases as the original CRISP-DM model, but also recommends specific activities for each of them. We also investigated how the process model need to be changed in order to fit the purpose of prototyping a data-driven application in the financial domain. Our research suggests that the process of prototyping is very similar to development of data-driven applications in general. The general approach when prototyping data-driven applications consists of three product complexity phases, i.e. a proof-of-concept (PoC), a minimum viable product (MVP), and lastly the final product, for which the PoC and MVP serves as prototypes.

Our findings can be used by developers in the domain of financial decision support tools as a concrete framework when developing data driven applications. For relatively inexperienced practicioners this can be especially helpful, as our detailed CRISP-DM provides concrete acitivities to conduct to ensure the sucess of each of the six phases of development.

Based on our findings, we outline the following directions for future work. First, is it possible to further detail the general approach of prototyping data-driven applications, presented in Section 7.5.1? Second, is the detailed CRISP-DM, presented in Section 6, applicable in more industries than mere financial decision support tools? Third, what requirements engineering methodologies may be used in order to effectively break down complex financial problems?

# References


[1] Fixer. https://fixer.io/. Accessed: 2021-10-20.

[2] Fortknox. https://www.fortnox.se/. Accessed: 2021-10-20.

[3] Swedish authority for privacy protection. https://www.imy.se/en/about-us/swedish-authority-for-privacy-protections-assignment/. Accessed: 2021-10-01.

[4] Swedish finacial supervisory authority. https://www.fi.se/en/. Accessed: 2021-10-01.

[5] Visma. https://www.visma.se/. Accessed: 2021-10-20.

[6] What is Data Mining? https://www.ibm.com/cloud/learn/data-mining. Accessed: 2021-10-20.

[7] What is erp? https://www.oracle.com/erp/what-is-erp. Accessed: 2022-01-09.

[8] Iso/iec/ieee international standard - systems and software engineering system life cycle processes. *IEEE Std 15288-2008*, pages 1–84, 2008.

[9] Hamza Hussein Altarturi, Keng-Yap Ng, Mohd Izuan Hafez Ninggal, Azree Shahrel Ahmad Nazri, and Abdul Azim Abd Ghani. A requirement engineering model for big data software. In *2017 IEEE Conference on Big Data and Analytics (ICBDA)*, pages 111–117, 2017.

[10] Söhnke M. Bartram. Linear and nonlinear foreign exchange rate exposures of german nonfinancial corporations. *Journal of International Money and Finance*, 23(4):673–699, Jan 2004. Article.

[11] Chris Becker and Daniel Fabbro. Hedging Instruments | RDP 2006-09: Limiting Foreign Exchange Exposure through Hedging: The Australian Experience. *Research Discussion Papers*, (December), 2006.

[12] Elizabeth Bjarnason. *Integrated Requirements Engineering - Understanding and Bridging Gaps in Software Development*. PhD Thesis, Lund University, 2013. https://portal.research.lu.se/en/publications/integrated-requirements-engineering-understanding-and-bridging-ga.

[13] Elizabeth Bjarnason, Franz Lang, and Alexander Mjöberg. A model of software prototyping based on a systematic map, 2021. In Proceedings of ACM / IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM), https://doi.org/10.1145/3475716.3475772.

[14] Elizabeth Bjarnason, Per Runeson, Markus Borg, Michael Unterkalmsteiner, Emelie Engström, Björn Regnell, Giedre Sabaliauskaite, Annabella Loconsole, Tony Gorschek, and Robert Feldt. Challenges and practices in aligning requirements with verification and validation: a case study of six companies. *Empirical software engineering*, 19(6):1809–1855, 2014.

[15] Z. Block and I. C. MacMillan. Milestones for successful venture planning. *Harvard Business Review*, 1985.

[16] Markus Borg. Agility in software 2.0 - notebook interfaces and MLOps with buttresses and rebars. *arXiv preprint arXiv:2111.14142*, 2021.

[17] John M. Borky and Thomas H. Bradley. Protecting information with cybersecurity. *Effective Model-Based Systems Engineering*, page 345–404, 2018.

[18] R. Budde and H. Zullighoven. Prototyping revisited. In *COMPEURO'90: Proceedings of the 1990 IEEE International Conference on Computer Systems and Software Engineering - Systems Engineering Aspects of Complex Computerized Systems*, pages 418–427, 1990.

