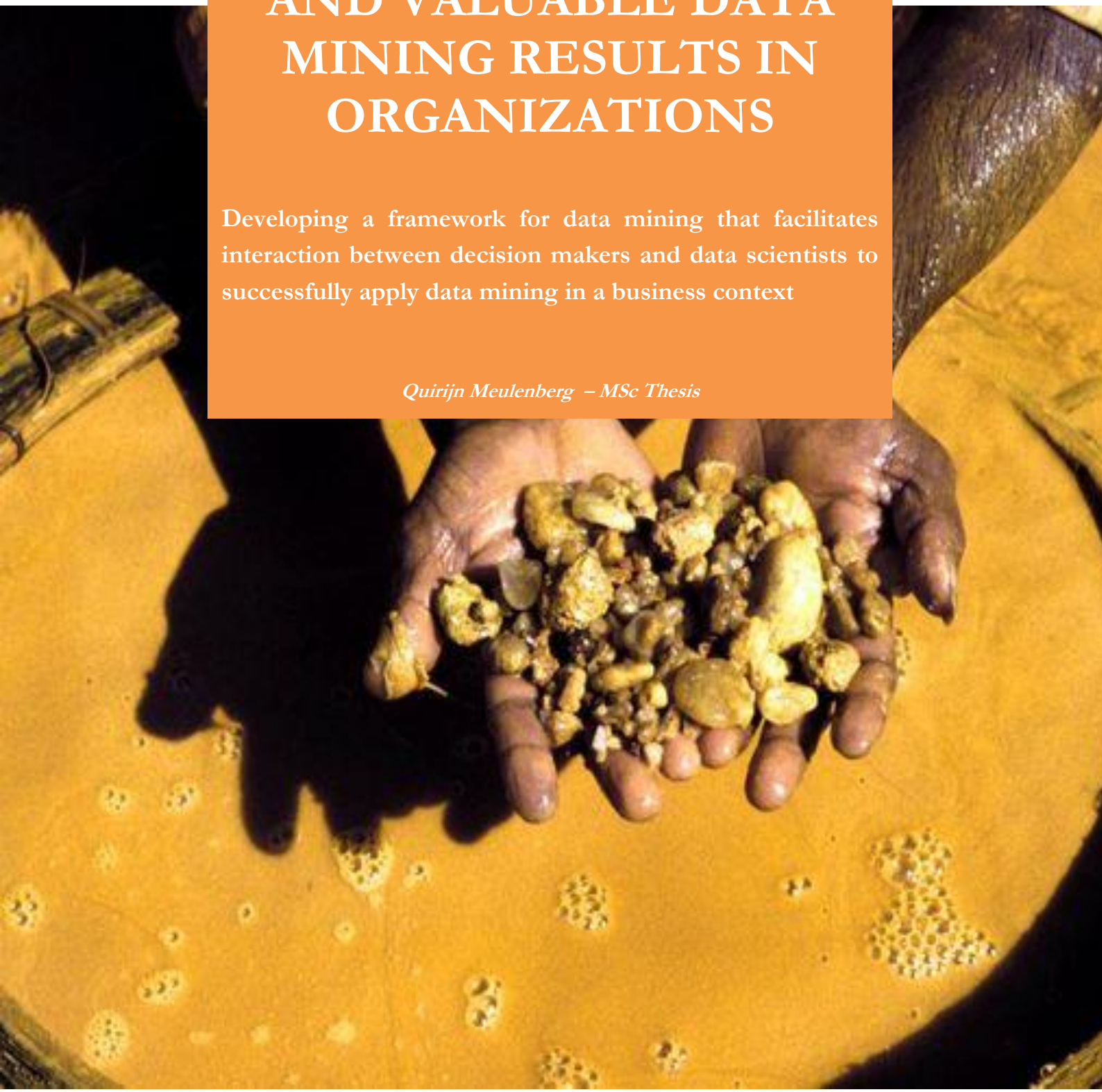
# TOWARDS MEANINGFUL AND VALUABLE DATA MINING RESULTS IN ORGANIZATIONS

**Developing a framework for data mining that facilitates interaction between decision makers and data scientists to successfully apply data mining in a business context**

*Quirijn Meulenberg – MSc Thesis*



TU Delft Delft University of Technology

[redacted]

# Towards Meaningful and Valuable Data Mining Results in Organizations

*Developing a framework for data mining that facilitates interaction between decision makers and data scientists to successfully apply data mining in a business context*

**PUBLIC VERSION**


By

**Q.J.J. (Quirijn) Meulenberg**


Report for the partial fulfillment of the requirements for the degree of

**Master of Science**

In Systems Engineering, Policy Analysis & Management

At Delft University of Technology,

To be defended publicly on August 18th 2015, 15:00 at the Faculty of Technology, Policy and Management of Delft University of Technology


**Quirijn Jan Jonathan Meulenberg**

1523112

MSc SEPAM

Systems Engineering, Policy Analysis & Management

Faculty of Technology, Policy and Management

Delft University of Technology


**Graduation committee**

Chair: Prof.dr.ir. A. Verbraeck *(TU Delft)*

First supervisor: Dr. M.A. Oey *(TU Delft)*

Second supervisor: Dr. H.G. van der Voort *(TU Delft)*

External supervisor: Dr. H. de Koning [redacted]

TU Delft Delft University of Technology

[redacted]

# Preface

This report is the final piece of the puzzle that I solved in 7 years at the Faculty of Technology, Policy and Management at Delft University of Technology. With this report, I complete the master program Systems Engineering, Policy Analysis and Management to proudly call myself an 'ingenieur'.

For this report, I conducted a literature study and a case study at [redacted] bank in the Netherlands. The goal of this research was to design a framework that could add to the successful application of data mining in organizations. It contains both technical improvements of data mining, as alterations in the way that stakeholders interact in a data mining project.

Although I am proud that my name is on the cover, this report is indebted to many people who, each in their own way, have contributed to the end result.

Alexander Verbraeck for being an outstanding chair who always focusses on solutions instead of the problem I was dealing with. This helped me greatly in the most problematic phase of my research. Haiko van der Voort for his positivity and his comments. This led me to consider data mining from a totally different perspective. This different perspective to me is one of the major added values of my work. Henk de Koning for giving me the freedom to shape my own research within a large and complex project. Besides, you adequately pinpointed the necessary improvements and kept patience in what sometimes was a struggle. Your patience allowed me to chase my academic challenge, while most of the challenges for [redacted] had already been dealt with. Last but certainly not least, I would like to thank Michel Oey for the thousands of questions we asked and answered together. Your interest for both the big picture and the tiny details made my thesis something to be proud of. Moreover, thank you for the regular appointments that always exceeded the scheduled time by at least 30 minutes and always gave me new insights, both on data mining as well as on orchestras. I now know that orchestras are complex worlds; maybe one day a framework will be developed that streamlines all their complex interactions.

My internship at [redacted] has been a true enrichment of my professional and personal life. Therefore, I owe a big 'thank you' to my colleagues at the [redacted] Blackbelt department, who made my internship enjoyable and fruitful. Other colleagues, working at the Big Data department and the team members from [redacted], helped me with their efforts and their enthusiasm, in order to secure the value for [redacted] of the insights of this research. I hope my research proves to stay of value for the seamless integration of the communication channels.

I would recommend every potential freshman to go and study in Delft. Living and studying in Delft for me proved to be the perfect recipe for inspiring and magnificent student days. I could have never succeeded in Delft without the many friends I have and got to meet here. Above all, I would like to thank my sister, brother, parents and Claartje. Without your patience to listen to stories on a subject that does not at all appeal to the imagination, you helped me strain and broaden my scope. With your leisure time, you gave me the energy to complete what sometimes was a challenge and to enjoy it all.

**Quirijn Jan Jonathan Meulenberg**

# Executive summary

## Introduction

Estimations show that an organization's productivity can increase by 6% if they can effectively use their existing data in decision making processes (Brynjolfsson et al., 2011). To exploit this potential, an organization should be able to successfully apply advanced data analytics in their decision making processes. The process of applying advanced analytics on large datasets is often referred to as data mining. When data mining is applied in business, the goal is to extract 'previously unknown, valid, and actionable information from large databases and then using the information to make crucial business decisions'(Cabena et al., 1999). From a high-level perspective, data mining converts data to information ('patterns') that can then be interpreted as knowledge.

## Research problem

However, in practice it is hard to interpret and implement the findings of data mining research. Data mining often delivers a large set of results, while not all results are interesting to the decision makers. Moreover, if data mining is deployed in organizations, decision makers find it hard to determine the meaning of a result, let alone that decision making on basis of data mining is possible (Han et al., 2007). These barriers prohibit that data mining in organizations can live up to the potential that has been identified. This lead to the following research question, that is the core of this research: "*How can data mining be integrated in organizations in such a way, that the results can help decision making in the organization?*"

## Research approach

The goal of this research is to design a framework for data mining in organizations. In this way, the function of data mining in business is captured. To this end, a design science research is carried out (Hevner, 2007; Hevner et al., 2004). Design science research is suitable for the design of a framework, as it focuses on the design of an artifact as a deliverable. Moreover, it provides an effective way to structure findings of both literature analysis and a case study. The requirements for data mining in organizations have been analyzed with an extensive literature analysis. In this approach, not only the internal complexities of data mining are taken into account, but also the complexities that exist between data mining and the stakeholders involved. In this way, both the technical aspects as the process of data mining can be improved. Besides, a case study at [redacted] bank for call reduction through web site improvement has been conducted. This case study was carried out as Participatory Action Research. The insights gained from the literature analysis and case study resulted in the a set of functional and non-functional requirements. This set of requirements is used to design propositions that can improve parts of the data mining process. Finally these propositions have been combined in one single framework for data mining in organizations in a participatory setting: the Convergence-Data mining-Evaluation-Framework (see Figure 1).

## Findings

From the literature analysis, the technical obstacles are caused by a lack of adoption of evaluation criteria. As a true data mining pattern should be valid, valuable and novel, in practice only validity evaluation measures are used. In this way, the value of a pattern to the decision maker is not taken into account, neither is the process aimed at searching for new knowledge. It seems to be hard to combine evaluation measures that comply with the three requirements for a data mining outcome. Currently, data mining is focused on objective outcomes, while the value of a patterns should be determined by problem-specific information.

The application of the data mining process in business is hard, because of the different perspectives that decision makers and data scientists use. While decision making uses deductive reasoning to reduce uncertainty, data mining is focused on inductive reasoning based on observations. Deduction is needed to come to actionable patterns, while the unique strength of data mining is the inductive search for previously unknown patterns. The value of a pattern, where the problem is converged to a known solution, and the novelty, the divergence of the observations to discover new insights, seem to contradict each other. Furthermore, the participation of decision makers in data mining is limited, creating a gap between the data mining phase and the interpretation phase. Since the data scientists are often separated from the decision makers in the organization, possible essential business knowledge is exempted from the process.

## The Convergence-Data mining-Evaluation framework

In order to solve these problems, five propositions have been designed: the *triple evaluation approach*, the *decoupling of the evaluation steps,* the *design of a convergence process*, *performing an attainability test* and to *consider evaluation phase as a boundary object.*

The triple evaluation approach explicitly takes into account three requirements for a data mining pattern: validity, utility and novelty. If patterns are evaluated along these three axis, true data mining patterns will result from the evaluation phase. Since the three requirements each measure something

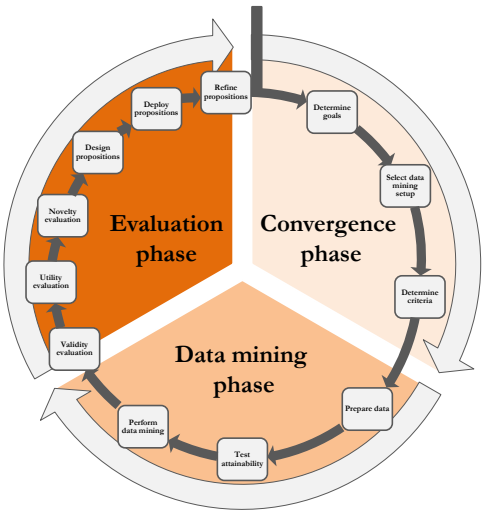



**Figure 1: the CDE-Framework**

totally different, the three evaluation steps are decoupled. In this way, objective evaluation is decoupled from the subjective evaluation, and induction for new knowledge is separated from the deduction for high-utility patterns. First, only statistically significant patterns are selected in the validity evaluation. Secondly, potential beneficial patterns are extracted from the valid patterns in the utility evaluation. Finally, the search for new knowledge is focused on providing a general rule for high utility patterns. In this way, the search for new information is focused on the aspects where it is possibly most relevant.

The evaluation criteria are determined beforehand, as the result of convergence between decision makers and data scientists. Through this convergence, the project can be better fitted to the problem situation. During this convergence, agreements on the data mining method and algorithm are made. To check whether this algorithm is suitable on the selected dataset, an attainability test has been designed to prevent delays further on in the process. Finally, to secure that the framework is applicable in organizations, the evaluation phase should allow decision makers to design propositions not only for valid, valuable and new patterns, but also for valid and valuable patterns. In this way, the most gain for the organization can be achieved. In order to specify how these propositions should work accordingly, a framework has been designed: the Convergence-Data mining- Evaluation Framework (CDE-F, see Figure 1).

## Applying the framework in business

As described, the framework facilitates convergence between the stakeholders in the data mining project. Furthermore, it describes three requirements for successfully securing the process in an organization. First, the information uncertainty can be reduced by delivering meso-level insights. In this way, the barrier for actually performing data mining will be lower. This aggregated insight can then be used to convince other (possible) stakeholders. Besides, since the process is quite intensive, a process manager may be needed to unify all stakeholders and stimulate decision making between parties involved. The way in which the framework should be applied in an organization, depends on organizational context it is deployed in. These context is defined by five aspects: the *centralization of decision making*, the *type of knowledge embedding*, the *external uncertainty*, the *dependencies between decision makers* and the *strategic nature of the problem*. Finally, it is crucial for the success of the framework that the valid and high-utility patterns can be deployed, even if they are not novel.

## Reflection

The unicity of this research is that data mining is approached from a technical and process perspective and that it focused on the decision maker instead of the data scientist. Given the limited time frame of the research, the generalizability of the framework was tested through though-experiments instead of multiple cases. While the framework was deemed generalizable by an expert, the generalizability is not tested in practice. The scalability of the approach is also not quantitatively proven, as the case study used a relatively small dataset.

## Future research

As the CDE-F is the starting point of considering data mining as a socio-technical system, it spikes several questions for future research. Since the data CDE-F framework is not validated in another case study but through thought experiments, the generalizability of the framework is not quantitatively determined. In order to prove that the framework is also valuable in very different settings of data mining, previous research can be conducted for the validity of the CDE-F.

A possible interesting comparison of the CDE-F is that with the storytelling skills of a data scientist. Although it has received little scientific attention, the benefits of solving the interpretation problem with soft skills is a potentially very simple solution.

This framework introduces the concept of novelty evaluation of patterns. In this report, it was used to gain additive insights by reflecting on the utility evaluation. However, the research and validation have also indicated possible other applications of novelty evaluation. Future research may focus on novelty evaluation, in order to strengthen its definition and its role in the framework.

# Contents


**Preface** ii

**Executive summary** iii

**Contents** vii

**1 Introduction** 1

1.1 An introduction to data mining 1

1.2 The data mining process 3

1.3 Problem statements 4

1.4 Data mining as socio-technical system 6

1.5 Data mining at the [redacted] call department 7

1.6 Research objective 9

**2 Research methodology** 11

2.1 Motive for research 11

2.2 Explorative research 12

2.3 Design science approach 12

2.4 Research overview 17

**3 Literature review** 18

3.1 Rule quantity problem 18

3.2 Rule quality problem 19

3.3 Misalignment between data mining and decision making 23

3.4 Non-functional requirements for data mining in business 31

3.5 Literature review conclusions 33

**4 Case study** 35

4.1 Data mining in the [redacted] organization 35

4.2 Data mining in the Prevent & Shift scrumteam 38

4.3 Carrying out the data mining project 40

4.4 Results 49

4.5 Insights gained from the case study 58

4.6 Case study conclusions 62

**5 Design propositions** 65

5.1 Overview of the design propositions for data mining framework 65

5.2 Triple evaluation approach 69

5.3 Decouple evaluation steps 72
5.4 Design a process for convergence and criteria determination 74
5.5 Perform attainability test 79
5.6 Allow freedom of interpretation to deal with friction 80
**6 Framework design 82**
6.1 The CDE-framework for data mining in organizations 82
6.2 Tailoring of the framework to the organizational context 87
6.3 Securing the CDE-F in organizations 91
**7 Framework validation 94**
7.1 Generalizability of the CDE-Framework 94
7.2 Expert validation 97
**8 Reflection 99**
8.1 Research outcomes 99
8.2 Research process 100
**9 Conclusions and recommendations 103**
9.1 Conclusions 103
9.2 Limitations 105
9.3 Recommendations 106
9.4 Future research 107
**Bibliography 109**
**Appendix A: Overview of data mining 120**
**Appendix B: Case study appendices 123**
**Appendix C Validation interviews 155**

# 1 Introduction

The increasing availability of computers with high computational power and the increasing storage capabilities and possibilities have been a very fertile soil for the development of large scale data collection and advanced statistical analyses. Currently, the world produces 1.7 quadrillion ($10^{15}$) bytes a minute, and the data sector experiences a 40% growth a year (European Commission, 2014). Estimations of the benefits show that an organization's productivity can increase by 6% if they are able to make use of this data in their decision making process (Brynjolfsson et al., 2011). To exploit this potential, an organization should be able to successfully apply advanced data analytics in their decision making processes.

## 1.1 An introduction to data mining

A term that is often used for advanced data analytics is data mining (DM) or knowledge discovery in databases (KDD). In this report, the term data mining is used. It is an umbrella term for a collection of methods and techniques.

### 1.1.1 Definition

These methods and techniques are then applied to datasets in order to translate or transform data for reasons of usefulness, compactness or abstractness (Fayyad et al., 1996). This is a rather broad definition, as there are many reasons for and many ways to apply data mining for many different purposes. For this research, a definition more pinpointed to the business application of DM is used: 'the process of extracting previously unknown, valid, and actionable information from large databases and then using the information to make crucial business decisions' (Cabena et al., 1999).

This definition shows that data mining's purpose is twofold: it needs to *extract information from data* and that *information needs to be used for decision making*. So, for data mining in business, the focus should not only be on how to translate or transform the data, but also on how to use the result from data mining in the organization.

### 1.1.2 Data, information and knowledge

In order to understand the aforementioned purpose, the distinction between data, information and knowledge needs to be explained. The three terms vary in their degree of abstraction and interpretation (Bhatt, 2001).

Data is the most simplified form of reality, as it contains raw facts of reality. In data, parts of reality that beforehand have been considered uninteresting are removed (Mitra, 2012). For data mining, data serves as the input of the process.

Information is the organized version of data, in such a way that it shows an understanding of relations between the data and the context (Light et al., 2004). In data mining, a significant (ir)regularity of multiple cases can be regarded as information.

Knowledge is the information embedded in people, so it is information that interacts with peoples' perspectives and beliefs and so forms a substantiation for human action (Drucker, 1989; Nonake & Takeuchi, 1995; Sveiby, 1997). So, data mining results can only represent knowledge if a user accepts the information and knows what to do with it.

### 1.1.3 Methods

Data mining offers a large toolbox of varying method to extract information from large databases. These methods all have different objectives. Their suitability depends on the data mining task and the type of information that is needed. In the standard reference work, Fayyad et al. describe 6 different methods of data mining: anomaly detection, association rule mining, clustering, classification, regression and summarization (Fayyad et al., 1996). These are shortly introduced below. Table 16 in Appendix A shows an overview of the literature study performed on the different types, with suggestions for further reading.

Anomaly detection has as goal to find non-frequently occurring exceptional patterns in large databases. This is often used for fraud detection or other malicious activity that is hard to detect (Chandola et al., 2009).

On the other hand, association rule mining is aimed at discovering frequent relationships between items in a database. Here the focus is not on exceptional cases, but the most common cases in a database. These analyses are typically used for a 'market basket analysis', to discover sets of products that are frequently bought together (Agrawal et al., 1993).

Clustering is searching and labelling of groups of cases that are similar to each other. The boundaries of such objects are then undefined and determined in the data mining process (Han & Kamber, 2006). This is called unsupervised learning. Here the focus is not to extract an exceptional or common relationship, but to aggregate the types of cases that are in the database. A common application is the identification of customer segments in customer relationship management (Ranjan & Bhatnagar, 2008).

Classification resembles clustering, but classification can be applied if the boundaries of a cluster are already known or should be determined deliberate. This is called supervised learning. The most commonly known application of classification is the way that search engines classify a search request to different topics. For instance, the search term 'windows' can belong to the class of computer software and to the class of actual windows, while the search for 'apple' will deliver results of fruit and electronics.

Regression is applied to gain insights in the relations between a set of variables and a selected target variable. This information is then used to compute the target variable if the independent variables are known. This is a classic method of statistical analysis, that can be used e.g. to predict the optimal stock size of ice cream in a supermarket by analyzing the weather forecast and the advertising budget spent (Fayyad et al., 1996).

Summarization is the compact representation of the data, such as an automated report, table or a graphic. The objective of summarization is to create insights in the data that is mostly used to explore the data. Common summarization techniques is the representation of means and standard deviations of variables (Fayyad et al., 1996).

### 1.1.4 Techniques and algorithms

As the DM method specifies the objective of the research, the DM technique specifies the way in which this objective is measured. A data mining technique specifies in what kind of data is analyzed in what way (Han & Kamber, 2006). If the objective of the data mining research is known, a deliberate decision on the method to satisfy the objective should be made (Berson et al., 1999). Appendix A shows a number of data mining techniques for every method and suggestions for further reading. It shows that each data mining method has a different goal. An example: in anomaly detection the goal is to find low frequent and surprising cases, while association rule mining focusses on relations with a high frequency. Since data mining is so diverse, it is hard for researchers to present a unifying perspective on data mining (Yang & Wu, 2006). In this report, all data mining methods are taken into account. The generalizability of the approach then has to be taken into account.

Every data mining technique can have different algorithms that can handle the data in the way the method prescribes. An algorithm can be regarded as the operationalization of the DM technique: it provides the code that specifically states the computational procedure to carry out the data mining technique (Cormen et al., 2009). Algorithms for the same technique differ in the way they compute and display the output. In this case, there is no best technique or algorithm (Nisbet, 2004). When the DM method has been decided, a deliberate decision has to be made on the exact algorithm that is used.

### 1.1.5 Applications

Since data mining has a large variety of methods and techniques, it can be used for many different purposes in many different domains. Data mining is not only applied in the field of information science, but also in more common domains. Literature reviews report of data mining applications in very different domains, such as healthcare, banking, retail, telecommunication and education (Kantardzic, 2011). Table 1 shows more concrete examples of data mining applications in these fields. The domains are very different from each other, and so it also differs what they want to get out of DM applications.

## 1.2 The data mining process

Although almost every data mining execution is unique, there is some generalizability in the way that the DM process is performed. Regarding data mining as a black box, every data mining research should facilitate two processes to be regarded as data mining: big data can be inputted in the process and knowledge can be extracted from the process (see Figure 2) (Fayyad et al., 1996; Nisbet et al., 2009). What separates DM from other methodologies that can perform these functions, is that it a DM application is applied in a semi-automatic or fully automatic way (Witten & Frank, 2005).



If the black box of data mining is opened from a technical perspective, a data scientist has to select the right data and transform this to be suitable for the right analysis. Then he has to perform the appropriate analysis and has to interpret and evaluate the outcomes. From a technical perspective, the iterative process then consists of five steps: data selection, data preprocessing, data transformation, data mining and pattern evaluation (Fayyad et al., 1996). The process is graphically presented in Figure 3. The function of each step in the process is briefly described in Table 1. A more extensive description of the data mining process is presented in Appendix A.

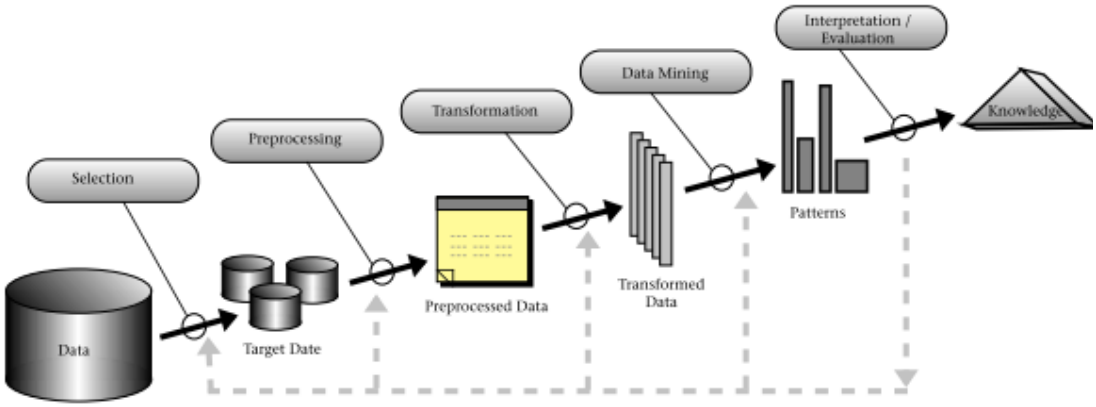



**Figure 3: the steps of the data mining process (Fayyad et al, 1996)**

**Table 1: a brief overview of the technical steps of the data mining process**

| Process step | Function |
|---|---|
| Data selection | Selection and isolation of the appropriate data for the analysis |
| Data preprocessing | Make data suitable for analysis in the data mining phase |
| Data transformation | Transform data or variables to more interpretable values |
| Data mining | Apply DM setup (method, technique, algorithm) to target data set |
| Pattern evaluation | Translate results into knowledge and action |

## 1.3 Problem statements

According to the definition of Cabena et al. , after the pattern evaluation phase the decision maker has valid, actionable and previously unknown knowledge that he can use to make crucial business decisions (Cabena et al., 1999). In practice however, it is hard to interpret and implement findings of DM research (Yang & Wu, 2006). Several sources conclude that pattern evaluation is too less studied (Berkhin, 2006; Köksal et al., 2011). The problematic interpretation of results has three main causes: the rule quantity problem, the rule quality problem and the misalignment between DM and decision making in organizations. Figure 4 gives an aggregated system view of data mining in organizations, and shows what elements are affected by the three problems.

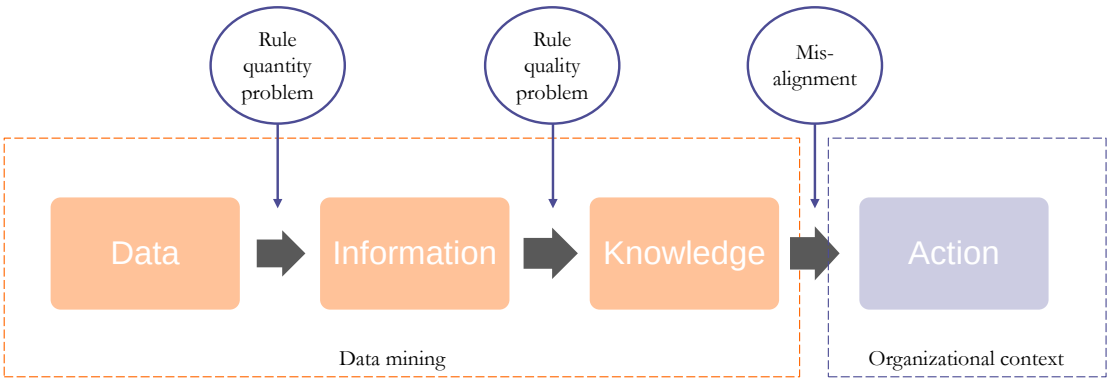



**Figure 4: the position of the three problem statements in the system of data mining n organizations**

### 1.3.1 Rule quantity problem

There are two main technical reasons why the interpretation of DM is troubled: the rule quality problem and the rule quantity problem (Guillet & Hamilton, 2007; Paul et al., 2014).

In data mining, large sets of data is analyzed statistically. To evaluate significant results, traditional significance measures are used, that are also used to find significant relations in small data sets(Boyd & Crawford, 2012). The rule quantity problem then occurs, because large sets of input data will then automatically lead to large sets of significant patterns (Kantardzic, 2011; Lenca et al., 2008). In practice, manual inspection of the delivered set of patterns is almost impossible and highly undesired (Natarajan & Shekar, 2005). Moreover, it also causes the rule quality problem.

### 1.3.2 Rule quality problem

The rule quality problem entails the usability of resulting patterns. If a DM application delivers so many results, not all results can and should be deployed by an organization. Besides the fact that this is an impossible operation, not all mined patterns are interesting (De Bie & Spyropoulou, 2013; Paul et al., 2014). However, the pattern evaluation phase is expected to select only the interesting patterns. What troubles this evaluation, is that there is no uniform definition of what this interestingness really means (Geng & Hamilton, 2006a).