[19] Pete Chapman, Julian Clinton, Randy Kerber, Thomas Khabaza, Thomas Reinartz, Colin Shearer, and Rüdiger Wirth. The crisp-dm user guide. In *4th CRISP-DM SIG Workshop in Brussels in March*, volume 1999. sn, 1999.

[20] Peter Chapman, Janet Clinton, Randy Kerber, Tom Khabaza, Thomas P. Reinartz, Colin Shearer, and Richard Wirth. Crisp-dm 1.0: Step-by-step data mining guide. 2000.

[21] Yi-Chein Chiang, Tung Liang Liao, and Tse-An Hsiao. Evaluating hedging strategies in the foreign exchange market with the stochastic dominance approach. *Applied Financial Economics*, 21(7):493 – 503, 2011.

[22] JJ Choi and AM Prasad. Exchange risk sensitivity and its determinants - a firm and industry analysis of us multinationals. *FINANCIAL MANAGEMENT*, 24(3):77 – 88, 1995.

[23] Roman Josue de las Heras Torres. 7 ways time series forecasting differs from machine learning, May 2018.

[24] Matthew Francis Dixon and Igor Halperin. The four horsemen of machine learning in finance. *Available at SSRN 3453564*, 2019.

[25] Rosie Dunford, Quanrong Su, and Ekraj Tamang. The pareto principle. 2014.

[26] Satu Elo, Maria Kääriäinen, Outi Kanste, Tarja Pölkki, Kati Utriainen, and Helvi Kyngäs. Qualitative content analysis: A focus on trustworthiness. *SAGE Open*, 4(1):2158244014522633, 2014.

[27] Eurostat. Methodological documents - definition of quality in statistics, 10 2003. Accessed: 2021-12-01.

[28] Brian Fitzgerald and Klaas-Jan Stol. Continuous software engineering and beyond: Trends and challenges. In *Proceedings of the 1st International Workshop on Rapid Continuous Software Engineering*, RCoSE 2014, page 1–9, New York, NY, USA, 2014. Association for Computing Machinery.

[29] Market Research Future. Global erp software market research report: Information by deployment (on-premise and cloud), function (supply chain and product management, sales and marketing, accounting & finance and human resource), vertical (manufacturing, retail, bfsi, it & telecom, aerospace & defence, education and others), organization size (small and medium enterprises and large enterprises) - forecast till 2027, Feb 2021.

[30] C. Giardino, M. Unterkalmsteiner, N. Paternoster, T. Gorschek, and P. Abrahamsson. What do we know about software development in startups. *IEEE Software*, 31(5):28–32, 2014.

[31] Stuart C. Gilson and Jerold Warner. Junk bonds, bank debt, and financing corporate growth. *Harvard Business School Working Paper*, 98(037), 1997.

[32] Shawkat Hammoudeh, Ahdi Noomen Ajmi, and Khaled Mokni. Relationship between green bonds and financial and environmental variables: A novel time-varying causality. *Energy Economics*, 92:104941, 2020.

[33] Jia He and Lilian K. Ng. The foreign exchange exposure of japanese multinational corporations. *The Journal of Finance*, 53(2):733–753, 2021/09/22/ 1998.

[34] M. Hernán and J. Robins. *Causal inference: What if*. Chapman & Hall/CRC, Boca Raton, FL, USA, 2020.

[35] Marc Hesenius, Nils Schwenzfeier, Ole Meyer, Wilhelm Koop, and Volker Gruhn. Towards a software engineering process for developing data-driven applications. In *2019 IEEE/ACM 7th International Workshop on Realizing Artificial Intelligence Synergies in Software Engineering (RAISE)*, pages 35–41, 2019.

[36] Stephanie Houde and Charles Hill. What do prototypes prototype? *Handbook of Human-Computer Interaction*, (2), 1997.

[37] Martin Höst, Björn Regnell, and Per Runeson. *Att genomföra examensarbete*. Studentlitteratur AB, 2006.

[38] Werner Jan. Risk and risk aversion when states of nature matter. *Economic Theory*, 41(2):231 – 246, 2009.

[39] Yuri Jolly. Data lineage: Data origination and where it moves over time | FSI. *Deloitte Netherlands*.

[40] Philippe Jorion. The exchange-rate exposure of us multinationals. *The exchange-rate exposure of US multinationals*, 63(3):331–345, Jul 1990.