When DM is deployed in a complex domain, resulting patterns may contradict the decision maker's domain knowledge or common sense. For a decision maker, it is hard to judge whether these patterns are useful to him or if they are spurious. This poses three threats (Vidulin et al., 2014). Firstly, if a decision maker draws a conclusion on a spurious relation, actions based on DM will not lead to the desired result. Secondly, if a decision maker wrongly judges that a contradictory pattern is spurious, new insights from DM are lost. Moreover, if a decision maker judges contradictory outcomes as spurious, he can distrust data mining or the selected data mining suite. This will have a negative effect on the adoption of data mining in business.

### 1.3.3 Misalignment between data mining and decision making

The rule quality problem indicates that it is possible to it is hard to translate information into knowledge, and eventually into action and value. These are the steps where the data, a simplified representation of reality, needs to be enriched again to reality. If DM is to be deployed in a business

context, decision makers find it hard understand what such a result really means (Han et al., 2007). DM is mostly performed by data scientist, and interpretation and decision making is mostly performed by other people that don't have expertise in DM (Harding et al., 2006; Kusiak, 2006; Wang et al., 2007). To improve data mining in a business context, the DM process and decision making processes in business have to be aligned.

One easy solution is then to change the required competencies in recruiting new managers so, that one person can be responsible for executing data mining and decision making. However, since the trend of advanced analytics is expected to keep increasing, a McKinsey report already predicted a large shortage of data scientists and data-savvy managers in the near future (Brown et al., 2011; Chen et al., 2012). Merging the competencies in one business role is then a utopia. The aim of the solution then must be that both schools-of-thought have to be aligned.

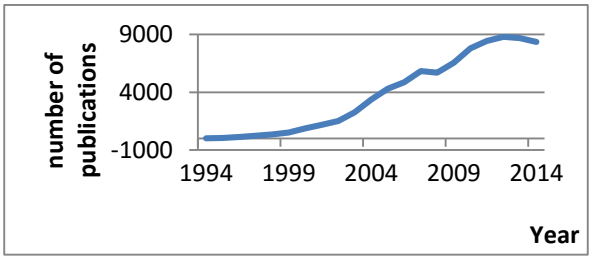



**Figure 5: Publications on data mining (source: Scopus)**

## 1.4 Data mining as socio-technical system

While the rule quantity and rule quality problem approach the actual data mining process, the misalignment problem describes complications with the way DM is treated in organizations. The problems described in the previous paragraph thus not only lie in the technical aspects of data mining, but also in the way that people deal with data mining in organizations. The pattern evaluation phase then is regarded as the interface between the technical system and the way that people handle the system.

By indicating a gap that focuses on the interface between the DM system and its actors, the idea that DM serves merely a technical function is rejected. In this research, DM then should be treated as a socio-technical system (Kroes et al., 2006). In that case, the function of DM is not only understood by their technical aspects, but also in the context of the human action DM eventually has to interact with. The way stakeholders work together then also becomes relevant for improving data mining (Baxter & Sommerville, 2011).

If data mining is regarded from a socio-technical perspective, it can be regarded as a radical innovation that needs to be secured in the current socio-technical regime of decision making in business. From 1995 onwards, new algorithms and advanced analyses developed quickly and the publications about DM did too (see Figure 5). Currently, the identified need for a unifying theory of DM and the problems of applying DM in business give a signal that organizations are trying to incorporate DM in their current way of working (Geels, 2002; Yang & Wu, 2006). If the current socio-technical system of decision making does not give a window of opportunity for DM to be incorporated, DM is unlikely to have long term impact on businesses and organizations through decision making. To be able to achieve the 6% benefit as mentioned by Brynjolfsson et al.(2011), obtaining a socio-technical view on DM is of added value.

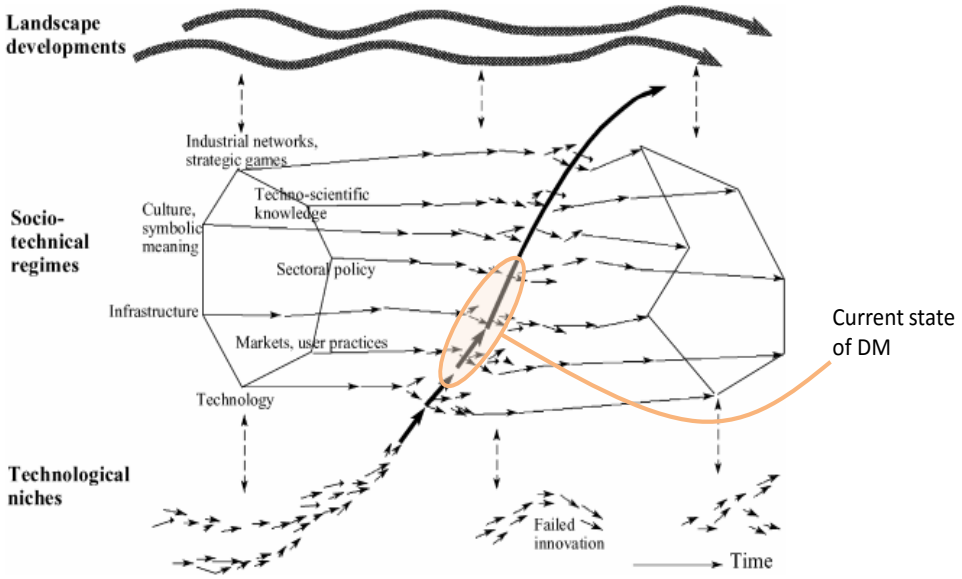



**Figure 6: Data mining is entering the socio-technical regime of business decison making**

## 1.5 Data mining at the [redacted] call department

The occasion that gave rise for this research, is a case study at [redacted] Bank Netherlands (from now on called [redacted] or [redacted] bank'). [redacted] A significant reduction of the calls will lead to a significant cost reduction for the call department. The goal of call reduction fits [redacted] overarching milestone that all of [redacted] customers perform all their daily banking business online.

Up until now, many call reduction strategies have been developed and implemented in the call center and with success. [redacted] The call department is therefore looking into advanced and sophisticated ways to further reduce calls, where decision makers have indicated great potential benefit for data mining too help them to make sophisticated solutions.

### 1.5.1 Channel shift

One of the potential solution directions is shift customers to another channel of service. Contact centers of large organizations offer their customers a broad variety of channels for providing service and making sales. Multi-channel customer strategies are "the design, deployment, and evaluation of channels to enhance customer value through effective customer acquisition, retention, and development" (Neslin et al., 2006). An [redacted] customer has mainly three options to get in touch with the bank: visit a branch, call the call center ('Call') or visit the [redacted] website ('Online', including app).

Organizations are constantly scanning for ways to reduce costs while minimally influencing other key performance indicators, such as customer satisfaction and sales. For companies that have large customer groups, providing service to such a large customer group is expensive. Offering the possibility to shift to another channel while retaining the service level can be an effective measure to both increase revenue and save costs (Capgemini, 2012). [redacted]

Especially for service providers, controlling costs of service channels and increasing efficiency can significantly improve the financial performance of the company. Earlier studies done for governments reported that the cost per contact can be reduced up to 98% when shifting contacts from call centers to online channels (PriceWaterhouseCoopers, 2009) [redacted]

[redacted]

### 1.5.2 [redacted] Call Reduction project

In this light, [redacted] started the Call Reduction program. This program consists mainly of two projects: Call Prevention and Call Segmentation. This program is graphically presented in .

The Call Prevention project has the main goal to [redacted] Therefore, three possible strategies have been devised:

1. [redacted]

[redacted] of calls is by installing an Interactive [redacted] To this end, call prevention program has several issues to tackle:

4. [redacted]

Simultaneously, the Call Segmentation project tries to [redacted]

## 1.6 Research objective

This research will focus on designing a data mining approach that can deliver valuable outcomes for decision makers. Since literature indicates that the problem lies within the pattern evaluation phase, the focus will be on improving pattern evaluation. These findings will be incorporated in a design artefact: the convergence-data mining-evaluation framework (CDE-F) for data mining in organizations.

The research will indicate the discrepancy between the both perspectives and how the integration of the perspectives is lacking in current data mining research. It also focuses on the organizational parts that are involved in data mining, to show that awareness of these involvements is essential for delivering valuable results. To test the design of the improved data mining process, a case study is performed at the [redacted] bank in the Netherlands. The main research question treated in this report is as follows:

*How can data mining be integrated in organizations in such a way, that the results can help decision making in the organization?*

Note that the word 'can' is used instead of the word 'should', as this research is of an exploratory nature.

There are two sub goals in this research: improving the evaluation of DM results by coping with the rule quantity- and rule quality problem, and an integration of the business perspective with the DM perspective in DM research. Therefore, first the current situation needs to be analyzed. This will lead to requirements for the improvement of DM. Consequently, these requirements need to be translated in concrete improvements for data mining projects. Thirdly, the combination of multiple perspectives in one framework should be investigated. Finally, the securization of all the improvements in the organization needs to be discussed, since multiple parties will be affected by the new design. This leads me to the following subquestions:

| |
|---|
| Analysis |
| 1. *How are data mining project carried out currently?* |
| 2. *How are data mining patterns evaluated currently?* |
| 3. *What are the requirements for successful decision making in data mining?* |
| Solution directions |
| 4. *How can the data mining process in organizations be improved?* |
| 5. *How can the evaluation of patterns be improved?* |
| Framework design |
| 6. *How can the data mining perspective and the business perspective be aligned in one hybrid data mining process?* |
| Securization |
| 7. *How should a data mining process be secured in the organization?* |

This report is structured as follows. In chapter 2, the research methodology for the design of the CDE-framework is described, following the design science research of Hevner and Hevner et al. (2007; 2004). Chapter 3 presents the analysis and the outcomes of the literature review. This literature review has been performed as the rigor cycle of the design science research. In chapter 4, the relevance cycle is described: a case study at [redacted] bank for finding propositions to improve [redacted] web site. Chapter 3 and 4 then result in a joint set of requirements to cope with the rule quantity and – quality problem and misalignment between data mining and decision making. Chapter 5 is a documentation and argumentation of the individual design propositions that satisfy the requirements. In chapter 6, the design propositions are combined in the Convergence-Data mining-Evaluation Framework for data mining in organizations. This framework is validated in Chapter 7. A personal reflection on the project and my learnings is presented in Chapter 8. Finally, the report leads to the conclusions and recommendations in Chapter 9.

# 2 Research methodology

As 1.3 and 1.4 show, there are problems with the interpretation of data mining outcomes. There is little focus on the how to translate patterns to action or value in business. Besides, data mining and decision making is badly aligned, due to the different nature of both schools-of-thought. Because data mining is regarded as a socio-technical system, improvements for data mining will be of both substantive and process-oriented nature. These different types of improvements need to be incorporated in one unified framework, where data mining and decision making goes hand in hand.

This chapter will present a systematic approach to cope with the problems described in Chapter 1. First a motivation for the research, with a focus on complex socio-technical systems, will be presented. Secondly, the explorative nature of the research will be described and the different explorative research steps will be introduced. Thirdly, the method for designing a new framework for data mining in business is described.

## 2.1 Motive for research

There is an academic and a societal motive to perform this research. From an academic perspective, the motive is to build desired knowledge on data mining in business. This research thus tries to find a way to combine the approach of data mining in business processes, so that it delivers desired outcomes from both perspectives. A successful combination is then to boost the use and value of advanced analytics in business.

The societal motive can be divided in a general motive and the case motive for [redacted] The general societal motive for this research is better opportunities for successful data mining applications for Currently, IT professionals describe advanced analytics as one of the major trends in the current decade (Chen et al., 2012). It is expected that data mining will be a more and more applied technique in business. Moreover, in a couple of years there will be a large shortage of managers that know how to apply and interpret these advanced data analytics (Brown et al., 2011). Considering this shortage, the previously described gap between data science and business knowledge is not expected to be bridged shortly by a change of competencies. Therefore, a process that could define the necessary interactions between the two parties, and can help the successful adoption of data mining techniques, would be of significant added value for an organization that wants to secure data mining.

From the perspective of the case study at [redacted] which will be further described in Chapter 4, the motive for the research is threefold. First, [redacted] Second, the case study will focus on finding [redacted] inally, the study will give insight in [redacted]

## 2.2 Explorative research

Because the integration of data mining and decision making is a clash of inductive and a deductive reasoning, the discussion will mostly take place on the methodological level. For this type of problems, exploratory research is best suitable to apply (Stebbins, 2001). Besides this, there is no research in the field of data mining how to bridge the gap of the discrepancy between data mining and business. This research tries to bridge this gap by combining different fields of study. Since not much literature has been written on this subject with this aim, the character of this research is best to be described as theory developing rather than theory confirming. The insights gained by combining data mining, decision making and organizations will be combined in a framework for data mining in a business context.

## 2.3 Design science approach

As already stated in the introduction, the artifact of this research is a framework for applying data mining in a business context. In order to do so, the research is focused on technical improvements for data mining, but also in the way that it is used in the organization. The research then examines the interface between technology (data mining), organizations (adhocracies) and people (decision making).

For this type of research, Hevner et al. have developed a design framework how to conduct, evaluate and present design-science research. They state two main requirements for design science: it has to have practical utility and it must be theoretically founded. Therefore they prescribe that design science research has to interact with the problem environment and the knowledge base. They secure this by combining design science with observations of the application environment and with valid theories, methods and processes. In this way, the design science research relies on the invention of new technical solutions that are founded by existing theories and knowledge (Hevner et al., 2004). Figure 11 shows the design framework.

The design science framework relies on three cycles. The main cycle in the framework is the design cycle. In this cycle, an artifact is designed and evaluated. This is where the innovative design takes place in multiple iterations to refine the design. In order to secure that the design cycle is useful in

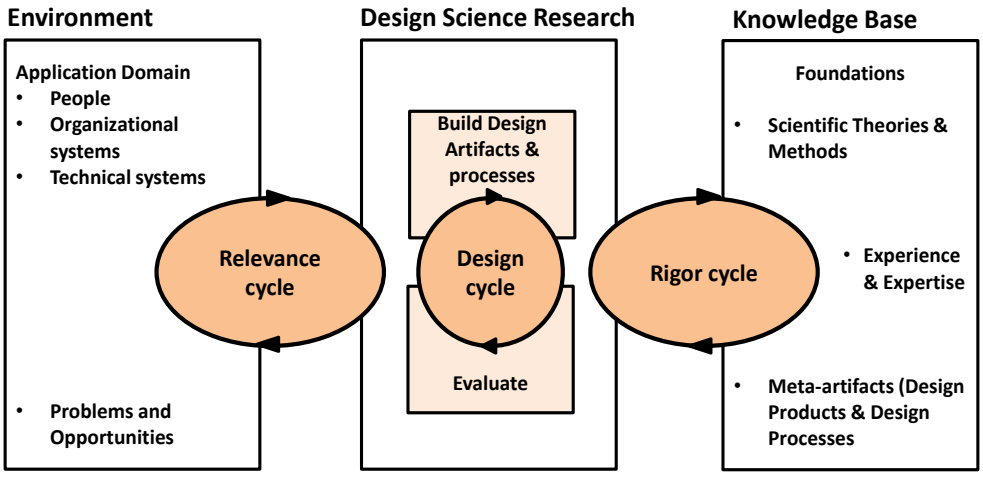



**Figure 8: Design science framework (Hevner, 2007)**

practice, the relevance and rigor cycle delivers propositions for the design from the application domain (Yin, 2014). Using propositions leads to a stronger focus on the output of the design cycles. This closes the gap between the analysis and the synthesis, leading to a clearer substantiation of the eventual design.

The rigor cycle secures that the design science is carried out in a scientific way. It provides prior knowledge for the design cycle, and also leads to propositions for the final design. This has three advantages. First, it provides foundation for the design science by use of existing theories, methods and concepts. Second, it prevents 're-inventing the wheel', the design did not exist before and thus is innovative. Third, lessons learned from the relevance cycle and design cycle can flow back as additional knowledge for scientific theories and methods.

### 2.3.1 Relevance cycle

The relevance of this project will be tested at [redacted] bank. There are mainly two problems there: [redacted] The people that are mainly involved are decision makers and data scientists. [redacted] The problem will be further elaborated in chapter 4.

#### Case study

In order incorporate and evaluate the relevance of the data mining framework, a case study is conducted at [redacted] Bank in The Netherlands. A case study is strong because it is a flexible research design, that allows to draw conclusions on empirical events in a real life context (Schell, 1992). Furthermore, when the boundary between the design and the context is unclear, a case study is very suitable (Yin, 2014). As the context analysis is of great importance in the new framework and is currently vague, a case study is very suitable to tweak and test the design.

A requirement of the case study that should be taken into account is that evidence from the case study is acquired through multiple sources. For this case study, therefor not only data will be gathered through comparison of results of traditional data mining and the proposed methodology, also interviews with the relevant stakeholders will be a source of input and evaluation for the framework.

A main critique on the case study is that in a singular case study design, the scientific generalizability of the proposed framework is not a given. This is also why the research is aimed at developing a theory instead of theory testing, as for testing multiple cases have to be performed. Theory development from case studies is proven to be a very powerful combination, since it is a powerful tool to come to creative, new insights from different fields (Eisenhardt, 1989). The creativity shown in a case study can then further be institutionalized in the design cycle and rigor cycle.

#### Participatory action research

Since the research is of an exploratory nature, the project has been carried out in many different setups. Each time, a new idea can be quickly tested, tweaked or removed from the experimental setup. In order to facilitate the best learning experience of both the researcher as the team Prevent & Shift, the case study is carried out as a Participatory Action research (PAR).

In action research, the researcher is enabled to spirally plan, implement and reflect on suggested improvements (Lewin, 1946). With advancing insights, the researcher can arrive at a design that is both new and tested.

Since the case study is applied in the team Prevent & Shift and the results are intended for further use in the organization, the project group played a role in the agenda setting, data collection and use of the outcomes. This makes the research design Participatory Action Research (PAR). An advantage of PAR is that through the involvement of the actual stakeholders, the research is likely to be relevant and is incrementally validated by the organization (Cornwall & Jewkes, 1995) . For Prevent & Shift, this means that they can influence the research in such a way that they can and want to use the research products after the research.

[redacted] PAR lowered the barrier for the organization to pick the project up themselves and to quickly make adjustments. This brought a sense of ownership to departments involved. The high involvement and low barrier enables the researcher to gain a lot of knowledge quickly. However, the downside of PAR then is that a lot of the knowledge is acquired in an informal way, and not through e.g. questionnaires or interviews. In order to cope with this downside, the insights gained by the researcher were double checked at other employees of the organization.

#### Validation

In action research, the influence of the participants holds advantages and disadvantages. Although many of the downsides are aimed at social research, some also hold for design science research. Most importantly, the project outcome may become influenced by hidden agendas or other strategic behavior of participants (Cornwall & Jewkes, 1995). The resulting framework can then become too much focused on the case study, possibly troubling future applications. Therefore, the generalizability of the new DM-framework needs to be critically evaluated.

To secure that the advancing insights of the case study can be strongly rooted in an academic context, the researcher must find a constant balance between interest for solving the problem and the interest of the research (Mckay & Marshall, 2001). This is why in PAR, the researcher should constantly go back and forth between analysis from the rigor cycle to the relevance cycle.

In order to check if both interests have been successfully integrated, the framework is validated in two ways: expert interviews and 'thought experiments' with previous project teams. The validation efforts of this research are mainly focused on the generalizability of the framework. The topic is explicitly discussed in an expert validation with Dr. Scott Cunningham from the faculty Technology, Policy and Management from Delft University of Technology. Moreover, the framework that results from the action researches is tested on other data mining cases at [redacted] Since the time window did not allow the possibility for multiple case studies, the new data mining framework is validated by analyzing its application on previous and current data mining projects at [redacted] These projects differ in problem context and data mining methods. The extent to which the framework is applicable in this different situation, codetermines the generalizability of the framework.

### 2.3.2 Rigor cycle

To make the system design rigor, the design has to be substantiated by existing concepts and theories. These theories can be either used to observe the artifact as a whole, or they can be used as a kernel

theory to analyze a specific part of the design (Markus et al., 2002). From a helicopter view, three types of literature are used: data mining literature, decision making literature and organizational literature. Data mining literature is used in order to scope the problem, to analyze the rule quality and rule quantity problem and to research the current pattern evaluation methods. Since the scope of the research is to apply DM in an organization, the process of data mining should be compared to the traditional processes in organizations. Decision making literature helps to map the normal way of working of organization: which people are involved and how they normally come to decisions? When comparing lessons from these two disciplines, the difference between data mining and decision making can be adequately pinpointed. Organizational literature helps to reflect on the influence of the organizational structure on the data mining problem. As Mintzberg describes, different organization types require different types of decision making (Mintzberg, 1993).

Besides these three main literature domains, some kernel theories are used during the analysis and synthesis. Kernel theories that are used to design parts of the framework are e.g. wicked problems. When a kernel theory is used in the design, this is stated explicitly.

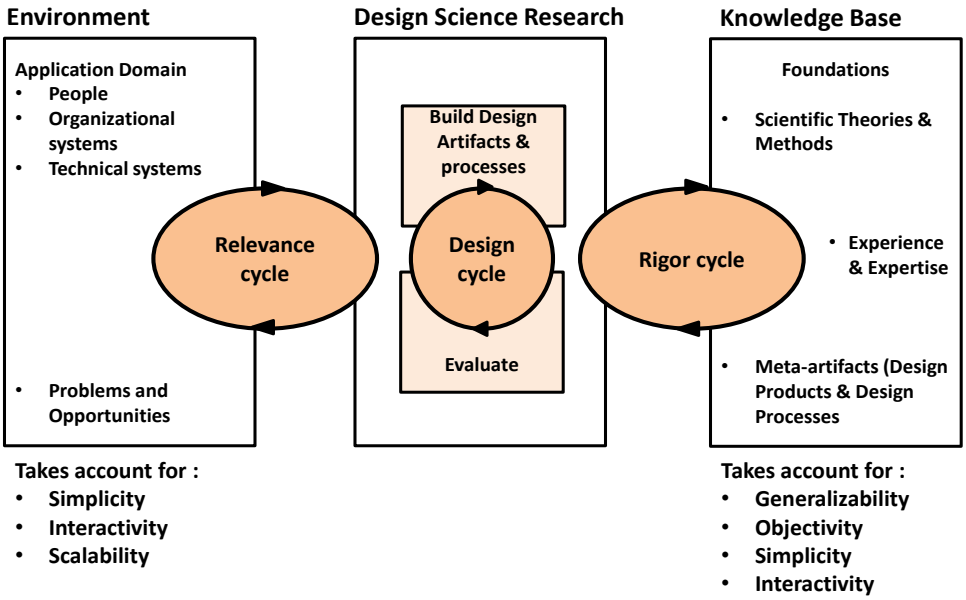



**Figure 9: Securing the fulfillment of non-functional requirements in the design cycles**

By building the framework from the rigor cycle, the generalizability of the approach is taken into account. Since DM consists of many methods and techniques, a generalizable approach is unlikely to follow from the relevance cycle with a case study. Literature and previous input should serve as an input to secure that the framework is applicable in different contexts with different DM methods and techniques. The tension that exists between objectivity and the interactivity of the DM process (previously described in 3.4) should be covered by evaluating existing methods to the extent in which they facilitate subjective, case-specific knowledge from decision makers, but still have an objective evaluation of the DM results.

### 2.3.3 Design cycles

In the design cycle, a framework for data mining is designed by combining insights from the relevance cycle and rigor cycle. This design is done in multiple iterations. In order to structure the design cycle, Hevner et al. determined a set of design guidelines for understanding, executing and evaluating the design. These guidelines are presented in .

**Table 2: Design guidelines for this research (Hevner et al., 2004)**

| Guidelines | | Operationalization for this research |
|---|---|---|
| 1 | Design as an artifact | The main deliverable of this research is a framework for applying data mining in a business context. |
| 2 | Problem relevance | Data mining is increasingly getting attention in businesses, but seldom delivers desired outcomes. The framework offers a solution to apply data mining, while taking the business context into account. It is societal relevant because it can be deployed in businesses. It is academically relevant because it can bring extra information on how to incorporate data mining in businesses (generalization) |
| 3 | Design evaluation | Observational: case study at [redacted] bank<br>Analytical: the structure of the framework is examined on the TIP design presented in table 1<br>Descriptive: provide sound line of argumentation for the design choices and validate the framework with expert interviews at TU Delft and [redacted] bank. |
| 4 | Research contributions | Artifact: methodology for improved application of data mining in a business context.<br>Foundations: novel methods for aligning business and data mining goals and interests in applied data mining. |
| 5 | Research rigor | Evaluation of generalizability of the data mining framework. |
| 6 | Design as a search process | The design cycle will be performed multiple times, with evaluation |
| 7 | Communication of research | Management-oriented communication is provided by analyzing the different design cycles in this report describing the securization of the framework within the organization. |

Since the design is a search process, the cycle has been carried out in multiple iterations in order to come to the final artifact: a framework for applying data mining in organizations. Not all iterations are documented. In Chapter 6, the final propositions of the design cycles are combined and presented.

## 2.4 Research overview

With the decision for the design science research methodology and the choice for an exploratory case study as participatory action research, the components of the research can be structured in an overview (see Figure 10).

The research consists of 3 parts: the analysis, synthesis and the evaluation. In the analysis, the first literature is analyzed in order to substantiate the first decisions in the case study. During the remainder of the analysis, propositions are the result of a constant back-and-forth impacts between the literature analysis and the case study. This eventually resulted in a set of design propositions. These propositions serve as the input for the synthesis. In the synthesis, the propositions are described in detail. In order to unify all the propositions, a design artifact is presented: a new data mining framework. This framework is validated through an expert interview and the previously described 'thought experiments'. Finally, the insights from the analysis and synthesis are combined in the evaluation phase. These results in conclusions on the improvements of data mining in organizations and recommendations for both the academic world and [redacted]

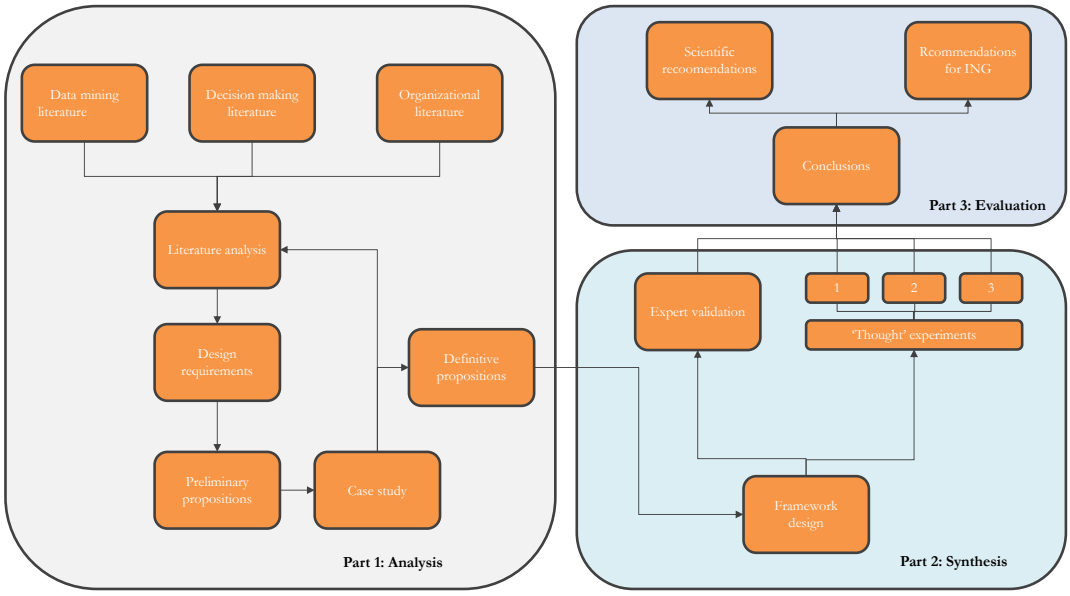



**Figure 10: Overview of the research phases**

# 3 Literature review

The two sub-goals of the research (improving pattern evaluation and integration of the DM and business perspective) describe both a technical problem as a process problem. This chapter examines the previous work done in the field of data mining, with respect to the three problem statements. In this way, the CDE-framework is substantiated by literature, so the research is rigor. This research results in an answer on the first two subquestions, and a direction for the third subquestion:

1. *How are data mining project carried out currently?*
2. *How are data mining patterns evaluated currently?*
3. *What are the requirements for successful decision making in data mining?*

## 3.1 Rule quantity problem

The rule quantity problem occurs when data mining delivers a large set of patterns. To be able to understand the rule quantity problem, the root cause for this problem needs to be clear. In previous research, the problem of the rule quantity problem has been acknowledged and its cause has been researched well. Furthermore, several solutions have been proposed. However, these solutions may trouble the rule quality problem, or may deliver insights on another level of detail. This level of detail may be not desired by the decision maker.

### 3.1.1 The issue of significance

Boyd & Crawford (2012) describe that the magnitude of the patterns delivered is influenced by the magnitude of the input data. This is because the evaluation of large data sets is done by applying traditional significance measures on the data set. These measures have been devised to be applied on smaller datasets, since large datasets were not available at that time. For example, traditional significance measures are able to find significant relationships from a population of 30. Since the input data for data mining is so much larger than is in traditional statistics, statistical significance is to be expected quickly (Boyd & Crawford, 2012; Hand, 1998).

### 3.1.2 Solution directions

In recent years, several solutions have been proposed to cope with the rule quantity problem. In their explorative study, Natarajan & Shekar (2005) describe solution directions to treat large quantities of patterns. However, they do distinguish between rule quality and –quantity in the analysis, but fail to make the distinction in their solution direction. For this research, there solution direction has been categorized by the two problems. Consequently, there are two solution directions to solve the rule quantity problem:

- Redundancy reduction
- Rule organization. Both solutions aim to generalize patterns, but apply generalization in a different way.

Redundancy reduction methods combine rules that contain more or less the same information. By pruning rules for the overlap in the information they have, fewer rules are presented while the

information is preserved in a generalized way. This leaves the decision maker with a helicopter view on the patterns in the dataset. Redundancy reduction is a less suitable method if the data set has a bias for generalization (Natarajan & Shekar, 2009). Moreover, it may increase the rule quality problem. Generalizing interesting rules may affect the interpretability or cause them to 'disappear' within a general, uninteresting rule. When applying redundancy reduction, the desired level of detail

Rule organization is also a generalization technique, but instead of generalizing rules, taxonomic factors are generalized. Previous examples of such taxonomic factors are product categories in a market basket analysis (Agrawal & Srikant, 1995). In this way, the pattern rule is described on a higher level. However, applying summarization and organization may only lead to knowledge that has been known before, because generalization on this level often leads to common information that does not provide insights. Moreover, if insights are explicitly demanded at a certain level (see the case study), organization is not an option to reduce the pattern set.

Regarding the previous, the conclusion can be drawn that dealing with rule quantity problem has an effect on the rule quality problem. While there are existing methods are effective in reducing rule quantity, they may increase the rule quality problem as a perverse effect. No solution is available that addresses the rule quantity problem without complicating the interpretation of data mining patterns.

## 3.2 Rule quality problem

While the rule quantity problem concerns the number of patterns, the rule quality problem concerns the meaning of a pattern. This paragraph analyzes why not all patterns are useful. Moreover, it offers a combined view on the criteria for meaningful patterns.

### 3.2.1 The issue of spurious relationships

Since the rule quality problem is related to the rule quantity problem, the rule quality problem also has a root cause in the significance measures. As described in 3.1.1, the significance measures cause that many rules follow when performing data mining. Besides causing a large set of patterns, it may also trouble the interpretation by decision makers because of spurious patterns. A spurious pattern is a pattern that is revealed just because 'enormous quantities of data can offer connections that radiate in all directions'(Boyd & Crawford, 2012; Hand, 1998). A very appealing example of a spurious correlation is given by Leinweber (2007), who proved that American stock market index strongly correlated with the butter production in Bangladesh. This relation is overfitted in the sample,

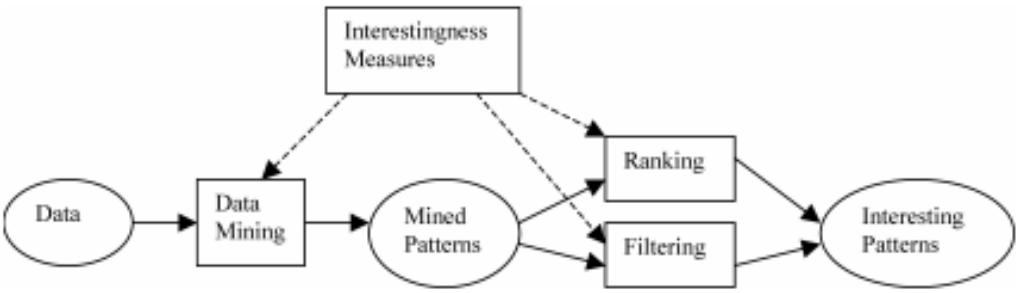



**Figure 11: possibilities for applying interestingness measures in the data mining process (Geng & Hamilton, 2006)**

consequently delivering '100% accuracy in sample, and 100% nonsense out-of-sample' (Leinweber, 2007)[1]. Not all patterns are equally useful or equally true, so that is why patterns need to be evaluated.

### 3.2.2 Pattern evaluation

The most common way to deal with the rule quality problem is by applying interesting measures in the evaluation phase. In order to make the interpretation and evaluation as objective as possible, interestingness measures are used in the evaluation phase to determine what information can be interpreted as knowledge (Natarajan & Shekar, 2009). Interestingness measure can be applied during the data mining phase, or as a ranking or filtering engine after the data mining phase (see Figure 11) (Geng & Hamilton, 2006a). Interestingness plays an important role in the data mining process, regardless from the data mining method or algorithm used.

Literature mainly distinguishes between two types of interestingness measures: objective and subjective measures of interestingness.

Objective measures are solely based on data and rely on statistics, probability theory or information theory (Han & Kamber, 2006; Natarajan & Shekar, 2005; Tan et al., 2004).

Subjective measures do not take into account the data that is used, but the user's beliefs and problem context that examines the data. These measures focus on the current knowledge of the user and try to prevent that already existing knowledge is rediscovered (Bie, 2013a; Geng & Hamilton, 2006b; Silberschatz, 1995).

Semantic measures represent the ability to translate the patterns into value for the user or the client. They should measure the utility or actionability of a result in order to evaluate the patterns from the user perspective and goals. For these three types of measures, Geng and Hamilton have designed nine criteria that can be used during the data mining exercise. These are presented in Figure 12.

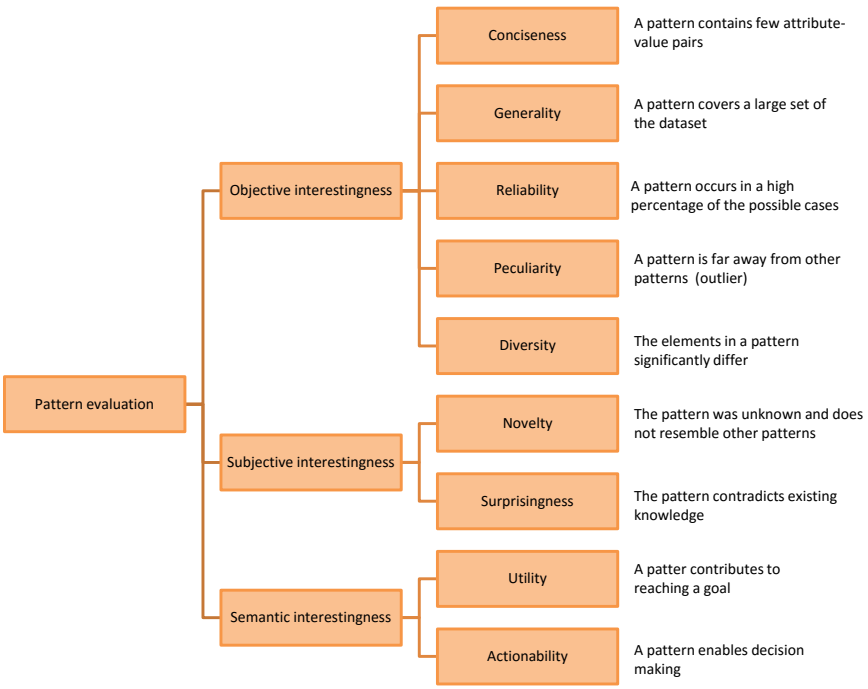



**Figure 12: overview of objective, subjective and semantic interestingness measures**

In recent years, high-utility pattern mining (HUP) gained more scientific attention. This school of thought is specifically focused on mining patterns that represent a large profit (Ahmed et al., 2011; Braynova & Pendharkar, 2005; Chaudhari & Verma, 2013). While it has become an accepted term for patterns that are most usable in organizations, it does not fully represent the true meaning of utility that Geng & Hamilton use (see figure 11). When you want to attain a goal, the pattern has to indicate a high benefit but should also indicate a possible action that can be undertaken. This is a combination of Cabena et al.'s concept of *actionability* and the *value* as indicated by researchers of HUP.

As literature disagrees on what represents a data mining result, the criteria for data mining patterns seem like a jumble. In an attempt to structure these views, this research will continue with a combination of the criteria for data mining results. These criteria merge the conceptual view of Cabena et al., the broadening view of Geng & Hamilton and the focused view of recent HUP developments into one view on data mining results. These criteria are *validity, utility* and *novelty* (see Figure 13).

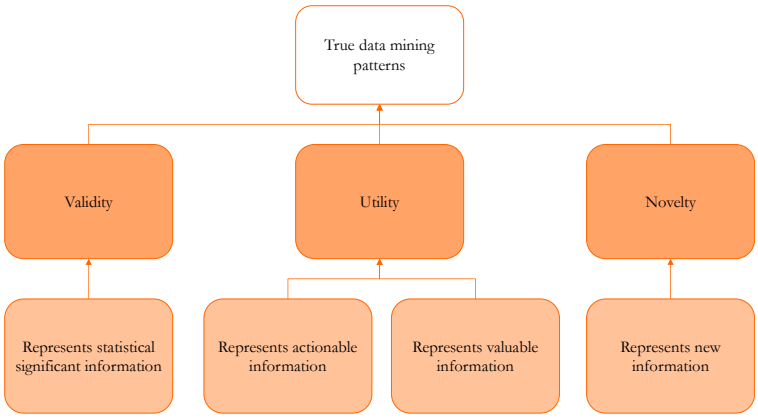



**Figure 13: The design of three holistic criteria for data mining results: validity, utility and novelty**

A data mining pattern should be valid because data mining relies on statistics. A data mining algorithm finds relations in a dataset, but these relations should be statistically significant to represent true meaning for the real world. The validity criterion combines the view conceptual view of Cabena et al. with the objective measures of interestingness from Geng & Hamilton.

The utility criterion combines the actionability of Cabena et al. and Geng & Hamilton with the focus on business value of HUP. The utility of a pattern consists of two elements: a pattern should contain information that enables a possible action and represents sufficient value for the organization. Organizations prefer patterns that indicates a possible action that they can take. However, they will only take action if they experience a feasible benefit from that action.

Finally, a pattern should represent new information for the organization. Data mining is an explorative study that uses induction gain insight in relations in data (Apté & Weiss, 1997; Freitas, 1997). One key benefit of data mining with respect to other statistical analyses is thus that it can help build new hypotheses and discover new knowledge. Therefore, a true data mining pattern should represent new knowledge for the organization. This combines the view of Cabena et al.'s *previously unknown* criterion and the subjective interestingness measures of Geng & Hamilton.

### 3.2.3 Support and confidence

With this broad availability of evaluation measures, how are patterns then currently evaluated? Traditionally, most data mining efforts (anomaly detection excluded) are evaluated with two statistical measures: confidence and support. A minimum-confidence threshold and minimum-support and threshold is specified, and results that satisfy both constraints are presented as patterns (Agrawal et al., 1993; Han & Kamber, 2006; Lenca et al., 2008; Mcgarry, 2005; Silberschatz et al., 1996; Tan et al., 2004). The support of a pattern measures the frequency of a pattern in the dataset, while confidence measures the accuracy of the pattern in the dataset. Mathematically, this is defined as follows:

$$Support = P(A, B, .., i)$$

$$Confidence = \max(P(B|A, .., i), P(A|B, .., i))$$

Where A and B are both items in the dataset. Perhaps both concepts are best explained with an appealing example of a market basket analysis (see example 1).

There are many other validity measures, such as leverage, coverage and lift (Kotsiantis & Kanellopoulos, 2006). Bayardo Jr. et al. showed that most applied measures, besides support and confidence, are monotone functions of support and confidence (Bayardo Jr & Agrawal, 1999). This is why the optimal rules of many DM efforts are often found near the support-confidence border (Tan et al., 2004). This leads to the conclusion that support and confidence dominate the results of the DM outcomes. Therefore, in the remaining of this report, support and confidence are treated as the key objective measures in DM.

A problem is that this dominance of support and confidence does not support decision makers effectively. Research on data mining, where the interesting outcomes were determined afterwards by decision makers, shows that the most interesting outcomes often do not score best on objective

**Example 1: beer and diapers**

Suppose you are a manager of a large supermarket. Your goal is to gain insight in the products that your clients often buy together, but the number of customers is too large to personally interview them. Because you have stored transaction data over the years, you adopt a data mining approach. After analyzing 100 transactions, you see the following subset of results:

| Product(s) | Frequency (n=100) |
|---|---|
| beer | 80 |
| diapers | 60 |
| beer & diapers | 40 |

The support is the number of transactions that contain a specific product, expressed as a fraction of the total transactions.

P(beer) = 80/100 = 0.8
P(diapers) = 60/100 = 0.6
P(beer,diapers) = 40/100 = 0.4

The confidence of a pattern is the chance that a product is bought, given that another product is already bought. Since the notion of sequentiality is not important in this case, the confidence is the maximum of two fractions: the number of transactions where beer is bought, knowing that diapers have been bought (expressed by P(beer|diapers)), or the number of transactions where diapers are bought, knowing that the client already bought beer(P(diapers|beer).

P(beer|diapers) = 40/60 = 0.67
P(diapers|beer) = 40/80 = 0.5

measures of interestingness (Carvalho et al., 2012). This supports the use of the utility and novelty criterion.

However, these concepts are not yet fully accepted as there is almost only theoretical work done in this area (Bie, 2013b). Previous research has indicated that it is hard to combine statistic-based evaluation and profit-driven decision making attributes (Choi et al., 2005). Most criteria, even the subjective and semantic measures, are therefore still concepts instead of measurable factors, and are almost never used.

## 3.3 Misalignment between data mining and decision making

As is elaborated in Chapter 1, it is important for the success of data mining that people in the organization know what to do with the patterns.

### 3.3.1 Knowledge creation cycle

The goal of data mining in organizations is to create knowledge that can be used for the benefit of the organization. If one is to design a process for data mining, it is wise to take into account lessons learned from knowledge creation processes in an organization. In their standard reference work for organizational knowledge creation, Lewin and Nonaka (1994a) describe four modes for knowledge creation that individuals sequentially have to go through to come to knowledge: socialization, combination, externalization and internalization (see Figure 14 for a graphic overview of the four modes).

In these four modes, two types of knowledge have been distinguished: explicit knowledge and tacit knowledge. Explicit knowledge is transmittable in formal, systematic language, while tacit knowledge resides in one's actions and experience and is therefore hard to transfer formally (Polanyi & Sen, 1967). Since tacit and explicit knowledge complement each other, the assumption is made that knowledge conversion can exist through interaction of stakeholders in a process.

In the socialization mode, tacit knowledge between individuals is transferred. This is done on a learn-by-seeing basis (e.g. on-the-job training). In combination mode, explicit knowledge is transferred

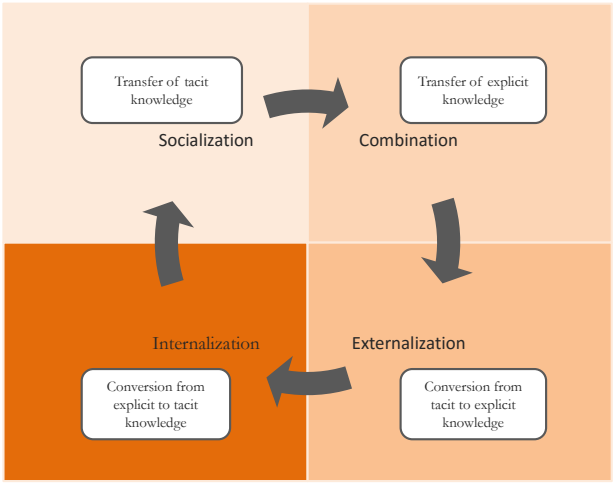



**Figure 14: the four modes of knowledge creation, adapted from (Lewin & Nonaka, 1994b)**

between individuals. Explicit knowledge can be exchanged via e.g. meetings, and can deliver new

knowledge through reconfiguration data or placing it in a different context. In externalization, tacit knowledge is converted to explicit knowledge, and in internalization it is the other way around.

Socialization and combination are characterized by repeated and intensive interaction between team members in an adhocracy. One of the major goals of these two types of knowledge creation is to create a common perspective between stakeholders in the process (Lewin & Nonaka, 1994b). A common perspective builds knowledge redundancy between the data scientists and decision makers. The goal here is not to completely understand each other, but to provide every stakeholder in an adhocracy insight to make crucial decisions in the wicked problem. The socialization and combination are then ways to create convergence in an adhocracy and transparency in the decision making process.

Relating this process to data mining in organizations, internalization should be regarded as the final goal of data mining: embedding new, explicit knowledge within decision makers. A data mining process should then not focus on only internalizing new knowledge, but should take the three preceding modes as a prerequisite for the success of data mining. Therefore, a good DM process incorporates all 4 modes of knowledge creation in the process.

### 3.3.2 Induction and deduction

When data mining is used in an organization, a connection has to be made between two systems: data mining and decision making. However, creating this connection has been difficult so far, since the two concepts differ in the way of reasoning. As decision making often aims for the most rational decision, traditional decision making is often deductive (Simon, 1979). Data mining, on the other hand, uses a large set of observations as input to induct rules and a theory for the delivered patterns (Apté & Weiss, 1997; Freitas, 1997). Example 3 gives an explanation of the difference in reasoning between induction and deduction.

**Example 3: induction and deduction**

Conventional decision making processes relies on deductive reasoning to come to an optimal decision. The difference between deductive and inductive reasoning is that deduction is logically valid and inductive reasoning is not. When deducing, the overall rule is known (Premise 1) and based upon a case observation (Premise 2), a logical conclusion is drawn.

Induction is exactly the other way around, where on the basis of empirical findings (Premise 3) a general rule is constructed. However this argument is substantiated, the general rule does not have to be true (in fact, the rule in the example is not true).

**Deduction**
**Premise 1:** all birds have wings
**Premise 2:** a swan is a bird
**Deductive conclusion:** a swan has wings

**Induction**
**Premise 3:** I have observed five swans, all five are white
**Inductive conclusion:** all swans are white

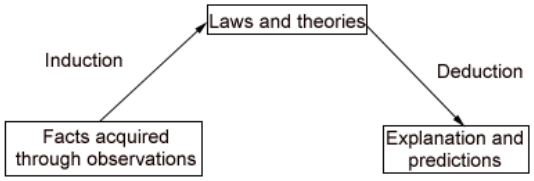



**Figure 15: difference between induction and deduction (Ali, 1998)**

When combining data mining in decision making, decision makers are confronted with another type of reasoning than that they are normally used to. Decision analysis is usually in rigid problem environments with decision prescriptions, while in data mining the search space may be much larger where it may be unclear how patterns add to attaining a business objective (Choi et al., 2005). As a consequence, a decision maker might feel that data mining might deliver false conclusions, as the rules are generated inductively. As a consequence, a decision maker might not comprehend or accept the outcomes of the data mining process (Boisot, 2004; Potes Ruiz et al., 2014). This then may lead to the decision maker distrusting the data mining execution (Alabdulkarim, 2013).

For successful data mining in business, deductive and inductive decision making should be balanced in a data mining framework. Deduction is essential to find actionable patterns, since DM should be able to logically understand the pattern and since human intelligence can specify business value (Choi et al., 2005; Witten & Frank, 2005). On the other hand, the added value of data mining in organizations is the inductive nature of the research. Because relations are not logically reasoned but empirically validated, previously unknown knowledge can be revealed (Bendasolli, 2013).

### 3.3.3 Friction between utility and novelty

Besides differences in the process approach, the deductive nature and inductive nature also causes friction between the utility and novelty of a pattern. With the converging view of decisions makers, a pattern should relate to the hypothesis or theory of the decision makers. Otherwise, a pattern would remain information and not become knowledge (see 1.1.2). However, this contradicts the search for patterns that satisfy the novelty criteria. Following the inductive approach, a stakeholder should form its theories on basis of the data mining patterns, and not the other way around. If a new insight comes up and it does not fit in the worldview of the stakeholder, that new insight will be rejected.

If a data mining process focuses on the attaining high-utility patterns, it is inevitable that the perspective of the decision maker needs to influence the project. On the other hand, this perspective should not get in the way of discovering new knowledge. Although it is the decision maker's main drive to look for actionable patterns, he should be open for receiving new knowledge. If the novelty of a pattern becomes of limited importance due to the inclusion of stakeholders in data mining, no true data mining patterns will result from the DM-process. Therefore, the data mining framework should adequately prevent friction between the utility and novelty of a pattern.

### 3.3.4 Data mining and decision support systems

When it comes to making decisions while balancing objectiveness and user preference, DM can learn from the best practices of data-driven decision making. To make decision making more substantiated, organizations often use data analysis to objectify the decision for an alternative. These models are therefore classified as data-driven decision making (DDDM) (Mandinach et al., 2006). Such a data driven model is called a decision support system. A decision support system (DSS) helps a decision maker to determine the norms for an effective decision and to offer decision makers knowledge to make substantiated decisions (Keen, 1980; Power, 2002). In order to cope with subjectivity in DSS, a decision maker exposes his values and assumptions and knows their consequences (Mitroff & Linstrone, 1993).

What DSS shows is that it is possible to allow a decision maker to input subjective information, while still remaining an objective decision making process. When a decision maker makes his preferences explicit, the outcomes of a DSS are subjective. However, the decision making process remains objective, while the inflow of subjective information is decoupled from the actual decision making process.

In an attempt to incorporate lessons of DSS to data mining, Bâra and Lungu (2012) designed a process to integrate DSS and DM. However, their proposed solution is questionable, since they propose to design solutions before going in to the DM phase. There are two reasons that this approach will not work in practice.

Firstly, thinking of strategies beforehand violates the inductive nature of DM. When thinking of and fixating on a number of solutions beforehand, the decision maker deductively thinks of possible strategies. In this way, he cannot gain insight in potential novel insights. DM is then merely a means to acquire data to substantiate predetermined strategies.

Secondly, in real DM problems it is often not possible to determine strategies beforehand. A DSS is suitable for simple, and structured problems (Courtney, 2001). However, DM problems are often unstructured: the problems and solutions are unclear and user-dependent (De Bruijn et al., 2010; Roman, 2010). In these types of problems, it is hard to determine the information and criteria to be used in the analysis, let alone that possible strategies can be easily designed beforehand.

### 3.3.5 Institutional context of data mining in organizations

The misalignment of data mining in business can also be explained from an organizational perspective. When deploying data mining in an organization, the organization has to decide in what way the process is embedded in the organization. This means decisions are needed on what parts of the organization will be involved and how they will interact. Mintzberg effectively distinguishes five different parts of an organization: the strategic decision makers, the middle managers, the support staff, the technical structure and the operational core of the company (Mintzberg, 1993).

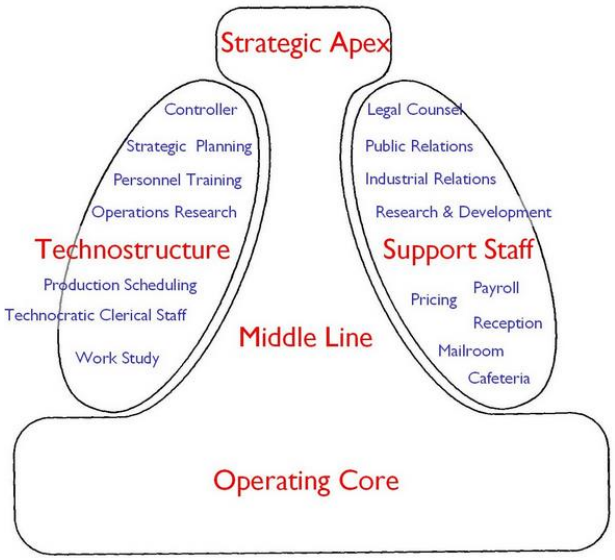



**Figure 16: the separation of the technostructure in an organization (Mintzberg, 1993)**

The middle part of Figure 16 is the backbone of the organization. The strategic apex of the organization determines the strategy, vision and goals of the company. The middle managers then have to implement that strategy in the operating core. The operating core itself produces value for the company by providing the core business to the customers. The support staff indirectly supports the core business of the company with other services, such as HR and procurement. The technostructure analyses the core business of the organization and tries to find improvements.

Data mining in an organization can be considered a part of the technostructure, as it tries to improve the core business through analysis. If data mining is to be successful in a business context, concrete improvements for the operating core should come from the analysis. Possible results of data mining will be implemented in the backbone of the organization. What type results they can use and value, is dependent of their goals, interests and mandate. As is stated in the previous, data mining does not succeed in effectively evaluating results on user specific goals and interest.

Because the gap between data mining and data mining in business has been acknowledged, previous research has been done on how to bridge this gap (Yang & Wu, 2006). The main contribution is the development of the CRISP-methodology, a process model that prescribes the data scientist to take account for the business goals in a data mining project (Sharma & Osei-Bryson, 2009; Sharma et al., 2012). The model consists of the following six phases: business understanding, data understanding, data preparation, modelling, evaluation and deployment. The model is presented in Figure 17. In the business understanding phase, the context of the problem where data mining is used for is sketched. This should give the data miner a clear understanding of the knowledge gaps that data mining should fill. A data miner should also know the business context in order to understand what the data actually means. Having a good business and data understanding is essential before performing the data mining itself: it gives a data miner insight in what data is necessary and possible for the analysis. With this information, he should choose the right data mining setup and evaluation measures. Unlike other data mining models, this model is the only process model that does not focus on the execution of data mining, but on the domain the research is executed.

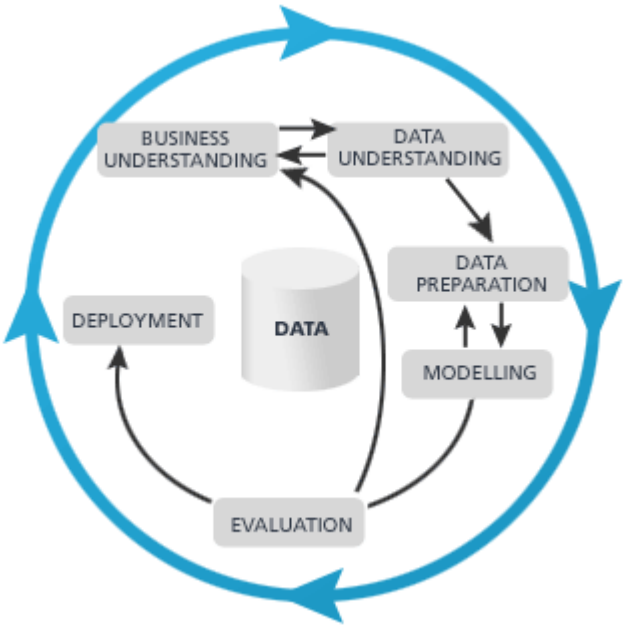



**Figure 17: the CRISP-DM model(Sharma & Osei-Bryson, 2009)**

Relating this to Mintzberg's organizations, the CRISP model forces the technostructure to think of the relevance of the project for the backbone of the organization. However, the data scientist may not be the right person to take account for the business relevance of a project. Regarding Mintzberg's organization structures, the technostructure is represented by the data scientists (Mintzberg, 1993). Their aim is to improve the core business through analysis. They are experts in carrying out analyses, and have less expertise on operation knowledge of the organization. The backbone of the organization (consisting of the strategic apex, middle line management and the operating core) has the most business knowledge, as they use this on a daily basis. They then should ensure that their interests are taken into account in the process. Therefore, they should be actively incorporated in data mining projects.

The participation of the backbone in data mining projects makes the interactions in such a project more complex. When multiple parts of the organization are involved in decision making, their organizational goals are likely to conflict. As Mintzberg describes, each part of an organization tends to pull the organization to a different direction (Mintzberg, 1980). While the technostructure largely wants to standardize outputs and processes in any way possible. This leads to a uniform data mining process. The operating core has the tendency to advocate for more trust in the professionalism of the employees. The degree to which this is facilitated in data mining determines the inflow of user specific information. A standardized process is easily deployed throughout the organization, but may leave valuable input of the user out of scope. A totally user specific process is likely to deliver results that the decision makers want to use, but makes it harder to deploy the process across the organization. The degree of standardization then determines the amount of preliminary work needed in a specific data mining project.