[41] S. Lauesen. *Software Requirements: Styles and Techniques*. Addison-Wesley, 2002.

[42] Mats Levander, Carl-Johan Rosenvinge, and Vanessa Sternbeck Fryxell. The Swedish derivative market, June 2021. Staff memo.

[43] Leszek Maciaszek. *Requirements analysis and system design*. Pearson Education, 2007.

[44] Chaudhary Wali Mohammad, Mohd. Shahid, and Syed Zeeshan Hussain. Fuzzy attributed goal oriented software requirements analysis with multiple stakeholders. *International Journal of Information Technology*, Jan 2018.

[45] Sergio Luján Mora. Why is important prototyping? *Human-Computer Interaction (course material)*, 2015.

[46] Hiroko Nagashima and Yuka Kato. Aprep-dm: a framework for automating the preprocessing of a sensor data analysis based on crisp-dm. In *2019 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*, pages 555–560, 2019.

[47] Deana R. Nance, Clifford W. Smith, and Charles W. Smithson. On the determinants of corporate hedging. *The Journal of Finance*, 48(1):267–284, 2021/09/22/ 1993.

[48] J Nielsen. Usability engineering. *AP Professional*, 1993.

[49] Sara Nilsson Tengstrand, Piotr Tomaszewski, Markus Borg, and Ronald Jabangwe. Challenges of adopting safe in the banking industry – a study two years after its introduction. In Peggy Gregory, Casper Lassenius, Xiaofeng Wang, and Philippe Kruchten, editors, *Agile Processes in Software Engineering and Extreme Programming*, pages 157–171, Cham, 2021. Springer International Publishing.

[50] Bank of England. Data quality framework, 03 2014. Accessed: 2021-12-01.

[51] N. Paternoster, C. Giardino, M. Unterkalmsteiner, T. Gorschek, and P. Abrahamsson. Software development in startup companies: A systematic mapping study. *Information and Software Techn*, 56(10):1200–1218, 2014.

[52] Neil Record. *Currency Overlay*. John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex PO19 8SQ, England, 2003.

[53] Per Runeson and Martin Höst. Guidelines for conducting and reporting case study research in software engineering. *Empirical Software Engineering*, 14(2):131 – 164, 2009.

[54] Jeff Saltz. Crisp-dm is still the most popular framework for executing data science projects. *Data Science Process Alliance*, November 30, 2020.

[55] Chhavi Saluja. Data preparation - a crucial step in data mining, Feb 2018.

[56] Benjamin Saunders, Julius Sim, Tom Kingstone, Shula Baker, Jackie Waterfield, Bernadette Bartlam, Heather Burroughs, and Clare Jinks. Saturation in qualitative research: exploring its conceptualization and operationalization. *Quality Quantity*, 52, 07 2018.

[57] Franziska Schäfer, Christian Zeiselmair, Jonas Becker, and Heiner Otten. Synthesizing crisp-dm and quality management: A data mining approach for production processes. In *2018 IEEE International Conference on Technology Management, Operations and Decisions (ICTMOD)*, pages 190–195, 2018.

[58] D. Sculley et al. Hidden Technical Debt in Machine Learning Systems. In *Proc. of the 28th Int'l Conf. on Neural Information Proc. Systems*, pages 2503–2511, 2015.

[59] Tripathi Shailesh, Muhr David, Brunner Manuel, Jodlbauer Herbert, Dehmer Matthias, and Emmert-Streib Frank. Ensuring the robustness and reliability of data-driven knowledge discovery models in production and manufacturing. *Frontiers in Artificial Intelligence*, 4, 2021.

[60] Mark Staples, Liming Zhu, and John Grundy. Continuous validation for data analytics systems. In *2016 IEEE/ACM 38th International Conference on Software Engineering Companion (ICSE-C)*, pages 769–772, 2016.

[61] Sara Nilsson Tengstrand, Piotr Tomaszewski, Markus Borg, and Ronald Jabangwe. Challenges of adopting SAFe in the banking industry - a study two years after its introduction. In *Proc. of the International Conference on Agile Software Development*, pages 157–171. Springer, 2021.

[62] Sigmund A. Tronvoll, Christer W. Elverum, and Torgeir Welo. Prototype experiments: Strategies and trade-offs. *Procedia CIRP*, 60:554–559, 2017. Complex Systems Engineering and Development Proceedings of the 27th CIRP Design Conference Cranfield University, UK 10th – 12th May 2017.