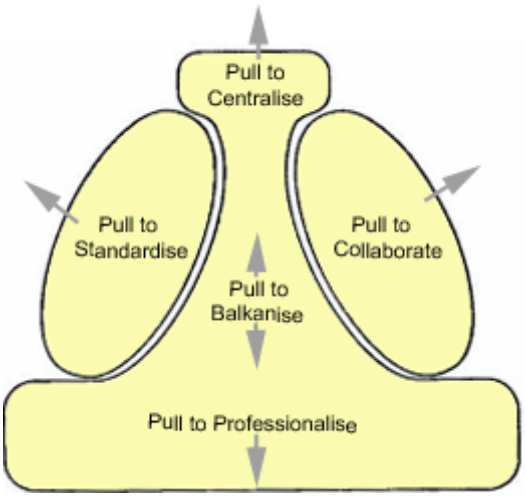



**Figure 18: the organizational pulls that conflict in the organization**

So, if the data scientists are unsuitable to incorporate business knowledge in data mining, who should? And in what way should they be incorporated in DM-projects? In their research for defining important institutional factors for decision making, Papadakis et al. have carried out an extensive study for important context factors for decision making in organizations (Papadakis et al., 1998). As in data mining projects, decisions concerns only a part of the organization, these factors are simplified to two driving forces: the *type of organization* and the *politicization of the playing field*. These driving forces can be characterized by 5 factors: the *centralization of decision making*, the *type of knowledge embedment*, *external uncertainty*, *dependencies between stakeholders* and the *strategic nature of the problem*. Figure 19 shows the institutional environment relevant for data mining in organizations.

Which part of the organization's backbone should be incorporated in data mining, depends on the type of organization DM is deployed in. As Mintzberg describes, organizations differ in the type of decision making. In a machine bureaucracy, the decision power is centralized, and the lower parts of the backbone are more focused on operations rather than decision making. In other organizations, such as an adhocracy or a professional bureaucracy, the lower parts of the backbone have more decision freedom. For data mining projects, persons with the right mandate to make decisions based on the data have to be incorporated in the project.

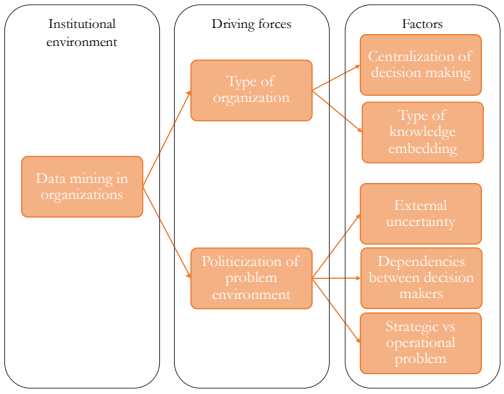



**Figure 19: context factors of the institutional environment of data mining in organizations**

While the previous focuses on who makes the decisions in organizations, how the decisions are substantiated is equally important. To explain this, a distinction between explicit and tacit knowledge must be made. Explicit knowledge is transmittable in formal, systematic language, while tacit knowledge resides in one's actions and experience and is therefore hard to transfer formally (Polanyi & Sen, 1967). In machine bureaucracy, the aim is to express and standardize existing knowledge as much as possible. Decisions are then made on explicit knowledge. However, in e.g. a professional bureaucracy, decisions are based on the considerations of an employee. Decisions are then substantiated with tacit knowledge. In order to fit data mining to the organization, the decision maker should take into account what substantiation is needed. As described before, Lewin and Nonaka prescribed four different modes to create knowledge in organizations (Lewin & Nonaka, 1994b). These modes differ in the type of knowledge that is inputted, and the type of knowledge that is created. Data mining prescribes the transformation of explicit knowledge to tacit knowledge: the internalization mode. This mode needs to be embedded in the data mining framework. However, the type of organization DM is deployed in should determine the need for and emphasis on the other modes of knowledge creation. For instance, in a professional bureaucracy, more attention should be paid to the transformation of tacit knowledge to explicit knowledge. On the other hand, in an adhocracy, effort must be done to share the existing knowledge before focusing on knowledge transformation.

To facilitate user interaction in data mining, one should prepare for possible perverse effects of the DM process. Besides the type of organization DM is deployed in, the politicization of the problem and its playing field needs to be taken into account. While the type of organization focuses on selecting the right parties to incorporate, the politicization of the environment influences how they interact with each other. A process works differently in a political environment than in a non-political environment, as actors may show more strategic behavior. The politicization of the environment is identified by three characteristics: the *external uncertainty*, the *dependencies between decision makers* and the *strategic nature of the problem*.

In pure data mining projects, data scientist may be influenced by their hinterland in the way that they execute their project. In an interactive project that includes decision makers, the hinterland of the project increases. Especially in a more political environment, stakeholders may experience pressure from external parties in the process. For not all stakeholders of the project outcome have to be included in the organization, a project team concerned with a decision can experience *external pressure* (Clegg et al., 1999; Walter et al., 2008). This can cause a reaction from stakeholders within a project, possibly leading to a different project scope or different project outcomes. In those cases, the project scope or -outcome may change in an unexpected way. The internal process may be strongly influenced by what happens outside the project. If a data mining framework should allow the influence of decision makers, it should provide ways to deal with this uncertainty.

Besides the external pressure on a data mining process, incorporating decision makers in the process may also lead to internal politics. When multiple stakeholders are involved, they may have different goals, interests and means. This may lead to internal conflicts in a project group. Stakeholders that are included in the process may have interdependencies, therefore they may show strategic behavior (De Bruijn et al., 2010). If a project has many *interdependencies between decision makers*, they may not only disagree with goals and means, but also with the steps that are taken. Therefore, they need to be aligned on the content of the project, but also the process steps need to be aligned.

When data mining facilitates interaction with decision makers, the decision makers may consider this project as a new platform for discussion and negotiation. If they do so on projects that have a high strategic character, decision makers may want to trouble the decision making process for their own goals. In strategic decision making processes, there is a much larger chance on the misuse of problem steps and process steps (De Bruijn et al., 2010). This type of behavior is new to data mining. If data mining is carried out in a strategic environment, the data mining process may have to be adapted to prevent stakeholders from blocking the data mining process. Therefore, the *strategic nature of the problem* is an important factor for the rules that have to be established to facilitate involvement of decision makers in data mining.

**Table 3: the boundaries of the organizational context of data mining in organizations**

| Institutional factor | Possible consequence |
|---|---|
| **Centralization of decision making** | **Centralized**: few stakeholders are involved, high parts of the organization have to be incorporated |
| | **Decentralized**: many stakeholders involved, multiple parts of the organization have to be incorporated to reach consensus |
| **Type of knowledge distribution** | **Explicit**: data mining results should comply with existing protocols or standards |
| | **Tacit**: data mining should be able to express tacit knowledge in the evaluation |
| **External uncertainty** | **High**: Project may be influenced by non-participating parties, project team should monitor the context of the problem |
| | **Low**: Project outcome is determined by internal dynamics of the project group |
| **Dependencies between decision makers** | **Many**: Project outcome may become the result of negotiation and bargaining |
| | **Few**: Stakeholder interaction is expected to have little influence on the project outcome |
| **Strategic content of the DM-problem** | **Strategic**: Project is more susceptible for misuse |
| | **Operational**: Project is unlikely to be influenced by strategic behaviour of the stakeholders |

## 3.4 Non-functional requirements for data mining in business

Whereas the functional analysis above can be used to improve data mining framework, it is important to take into account how the framework will be put in operation during before designing it. Therefore, non-functional requirements are defined that can be used to evaluate the quality of the DM framework.

### 3.4.1 Interactivity

A more or less obvious way to reach convergence between the different perspectives, is to design interactivity between the perspectives in the process. In their exploratory study for convergence between multiple perspectives in data mining, Blumenstock et al. found that interactivity is one of the main success criteria to include background knowledge (Blumenstock et al., 2010). Interactivity works bilateral: it facilitates the inclusion of background knowledge in data mining and it leads to a larger data understanding of business experts (Ankerst et al., 1999). The new framework for data mining should therefore enable the business perspective to influence the data mining outcome in an appropriate way. This is further described in 2.4.6.

### 3.4.2 Simplicity

Normally, the usability of the designed system would be included as a non-functional requirement of the DM framework. However, previous research has given an indication that the usability DM can be raised by analyzing the simplicity of the process. Therefore, the term simplicity is used as in stead of usability.

In their explorative study for creating convergence between domain knowledge and data mining, Blumenstock et al. concluded to keep the solution simple. Instead of letting domain knowledge making state-of-the-art algorithms even more complex, the solution for aligning the business with DM should at least be a simple one (Blumenstock et al., 2010). The real meaning of a result will be even more difficult to interpret if the DM process is not a simple one (Han & Kamber, 2006). If the DM process is able to create convergence and evaluate patterns in a simple way, DM will be more understandable for decision makers. This will lower the possibility of distrusting the outcomes and can lead to a higher acceptance of data mining results. This can in turn lead to an increase in adoption of DM applications (Alabdulkarim, 2013).

### 3.4.3 Generalizability

Since data mining is a large collection of methods and techniques, large differences can exist between two DM applications. A data mining method becomes stronger and will be more accepted if it can handle different types of data (Chen et al., 1996). For the DM process in turn, the difficulty is to unify all these different approaches in one single framework. This has been indicated as one of the major research problems in the field of data mining (Yang & Wu, 2006). Since in this report a DM process is designed for a specific case, it is not a given that the process can be implemented in a different setting. The scientific value of the DM framework will increase if it is designed and evaluated on its generalizability.

### 3.4.4 Scalability

In the current efforts of facilitating interactivity in the DM process, authors reported on using domain knowledge to let domain experts design decision trees (Ankerst et al., 1999)(Vidulin et al., 2014). This lead to an interactive process but was very time consuming. Although the interactivity of the process leads to accurate results, the approach is not able to handle large sets of data. The approach demanded a lot of effort of domain experts, making the method only feasible for small datasets. Since the basic quality of DM is that it can handle large quantities of data, the delivered method should be scalable for large quantities of data.

### 3.4.5 Objectivity

In an attempt to facilitate interactivity in the data mining process, Vidulin et al. designed a process that included human evaluation to eliminate less-credible solutions. Moreover, they argue that in a DM application in a domain, the less credible solutions must be exempted from the created DM models (Vidulin et al., 2014). This can be substantiated by the claim (see section 2.1.3) that in order to gain knowledge from DM applications, a decision maker should be able to relate to the results and accept them. However, in the way that user preference is facilitated in the existing efforts, the method can easily become subject to the 'winner picking' (Emmett, 2010). This occurs when the selection efforts are focused on choosing one outcome that the decision maker finds desirable. This conflicts with the objective nature of data mining and with the novelty requirement for DM results. In order to secure that the best results are picked instead of the most desirable, the process should also be objective.

## 3.5 Literature review conclusions

Concluding on the literature review and related work, the first subquestion can be answered and the second answer can be answered partly.

1. *How are data mining projects carried out currently?*

In current data mining project, the role of the data scientist is much emphasized. Current data mining research focusses on the technical aspects of the project, rather than the environment in which the project is carried out. Due to this focus, there is a gap between data mining and the organization itself. This gap is characterized by two elements. Firstly, data scientists are housed in a different part of the organization (technostructure) than the decision makers (strategic apex, middle management or operating core). Moreover, the gap is also illustrated by the difference in reasoning between data scientists and decision makers. Some effort is done to bridge this gap. The widely accepted CRISP-DM framework prescribes the data scientist to take its environment into account. However, as data scientists are separated of the actual organization, they are not the most likely group to adequately take into account the business aspects of a data mining project.

2. *How are data mining patterns evaluated currently?*

Theoretically, there are three types of patterns: objective, subjective and semantic evaluation measures. In practice, nearly every case only adopts objective measures: support and confidence. These measure the statistical validity of an outcome. However, the most statistically valid result is not necessarily the best result .The measures used in the pattern evaluation phase are not sufficient to satisfy all elements of desired data mining results. While objective measures describe the validity of results, they do not describe the actionability of a pattern and it does not guarantee that the knowledge was previously unknown. Subjective and semantic measures do this, but are seldom applied. It seems as the different types of measures are difficult to apply simultaneously.

The lacking application of subjective and semantic measures has mainly two origins. First, there is a hesitance in data mining to incorporate the user's preference, as user dependent outcomes may violate the objective nature of data mining. Second, decision making and data mining respectively use deductive reasoning and inductive reasoning. Deduction is needed to come to actionable patterns, while the unique strength of data mining is the inductive search for previously unknown patterns. It is hard to incorporate both ways of reasoning in one single process.

3. *What are the requirements for successful decision making in data mining?*

The main quality criterion current literature uses, is the term 'interestingness'. However, this term does not cover the requirements for DM patterns. For a pattern to represent discovered knowledge, it must be valid, actionable and previously unknown. This is why patterns should not only be evaluated on their statistical relevance, but also on their utility and novelty. Without the freedom for stakeholders to influence data mining with problem-specific information and user-specific information, data mining will not deliver relevant outcomes for the problem.

The actionability of a pattern largely depends on the user that is going to use the patterns to create value for the organization. Data mining takes place in the technostructure of an organization, while the findings have to affect the core business of the organization. The separation of data mining from the core business causes that data miners have different goals than decision makers. For an analyst,

the goal is to understand the data, while a decision maker wants to make an optimal decision. To be able to come to actionable patterns, the data mining process should facilitate the user to explicitly incorporate problem-specific information and his preferences in the decision making process.

Currently decision making is focused on objective outcomes. When data mining is applied in business, this is why there exists a tension between the objectivity of an outcome and the actionability of an outcome. When decision making on data mining moves to process-oriented outcomes, subjective decisions can be made in an objective decision making process. Moreover, the friction between utility and novelty of a pattern should be prevented.

Since DM is often carried out in adhocracies, there are often multiple stakeholders that have multiple goal and interests. In order to bridge the gap between data mining and decision making, these stakeholders should participate in data mining projects. The DM process should be able to create consensus between these stakeholders. Agreements on the decision making process should be made through convergence. Besides this, the data mining framework should secure that data mining is an interactive process that provides simple solutions for complex decision making issues. The framework should be generalizable to be applicable for different methods and techniques. Moreover, the DM process should be scalable to handle large data sets. Finally, the process should be objective to ensure that the best results are selected.

The requirements that are distilled from the literature review are presented in Table 4. Here, a distinction has been made between functional requirements (what the framework should do) and non-functional requirements (how the framework should work).

**Table 4: Design requirements resulting from the literature analysis**

| Functional requirements |
| --- |
| • Perform validity evaluation<br>• Perform utility evaluation<br>• Perform novelty evaluation<br>• Subjective, problem specific information should be able to influence the DM process<br>• Prevent friction between utility and novelty<br>• Create a common perspective for all stakeholders in the data mining project<br>• Secure the 4 modes of knowledge creation |
| **Non-functional requirements** |
| • Interactive process between data scientists and decision makers<br>• Provide simple solutions for complex decision making issues<br>• Generalizable framework to apply in different contexts<br>• Scalable process to apply on large datasets<br>• Objective pattern selection and pattern evaluation |

# 4 Case study

The case study that has been performed at [redacted] bank follows from the business desire to shift customers from the call center to the [redacted] web site. [redacted] In this chapter, the case study performed at [redacted] Bank is described and the main scientific findings will be discussed. This will result in an additive insights for the following research question:

3. *What are the requirements for successful decision making in data mining?*

And an answer on the following research questions:

4. *How can the data mining process in organizations be improved?*
5. *How can the evaluation of patterns be improved?*

The recommendations for [redacted] bank that follow from the case study, will be treated in Chapter 9. This research described below was performed as an exploratory participatory case study. This means that the process improvements and the framework that is proposed in Chapter 6, has been established with an active participation of both researcher and the relevant organizational parts that are affected by the problem (Tedmanson & Subhabrata, 2010).

## 4.1 Data mining in the [redacted] organization

As described in 3.3.5, the organizational structure of the problem environment is of importance for applying data mining in that environment. As described by De Bruijn & Ten Heuvelhof, the internal organizational structure impacts the way in which decisions are made (De Bruijn & Ten Heuvelhof, 2008). In order to fit data mining to decision making in projects, the organizational structure relevant to the Call Reduction Program is analyzed. The organization is represented in Figure 20.

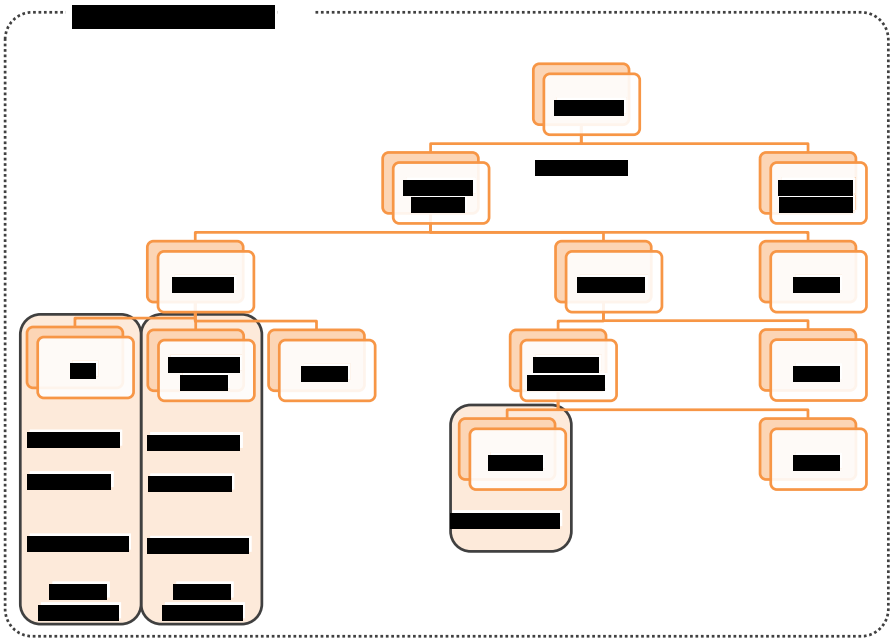
### 4.1.1 [redacted] organizational structure

[redacted] is an organization that is in rapid development. In order to cope with innovations and transitions in banking, its organizational structure is being changed to [redacted] In this part, the original structure is described first. Secondly, the change trend is described. Thirdly, the implications for the program Call Reduction case are presented.

#### [redacted] original structure

From the perspective of the total organization, [redacted] as a whole may be best considered [redacted]

For the case of Prevent and shift, the [redacted] When a customer wishes to contact the bank, [redacted] offers him 3 ways to do so: call to the



call center, visit the web site (including the app) or visit a branch.

**The change trend in the organization:**

In order to cope with the rapid technological developments in banking, and for reasons of cost reduction, new strategy is to . In order to establish this, the channels should provide the same service. For the organization, among other reasons, the most important reasons to change the organization are (Molenaar, 2014):

-
-
-
-

[redacted]

In business scrum, a project team with representatives from multiple disciplines is formed to deliver certain products or solutions. These teams are more flexible and autonomous in the projects they start and end than in the original situation. A team member does not have a predefined role in the team, but has his own expertise. This team is established with a purpose, but have freedom in how they fulfill this purpose. The involvement of higher-level decision makers in a scrum team is low. The scrum method originally has a leader, the Product Owner. At IN [redacted] Decision making in a scrum team is therefore *decentralized*, and has to occur through alignment of the team members (Mintzberg & McHugh, 1985).

As can be derived for the most important reasons for change, scrum teams have been established for new innovations with less bureaucracy. Innovations can be developed quickly in the scrum team, without a great deal of substantiation. Tacit knowledge can be used in the scrum team. Scrum teams may work according to professional knowledge and experience, the procedures for changes in the organization are still the same as in the original situation. For instance, a change to be made in a communication channel at [redacted] needs to be scored on several points: [redacted] To implement a change, the benefit of the innovation needs to convince other parts of the organization via these criteria. Therefore, *knowledge is and should be embedded explicitly*.

Scrum teams plan their projects in sprints: 3-week timeframes with clear cut deliverables. Instead of reporting to higher parts of the organization, the project teams organizes a demo session with the deliverables of the sprint. This demo is the official moment for the parties outside the scrum team to influence the scrum team's trajectory. Outside these official moments, external parties can influence the representing team members. However, since the important parts of the organization are (temporarily) included in the scrum team, the *external uncertainty* of the case study is deemed to be low.

Within scrum teams, [redacted] Therefore, in scrum teams there are *many dependencies between decision makers*.

Finally, working in scrum teams offers the representatives a new platform for interaction and negotiations with other stakeholders. If the problem is of a strategic nature, stakeholders may use the scrum team for strategic decision making. The process may then not only be used for data mining, but also to settle strategic issues between the parties. Since the solution [redacted] to the problem [redacted] d), the project can be considered more operational than strategic. Moreover, the employees incorporated in the project all come from the more operational parts of the organization, so *strategic behavior is not to be expected* in the project.

[2] Since the scrum method is being increasingly adopted by several pars of organizations, literature is starting to make a distinction between 'business scrum' and 'IT scrum'. In this report, 'IT scrum' is out of scope. For the sake of simplification, 'business scrum' in this report is mentioned as 'scrum'.

[redacted]

### Conclusions on the institutional context

In order to apply data mining in a proper manner at [redacted] the institutional context needs to be evaluated. The two important driving forces in the institutional context are, as described in 3.3.5, the organization type and the politicization of the problem environment.

Concluding on the organization, [redacted]

Therefore, insights gained from data mining have to be made explicit.

Scrum teams combines people from several disciplines on one project. By doing so, several expertises, opinions, goals and interests are combined in one decision making process. In order to streamline working in a scrum team, employees from several disciplines have to agree on the goal and interests of the whole team. Therefore, the scrum team [redacted]

## 4.2 Data mining in the [redacted] scrumteam

As a case study for the design of a framework for data mining in organizations, data mining has been applied in the Prevent & Shift scrumteam. This paragraph presents the data mining project and its outcomes. First, the organizational parts of the scrumteam have been analyzed, to identify gaps and similarities between the different stakeholders involved. Furthermore, the way in which the data mining project is carried out is described briefly (and extensively in Appendix B). Finally, the results and deliverables for the data mining project are presented.

In the scrumteam, [redacted]

### 4.2.1 Call

The Call department is represented by 5 representatives in team P&S. Call is the main initiator of the project [redacted], [redacted] The [redacted] Therefore, the Call department looks into other, more sophisticated opportunities of call reduction.

In order to implement changes, Call has the mandate for to change practically everything within Call. [redacted]

[redacted]

### 4.2.2 Internet

The Internet department is represented by 1 representative in team [redacted] other representatives have been temporarily involved in the project. The goal of Internet is to maintain the web site's quality and security, and to improve the functionality of the web site. [redacted] In order to decide if the measured value is large enough to propose an improvement, Internet uses a Cannel Improvement Decision Matrix. [redacted]

### 4.2.3 Big data

However not a part of team P&S, for the case study a data scientist was temporarily involved. [redacted]

[redacted]

**Figure 21: The formation of the P&S scrumteam across the channels**

[redacted]

## 4.3 Carrying out the data mining project

In order to set up a data mining project, decisions have to be made on scope of the project and the way in which it is carried out. According to the CRISP methodology, this is done by the data scientists in the business understanding phase. However, this report presents the argument that this needs to be done in collaboration between data scientists and the involved parts of the organization. Therefore, team P&S and the data scientists had to create consensus on the goal and scope of the problem and the set-up of the data mining research.

### 4.3.1 Goal of the project

With the interaction between Call, Internet and Big Data in the P&S project, two types of consensus was needed. [redacted]. This is a case of alignment between the technostructure and the operating core of an organization.

To align the problem perception between Call and Internet, both teams need insight in each other's perception. For this, Nonaka & Takeuchi propose to use socialization and combination as modes of knowledge transfer. Working in a scrum team facilitates more interaction between Call and Internet. In that way, socialization (transfer of tacit knowledge) is easier achieved. For Call, this meant that they had to explore the functionalities of the web site of [redacted] as well as to learn its architecture. For Internet, this meant that they had to meet with call agents in order to grasp the dynamics in the call center. Combination (transfer of explicit knowledge) can be achieved by introducing the stakeholders to the procedures of the other stakeholders. For Call, this meant the exchange of the Internet 'Funnel' Dashboard, while Internet had to gain insight on the KPI's in the call center and how they are constructed.

The main goal of Call in P&S is to gain insight in the volume of calls that can be prevented or shifted to the online environment. This is a very difficult problem, since there is no definition or consensus on when a call can be prevented or shifted. However, the clearest group of customers that can be distinguished, overlaps with goals of Internet. Within the group of customers that call, there is a large group of customers, whose purpose for the call originated online, because they tried to find information or perform a task online but were not able to do this. They then have to contact the call center because the online environment did not fulfill their needs sufficiently. The customer has to take an extra action in another channel to complete his tasks, while he first preferred to do complete his tasks online (Oppewal et al., 2013). For Call, this is the easiest target group of customers. These customers intended to do tasks online, and therefore are most willing to shift to another channel.

These customers also add to the business goal of the Internet department. [redacted] For Internet, it is interesting to investigate what these customers that shift to Call have been doing online. [redacted]

While the alignment between Call and Internet is of a substantive nature, data scientists have interest in the way the project is carried out. With their incorporation in the project, the operating core (Call

and Internet) have to come to agreements with the technostructure (data scientists) on the way the problem is analyzed. For the technostructure, there is potential to review and improve the way in which both channels analyze their performance. In the initial situation, there were some indicators that for improvements. was adequate enough for Internet to gain insight in its performance, but showed curious results. For the technostructure, this served as an indication that the performance analyses could be improved.

Besides the general improvement of analyses, .

### 4.3.2 Scope of the project

In order to unify all the business goals in one project, the scope of the research had to entail the desire of Call to lower the call flow sophisticatedly, the desire of Internet to improve the web site and the desire of the data scientists to improve the performance evaluation and to test the possibilities for integration in channel performances. Therefore, the scope of the project is on:

- Calls that have been made by clients that wanted to use the web site, but did not succeed online
- The preceding internet activity of those clients on the web site
- Without the limitations of the pages that belong to Internet's funnel

Since the project has been the first step of collaboration between Call and Internet and explored the possibilities for data mining in the organization, the decision has been made to position the research in the organization as a Proof-of-Concept (PoC). Therefore, the choice has been made to limit the research to one process. During the selection of the case study, the internet department focused on improving the process of retransferring direct debits ('terugboeken incasso' in Dutch). Therefore, the focus of the case is on calls that considered claiming back a direct debit. he scope of the target population then was:

- consumer banking customers,
- who have successfully retransferred a direct debit through the call channel,
- during the period of August and September 2014,
- but have been online in the preceding 24 hours before the call
- on the web site, both in the open and closed environment

The scope has been limited to August and September 2014 in consultation with IT architects at the Internet department. As the web site of changes constantly, some significant changes in logging

took place in July. In order to take consistent data as input, the advice was given to only focus on August and September.

### 4.3.3 Data mining setup

The setup of a data mining project is largely influenced by two factors: the parties involved and the data available. The previous paragraph entailed the alignment of the parties involved, which resulted in the desired contents of the project. These desires need to be aligned with the possibilities given by the available data. During the case study, this happened simultaneously with the alignment of the business goals. Below, the business goals are translated to an actual setup. First, the assumptions that are made explicit. These assumptions need to be accepted to accept the results of the analysis. Secondly, the data that is used is presented and described briefly. Thirdly, the data mining setup that is used for the analysis is presented. Finally, the criteria used for the evaluation of the patterns are described.

#### Assumptions

Since data is a reduced representation of the actual situation, assumptions have to be made to fit the data used to the problem situation. In order to successfully target, isolate and preprocess data for data mining, some assumptions have had to be made. In general, these are the four assumptions that will be discussed in the rest of the paragraph:

- There is a relation between call activity and internet activity of a customer
- Internet activity is influential when it has occurred less 24 hours before the call
- The most recent visit of a customer that precedes the call is the call that gave rise to the call
- [redacted] web site visits from the same device are highly likely to be done by either the same person, or the same financial entity

Figure 22 presents a fictive example of the effect of the assumptions on the data that is selected for further analysis. This example represents the internet and call activity of one customer. In this case, the internet activity of Wednesday morning is selected for further analysis. In the following, the assumptions are elaborated.

If the analysis of internet activity preceding a call is of any use, one must support the theorem that the call is in some way the result of the internet activity. If this relation is assumed, then another theorem holds that a customer's first chosen channel is his most preferred channel (Oppewal et al., 2013).

For the analysis, a timeframe of 24 hours has been elected as a maximum duration between the internet activity and the call. If the time span between a web visit and the call is more than this, it becomes less likely that the internet activity and the call are related.

It may be possible that within these 24 hours, a customer has paid multiple visits. If this was the case, the choice has been made to only take into account the most recent internet activity before the call. If more visits could have been selected, some customers' activity would weigh heavier in the analysis then others.