[63] Matt Trotter. Machine learning deployment for enterprise, Nov 2021.

[64] Zsuzsa Varvasovszky and Ruairí Brugha. A stakeholder analysis. *Health Policy and Planning*, 15(3):338–345, 09 2000.

[65] Andreas Vogelsang and Markus Borg. Requirements engineering for machine learning: Perspectives from data scientists. *2019 IEEE 27th International Requirements Engineering Conference Workshops (REW), Requirements Engineering Conference Workshops (REW), 2019 IEEE 27th International*, pages 245 – 251, 2019.

[66] Krzysztof Wnuk. Involving relevant stakeholders into the decision process about software components. In *2017 IEEE International Conference on Software Architecture Workshops (ICSAW)*, pages 129–132, 2017.

[67] Robert K. Yin. *Case study research : design and methods.* SAGE, 2014.

[68] Xinbo Zhang. Multinational companies' hedging effectiveness of foreign exchange risk: A quantitative comparison study. *Fudan Journal of the Humanities and Social Sciences*, 14(2):285–302, Jun 2021.

INSTITUTIONEN FÖR DATAVETENSKAP | LUNDS TEKNISKA HÖGSKOLA | PRESENTERAD 2021-06-03

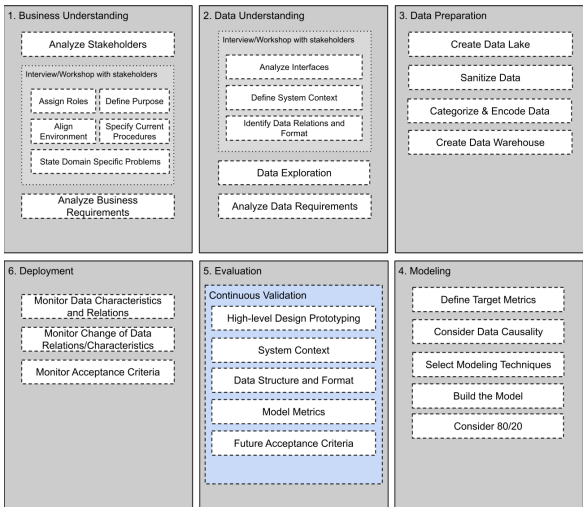
**EXAMENSARBETE** Concretizing CRISP-DM for Data-Driven Financial Decision Support Tools.
**STUDENT** Joel Järlesäter and Simon Grimheden
**HANDLEDARE** Markus Borg (LTH)
**EXAMINATOR** Elizabeth Bjarnason (LTH)

# A more detailed process model for prototyping data-driven applications

POPULÄRVETENSKAPLIG SAMMANFATTNING **Joel Järlesäter and Simon Grimheden**

While most people are aware that the use of machine learning and artificial intelligence is rapidly growing, few teams have a clearly defined methodology of prototyping and implementing applications using these technologies. By utilizing this detailed process model, teams might very well have a better success ratio in future projects.

Artificial Intelligence and Machine Learning are topics which have been researched for many decades, but only in recent times have computing power become sufficient to make these technologies available to the general public. The majority of projects within the area of data-driven applications, run by less practically experienced developers, are usually done in an exploratory and unstructured way. To support the development of these types of applications, development process models and frameworks such as Cross Industry Standard Process for Data Mining (CRISP-DM) have been created. With it's six phases, it gives a high-level structure of what the development process should include. However, it does not provide specific recommendations for each phase. In this study, we've interview five developers/data scientists active in the industry in order to create a detailed and more helpful process model to guide teams when implementing, but mostly prototyping, data-driven applications. These interviews are based on the problem domain of financial decision support tools and the challenges posed by the domain, namely (1) a large amount of interfaces, (2) uncertainty of data, and (3) difficulty in defining the main purpose of the data-driven modules included in the applications.




Our findings resulted in a new detailed version of CRISP-DM with concrete activities for use when developing and prototyping data driven applications. This detailed process model provides both experienced and inexperienced developers with a rigid framework to use to ensure a successful development cycle. The detailed process model is based on the same six development phases as the original CRISP-DM, however the evaluation phase has been modified in order to encourage continuous validation, resulting in more efficient development.