[redacted]

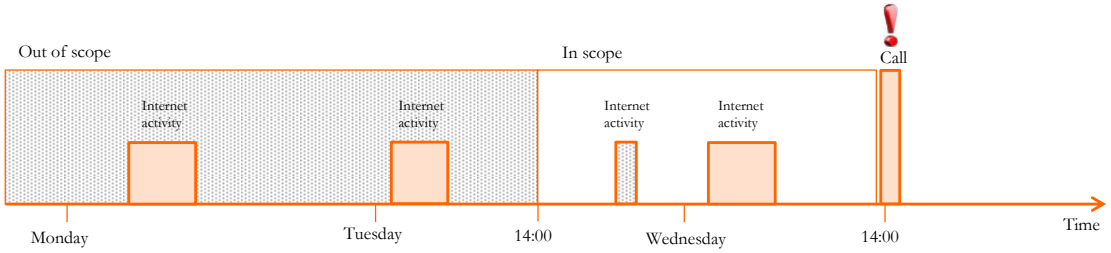



**Figure 22: example of the effect of the assumptions on the selected data**

### Data

In order to be able to carry out the research, the data must be suitable for the data mining setup that is selected. The requirements for the data mining setup (the data mining setup will be elaborated in the next paragraph) are presented in Table 5. Distinction has been made between minimum requirements (requirements that must be fulfilled in order to be able to mine patterns) and additional requirements (requirements that provide extra options for research). Table 6 presents the essential data that has been selected from both databases. Appendix B offers a more extensive description of the data.

In order to gain insight in the contact flow between Internet and Call, call data and internet data have to be linked to each other. The combination of these two types of data is possible, [redacted] This resulted in the case study in many loops of clients repeatedly clicking on the same page. Together with employees at the Internet department, the

pages below the generalized pages have been identified. As this was not a desirable situation, Internet indicated that this did not violate the minimum requirements.

**Table 5: the minimum data requirements to make sequence mining feasible**

| Call database | Met? |
|---|---|
| [redacted] | |
| [redacted] | |
| [redacted] | |
| [redacted] | |
| [redacted] | |
| [redacted] | |
| **Internet database** | |
| [redacted] | |
| [redacted] | |
| [redacted] | |
| [redacted] | |
| [redacted] | |
| [redacted] | |
| [redacted] | |
| [redacted] | |
| [redacted] | |

To gain insight in the online behavior of the customer, [redacted] This relates to the last assumption in the previous paragraph. To make this possible, some extra combination of data was required. [redacted]

**Table 6: Overview of essential data needed to create dataset**

| Call database | |
|---|---|
| [redacted] | |
| [redacted] | |
| [redacted] | |
| **Internet database** | |
| [redacted] | |
| [redacted] | |
| [redacted] | |
| [redacted] | by the custome |
| [redacted] | |

[redacted] **setup**

The data mining setup consists of the setup how the data mining goals will be attained. In order to do so, team P&S needs to select the methods, techniques and algorithm they are going to analyze the data that is selected in the convergence phase. This results in the experimental setup of the research.

[redacted] Pages that are frequently visited together are investigated. To this end, the appropriate data mining methods is association rule mining.

The market basket example treated in Chapter 3 is a classic example of association rule mining. However, analyzing web page visits is a more complex version of a market basket analysis, since it consists of longitudinal data. For the analysis of pages, the order in which pages are visited is important. So, a technique is needed to structure, order and analyze case-dependent and time-dependent data. It can take the same input as traditional association rule mining techniques, but its output is that it creates sequences instead of baskets. Figure 23 shows how the input data is treated differently from traditional association rule mining. Traditionally, the input data is transformed into baskets, where duplicate values are removed. The order of the items in the baskets does not matter for the further analysis. For the case study, the input data needs to be transformed in sequences of pages, ordered from earliest to most recent. Moreover, the duplicate values still appear in the sequences (see 313417911002 for instance). In this way, 'baskets' (from now on called 'sequences') can become much more complex than traditional baskets.

The data mining technique suitable for this type of association rules is sequential pattern mining (or sequence mining) (Brzinsky-fay & Kohler, 2006). Since it is a frequently used method, the expectation was that it was the most applicable option. Besides this method, two different types of Markov Chains have been investigated: the anomaly detection Markov model and the hidden Markov Chain detection. While the anomaly detection method was too rigid, the hidden Markov model was deemed possible. However, because the choice was made to focus on frequency instead of probability distribution, sequence mining was preferred.




**Figure 23: example of the difference in transformation between association rule mining and sequence mining**

When sequence mining proved to cause some problems delivering valid results the option for Markov modelling was re-investigated. It then turned out that this would not deliver a valid model, as the number of cases needed was impossibly high. This all is further elaborated described in Appendix B.

Selecting the data mining algorithm is largely done by data scientists. [redacted] To be able to specify his choice, he has to argue on a set of requirements for the algorithm.

In order to select the proper data mining algorithm, the algorithm needs to be easy to use, applicable, accurate and scalable (Nisbet, 2004). Because of its greater ease of use and applicability, the decision has been made to use the TraMineR package for R. This decision is elaborated in Appendix B.

### Criteria

For this case study, both statistical criteria have been defined as well as criteria that measure the utility and the novelty of a pattern. The rise for these three types of criteria are described in chapter 3 and further elaborated in chapter 5. In this part, the completion of these three types of criteria are discussed.

#### Validity criteria

The statistical criteria consisted of frequency criteria. For Call, high frequent patterns are more desirable, as they show a significant pattern and also indicate more reduction potential. After the data exploration, it became clear that the data in the dataset is very diffuse. This means that there is no clear pattern with a high frequency. [redacted] A possible critique here is that this threshold is way too low to indicate true reduction potential. However, as the learning experience was deemed to be more important than business success on the short term, the decision was made to accept this barrier for further analysis.

#### Utility criteria

Besides the frequency criteria, utility criteria have been identified in order to rank and filter the statistical significant patterns according to their meaning and value for the organization. Here, there was a discrepancy between the criteria of the different parties. Whereas a high frequent pattern indicates a large reduction potential for Call, meaningful patterns for Internet are patterns that differ significantly from what a customer should actually do.

[redacted] This lead to the construction of two pattern quality definitions:

1. [redacted]

Together with the data scientist, these definitions have been deducted to measurable utility criteria to evaluate the patterns on (also see Figure 24):

[redacted]

The implications of these utility criteria on the project is that [redacted]. Appendix B holds more information on this, as well as an extensive description of the process of criteria determination Appendix B.

### Novelty evaluation steps

To focus on finding new knowledge, a separate evaluation step has been designed. In the case study, three possible types of new knowledge have been identified:

- Patterns that not have identified before and may contradict existing hypotheses or form new hypotheses
- New knowledge that can be obtained by reflecting on the insights, gained from validity and utility evaluation, from a different perspective
- Patterns that can be obtained by reflecting on a large set of high-utility patterns (meta-novelty)

Meta-novelty was deemed unsuitable for the case study, since there were too little input cases to perform a valid evaluation. It can be regarded as a big data project on outcomes of other big data projects. As it is an important type of novelty, it is not described here but in 4.5.4.

**Figure 24: Breakdown scheme of the translation of the business goal to evaluation criteriria**

With the first type of novelty, a data scientist looks in the initial dataset for patterns that may contradict the knowledge of the decision maker. These patterns provide new knowledge on the same level of abstractness as the high-utility patterns do. A criticaster of the utility evaluation may pose that new knowledge may not be found if you enable too much influence of the decision maker. However, as is described in the literature review, a pattern is interpretable if a decision maker can relate to the information it contains. Therefore, it is very hard to discover new knowledge in the initial dataset without any guidance. And by determining utility criteria before data mining, winner-picking is prevented as much as possible. Therefore, utility evaluation may show new knowledge as well. In the case study, it proved to be very hard to look for new knowledge on the same level as you look for valuable knowledge.

The second type of novelty takes the lessons learned from the previous validation steps as starting point. The starting point then is a set of valid patterns, all with a utility score (step 1 in Figure 25). In order to find new knowledge in this set of patterns, the scrum team had to make a cut-off point in order to create two sets: a set of high-utility patterns and a set of low-utility patterns. The data scientist can review the two sets of patterns from a totally different perspective (step 2 in Figure 25), in order to see if there is a general rule that can help the decision makers build new hypotheses (step 3 in Figure 25). In this case study, the new perspective was to look at the two sets of patterns from a customer perspective, rather than a channel perspective. [redacted] For instance, the age of the clients may matter, or the amount of direct debits they have on their payment account.

The search for new knowledge that can be obtained by reflecting on the insights from the preceding evaluation steps offers opportunities for further use. While the search for patterns with new knowledge

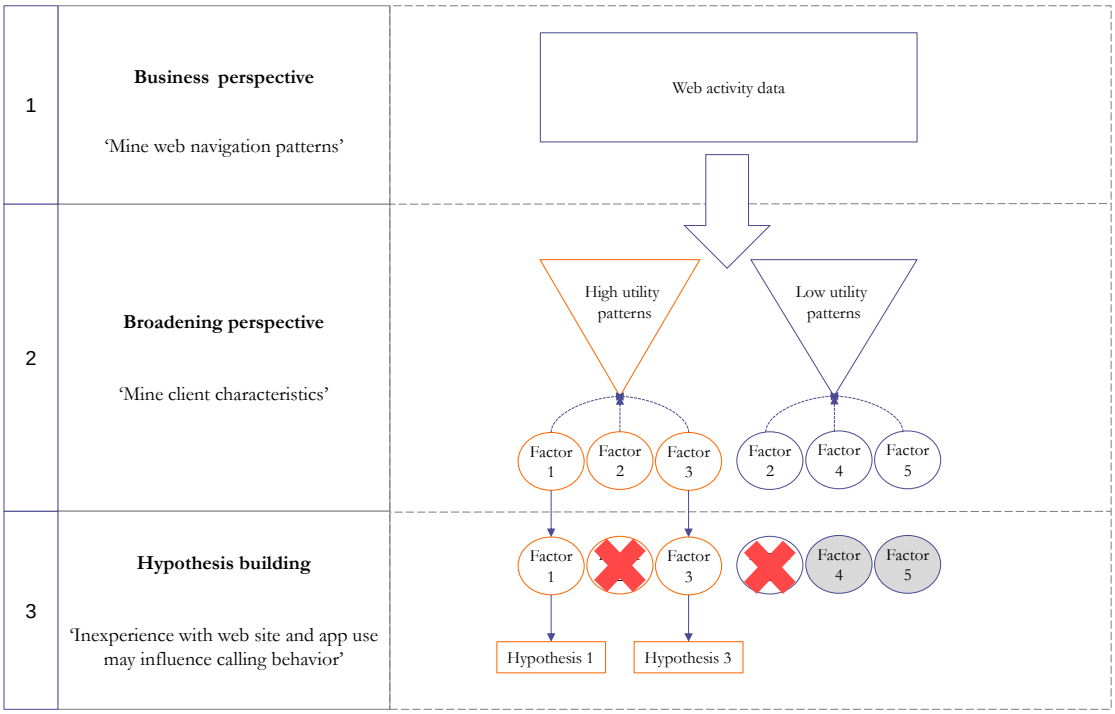



**Figure 25: schematic overview of how novelty evaluation (step 2 and 3) can be conducted to gain additive knowledge**

in the initial data set is rather unstructured, the search for patterns from a new perspective builds on the lessons learned in the previous evaluation steps. In that way, the focus is on finding new explanations why decision makers find one pattern useful and the other not useful. [redacted]

By pinpointing the search for new knowledge on the difference between high-utility patterns and low-utility patterns, there is a greater chance that new and useful information can be found. Besides, by facilitating a new perspective, the horizon of the decision maker may be broadened or questioned.

In the case study, the search for latent factors was done by an exploratory factor analysis with multiple groups. With exploratory factor analysis, the common variance of variables is analyzed. If two variables have a large common variance, they may represent a higher order factor, that is not present in the dataset (Costello & Osborne, 1994).

While the validity and utility evaluation analyzed online activity of clients, the novelty evaluation was focused on giving an explanation from the perspective of clients[redacted]s. As applied in the research, the extent to which this novelty evaluation is truly exploratory is open for discussion.

## 4.4 Results

In the previous, the decisions made in the case study have been elaborated. In this paragraph, the results are presented for the four most important steps: the set-up, the validity evaluation, the utility evaluation and the novelty evaluation step. In each step, the distinction has been made between results for the case study (lowering the call flow by improving the web site) and for the research (improving data mining in organizations).

### 4.4.1 Project setup

In the setup of the project, the decisions were made on how to conduct the project. This has influence on the data that is collected, and the way in which stakeholders communicate during the project.

### Case study results

The most important result of the project setup was the increased collaboration of two departments that have been separated before: Call and Internet. During the project, both departments converged to a view on a problem. It leads to a situation where both departments feel more responsible for the other's performance. More concrete, the collaboration lead to the connection between the call and internet database. Traditionally, Call and Internet were separated and kept their data separated. For the case study, they converged view lead to the need for coupling Call and Internet data. This had never been researched before, and proved to be possible. This coupling opens up possibilities for an even more intensive collaboration.



**Figure 26: the meso-level insight; [redacted] of the clients that retransfer a direct debit through call, have been online in the preceding 24 hours**

The coupling of the data lead to an interesting insight that Call and Internet both could use in their performance evaluation. With the first connection between the data, Internet could investigate the outflow of their clients to Call (see Figure 26). In the left bar, the separate channel performance information is provided. This information provides the information that Internet is responsible for the bulk of the retransfers. [redacted]. However, when the focus is on the [redacted] of the retransfers that happen through Call, the insight arises that more than three quarters of those calls could have been prevented (if you support the assumptions). [redacted].

With the data of [redacted] 2014, a dataset of [redacted] sequences was selected. After the first cleaning steps (removing sequences only one click, and removing the main outliers), a dataset of 6901 cases remained. The summary of the data, which was requested through the TraMineR package, showed that 4721 unique cases remained, where [redacted] unique pages have been visited. This means that [redacted]% of the cases is unique, which leaves us with a relatively scattered dataset.

Case study set up in numbers

[redacted]

**Figure 27: Important numbers that indicate the complexity of the case study**

The intended setup lead to a complex situation. The intention was to use a narrow scope (direct debit process), but mine with a great level of detail (do not aggregate pages). Due to this decision and the fact that the selected data contained 1448 unique pages, the sheer size of possible combinations is almost infinite. Under the assumption that all pages are directly linked to one another (so every page is just one click away), and with an average sequence length of 11 (see Figure 46 in Appendix B), the number of possible patterns is theoretically $1.43 \cdot 10^{27}$. With the technical limitations of the algorithm, a maximum sequence length of 4 could be analyzed. This drastically brought the theoretical amount of patterns to $182 \cdot 10^{9}$ (182 billion) patterns[3]. As the initial setup was too complex to carry out, later iterations of the project offered more mandate for the data scientist to judge the complexity of the intended setup.

### Research results

With respect to formalizing a data mining process with stakeholder participation, the way in which agreements have been made are the result of this project phase. Since previous literature assigned the data scientist to determine the scope of the research, this case study was a first attempt to do this in a participatory way. In order to do so, insights on convergence of stakeholders and merging the objective data science with subjective business decision making are needed. These are all soft results and have not been measured.

[3] This is calculated by the $\sum_{k=1}^{4}$ [redacted]$_{k}$) combinations of pages (following the nCr principle).

With respect to the convergence of stakeholders, an early participation of the participants worked out well. When stakeholders are incorporated early in the process, there is time to discuss the problem and to come to shared goals. When the business stakeholders determine the problem for the organization as specific as possible, data scientist can translate that problem to a data mining setup. This leads to a better business understanding for the data scientist, and a better data understanding for the business stakeholders. This participation caused the fact that unintended insights (the meso-level insight in Figure 26) can be valued by business stakeholders. If it was not for the early participation of those stakeholders, this insight would not have created the amount of enthusiasm and cooperation from that point on.

Furthermore, it offers possibilities to objectify subjective information. For instance, the utility criteria have been determined before the data mining phase. When the criteria are not directly linked to the results, subjective results can be obtained in an objective way.

One downside of the participatory setting, was that the business perspective on the problem overruled the possible complexity issues for the data mining phase. In a participatory setting, a data scientist may focus too much on convergence instead of critically reviewing the proposed decisions from his own perspective.

### 4.4.2 Validity evaluation

In validity evaluation, the patterns that have been mined are evaluated on the validity criteria. Since the early exploration of the data showed that the patterns were very scattered, the frequency criterion was set relatively low [redacted]

#### Case study results

The validity evaluation of this research was an internal function of the TraMineR package of the data mining suite. As not all theoretically possible patterns are in the dataset, we cannot say with certainty how large the reduction in the validity evaluation really is. The theoretical possibility is used as an approximation, because TraMineR could not process an unconstrained dataset.

After the validity evaluation on the patterns that resulted from the setup, 59.063 valid patterns remained in a new dataset. This dataset was exported, so it could serve as input for the utility evaluation. Compared to the theoretically possible number of patterns, less than 1% of the possible patterns remained.

#### Research results

Firstly, the main insight is that validity evaluation can be used to significantly reduce the set of patterns that flows in to the utility evaluation. Even with a low validity threshold, many of the theoretically possible patterns are exempted from further research. Validity evaluation is therefore essential to cope with the rule quality problem. Of the three evaluation steps, validity evaluation is the most common one and is therefore highly likely to be automated in a data mining suite. The data scientist should be the leader of this process step, as he has the most expertise in this field. Two typical decisions that he should make or reflect upon, is the type of statistical measure that is used (frequency- or probability measure) and the retainment of the validity threshold.

During the validity evaluation, the data scientist has the much influence on the decisions made. While the business stakeholders can have influence in the preliminary phases on the project, a data scientist

should make the judgement to retain the validity measures after the first data exploration. As in this case, the data was very widespread, the data scientist had to make a tradeoff between the rigidness of the statistical measures and the workable outcomes after the validity evaluation. This problem arose because of the difficult scope that was chosen: the frequency of patterns should be high enough to represent value, but low enough to not only take into account regular web site activities (e.g. doing a normal bank transfer).

### 4.4.3 Utility evaluation

The 59.063 patterns that result from the validity evaluation should be evaluated on the value and information they hold for the stakeholders of Team Prevent & Shift. To this end, the utility evaluation focused on two aspects:

- pruning the patterns on their existence in a dataset (also 10.000 patterns) of successful online retransfer requests
- ranking the patterns on a utility function, that consisted of predefined criteria

#### Case study results

The utility evaluation was meant to identify possible web site improvements for Internet. Therefore, the level of detail for this analysis was high. From the utility criteria determined in the project setup, a utility function was constructed. This utility function focused on 4 aspects that are described in 4.3.3.



**Figure 28: the 50 best-ranked patterns resulting from the utility evaluation (without normalized scores visible, bur ranked with normalized scores)**

Each score on the factor was normalized, in order to make the factors comparable. This utility function resulted in a utility score, a dimensionless score that facilitated the ranking of the patterns in the

dataset. This resulted in a ranking of all 2507 valid patterns, of which the 50 best-ranked patterns are presented in Figure 28.

One might suspect that the interpretation of those patterns scores is still a difficult task, and it is. Possible improvements such as the aggregation of patterns was unsuitable for TraMineR, while no automated outranking method was found to facilitate the outranking of this amount of cases.

The interpretation of the patterns largely happened through a joint discussion. Eventually, the high ranked patterns have been aggregated to two main archetypes: clients that are on the direct debit-page ('incasso:overzicht' in Figure 28) but are not able to start the retransfer process, and clients that see a direct debit on their account balance, but do not manage to start the retransfer process (indicated in Figure 28 with a blue dot, backwards engineered in Figure 29 after discussion on interpretation).



**Figure 29: two typical types of high-utility patterns that are interpretable**

In the first archetype, This archetype scores well because of the relatively high confidence ratio. This means that (see Figure 30). It was not possible to identify if those clients navigated themselves or were redirected from a search engine (such as Google).



**Figure 30: an overview of the clickthrough rates of 'Call retransfer'- and 'Internet retransfer'-clients**

This pattern gave Internet the insight that the first step before their funnel starts [redacted] redirects clients very reliable to the retransfer process. The insight that this archetype offered was that Internet did not need to make many changes on the pages that only concern direct debits, since people on that page often succeed in a retransfer. Although Internet's efforts were focused on improving processes within the funnel, it seems as though their scope needs to be broadened.

The second archetype gives an explanation of how people discover a direct debit they want to retransfer. [redacted]. The click-through rate of customers that end up calling for the retransfer is much lower than the rate of those who succeed online (see Figure 31).

This pattern gave the insight to Internet that clients do not always intend to retransfer a direct debit when they are online, but may stumble upon an unwanted debit when they do their banking business. These type of customers need to be taken by the hand when they may stumble upon that information, and should be better directed to the retransfer overview page.

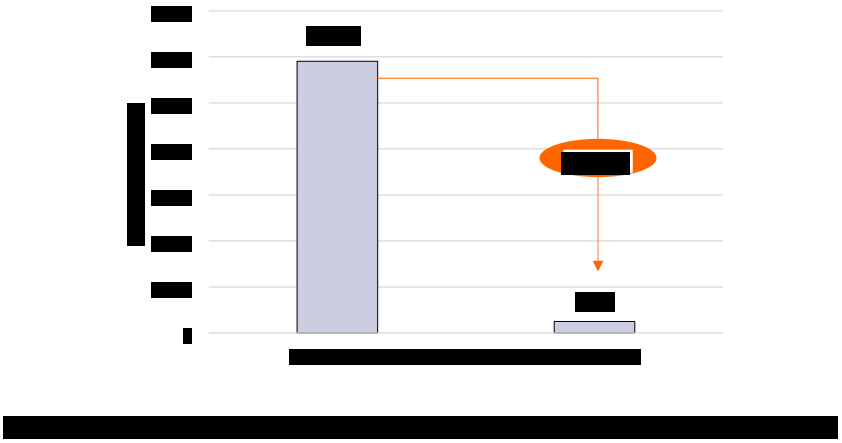


**Figure 31: the clickthrough rate of the [redacted] to the [redacted]**

### Research results

The addition of a utility evaluation in the data mining process proved to be very valuable. In this case study, the (sub) patterns were pruned on their existence in the dataset of cases where the retransfer of a collection was done successfully. This was done to evaluate if a pattern was significantly different from a 'successful' pattern or (if it did not exist in the successful dataset) if it did not have influence on retransferring a direct debit. As a result, the initial 59.063 valid patterns were pruned to a set of just 2507 patterns, a decrease of nearly 96% (see Figure 32). In this way, utility evaluation successfully tackles the rule quantity problem.

Furthermore, the resulting 2507 patterns were ranked on their score on the utility criteria. In this way, the patterns that were discussed were structured on their level of importance. The ranking of the patterns helped the discussion of the results and improved the interpretability of the patterns. In this way, ranking tackles the rule quality problem.



However, handing over a ranked list of patterns to a decision maker, that did was not acquainted with the research, did not result in more understanding of the decision maker. While the interpretation of results is improved through utility evaluation, the designed solutions still make the interpretation the job of a person. If that person is involved in the early stages of the process, he will understand the patterns better. In this research, guidance for external parties or for computer evaluation was very little to no existent. Aggregation of patterns proved to be technically infeasible, while the search for an automated outranking method was not fruitful. Both elements can serve as a way to improve the data mining process, and are therefore discussed in 9.4.

### 4.4.4 Novelty evaluation

As is described in 4.3.3, this case study focused on finding a latent factor that could explain the high utility patterns from the client perspective. This was done by inputting client data in an exploratory factor analysis setup.

#### Case study results

For the exploratory factor analysis, the first 50 patterns were deemed as high utility patterns. These were all patterns that resemble one of the two archetypes (discussed in 4.4.3).

After performing the exploratory factor analysis, the high-utility patterns ended in a 4 factor solution and the low-utility patterns in a two-factor solution. These are presented in Table 7.

As the factor analysis shows, the factors of the low-utility factors are exactly the same as two factors found in the high-utility pattern set (Factor 1 and A, Factor 2 and B). This means that these two factors do not explain the difference between high-utility and low-utility, and are therefore removed from the factor solutions. In this way, two factors seem to have influence on a pattern being of high-utility. These factors are the [redacted] and [redacted]

The [redacted]e of customers has a positive effect on patterns being of high utility. The meaning that team P&S gave to this factor, is [redacted]

The second characteristic factor in the high-utility group is the [redacted]. [redacted]

[redacted]

**Table 7: the factors of the both sets of utility patterns**

| Variable | High-utility factors | | | | Low-utility factors | |
|---|---|---|---|---|---|---|
| | Factor 1 | Factor 2 | Factor 3 | Factor 4 | Factor A | Factor B |
| [redacted] | + | | | | + | |
| [redacted] | | | + | | | |
| [redacted] | + | | | | + | |
| [redacted] | + | | | | + | |
| [redacted] | | | - | | | |
| [redacted] | | | | + | | |
| [redacted] | + | | | | + | |
| [redacted] | | + | | | | + |
| [redacted] | | | | + | | |
| [redacted] | | | + | | | |
| [redacted] | | + | | | | + |

### Research results

Perhaps the most important research result for novelty evaluation is not to be found in the results of the case study, but in the preparation of it. The indication of different types of novelty can be seen as a new way to look at the novelty criterion of data mining patterns.

Moreover, the conclusion that novelty is an additive criterion for decision makers is an insight in the discrepancy in opinion between the participants. As a data scientist it is important that all three criteria have been met, the decision makers are more likely to value new insights as unimportant (at least, when they have insights with sufficient utility). The idea of forming new hypotheses may be too abstract for decision makers, who seem to value short term value as most important. In this way, the decision makers were less enthusiastic about the idea of imposing a different perspective on the problem. Therefore, if novelty evaluation is positioned as an additive learning experience on the insights that already have been gained, it is most accepted by the organization.

In order to create a high-utility and low-utility group, team P&S needed to decide upon the cut-off point for high-utility patterns versus low-utility patterns. Here, it is most important that the high-utility set is as conclusive as can be. However, for the sake of validity, the number of variables that can be inputted in a factor analysis depends on the amount of cases to be analyzed (Costello & Osborne, 1994). This was a trade-off that certainly had its consequences. Because the conciseness of high-utility patterns was valued highly, concessions had to be made on the amount of variables that were inputted in the factor analysis.

### 4.4.5 Deliverables of the data mining project

[redacted]

## 4.5 Insights gained from the case study

The previous briefly presents the contents of the case study and the way the case study was executed. With respect to the final deliverable of this research, design of a data mining framework, the new insights are discussed below.

### 4.5.1 Stakeholders should be early participants in the project

As became clear in the early phases of the project, it takes the support multiple stakeholders to conduct a successful and meaningful data mining research. However, these stakeholders have different interests and have different mandate. Previous DM-frameworks prescribe the data scientist to take the business perspective into account, besides his responsibility for a sound analysis. One of the main findings of Chapter 2 was that taking account for the business perspective does not fit the role of a data scientist. Therefore, this report advocates for active participation of business stakeholders in the data mining process. Lessons from the practice support this, because an early collaboration between stakeholders brings two main benefits. Firstly, the decision makers are on board very early in the process. This means that the business goals can be discussed by the people who actually have to turn the potential

insights into action. This leads to a good fit of the DM-project to the business desires and it creates ownership among the business stakeholders. Secondly, the gap between data mining and decision making is bridged. As the decision makers are incorporate in the early stages of the process, the discussion on the future actions that have to be taken can start as well. In this way, pattern evaluation and decision making can be combined. When decision makers are active participants of a data mining project, their focus is more on the information that is needed to make a good decision. In this way, participants in a data mining project are more forced to think about the actionability of patterns. No handover of knowledge is needed after the patterns are specified, leading to an integration of data mining and decision making.

One of the potential downsides of this participatory approach is that with the inclusion of multiple stakeholders so early in the process, the process may become too subjective and may become part of politicization. While in the case study, there was little risk on politicization, the threat of too much subjectivity in the evaluation was coped with by decoupling the subjective evaluation from the objective evaluation. Moreover, the design of subjective evaluation criteria had to be done even before the data mining phase, in order to prevent 'winner picking' as much as possible. In this way, stakeholder participation can be facilitated without violating the objective nature of data mining.

### 4.5.2 Utility is very case-specific

The literature study gave a clear indication to incorporate a utility evaluation step in the data mining process. The initial goal of the case study on this point was to find aspects that define the utility of a DM-pattern. However, as became clear in the case study, utility cannot be universally expressed in one general function. The actual meaning of the utility of a pattern is case-specific and influenced by the participants in the project. In this research, it proved to be impossible to define a general utility function. Therefore, decisions makers should be allowed more freedom in the data mining process to determine what information is relevant to them. After all, they know what information is needed to come to a decision. The implications of this finding is that for every data mining project, a new utility function has to be defined.

In previous data mining frameworks, the data scientist was expected to take account for the business perspective and come to measurable criteria in an objective way. If decision makers are allowed freedom for determining their own criteria, the evaluation becomes more subjective. However, decision makers generally do not have expertise in the data and construction of objective measures. This is where the skills and expertise of the data scientist is of use. In order to maintain an objective process with subjective information, the data scientist can translate the subjective information from the decision makers to well-constructed measures for pattern evaluation. In this way, data scientist have control to ensure that the subjective information is incorporated appropriately. The data scientist should focus on the consistency and measurability of the evaluation criteria, in order to translate the business utility to a correct utility function.

### 4.5.3 Multiple evaluation criteria decrease the set of resulting patterns successfully

As is described before, the participation of decision makers causes a situation in which more emphasis is laid on the utility of patterns. Utility evaluation can be used as a ranking or filtering mechanism. This results respectively in a prioritization of the patterns for decision makers to interpret, or in a trimmed set of patterns for interpretation by the decision makers. When the set of patterns are organized

according to their value for the decision maker, it becomes easier for the decision makers to discover actionable patterns out of the total set of valid patterns.

In the case study, the validity evaluation pruned nearly all theoretically possible patterns, while the utility evaluation reduced the set of valid patterns with nearly 96%. While this exact pruning rule is not generalizable (since utility is case specific), it indicates the benefit of successfully incorporating utility evaluation in data mining. Moreover, the ranking of the 2507 patterns on their utility value gave the decision makers an overview on the high utility patterns. This helped to bridge the gap between the patterns and the interpretation of those patterns.

### 4.5.4 Novelty is multi-interpretable

During the case study, a first attempt to stimulate to find new knowledge has been tested. However, the term novelty seemed the one that was hardest to grasp by both the decision makers as the data scientist. Eventually, two types of novelty were identified: case-specific novelty and meta-novelty.

Case-specific novelty is the search for new information within a data mining project. Traditionally, a data scientist would test multiple data sources and multiple angles to try to gain insight into a problem. He then creates multiple setups of a data mining project, resulting in a number of iterations during the project. This ongoing process of learning and applying the new lessons is one of the main benefits of data mining for knowledge discovery. However, when decision makers have the freedom to influence the scope of the project along their perspective, the results become a result of that perspective. In this way, only insights from that perspective are desired and recognizable as patterns that have utility. Thus, the pitfall of stakeholder participation in data mining projects is a too narrow scope to discover new knowledge. In order to cope with this problem in a participatory setting, a totally different setup of the data mining project needs to be explicitly mentioned in the data mining process. While a true data miner would view this as a new iteration of the search for valuable and meaningful knowledge, this new iteration is not a given in a participatory data mining process.

Meta-novelty is the search for higher level knowledge. When data mining is performed a number of times on a number of processes, two large sets of patterns can be identified: a set of high-utility patterns and a set of low utility patterns. When sufficient patterns are acquired, it becomes interesting to reflect on the patterns. In this way, a general rule can be discovered that can explain the distinction between high-utility and low-utility. It therefore offers a reflection on the user perspective in the previous data mining executions. To do this, the initial discovered patterns (that result from the utility evaluation) serve as input for a higher-level data mining project. Both business as the data scientist acknowledged the potential benefit of this exercise. Unfortunately, due to time constraints and the lack of available patterns, this meta-evaluation has not been possible to carry out in this case study.

### 4.5.5 There is a difference in opinion on the minimal context for action

As is described in the 3.3, a data mining process should try to find a balance between inductive reasoning and deductive reasoning. This imbalance on the process perspective is also expressed in technical aspects. While data scientists indicate the novelty as a hard constraint for a data mining pattern, the scrum team judged pattern validity and –utility as a minimal context to substantiate action. This caused a situation where the stakeholders wanted to proceed to design propositions to attain the goals of the project, while the data scientist dove deep in the novelty evaluation. If this situation arises, it causes a threat for the project team to misalign and fall apart. If a framework successfully creates convergence between stakeholders, it should also effectively deal with this misalignment.

### 4.5.6 Data mining projects are not always attainable

This case study is an example of time-dependent association rule mining. This means that the DM-method resembles that from the market-basket example in Chapter 3, but with an added complexity because the sequence of the visited pages matters. It is this complexity that makes this problem less suitable for the widely accepted algorithms for association rule mining. Sequence mining can therefore be regarded as a state-of-the-art mining method, and should still be developed further. The most suitable sequence mining package, TraMineR, has been developed in 2011 and the algorithm is still in development (Gabadinho, Ritschard, Studer, et al., 2011) (Gabadinho et al., 2015). There were some problems with the settings of the analysis that caused a significant delay in the project. Other possible DM-suites also seemed problematic (which is further described in Appendix B, in the following the complexity with the TraMineR package is discussed).



In the initial agreements on how to mine and analyse the data, the agreements were made to include every case and every page of that case. With this dataset, the model was able to transform the input to sequences, but was not able to analyse those sequences. After a histogram plot, some outliers were detected and removed. However, the sequences were still too long to analyse the subsequences. With a mean of 11 clicks per session, the model did not deliver any results. Therefore, the complexity of the input data needed to be decreased. After removing loops (the same visited multiple times in a row) and aggregating pages that have an arguably different content than direct debits (mortgages,



| Subsequences up to length 5 | | | | |
|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 |
| [redacted] (H) | H-L | H-L-B | H-L-B-T | H-L-B-T-C |
| [redacted] (L) | L-B | L-B-T | L-B-T-C | L-B-T-C-O |
| [redacted] (B) | B-T | B-T-C | B-T-C-O | |
| [redacted] (T) | T-C | T-C-O | | |
| [redacted] (C) | C-O | | | |
| [redacted] (O) | | | | |

**Figure 34: Example of the possible subsequences that can result from TraMineR**

investments, etc.), the complexity of the dataset was reduced so that subsequences up to 5 clicks long could be analysed. With a mean of 11 clicks per session, most of the sequences could not be entirely analysed at once, but the sequences are then broken up in parts (see figure 22 for example). These parts would have appeared in the results as well, but now the 'mother-sequence' cannot be shown entirely. The limitation that subsequences can have a maximum length of 4 is due to the complexity of the data and the maturity of the algorithm used. The implications for the results are not severe, as the main focus was already on subsequences.

However, this complexity delayed the case study significantly. Looking back on the project, team P&S felt that with a proper focus on the attainability of the DM-suite, some challenges could have been overcome earlier on in the project. If the proposed setup would be approached with some realism earlier in the DM-process, these delays may be prevented in future research.

## 4.6 Case study conclusions

In this part, the lessons learned from the case study serve as the basis for the answer on the subquestions. These insights will then lead to a revised set of requirements that will be used to design propositions for the improvement of DM in organizations.

### 4.6.1 Answers to the subquestions

This case study was conducted as the relevance cycle for the design science approach of improving data mining in organizations. This cycle serves as input for the design of propositions for a new data mining framework. Before the requirements table is reviewed, two other subquestions will be answered:

4. *How can the data mining process in organizations be improved?*

When data mining is performed in organizations, performing a solid and adequate analysis is not the only concern of the project. Moreover, the results must offer a decision maker sufficient knowledge to make a decision or propose a change. While the quality of the analysis is the concern of a data scientist, the usability of the results in the organization could be better secured by the organization's stakeholders. These departments need to participate in the beginning of the process, in order to get acquainted with DM and to determine the scope of the project. Moreover, they should determine evaluation criteria that measure the value of a pattern for the organization. In this part, the data scientist plays the role of a facilitator and makes sure that the business criteria are deducted to consistent, measurable variables.

The early incorporation of these parties leads to a situation in which there are multiple parties with different goals and interest. Therefore, the begin of the data mining process needs to emphasize on convergence between the stakeholders. Agreements need to be made on the problem, the data mining approach and the way in which resulting patterns are treated. By the early participation, decision makers can objectively determine beforehand what type of information for them is valuable and actionable. In this way, the gap between data mining and decision making is bridged.

5. *How can the evaluation of patterns be improved?*

In order to improve data mining in organizations, the pattern evaluation phase should focus on finding patterns that are valid, have a high utility and represent new knowledge.

This can also be done by taking the business perspective more into account. As is described in the answer on the previous subquestion, the primary concern of the data scientist is to make a sound analysis. Therefore, he focuses mostly on attaining valid results. While validity is an important characteristic of meaningful patterns, the results should also be valuable and new.

The validity of a pattern should be measured with frequency- or likelihood measures. The value of a pattern is most relevant to decision makers, and should represent the benefit that a stakeholder can gain from the information the pattern contains. Therefore, a utility evaluation can be performed to rank or prune valid data mining patterns. Since utility is case-specific (and dependent on the stakeholders included in the process), no general utility function can be presented. A guideline for the design of a utility function is that is consists of actionability, interpretability and the value of a pattern. The novelty of a pattern should enforce a different perspective on the data mining project. This may sound logical for the data miner, but is not a given for a decision maker. On the other hand, novelty can serve as a meta-analysis to indicate a new possible pattern within the high-utility patterns that does not exist in the set of low-utility patterns. In general terms, novelty measures the generalizability of the knowledge that has been gained through data mining.

### 4.6.2 Framework design requirements

With insights from the literature review, some solution directions could be identified. These directions were tested in the case study in several iterations. The case study entailed the improvement of the [redacted] website for the functionality of retransferring a direct debit. The insights from the case study can serve as an extra input for the design requirements for the data mining framework, in order to secure that the framework has a theoretical and practical foundation.

Many of the requirements identified in the literature review still hold after performing the case study. For instance, utility and novelty evaluation have been tested and seemed beneficial for the project outcomes. Although the case study provided extra information on the process of determining and designing these criteria, no new requirement will be designed for this. These extra insights will fall under the already identified requirements for utility and novelty evaluation. The requirements that have not been substantiated through literature, but proved to be of significant importance in practice are the following: *investigate the attainability of the approach before data mining* and *determine the minimal context for action*. The now complete overview of requirements is presented in Table 8.

Firstly, since the data mining itself caused some problems during the cases study, a data mining framework should force a critical view on the possibility of applying data mining and a specific data mining suite. Since the current state-of-the-art of data mining was not as far developed, it had problems handling the data on the initial desired level of detail. This was an unforeseen outcome, and no scientific guidance here has been found. The possible back-up method proved to be in any case invalid, since it was practically impossible to acquire a valid population size. This could have been discovered earlier on in the process. If that step is secured in the data mining framework, possible adaptations can be applied or expectancies can be managed.

Secondly, during the case study, the decision makers treated the novelty of a pattern in a different way than the literature describes. Data scientists treated novelty as a hard constraint. On the other hand, decision makers indicated that DM should not necessarily deliver a new insight, but that a valid and high-utility pattern is also sufficient to design propositions for. Therefore, the framework should in some way allow that data scientists and decision makers can see different things in the patterns. This

freedom for different interpretations should however not conflict with the convergence that is discussed before.

**Table 8: Overview of design requirements**

| Functional requirements |
|---|
| • Perform validity evaluation |
| • Perform utility evaluation |
| • Perform novelty evaluation |
| • Subjective, problem specific information should be able to influence the DM process |
| • Prevent friction between utility and novelty |
| • Create a common perspective for all stakeholders in the data mining project |
| • Secure the 4 modes of knowledge creation |
| • Investigate attainability of the approach before data mining |
| • Determine the minimal context for action |
| • **Non-functional requirements** |
| • Interactive process between data scientists and decision makers |
| • Provide simple solutions for complex decision making issues |
| • Generalizable framework to apply in different contexts |
| • Scalable process to apply on large datasets |
| • Objective pattern selection and pattern evaluation |

# 5 Design propositions

Following the design science framework of Hevner et al. (2007; 2004), a design can be made with the input of the relevance and the rigor cycle. Both cycles resulted in functional and non-functional requirements for a framework for data mining in organizations. This chapter takes these requirements as a starting point and presents the propositions that satisfy the requirements. Firstly, the propositions are shortly described. After that, a more detailed analysis of every proposition is given. These propositions will serve as the input for the eventual framework for DM in organizations.

## 5.1 Overview of the design propositions for data mining framework

In the literature review and the case study, several existing problems have been identified why data mining is hard to apply in business. These resulted in functional requirement and non-functional requirements to be incorporated in a single data mining framework. These design propositions fulfill the functional and non-functional requirements, and should thus help to solve the problems that have been identified.

Five design elements are proposed that solve parts of the problem. Table 9 shows the design propositions and their origin in the design science research framework. Below, a short introduction of the proposition is given. A detailed and substantial description is described from paragraph 5.2 and onwards.

**Table 9: Design propositions from the relevance and rigor cycle**

| Literature analysis | | Case study |
|---|---|---|
| - Triple evaluation approach<br>- Design a process for convergence | | - Decouple evaluation steps<br>- Perform attainability test<br>- Allow freedom of interpretation to deal with friction |

### 5.1.1 Triple evaluation approach

The triple evaluation approach deals with the rule quality and rule quantity problem. These problems originate from the fact that patterns are only evaluated statistically (see paragraph 3.1 and 3.2) It explicitly takes into account measures for the validity, utility and novelty of a data mining result, based on the definition of a data mining pattern by Cabena et al. (1999). This allows the decision maker to specify potential value of a pattern in the evaluation.

### 5.1.2 Decouple evaluation steps

By decoupling the evaluation steps, not all evaluation criteria are used at once. This allows the opportunity to use one criterion as a hard constraint and others as a ranking mechanism (Geng & Hamilton, 2006a). In this way it is made possible that the three types of evaluation serve a different goal, which they are supposed to. Validity evaluation is used to filter statistical invalid patterns. Utility

evaluation is then used to rank the patterns according to their business value. Finally, novelty evaluation is applied after the utility evaluation to investigate differences between high-utility and low utility patterns.

Decoupling the evaluation steps partly solves the third problem identified in the literature: the misalignment between data mining and decision making. If the process steps are decoupled, the opportunity arises to vary between inductive and deductive reasoning. Furthermore, it offers the opportunity to search for new information with a stronger focus.

### 5.1.3 Design a process for convergence

The new DM process should facilitate the stakeholders of a DM project to create a common perspective on the problem. As is described in 3.3, participators in DM have different goals, means and competencies to solve them. In data mining in organizations, these perspectives should help each other to come to valid, valuable and new solutions for the problem. Therefore, the process design focusses on convergence and making agreements on the evaluation with active involvement of decision makers. In this way, the socialization and combination mode of the knowledge creation cycle (see 3.3.1) are secured. Externalization is accounted for by determining evaluation criteria in an open and transparent process. By determining utility criteria beforehand, a decision maker is able to specify his preference without violating a core value of the data scientists: an objective pattern evaluation.

### 5.1.4 Perform attainability test

The attainability test consists of two types of tests: a complexity check and a minimum observations check. The goal of this proposition is to make the attainability of the project clear earlier on in the process. In this way, significant delays can be prevented.

Since data mining is relatively new and in rapid development, not all algorithms have been thoroughly tested on different types of datasets. During the case study at [redacted] the data appeared to be too complex to carry out the sequence analysis with the compatible algorithm. This cost a significant delay in the project. In order to prevent this from possibly happening in other projects, the complexity of the data needs to be assessed. The attainability can then be tested by performing a quick run of the most complex data. However, since there is not much scientific guidance on this topic, it also comes down to trust in the professional judgement of a data scientist.

The minimum observations check is performed to investigate if the created dataset is suitable to validly perform data mining[4]. To this end, the number of variables and the number of observations needs to be taken into account to see if there are sufficient observations available. If this is not the case, the adhocracy must judge if they change the experimental setup, aggregate the data or stop the data mining project.

### 5.1.5 Allow freedom of interpretation to deal with friction

While the business perspective and the data mining perspective have to converge to a common perspective for determining evaluation criteria, the way in which the patterns are treated after the evaluation does not necessarily need consensus. Here is where there is friction between the utility and novelty of a pattern. In order to stimulate the deployment of the results, the interpretation of a pattern

[4] note: validity here concerns something different then the validity in 5.1.1. Here, it concerns the validity to perform data mining, not the validity of a pattern that results from data mining.

should not be challenged, as long as it leads to a system or artifact. Based on the outcomes of that system, the interpretation can be challenged and adjusted. This creates a constant refinement of the system and stimulates the internalization of data mining results.

### 5.1.6 Fulfilling the requirements

In chapter 3 and 4, the rule quantity problem, rule quality problem and the misalignment between data mining and decision making has been analyzed and transformed to requirements: elements that a good data mining framework should contain. These requirements need to be satisfied by the design propositions, in order to provide a good data mining framework. Table 10 indicates what functional requirements are fulfilled by the propositions, with a short explanation. A more detailed description is given in next paragraphs.

**Table 10: Fullfillment of the requirements by the propositions**

| | TRIP | DEC | PRO | ATT | FREE | Explanation |
|---|---|---|---|---|---|---|
| **Functional requirements** | | | | | | |
| **Perform validity evaluation** | | | | | | TRIP explicitly takes into account the validity of a pattern |
| **Perform utility evaluation** | | | | | | See above, but then for utility |
| **Perform novelty evaluation** | | | | | | See above, but then for novelty |
| **Subjective, problem specific information should be able to influence the DM process** | | | | | | Utility evaluation is decoupled from validity evaluation to facilitate subjectivity in an objective way. |
| **Prevent friction between utility and novelty** | | | | | | Decouple utility and novelty evaluation, novelty as additive insight |
| **Create a common perspective for all stakeholders in the data mining project** | | | | | | Reaching agreements on problem, goal and data mining suite in earlier stages of the process |
| **Secure the 4 modes of knowledge creation** | | | | | | Socialization and combination in the convergence phase, externalization by AHP, internalization through crystallization |
| **Investigate feasibility of the approach before data mining** | | | | | | Operational and statistical attainability is tested |
| **Determine minimal context for action** | | | | | | Allow crystallization of valid and high-utility insights |
| **Non-functional requirements** | | | | | | |
| **Interactivity** | | | | | | Process design forces early involvement of decision makers |
| **Simplicity** | | | | | | Decoupling and freedom for interpretation provides simple evaluation |
| **Generalizability** | | | | | | Process design provides process-oriented decisions instead of outcome-oriented |
| **Scalability** | | | | | | Evaluation is automized as much as possible, would be improved when incorporated with outranking mechanism |
| **Objectivity** | | | | | | Objective evaluation is decoupled from subjective information, winner picking is prevented through AHP |

## 5.2 Triple evaluation approach

The problem with subjective and semantic interestingness is that both are broad terms and offer a lot of freedom for the user to define semantic and subjective information. However, the freedom may not be beneficial for the data mining outcomes, since they are often not used in the pattern evaluation. Too much freedom then has a negative effect on the adoption of those measures. Interesting then is not the right term to use in pattern evaluation. In data mining, the objective is not to look for interesting information, but for information that is valid, actionable and unknown.

In the previous, it seemed hard to combine the three types of evaluation measures in one process, and no description has been given on how to do this. To reduce complexity, I argue to pinpoint the evaluation more to the definition of data mining. Therefore three concepts of evaluation are introduced: validity, utility and novelty. In this way, the evaluation approach secures that patterns will be tested on all requirements for DM outcomes (see ). By explicitly securing these concepts in the data mining framework, the adoption of semantic and subjective information becomes a part of DM. In this way, the rule quality and rule quantity problem are properly addressed.

The introduction of validity, utility and novelty evaluation can be regarded as a prescription for the outcomes on the meso-level. Lower-level prescriptions (e.g. specific evaluation measures) for these evaluation steps conflict with the generalizability of data mining. Since DM consists of a broad variety of methods and techniques, it is unlikely that a single evaluation measure for every type of DM exists. On a case level, consensus on should be reached to translate the three evaluation concepts to concrete measures (more on this in 5.4).

As is elaborated further on this chapter, the three evaluation steps rely on different methods. Validity evaluation relies on statistics, utility evaluation is applied with the Analytical Hierarchy process and novelty evaluation is done by applying exploratory factor analysis in two groups.

### 5.2.1 Validity

Validity is the label of the traditional methods to evaluate DM outcomes. Validity measures, as described in chapter 2, are often based on frequency values. By incorporating validity measures, data

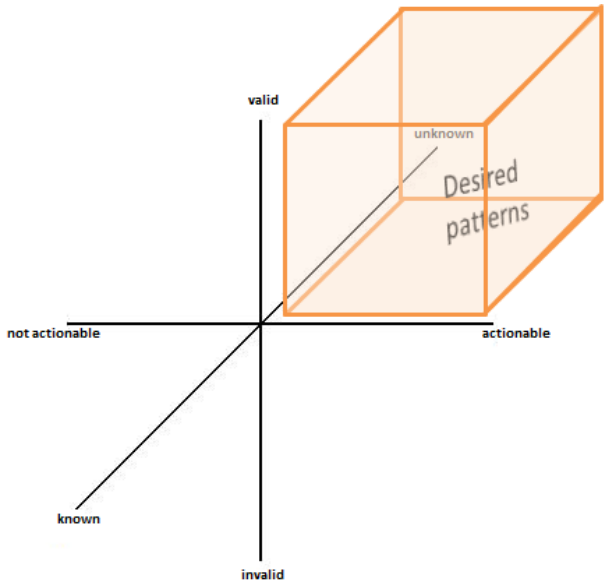



**Figure 35: Demarcation of desired patterns: valid, unknown**

mining delivers patterns that are significant, so that the researcher can make conclusions based on the patterns. Data mining outcomes are then scientifically acceptable.

From both the data mining perspective as the business perspective, validity evaluation is a logical step. Validity does not include user- or problem specific dynamics in the evaluation, so the results of the evaluation phase are objective observations. Validity evaluation is used as the first filter of patterns and is treated as a constraint. If a pattern does not satisfy the validity constraint, it is not taken into account for the consequent phases.

The validity evaluation step is not a revolutionary step, since it is always used in data mining. This step alone does not solve the rule quantity or –quality problem. For that reason, utility and novelty evaluation are introduced.

### 5.2.2 Utility

The main goal of utility evaluation is to highlight patterns that are relevant to the operational goal of the stakeholders involved. Examples of utility goals are e.g. business value or actionability of a pattern. In previous DM processes used, use of these types of seemed hard to incorporate in the objective nature of data mining. In this framework, the notion that merely statistic measures should be used is released, resulting in more effective ways to define utility (Choi et al., 2005). To this end, the utility evaluation is deliberately decoupled from the validity evaluation of patterns.

In the utility evaluation phase, patterns are judged on their potential contribution to attaining a certain business goal that is determined in the converging phase of the DM process. To this end, a process that resembles high utility pattern mining is embedded in the DM framework. High utility pattern mining is a niche development of DM, explicitly taking into account other values than statistical validity (Jagannath, 2003)(Braynova & Pendharkar, 2005). This niche only takes into account the utility preference of patterns (so ignores statistical validity), and therefore develops algorithms only for utility evaluation. It then uses utility as a filtering constraint for patterns. Since utility is not the only requirement for a DM pattern, the new framework proposes to use utility more as a ranking mechanism than only as a constraint. If utility is used for both filtering and ranking, the resulting set of patterns is smaller in size and ranked to their potential to fulfill business goals. In this way, the rule quantity and rule quality problem are adequately addressed (Yen et al., 2005).

Since utility guided pattern mining is rather new, most in-algorithm utility filtering techniques only are able to capture basic utility considerations such as profits. In more complex DM exercises, such as in the case study at [redacted] utility is not easily captured by constraints as profit. To release decision makers from the limits that this in-algorithm evaluation poses, the proposition is made to considerate utility as a pruning and ranking mechanism outside of the DM suite.

Ideally, utility guided pattern mining is combined with an alternative outranking method to choose the best alternative. It would add to the objectivity, as patterns are explicitly ranked on their criteria scores instead of an overall utility approach. A possible application of this outranking method is ELECTRE, an easy way of measuring the robustness of an alternative (Roy, 1991). However, ELECTRE is not suitable to apply on large sets of outcomes, since it would become a very time-consuming method. Therefore, the outranking of alternatives is still problematic in data mining. Ironically, if it would be possible to automate the outranking step, the scalability of the framework would be increased. The

framework then can automatically outrank a lot of patterns. Further research can be conducted in how this can be integrated in DM.

### 5.2.3 Novelty

A possible critique on the utility approach could be that it is too much dominated by the decision maker's ideas, and is therefore not inductive enough. Possible novel information is then left out of scope of the project, making it questionable if one is still performing DM. In order to deal with a possible lack of novel information, the novelty evaluation phase is introduced as a targeted way to look for novel information.

When performing utility evaluation, patterns are selected that are interesting according to the perspective of the decision maker. Although performed with AHP to balance deduction and induction, utility evaluation does not necessarily provide previously unknown insights. However, previously unknown insights that are not valuable for the decision maker are not desirable. That is why the explicit search for novel information is best performed after the utility evaluation phase.

In the novelty evaluation phase, the scope that has been narrowed down by the utility evaluation phase is again broadened (this is substantiated in 5.3). After the utility phase, a decision maker has patterns distinguished in two categories: relevant and not relevant. This is already information on which a business decision can be based. However, other possible interesting explanations lay in this distinction.

In principle, why some patterns represent value and some do not, is because to score different on the predefined criteria. However, when looking inductively at these observations, possible other explanations may be a root cause for a pattern being valuable or relevant. With DM, context variables of the observations can be examined to look for regularities of interesting patterns. If the context variables give rise to expect regularity, a more general theory on why a pattern is relevant can be developed. Example 3 gives an example for the imaginary case of the supermarket manager.

In novelty evaluation, a general theory is assumable if the context variables of the high-utility patterns are significantly different from the variables of the low-utility patterns. To this end, the parties in the adhocracy should make a deliberate decision on where to place the cut of high vs low utility. This is likely most influenced by the decision maker, since he has had the most influence on the utility considerations. Context variables of the high-utility patterns are examined to discover an unknown variable that may explain why a pattern is interesting. This needs to be done without a pre-existing concept or idea. However, this latent variable needs to significantly differ from the latent variable in the low-utility pattern set.

The setup of the analysis is an exploratory factor analysis (EFA) with difference between groups (Costello & Osborne, 1994). Exploratory factor analysis is a variable reduction method that identifies the number of latent constructs and the underlying factors of the input variables. The methods does so by taking account of the common variance of the variables in the data set (Suhr, 2006). In more practical terms, EFA groups variables that together seem to measure the same object that is not specified in the dataset.

EFA consists of several steps to determine factors from a dataset (Suhr, 2006). First, the initial number of factors needs to be determined. This is done statistically. The factors need to be uncorrelated, and their eigenvalues represent the amount of variance accounted for by each factor. Consequently, the

modeler determines the number of factors that he will take into account for the following analysis. He should do so in two ways: investigate the variance of the factor and check if the factors are interpretable. A factor is interpretable if the variables have in some way a shared meaning. If this is not the case, it is impossible label the factor, so no hypothesis can be made.

When performing EFA in multiple groups, difference in factor variances and covariances are allowed along the groups. For the case study in this report, it means that the factors and loadings the loadings of variable (i.e. contribution of a variable to the factor) in the group 'high-utility patterns' may possibly differ from factors and loadings in the group 'low-utility patterns'. Any difference between the factors of the two groups may indicate an explanation for the distinction between high- and low-utility patterns.

## 5.3 Decouple evaluation steps

Since the three evaluation steps all serve a different goal of data mining, it is advisable to decouple the evaluation in three different phases. Where validity evaluation aims at delivering statistically valid patterns, utility evaluation focusses on action and novelty evaluation focusses knowledge discovery. It is not logical to combine all the three criteria to a single pattern score. When evaluation steps are decoupled, it gives the decision maker insight how a certain patterns scores on every aspect. This leads to a simpler and more adequate process. Moreover, validity and novelty evaluation is dominated by the DM-perspective, while utility evaluation is dominated by the business perspective. By decoupling the evaluation step, the friction between utility and novelty is minimized. The deductive step of utility evaluation is detached from an inductive step: novelty evaluation. This leads to the following order of evaluation steps (see Figure 36):

1. Validity evaluation to filter statistically invalid patterns
2. Utility evaluation to discover the most valuable patterns from a business perspective
3. Novelty evaluation to discover a possible regularity in high-utility patterns

This order of evaluation steps is deliberately chosen. Firstly, validity of a pattern should in all cases be treated a hard constraint. Non-valid patterns should be excluded from further analysis, so only patterns are selected that are sufficiently supported by the data. If validity of a pattern was not a hard constraint, a decision maker can design improvements while not supported by the data. Validity evaluation is a form of inductive reasoning, as it aims to find regularities based on observations. In this case, the outcomes are not yet influenced by the decision maker's preferences.

Utility evaluation is used as a way to reduce the rule quantity and –quality problem, by focusing on the patterns that are valuable for the business. The patterns resulting from the validity evaluation are pruned or ranked according to the predetermined criteria of business value. This leaves a set of high-utility patterns and a set of low-utility patterns. By using criteria, a decision maker is able to deduct patterns that are relevant for him. Since the criteria are predetermined, the user preference does not violate the objective nature of pattern evaluation.
Novelty evaluation should take into account both the results from the validity and utility evaluation. In organizations, a valid and novel insight is still useless if it has a low utility value. In that case, the analysis for new information needs to be more focused than just looking for new patterns in the total database. Therefore, the goal of novelty evaluation is that it should be an additive value to the outcomes of the previous analyses.

In the new framework, novelty evaluation is aimed at providing a possible explanation for the distinction between high-utility and low-utility patterns. The distinction between these two categories has been made deductively. In novelty evolution, induction can be used to form new hypotheses on why some patterns have a higher utility value than others. As is described in 5.2.3, this is done by analyzing context factors concerning the patterns.

By narrowing the scope of novelty information by incorporating the insights from the validity and utility evaluation, data scientist and decision makers are able to form hypotheses concerning the previous outcomes. In this way, no effort is wasted on finding new patterns that will be excluded beforehand. Moreover, the new information is scoped in such a way that it helps the decision maker get an extra understanding of the insights that he has already gained.

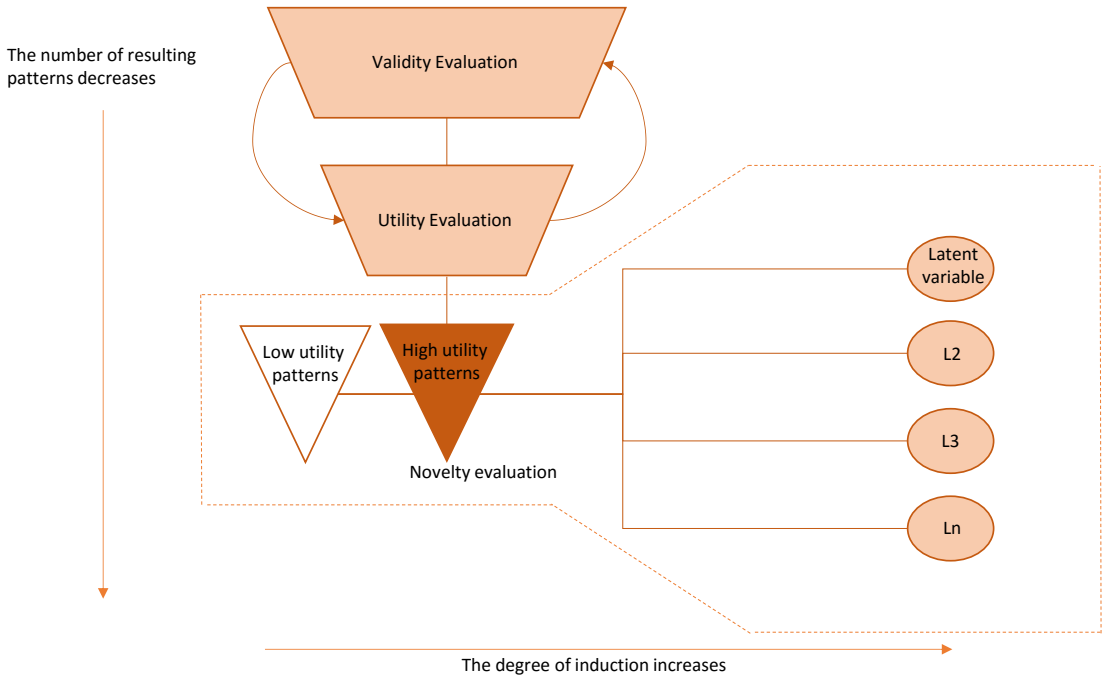



**Figure 36: Decoupling of the three evaluation steps**

## 5.4 Design a process for convergence and criteria determination

One of the main improvements for data mining in organization is to make it interactive by incorporating business stakeholders. As is described in 3.3.4, this creates a multi-perspective environment, where parties have a different view on the problem and approach. Before data mining is carried out, time should be spent on unifying these perspective to one project goal. Especially in decentralized teams (such as scrum teams), the participants of the project need to come to agreements on the problem, scope and evaluation criteria of the data mining project. Since the participants are all professionals of data mining or decision making, the problem of convergence is not that the participants do not have enough expertise to discuss the content of a data mining project. It is the way in which this discussion will take place that is important for the convergence of stakeholders. Therefore, it is wise to obtain a process-oriented approach. (Zeleny, 1998).

In order to cope with these (for data mining new) dynamics of interactivity, process management thinking can be adopted. The strength of process management is the creation of a demarcated playing field for discussion and negotiation of multiple stakeholders (De Bruijn et al., 2010). If such a playing field is created in data mining, stakeholders can interact with each other and form a common perspective on the problem field and the evaluation.

A good process consists design secures four elements: openness, protection of core values, progress and substance (De Bruijn et al., 2010). In order to secure this, several design principles have been translated in the context of data mining in Table 11.

Not all design principles should be adopted in every situation. To facilitate a good field for discussion, it is wise to always adopt principles for openness and the protection of core values. When there is transparency in the goals and interests of stakeholders, a common perspective can be found more easily. The protection of core values should secure a balance between utility thinking and novelty thinking, leading to true data mining patterns. In less strategic environments, progress and substance should normally be secured already. Increasingly with the strategic nature and the politicization of the organizational context, data mining can be confronted with unintended dynamics: sluggish decision making and strategic behavior. In that cases, the process design should incorporate more design principles for progress and substance. These design principles will be discussed throughout the paragraph, but Table 11 shows how the principles are fulfilled in the framework.

In general, the process design should focus on two elements of the data mining process: convergence between stakeholders and criteria determination. The possible design principles for these elements are also listed in Table 11, and will be elaborated in the following subparagraphs.

**Table 11: Process design principles in the DM process**

| | | Part of process design | Operationalization in process |
|---|---|---|---|
| Openness | Involve both decision makers and data scientists in decision making | Convergence between stakeholders | Decision makers and data scientist jointly work on problem definition and evaluation criteria |
| | Explicit specification of evaluation measures before performing DM | Criteria determination | AHP forces decision maker to reveal preferences before knowing which patterns are delivered |
| | Transparency in pattern evaluation | Criteria determination | Open discussion on criteria, agreements needed to use criteria |
| Protection of core values | Integrate deduction and induction | Criteria determination | Deduction facilitated with AHP, which is compatible with induction |
| | Agreements on evaluation quality rather than pattern quality | Convergence between stakeholders | Agreements do not secure that good patterns are found, but secure a good DM-process |
| | Objective validity evaluation | Criteria determination | Validity evaluation secures that only statistical significant patterns will be taken into account |
| | Incorporate user preferences | Criteria determination | Utility evaluation allows decision maker to focus on desired results |
| | Novelty evaluation | Criteria determination | Novelty evaluation provides extra options for new insights |
| Progress | Early participation of decision makers | Convergence between stakeholders | Decision makers participate from the beginning of the process |
| | Decision on attainability | Data mining | If the setup does not seem promising, a party can decide not to participate |
| | Clear outcomes of process steps | All | Clear defined deliverables before proceeding to next phase |
| | Quick insights in DM potential | Convergence between stakeholders | Clear problem definition enables quick scan in data |
| Substance | Bundling and unbundling of decision makers and data scientists | Convergence between stakeholders, design as boundary object | Bundling in problem definition and criteria determination, unbundling in evaluation |
| | Staging of discussion and selection in information need | All | Each phase has different information need and ends with decisions on that need |

The process that is designed takes into account the 4 modes of organizational knowledge creation. Socialization and combination will be used to create convergence in an adhocracy, externalization is used to facilitate user preferences of decision makers in an objective process. This makes it easier to interpret the patterns i.e. the internalization phase.

### 5.4.1 Convergence between stakeholders

Data mining often deals with problems that Rittel and Webber label as a 'wicked problem' (Rittel & Webber, 1973). Wicked problems are so complex and interdependent, that it is impossible to find one true answer. Even if a single true solution would exist for these problems, the analyst does not have

criteria that give an indication that the true solution has been found (unlike e.g. solving a math problem).

Instead, the goal should be to come up with good solutions. The goal of data mining in organizations is not to find the only true answer, but to find good solutions to business problems. In order to come to a good solution, the different stakeholders need to come to consensus on what a good solution looks like in their perspective. This is the convergence that Mintzberg & McHugh mention (1985). In wicked problems, multiple interpretations of the problem exist. . So it is up to the analyst to delineate the boundaries of the solution space. To do so, the problem and its characteristics have to be well-defined. In this way, the complexity of the problem is adequately dealt with.

In the CRISP model, this is the responsibility of the data scientists, taking account for the business problem and understanding the data. However, this does not reach consensus between the stakeholders involved. Therefore, the understanding phases also have to be interactive processes. This is supported by the knowledge creation cycle, since socialization and combination have the goal to form a common perspective on the problem (Lewin & Nonaka, 1994b).

The implications of the proposed convergence is that before actually performing data mining, the participants in an adhocracy form a common perspective on the problem that needs to be tackled. Since the participants come from different parts and departments of an organization, they have other views, interests and competencies. Through sharing tacit and explicit knowledge, an open discussion can be held on the opportunities and limitations of the data mining project.

For decision makers, it means that they will have to get acquainted with the data and data mining. The increased data-savviness of a decision maker tacitly manages the expectancies he has: when gaining insight in the possibilities of data mining and the data available, a decision maker learns what types of patterns can be delivered. With this increase in knowledge on data, a decision maker knows better what knowledge and what decisions are expected from him. This experience enables a decision maker to reason on the first design choices he has to make: the data mining setup and the data mining goal. This decision can now be made from the common perspective, which improves the trust and the commitment of the decision makers to the project. These decisions can now be influenced not only by the data mining goal, but also from the business goals of the project.

For data scientists, the knowledge redundancy means that they will have to gain insight in the current processes of the operating core of the company. This resembles the business understanding phase of the CRISP model. This phase is targeted at 'translating the business goals to data mining goals' (Sharma & Osei-Bryson, 2009). Whereas in the CRISP model, this is a sole function of the data scientist, this should be a shared responsibility of the adhocracy. Since the data scientist is embedded in the technostructure of the organization, it is strange to expect him to make a clear business problem description. Turning this into a shared responsibility raises the commitment and trust in the consequent steps of the DM process. Collaboration in the first phases of the project is therefore beneficial. For the data scientist, the process of convergence helps to gain insight in the current problems to be solved. He can than make a good decision on the relevant elements of data mining to explain to the business perspective. Decision makers and data scientists then can jointly come to a decision on the data mining setup and the relevant data that should be used in the following.

The convergence phase ends with consensus on the following elements:

- A demarcated problem statement, together with used assumptions
- A desirable end product or final insight
- Data deemed relevant for the problem
- Data mining suite

After these modes, the data scientist has sufficient inflow of the business perspective to perform an attainability test. This test is described in 5.5. The attainability test offers a data scientist a possibility to protect their core values and an exit option. The attainability test results in a decision if valid outcomes are to be expected and if the desired end product is realistic to expect. Outcomes of the attainability test offer the data scientist leverage to force changes in the convergence phase, in order to reach a valid DM execution or a valid end product. Besides, outcomes of the attainability test help the data miner to step out of the process if the tests show that reasonable outcomes are not to be expected. These tests make the attainability judgement of the data scientists transparent for the decision makers, so they too receive information to negotiate the steps to be taken.

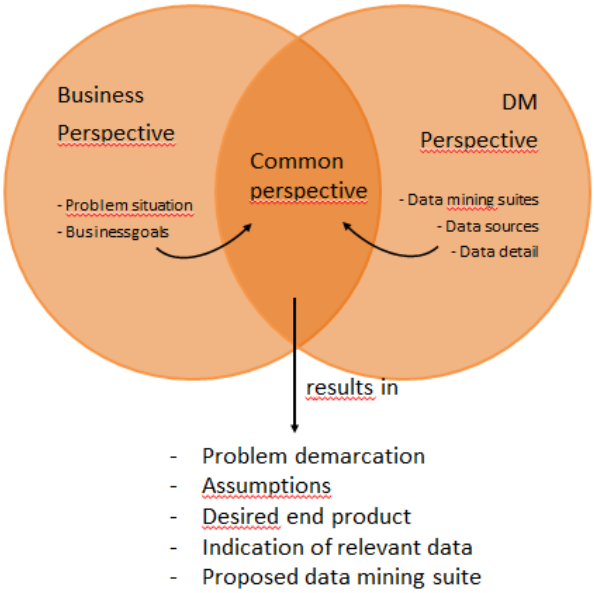



**Figure 37: convergence between business and DM-perspective**

### 5.4.2 Criteria determination

As is mentioned 5.2.2, the decision maker should get explicit influence in determining what patterns are relevant for him. In this way, deductive reasoning is applied on the patterns. The decision maker can incorporate existing ideas in the evaluation phase. However, this poses a threat to discovering new knowledge. In order to secure that *utility* and *novelty* do not conflict each other, a process has to be designed that make the two go hand in hand. This process is based on different decision making rounds for every type of criteria that has to be agreed upon.

While the main reason for allowing user specific information is finding actionable patterns, it needs to be prevented that winner picking will take place. A decision makers should not relate to the set of existing patterns (as was the case in the previous research in 3.4.5), because then he relates to a finite set of possible solutions (Wang et al., 2014). Instead, he should obtain a theoretically infinite set of

solutions, by specifying the characteristics of patterns that might be interesting for him. In this way, a decision maker is forced to specify beforehand what the definition of utility is to him.

The Analytical Hierarchy Process (AHP) is fit to be applied in the DM process. AHP is a decision making process that combines inductive and deductive reasoning in order to structure problems and solutions (Wang et al., 2014). The key concepts of AHP are determining objectives, criteria and weights, normalization of scores and ranking (Saaty, 1990). It deals with subjectivity and preference effectively by defining criteria before creating patterns in DM. This results in an objective pattern selection and –evaluation phase, while still facilitating subjective information. It is a broadly accepted and applied decision making theory, therefore it is easily applicable in organizations. It is a theory that can be applied for different types of criteria, as it can handle both linear and nonlinear scales.

When applying AHP in data mining, the following subsequent steps have to be taken (Saaty, 2008):

- Problem definition and goal setting
- Concretize goal by low level, measurable criteria
- Prioritize criteria
- Measure alternatives along the criteria

In more traditional decision problems, AHP is used when the alternatives to a problem are already clear. In the case for DM, the alternatives are not known beforehand, but the patterns generated will be evaluated on basis of their utility. This results in an objective evaluation of patterns with subjective criteria.

In the DM framework, the criteria for the evaluation should be determined after the convergence phase, when the problem and goal has been made clear. The goal is then deduced to measurable criteria. These measurable criteria should specify ideas that the decision maker has about patterns that can be used for the benefit of the organization. In this way, the criteria are dependent of the perspective of the project but not biased towards a certain pattern. Subsequently, a decision maker needs to prioritize the criteria before measuring the results along the criteria.

Because AHP focuses on criteria selection, it offers a basis for a transparent, substantive discussion on preferences between participants. A decision maker is forced to specify his or her preference in smaller, measurable elements that can be questioned, making the discussion on preference more tangible. Because AHP forces criteria selection before data mining, it provides a solid basis for discussion and negotiation, as decision makers and data scientists have to break down their preference in measurable elements. When criteria have been determined, both parties can feel this as a quick win and become stimulated to be cooperative in the remaining of the process because there is a prospect of gain. Finally, AHP also provides the opportunity to approach utility measurement as a multi-criteria problem. This gives the opportunity to negotiate on two levels: the criteria to be used and the weight of the criteria.

Although AHP effectively incorporates the user preference in the data mining process, there are some undesired effects of this decision making method. The largest problem is that the project outcome becomes sensitive for the scale of the weights and the weights of the criteria (Kasperczyk & Knickel, 2004). A decision maker does not have to substantiate his choices for weight factors substantively, but

the criteria air compared pairwise. Using a different scale or a slight alteration of the weights may lead to totally different results. Therefore it is important that the sensitivity of the results is checked.

## 5.5 Perform attainability test

The convergence phase results in agreements on desired insights and the data mining suite. In that stage, the focus is on merging the perspectives and creating a common goal: the preparation of data mining. This phase is not limited by the possibilities of data mining. To relate the desired setup to the possibilities of DM, attainability tests should be performed. These tests should be regarded as a reality check on those agreements. Performing these checks may prevent future projects from significant delays later on in the process. The attainability test results in a decision to proceed with data mining, a change in the project goal, a change in the data mining setup or project termination.

Two factors that determine the attainability have been derived from the case study: a complexity check and a minimum observations check.

### 5.5.1 Complexity check

During the case study, a new problem came to light: the dataset was too complex to be totally carried out with the algorithm suitable for extracting the patterns. Despite the algorithm being state-of-the-art, the sequence patterns were too long for the algorithm to validly analyze subsequences (for more detail, see Chapter 4). This caused quite some delay in the project, since previous use of this algorithm did not report on any issues of this kind. In other cases, complex data may lead to unnecessary computation time or overfitting (Fahrnkopf, 2015).

In order to examine if the application of the data mining setup on the prepared dataset is compatible, some tests must be done to assess the complexity of the dataset. In the best form, this test presents the outcome in the form of a statistical measure. To the knowledge of the author, these measures however only exist for classification problems (Okun & Priisalu, 2007)(Duin & Pękalska, 2006). For other methods, the complexity decision largely depends on the professional judgement of the data scientist. On basis of the case study, the conclusion can be made that complexity of the data consists of the interdependencies in the data (e.g. the time-sequencing of click behavior) and the magnitude of a case (e.g. the amount of clicks in a web site visit). Future research could go into a more precise definition of complexity and the determination of adequate complexity measures.

If a dataset is too complex to apply the intended data mining setup on it, the project group can in general do two things: reduce the complexity of the data set or choose a different data mining suite. The complexity of the dataset can be reduced by bringing back the number of levels in the dataset. This can be done in the following ways (Fahrnkopf, 2015):

- Eliminating all records with rare levels.
- Automatic clustering
- Domain-expertise clustering

Any of the two clustering methods may lead to aggregated results. If this is incompatible with the predefined goal, then DM setup must be changed or the project can be terminated.

### 5.5.2 Minimum observations check

Complex problems often pave the way for the design of complex constructs, also in data mining. However, if the sample size is too small, the effect of that complex constructs may be overestimated (Duin & Pękalska, 2006). After selecting the data mining suite, the minimum amount of observations must be met to make a valid model. However, previous research states that it is very hard to define rules for determining the minimum sample size, so the practical limits often determine the sample size (Wielenga, 2007).

However, in some cases the minimum observations can be measured by the amounts of variables taken into account and the number of steps. To this end, the power of a method must be taken into account. Power measures originate from traditional statistics and their applications varies for different methods (Parry, 2010). For instance with sequence analyses, Vanvoorhis & Morgan (2007) used the rule of thumb that per variable taken into account, at least 10 cases need to be taken into account. This means that for the 1600 [redacted] pages, at least 16 000 cases needs to be taken into account.

## 5.6 Allow freedom of interpretation to deal with friction

One of the main improvements of the data mining process is to create convergence between the different stakeholders in the process. By creating a common perspective, shared goals can be determined and agreements can made about the following stages of the DM process.

However, as became clear in the case study and the literature review, there exists a friction between the utility of a pattern and the novelty of a pattern. In the case study, it became clear that the novelty of a pattern is not a hard constraint for decision makers to undertake action. On the other hand, data scientists regard novelty as an essential part of a DM pattern. At this point, there is also friction between the data mining perspective and the business perspective.

A solution for this problem is to allow freedom among stakeholders for a different interpretation of results. From an organizational perspective, the adhocracy must be seen as a Kantian organization. A Kantian organization encourages the development of multiple interpretations of data (Courtney, 2001). In practice, this means that it is okay that a decision maker's minimum requirements for interpretation are validity and utility, while data scientists value the novelty of a pattern as a constraint. If a pattern is valid and has high utility, it would seem logical that a decision maker would take action to pursue that utility value. The minimal context for action can then be defined as validity and utility. On the other hand, the decision makers should value the other view on the data by the data scientists.

From a theoretical point of view, this proposition is supported by Lewin & Nonaka's knowledge creation cycle, as described in 3.3.1 (Lewin & Nonaka, 1994b). According to them, the transfer of concrete knowledge to tacit knowledge is final when it leads something concrete: a system, concept or artifact. This internalization is a process of trial and error. Other stakeholders can then challenge the applicability of the artifact, which again leads to refinement of that artifact. This process is called crystallization. According to that methodology, a data scientist should not stop a decision maker from undertaking action on a valid and high-utility pattern. He is able to test the real applicability of the pattern after it has resulted into action. In this way, the artifact can be improved incrementally. Figure 38 shows how crystallization leads to a process of constant refinement of the artifact.

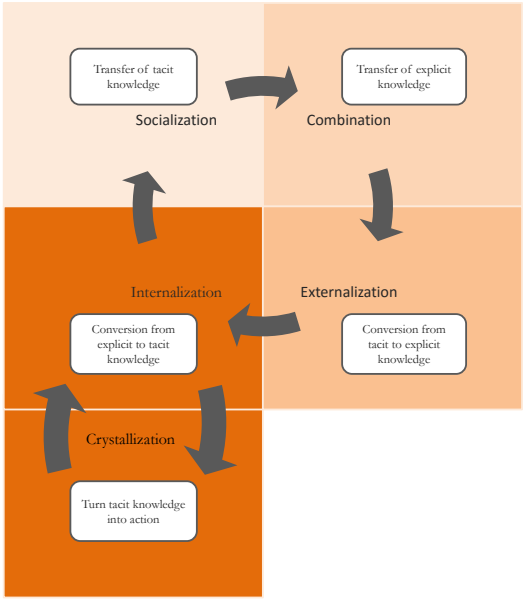



**Figure 38: Allowing freedom for interpretation leads to a process of constant refinement**

The process of crystallization is especially suitable for scrum teams, as they have a nature of trial-and-error. However, it is also applicable in more centralized organizations, as the mandate to undertake action often lies with the decision maker. However, after that action results in a system or artifact, a data scientist or other stakeholders can test and challenge the outcomes of that artifact. A decision maker may be more eager to listen to the concerns if that leads to a refined version of his primary action.

# 6 Framework design

The previous chapter presented the propositions that solve the three main problems of data mining: the rule quantity problem, the rule quality problem and the misalignment of data mining and decision making. Each of the individual solutions improves data mining, but it still remains open how these solutions should be integrated. In design science research, the aim is not only to design improvements but to design an artifact as a deliverable, all the improvements have been formalized in a framework: the Convergence-Data mining-Evaluation Framework (CDE-F).

## 6.1 The CDE-framework for data mining in organizations

The relevance and the rigor cycle propose a number of fundamental improvements of the existing data mining framework (e.g. the CRISP-DM framework). The propositions from the relevance and the rigor cycle can be merged into a single, improved framework for data mining in organizations. This framework is presented in Figure 39. On a high level, this framework identifies three main phases of a data mining project: the convergence phase, the data mining phase and the evaluation phase. This is why the framework carries the name *Convergence-Data mining-Evaluation-Framework (CDE-F)*. In the following, the operational process steps will be shortly described. These three phases have been substantially described in the previous. In the following, an operational description is presented for applying this framework in organizations. An overview of the operational CDE-F is presented in Figure 40.

### 6.1.1 Convergence phase

The convergence phase entails the improvements described in 5.4.1. The main goal of the convergence phase is to merge the data mining perspective and the business perspective in a common perspective. Consensus should be reached on the problem definitions and evaluation criteria to be used. After the convergence phase, all the necessary decisions to start mining the data have to be taken. To this end,

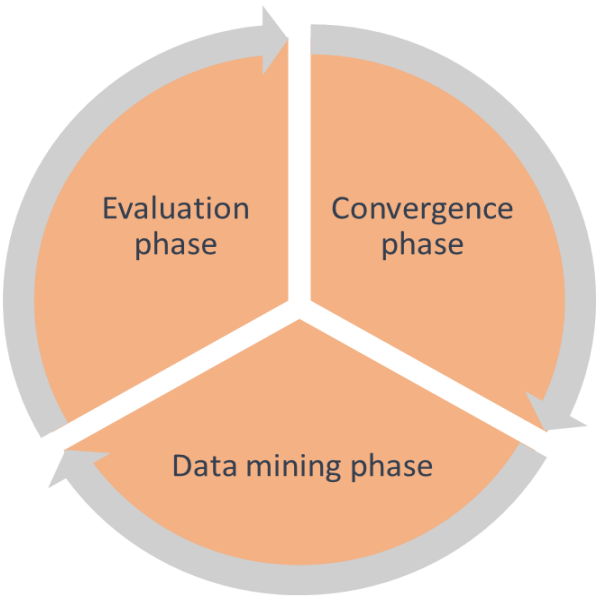



**Figure 39: The three phases of the new DM model**

this phase should be executed in 4 steps: problem formulation, goal determination, criteria determination and selecting the data mining suite.

### Determine goals

In the goal determination step, both the business goal and the data mining goal have to be described(Sharma & Osei-Bryson, 2009). The business goal is derived from the concise problem statement in the first step. The data mining goal describes the method in which the required insights for the business goal will be attained. Besides, this goal also specifies the size of the test group. The data mining goal is dependent on the available data, so a quick scan of the data has to be made. Besides, possible variables for novelty evaluation need to be investigated.

### Select data mining setup

The data mining goal and the business goals help to narrow down the search space for a suitable data mining suite. In this step, the expertise of the data scientist makes that the decision largely determined by what the data scientists thinks is the best DM-suite. In order to create transparency in the decision for the suite, data scientists need to substantiate the decision with the selection criteria as defined by Nisbet (2004):

- Ease-of-use
- Applicability
- Scalability
- Accuracy

In this way, decision makers gain more insight in why a decision has been made. Moreover, the selection criteria substantially structure the discussion, so that it is easier for the decision maker to participate. This step finally ends in a data mining method, technique and a specific algorithm.

### Determine criteria

Now that the problem, size of the data set and the data mining setup are determined, a substantive discussion should be held on how the patterns should be evaluated. This discussion is structured in three types of criteria that have to be defined:

- Validity measures, that specify that the patterns actually exist in real-life
- Utility measures, that analyze the value of the pattern for the organization
- Novelty measures, that check if there is a general regularity between high-utility and low-utility patterns.

In this step, the expertise of data scientists and decision makers needs to be combined. For e.g. validity evaluation, this means that a data scientist can advise the type of evaluation measure, while the decision maker determines the validity threshold. Utility measures are defined by deducting the business goal to lower-level, measurable criteria. Besides that, the criteria are compared pairwise to create weight factors. These are the first steps of the Analytical Hierarchy Process (Saaty, 1990, 1999, 2008) Novelty evaluation is prepared by determining variables that could be of interest to explain the difference in utility of patterns. These types of variables have been investigated in the goal determination step, and are now decided.

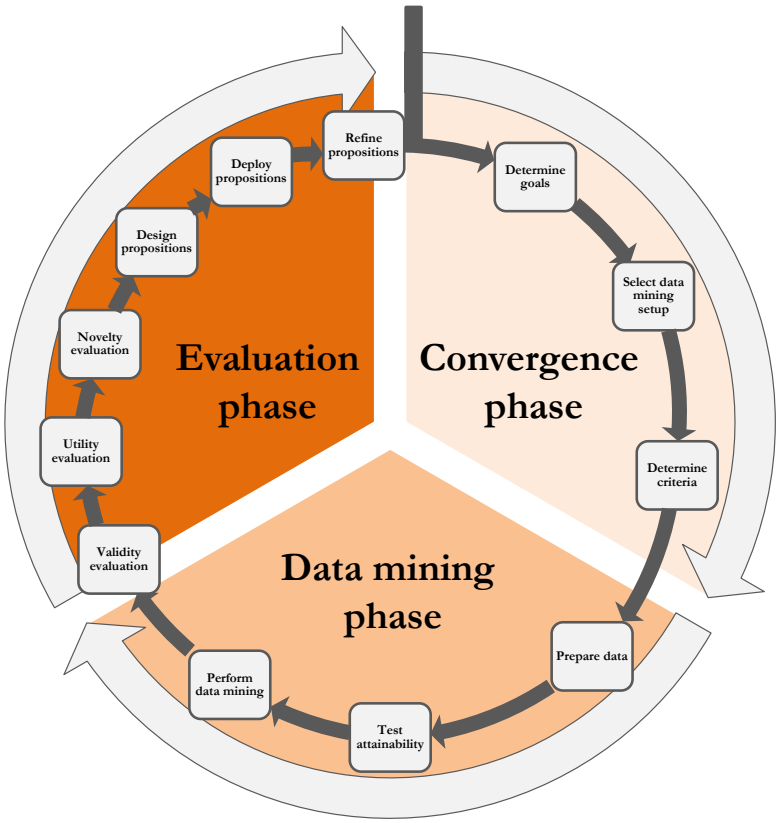



**Figure 40: CDE-F with process steps**

These three sets of criteria result in an evaluation framework for the patterns. To optimize the application for the business, the role of each evaluation step (e.g. hard constraint, pruning mechanism, ranking mechanism, additional evaluation) has to be determined. Validity should be treated as a hard constraint, since it is unlikely that a business decision will be based on an invalid pattern.

### 6.1.2 Data mining phase

The problem definition is used in the data mining phase to explore the data and determine the data mining setup for the analysis. The data mining is the phase that resembles most of traditional data mining, where data is transferred into patterns.

#### Prepare data

In this step, the data elicited in the convergence phase is isolated, combined and prepared for the analysis. For data preparation, the goal is to transform the data in such a way that it can be handled by the DM setup in an efficient way. This means that data that is irrelevant for the analysis is exempted from the data set. Other transformation steps that can be taken are clustering, classification or outlier removal. Data preparation has been more extensively described in data warehousing literature, such as Devlin & Cote, Kimball & Ross and Luj & Trujillo (1996; 2011; 2005).

#### 1.1.1.1 Test attainability

When the data is prepared, the data can be explored to check if the dataset fulfills the desired size and contains the desired variables. Moreover, some checks can be done to see if the desired insight could be validly attained from the selected data. If the data set does not pass the attainability test, four options arise:

- Change the DM-suite
- Prepare the data differently (e.g. aggregate, cluster, outlier removal)
- Manage expectations
- Terminate the project

#### Perform data mining

In this step, the data is applied in the data mining setup to create patterns and possibly analyze them even further. The resulting patterns are then transferred to the evaluation phase.

### 6.1.3 Evaluation phase

In this phase, results are evaluated according to the triple evaluation approach. These steps are decoupled for the sake of simplicity and to position novelty evaluation as a step for additive insights.

#### Validity evaluation

Firstly, the patterns need to be controlled on their statistical validity, as is determined in the convergence phase. Many DM suites offer possibilities to easily specify validity constraints in the algorithm. If this is the case, the validity evaluation is performed simultaneously with the actual data mining. If more complex validity measures are used in the project, a separate validity evaluation might be needed.

Since the validity evaluation serves as a hard constraint for DM patterns, the outcome of this step is a pruned set of statistically valid patterns.

#### Utility evaluation

Consequently, the set of statistically valid patterns are analyzed on their utility value. This is according to the final steps of the AHP (Saaty, 1990, 1999, 2008) To this end, the criteria and weight factors from the criteria determination step are used. The decision maker can choose to manually select the patterns with high utility, or he can specify a minimum utility threshold for patterns. To check the results for the influence of the weight factors, a sensitivity analysis is conducted. If the patterns prove to be sensitive for those factors, a decision maker can choose to go for the patterns with higher utility or for the most robust patterns.

The utility evaluation phase ends with a set of patterns that is deemed as valuable by the decision maker.

#### Novelty evaluation

For the novelty evaluation, the patterns from the utility evaluation are divided in two groups: high-utility patterns and low utility patterns. High-utility patterns are results that the decision maker values enough to design propositions for the organizations, low-utility patterns are results that are deemed to have too little potential for the organization.

With the set of context variables that has been determined in the convergence phase, an exploratory factor analysis with multiple groups is performed to check whether the high-utility patterns show a regularity that is not found in the low-utility patterns. If new insights follow from this exercise, they may substantiate new propositions.

### Design propositions

In the final step of data mining, the results are compared to the initial problem statement and business goal. This is where the actual interpretation of the results take place. The results from the utility and novelty evaluation steps can be translated into proposition that are beneficial for the organization. Besides, as DM is an iterative process, the new insight may give rise for a new iteration for the problem. In this process step, the interpretation of patterns may differ among stakeholders. The CDE-F offers freedom for stakeholders to differ in their interpretation. For instance, a decision maker may turn a valid result with a high utility into action. He then should be able to do so.

### Deploy propositions

The freedom of interpretation of results is offered under the restriction that the interpreted data must lead to something tangible: a system or other artifact. In order to do so, the decision maker should explicitly deploy the insights gained in the data mining research in the organization. In this way, the knowledge is eventually turned into action. In this way, this step secures that the added value of data mining is expressed in the organizations.

### Refine propositions

The freedom that participants have in the interpretation of patterns implies that stakeholders do not need to converge to a shared view on how to deploy the results in the organization. A discrepancy in the interpretation of patterns should not stop a stakeholder from turning his interpretation into action, as long as that insights leads to an artifact. This enables a discussion on the applicability of the artifact. The actual added value of the insight in the organization can be evaluated, and this offers a platform for stakeholders to discuss their differences in interpretation. This discussion can then lead to a constant refinement of the resulting artifact.

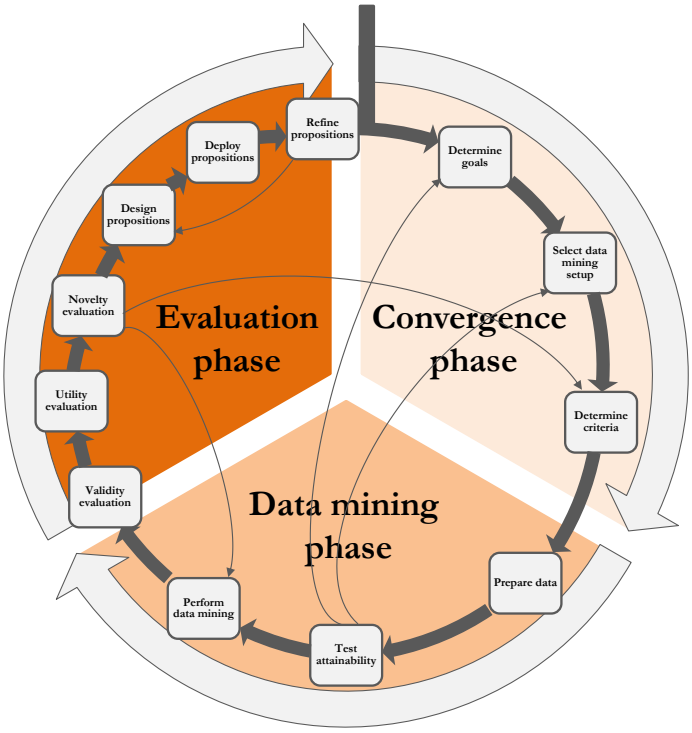



**Figure 41: Loops in the CDE-F**

### 6.1.4 Loops in the CDE-F

As Figure 40 indicates, the CDE-F is a constant loop, meaning that a data mining project is carried out in multiple iterations. However, insights during the data mining project may cause the need to immediately revert to an earlier step in the process. These loops in the framework are presented in Figure 41, and briefly described below.

The first three loops occur after the attainability test. The attainability test can be regarded as a go/no go moment to continue with the data mining project. If a project is unattainable in the current form, either the goal is unattainable (e.g. in the case study, insights on a too detailed level were desired), or the data mining setup needs to be changed (like in the case study, the feasibility of a Markov model was investigated) or the data needs to be prepared differently (like in the case study was the case for the aggregation of pages and removing the loops in click patterns).

Another loop can be found in the actual data mining. After evaluating the patterns, a new iteration may be started with slightly different data. The adjustments of this new setup are not rigorous, so there is no need to return to the convergence phase. This is the iterative approach of data mining from the technical perspective, as is described by Fayyad et al. (1996).

If the novelty evaluation provides a possible new insight, that insight might be worth the trouble to directly incorporate as a pattern evaluation criterion. Therefore, the novelty evaluation may cause the project to jump back to the criteria determination step.

After the data mining results are turned into action, the system is constantly refined because other stakeholders may challenge the applicability of the system. This is a constant process of trying to internalize knowledge. In order to refine the system, the convergence and data mining phase can be skipped. In this way, a direct link exists between the refinement of proposition and the design of a proposition.

## 6.2 Tailoring of the framework to the organizational context

As 3.3.5 points out, the institutional context of the organization has an influence on the way that data mining should be deployed in the organization. For instance, in a centralized organization the goal and the scope of the project are established quite easily. On the other hand, if the process is susceptible for politicization and strategic behavior, goals will be extensively discusses and negotiated. In order to guide the adoption of the CDE-F in different types of organizations, three archetypes of the organizational context have been determined: a *centralized organization,* a *scrum team* and a *political environment.* These archetypes are used to discuss how the framework can be tailored to the organization. Moreover, it indicates the important task divisions between three possible roles in the framework: a data scientist, decision maker and process manager. For the process, four important elements of the CDE-F have been selected: the *project scope, data mining, evaluation* and *interpretation.*

**Table 12: Task division for CDE-F application in a centralized organization**

| Centralized organization | | | |
|---|---|---|---|
| | Data scientist | Decision maker | Process manager |
| Project scope | Challenge scope of decision maker<br><br>Translate business scope to project setup<br><br>Novelty evaluation perspective | Determine business goals<br><br>Planning and resources | n.a. |
| Data mining phase | Select, transform and explore data<br><br>Attainability test<br><br>Mine patterns | Monitor progress<br><br>Change scope if needed | n.a. |
| Evaluation | Validity criteria<br><br>Translate utility criteria to utility function<br><br>Novelty evaluation | Determine utility criteria | n.a. |
| Interpretation | Structure results | Investigate results<br><br>Evaluate new insights<br><br>Decision on new iteration or follow-up<br><br>Design propositions | n.a. |

### 6.2.1 Centralized organization

A centralized organization has similarities with a machine bureaucracy. A centralized organization has a high-level, strong decision maker. While this may be in a political environment, the political aspects are outside the project scope. The data scientists represent the technostructure of the organization and have little interaction with business stakeholders. The organization thrives on explicit knowledge.

In a centralized environment, the interactivity of the CDE-F is limited. Therefore, a process manager is not required. Projects start from a business desire, and the decision maker determines the scope of the project. The data scientist should not boldly accept this scope, but should challenge the scope of the business. He does this from the data mining perspective, in order to secure that the business goals are translatable to a project. Furthermore, this challenge should lead to a broader scope, with the incorporation of novelty evaluation. The data scientist translates this the scope to a project setup, while the decision maker makes a planning and makes resources available.

The data mining phase is the responsibility of the data scientist. The decision maker keeps an eye on the project (e.g. through steering committee meetings), and changes the scope of the project if needed. The evaluation phase is a lot more interactive. The decision maker should determine the utility criteria, and the data scientist translates this to a utility function. This requires collaboration in the convergence

phase. Eventually, the data scientists structures the results (high utilty results and insights from the novelty evaluation). On the basis of these results, the decision makers designs propositions and makes a decision to extend or terminate the project.

**Table 13: Task division for CDE-F application in a scrum team**

| Scrum team | | | |
|---|---|---|---|
| | Data scientist | Decision maker | Process manager |
| Project scope | Agree on project scope<br><br>Translate business scope to project setup<br><br>Agree on data mining setup | Agree on project scope<br><br>Agree on business scope<br><br>Agree on data mining setup | Select stakeholders<br><br>Convergence between stakeholders<br><br>Planning and resources |
| Data mining | Select, transform and explore data<br><br>Attainability test<br><br>Possible scope changes<br><br>Mine patterns | Help with data selection<br><br>Possible scope changes | Monitor progress<br><br>Manage external influences<br><br>Planning and resources |
| Evaluation | Determine validity criteria<br><br>Facilitate discussion on utility criteria<br><br>Translate utility criteria to utility function<br><br>Novelty evaluation | Determine utility criteria<br><br>Novelty evaluation | Monitor progress<br><br>Manage external influences |
| Interpretation | Structure results<br><br>Investigate results<br><br>Evaluate new insights | Investigate results<br><br>Evaluate new insights<br><br>Decision on new iteration or follow-up<br><br>Design propositions | Monitor progress<br><br>Manage external influences |

### 6.2.2 Scrum team

A scrum team is a flat organization with less explicit knowledge and much freedom for professionalism. In a scrum team, there is more overlap between the responsibilities of the participants in the project. The project is carried out in an interactive setting and decisions are made through convergence of the participants. Although the responsibilities are carried collectively, every participant can be utilized on his or her expertise. In this way, a task division is established. Depending on the number of stakeholders involved and their interrelations, the choice can be made to assign a separate process manager. In most cases, the product owner of the scrum team can fulfil this role. In a less

strategic environment, the product owner can be regarded as a decision maker, but still has to be concerned with the project planning, as well as the substance and progress.

The preliminary phases of data mining in a scrum team are very intensive. The multiple goals, perspectives and expertises need to be aligned to one shared project scope. If there exists any political tension, the process manager should secure substance and progress by determining the rules of the process (a selection of, or all process rules as described in 5.4). Through knowledge sharing, business stakeholders become more data-savvy and data scientists gain insight in the desired outcome of the project. Through mutual understanding, the gap between the business goal and the data mining setup becomes smaller. While the data mining phase is still dominated by the expertise of the data scientist, the decision makers and a possible process manager stay up to data through frequent interaction and can adjust the scope where needed. Because of this commitment, a process manager is able to inform external parties (that are e.g. needed for deployment of results) and can make a better informed decision to change the resources and their allocation. With respect to the evaluation, the data scientists facilitates the discussion on utility criteria. This guides decision makers in the translation of their business goals to utility criteria. Eventually, the data scientists are responsible for the quality of the utility function. Finally, the resulting patterns are collectively interpreted. The data scientist focuses more on improvements of the analysis, while the decision makers focus on the design of propositions.

### 6.2.3 Political environment

In a political environment, the presence of a process manager is very important. The data mining projects is carried out with stakeholders that have conflicting interests or a strategic agenda. If this is not managed correctly, the process may become sluggish. Moreover, the project can change due to strategic behaviour. This led to a beneficial outcome for one party, but may lead to unintended outcomes for the organization.

In order to prevent unintended outcomes, the process manager plays an important role in the initial phase of the project. Unlike the product owner, the process manager is objective and does not have influence on the final result of the project. However, his objectives are to design a transparent process, decouple the process steps and indicate the minimum consensus needed to proceed in the process (De Bruijn et al., 2010). Substantive discussions are left to the data scientists and decision makers, who have to build consensus on the project goal and the setup. Therefore, the roles of the data scientist and the decision maker do not differ significantly from their role in a scrum team. Of course, the exact role of these parties is also subject to the outcome of the discussion on the division of tasks.

The process manager designs and manages a process in which both data scientist and decision makers can have substantive discussions. This makes him both a process manager as a process architect. In the preliminary phases of the project, the process manager actively builds consensus by facilitating discussions on the project scope. He ensures that the discussion on evaluation criteria is decoupled, so that data scientist can have more mandate for the validity criteria and the business has more influence on the utility function. He oversees that the determination of the project scope and evaluation criteria is transparent, e.g. by applying the AHP to define the subjective utility assumptions (see 5.2). This transparency builds trust and consensus. In case no consensus can be reached, a process manager supervises the exit options of the process.

**Table 14: Task division in a political environment**

| Political environment | | | |
|---|---|---|---|
| | Data scientist | Decision maker | Process manager |
| Project scope | Agree on project scope<br>Agree on project goal<br>Translate business scope to project setup<br>Agree on data mining setup | Agree on project scope<br>Agree on project goal<br>Agree on business scope<br>Agree on data mining setup | Select stakeholders<br>Agreements on process steps<br>Enable discussion on role division<br>Determine exit options<br>Determine decision moments |
| Data mining | Select, transform and explore data<br>Attainability test<br>Possible scope changes<br>Mine patterns | Help with data selection<br>Possible scope changes | Monitor progress<br>Control process on process agreements |
| Evaluation | Determine validity criteria<br>Facilitate discussion on utility criteria<br>Translate utility criteria to utility function<br>Novelty evaluation | Determine utility criteria<br>Novelty evaluation | Monitor progress<br>Control process on process agreements<br>Facilitate transparency in utility criteria<br>Facilitate novelty evaluation discussion |
| Interpretation | Structure results<br>Investigate results<br>Evaluate new insights | Investigate results<br>Evaluate new insights<br>Decision on new iteration or follow-up<br>Design propositions | Monitor progress<br>Control process on process agreements |

## 6.3 Securing the CDE-F in organizations

In order to embed data mining successfully in the organization, the CDE-F has been designed. In order to embed the CDE-F in organizations, some contextual factors need be taken into account. After evaluating the CDE-F with the team Prevent & Shift, three prerequisites for embedding came to light: lower the information uncertainty, process management competencies and the minimum level of substantiation for action.

### 6.3.1 Information uncertainty

The literature review and the case study showed that decision makers have trouble dealing with the inductive nature of data mining. Even though a problem exists, data mining is more uncertain about the outcomes than decision makers usually tend to allow in their decision making processes. Because of this uncertainty, a decision maker may have his doubts about applying data mining.

In order to decrease the uncertainty of data mining outcomes during the process, project teams wanting to carry out data mining should decrease the uncertainty of data mining outcomes by gathering context information. This information focuses on important decision making factors, such as the magnitude of the problem and the potential benefit. The meso-level insight of the case study, that lead to the design of the integrated channel performance dashboard, is a way of reducing the information uncertainty. Since the dashboard will create an overview on the call flow between Call and Internet, the potential benefit of the micro-analysis can be more easily determined When this information is gathered, a decision maker can consider if the potential gain outweighs the remaining uncertainty. Besides, a more substantiated decision can be made on the focus of the project.

Applying this on the case study, the information gathered to lower the information uncertainty led to an integrated channel performance dashboard between the Call department and Internet department. Initiating the data mining project, the project team had two main concerns:

- The uncertainty of how the outcomes would look like
- The potential gain of the solution

Since the first problem is inherent two data mining, this cannot be prevented. However, the second problem originated from the separated performance evaluations of both the Call and Internet department. Before this project, no information was available on how many calls originate from internet activity. In order to lower the border to undertake a complex task as data mining, insight was needed on the magnitude of this flow on an individual process level. The integrated channel performance dashboard gave this insight. With this dashboard, the project team could create enthusiasm among other stakeholders. This significantly lowered the border for decision makers to engage in the data mining process. Currently, the dashboard is constructed on a regular basis.

### 6.3.2 Competencies of process manager

When data mining is performed in an organization, multiple parties with different interest have to be aligned. This does not only limit aligning the group of decision makers and the group of data scientists. Additionally, intra-group alignment may be needed as well. For the case at [redacted] bank, decision makers from both Call and Internet had to come to agreements before performing the actual data mining.

Because working with the CDE-F is not a regular activity for the parties involved in the process, one should not expect that the process will be carried out according to the CDE-F without any effort. The effort it takes to align the stakeholders and come to a unified project scope is dependent on the organizational context of the project (see 3.3.5 and 6.2). Especially in a strategic environment, the presence of a process manager is desired. The goal of the process manager should be to unify all stakeholders, while providing a good structure for goal alignment and stimulate decision making. In this way, two important factors of process management are secured: substance and progress.

### 6.3.3 Minimum level of substantiation for action

A limitation of the CDE-F is that it does not provide guidance in the deployment of the results of data mining (this is further elaborated in 9.2). However, as described in5.6, some freedom must be allowed for parties to interpret data in their own way. With respect to the deployment of insights, the minimal context to judge a pattern differs between the data scientists and decision makers. As data scientists truly value a pattern when all three requirements are met, a decision maker can substantiate actions for patterns that are valid and have a high utility.

In order to cope with this difference, the deployment of results from the CDE-F must be approached with a certain sense of realism and pragmatism. Since the overarching organizational goal of both the technostructure and the operating core is to create value for the company, patterns that have a high utility and are valid should not be neglected just because they are not new.

# 7 Framework validation

Data mining is an overarching term for a large toolbox of very different tools that can be used to solve complex projects with advanced analytics. Data mining projects can thus be executed in very different ways for very different purposes. As a hammer is suitable for hitting nails in the wall and unfit to saw planks, the data mining approach needs to be tailored to the organization.

One of the non-functional requirements of the data mining framework is the generalizability of the framework. To stay in the metaphor of the toolbox: it must be applicable in chars where a hammer has to be used, as well as for chars that require a saw. This framework is co-established through insights of a single case study. Ideally, these insights are then tested in other case studies to test the generalizability of the framework. Due to time- and resource constraints, this was not possible. In order to see whether these insights are also applicable in other cases, the CDE-Framework is validated in a different way. To test the generalizability, the framework is applied on recently finished projects or pending projects at [redacted] Furthermore, an expert validation is conducted through an interview with Scott Cunningham, associate professor at Delft University of Technology and uncredited involved in the development of the CRISP-DM. This expert interview focused on the total acceptability of the CDE-F, not only on the generalizability.

## 7.1 Generalizability of the CDE-Framework

Ideally, the generalizability of the CDE-F would be tested in what Hevner et al. call a field study (2004). Multiple case studies of varying data mining setups would then be tested in multiple organizational contexts. However, this proved to be not possible in the given timeframe and with the available resources. To fully test the generalizability, all variants of data mining setups need to be tested in various organizational contexts. With the current validation setup, the hard conclusion that the framework is generalizable cannot be drawn, but it does shine a light on the most important generalizability issues of this framework.

In order to check the generalizability of the framework, interviews were held with business stakeholders with checks from the data scientists. In this interviews, the main focus on the generalizability of the framework was on the key elements of the framework: the incorporation of stakeholders in the convergence phase and the implementation of a triple evaluation approach. The extent to which the framework can be blueprinted on these projects gives an indication of the generalizability of this framework.

The projects selected for this validation step differ in the organizational context and data mining suite. The goal, DM-method and the way in which the CDE-F can be blueprinted on that project are presented in Table 15. The most important elements will be discussed below. Complete descriptions of the project, include answers on additional questions are presented in Appendix C.

### 7.1.1 Clustering

The goal of the CandY project was to [redacted]

**Table 15: Overview of the projects used as a thought-experiment for the validation of the CDE-F**

| | CandY | [redacted] | Big Data @ [redacted] |
|---|---|---|---|
| Goal | [redacted] | [redacted] | [redacted] |
| Method | Clustering | Anomaly detection | Random forest classification |
| Organizational context | More centralized decision making | Participatory setting | More centralized decision making |
| Convergence | [redacted] | [redacted] | [redacted] |
| Validity evaluation | [redacted] | [redacted] | [redacted] |
| Utility evaluation | [redacted] | | |
| Novelty evaluation | [redacted] | [redacted] | [redacted] |

[redacted] After a first unsuccessful attempt (the project delivered insights but no value for the operation), the project has been carried out again with the incorporation of business stakeholders. Just as the case study, the second attempt was a participatory setting. However, as a steering committee took charge in this project, the goal of the project was delivered top-down to the participants: [redacted] In order to do so, the stakeholders indicated two important factors to discover: [redacted] Relating this to the CDE-F, the [redacted] secures the actionability and [redacted] secures the value of a result. In this way, these two element represent the utility of patterns. After these insights, propositions were formed that in the end [redacted] As novelty evaluation, the project team (without the steering committee) tried to incorporate a longitudinal factor, in order to check if some clusters had a regularity in there transfers. Furthermore, this allowed the team to evaluate the likelihood that customers move to another cluster. This new perspective can be regarded as novelty evaluation.

This project had a clear focus on actionable and valuable results, just as the CDE-F framework. It is therefore that the framework could be blueprinted almost entirely on this project, although it concerned a different data mining method in a different organizational context. The different organization context proved to

### 7.1.2 Anomaly detection

The [redacted] case was a very difficult case. The goal was to [redacted] r. After that, anomaly detection could be performed on the whole dataset, and anomalies could be evaluated with the two specified rules. There were some difficulties that made it impossible in the current setting to come up with a working model.

Firstly, for the business there is an enormous gap between the recognition of a potential [redacted] case and taking the actual measures [redacted] After a potential [redacted] case comes up, commercial banking employees have to [redacted]. This costs a lot of time, effort and thus costs. Moreover, [redacted] A potential benefit that this project could have had from the CDE-F is to evaluate the cases, that have been detected as possible [redacted], on the magnitude of the [redacted] This amount could be compared to the average cost and success rate of the process of [redacted] With this information, a commercial banking employee could make an informed decision whether or not is feasible to start [redacted] This is a good example of how utility evaluation could help to focus more on the results rather than the input.

As stated in the previous paragraph, [redacted] e is time- and cost expensive. Moreover, the success rate of such cases is relatively low. Therefore, there were not too many [redacted] cases that could serve as the input to specify a general rule [redacted]. It is therefore that the learnings of the project could not come to fruition in the form of an automatic fraud detection mechanism. While the eventual learnings of the project should not be underestimated, the initial project goal was unattainable. If the CDE-F is blueprinted on this project, the step of testing the attainability of the project would likely have indicated this problem earlier on in the process. When the project team has this insight in the beginning of the project, they could have managed expectations, adjusted the project or terminated the project earlier on.

### 7.1.3 Classification

The goal of the 'Big Data @ [redacted]' project was t [redacted]. This was done by analyzing a set of verified [redacted] rospects, and classifying every regular customer either in or out the set of [redacted] prospects. The eventual deliverable should be a model that regularly makes these classifications and learns from the result of the outbound calls that are made. This model had a large set of client characteristics as input variables and a set of valid variables with their coefficients as output. [redacted]

An important notion that has to be made here, is that this project is a form of machine learning. The critical difference between machine learning projects and supervised learning, the model outcomes are not evaluated by human intervention, but the model evolves through the input of previous model outcomes. One might say that machine learning is the purest form of data mining and the working of a machine learning model is violated by human activity other than data selection. To clarify this, one can compare the evolution of a machine learning model with normal evolution. In nature, natural

[5] [redacted]

selection takes place and offers the strongest breeds the most chances on survival and reproduction. If humans intervene in nature, the result of evolution will also be different.[6]

Although this human intervention not necessarily affects the model outcome negatively, a possible combination of the CDE-F will cause some friction. For the case 'Big Data @ [redacted]a' , this was expressed in difficulties with a possible adoption of utility evaluation. The utility evaluation in the CDE-F is the step with the most subjective information, which has a large influence on the resulting patterns. In machine learning, the main conception is that patterns should only be evaluated on statistical significance. The constant updates of the model output will eventually lead to a situation where the model output will have a high utility for the organization. Using business knowledge in a utility evaluation would be too static for machine learning.

However, business knowledge can be used in machine learning projects. All the process steps before the data mining phase can benefit from the knowledge and interests of participating stakeholders. In the 'Big Data @ [redacted] -project, this was mainly focused on determining the input data. In this way, the information that stakeholders find relevant (and may wanted to use as utility evaluation) can serve as input data for machine learning.

The friction between machine learning and utility evaluation is a newly discovered one and deserves more attention. Future research may focus on the integration of machine learning and utility evaluation. Concepts of that research could be that a definition is given for input data and evaluation criteria. When does something serve as input and when does it serve as evaluation?

## 7.2 Expert validation

To broaden the scope of the validation from the generalizability to its overall acceptability in both organizations and science, an expert validation has been conducted with Scott Cunningham, associate professor at the TU Delft. He published both in the field of decision support systems as data mining.

The motive for the research is fully endorsed by Scott Cunningham. While possible solutions have been designed for the rule quantity and –quality problem, he judges the misalignment problem an interface problem between data mining and decision making, and this received too little academic attention. In this light, this research certainly delivered interesting insights.

Overall, Scott Cunningham rated the CDE-F as a generalizable framework that will be accepted both from the perspective of data mining as decision making. Approaching data mining from multiple perspectives than just a technical one certainly has its added value, as it offers solutions for the misalignment problem. Moreover, it may just have impact on a fundamentally unsolved problem: the model selection problem.

With respect to machine learning, minor problems can be expected with the incorporation of the user perspective (as shown in the previous paragraph). However, these problems can be solved and do not affect the generalizability of the approach. Possibly Bayesian learning can add something to this.

[6] Take for instance the peppered moth evolution during London's industrialization (for more info, visit https://goo.gl/CaqFIm)

With regard to the novelty evaluation, Mr. Cunningham pointed out that novelty can also be used to evaluate the validity of the model. To make a valid model, it is inevitable to make assumptions. Novelty can evaluate the validity of the model on the number of assumptions is needs: a cost-benefit insight. How valid is the model, but at the cost of how many assumptions.

Positioning this research in the light of current developments in data mining, an interesting contrast becomes clear. Although it is still a vague concept and not scientifically treated, a new idea exists that the data scientist should develop soft skills to translate between data mining and domain knowledge ('Storytelling'). It demands more capabilities of the data scientist. While the CDE-F offers an interactive solution to the misalignment problem, the 'data scientist as storyteller' keeps the unilateral view on the misalignment problem. In essence, the idea is that besides advanced analytics capacities, a data scientist should develop enough understanding of the organization. Consequently, a data scientist should construct an effective communication method to secure that the results from data mining projects are used in the organization. This direction has the potential to be a simpler solution, as there is less participation required of the stakeholders. This should make the process robust for politicization and prevents sluggish decision making. On the other hand, the CDE-F secures ownership of the results through active stakeholder participation, and offers more concrete improvements of the data mining process.

Possibly, the CDE-F could be improved in a way that in projects where the chance on politicization exists, the data scientist gets more influence and decision power. This will lead to a simpler data mining project. On the other hand, Storytelling can adopt lessons from the CDE-F to ensure meaningful and valuable results that are supported by the stakeholders.

# 8 Reflection

This research has been carried out in order to receive the degree Master of Science at Delft University of Technology. In this paragraph, I will reflect on the course of this graduation project. This is thus a personal reflection, to inform possible future graduation students or [redacted] interns on my findings and lessons. I will reflect on two following parts of the graduation: the research outcomes and the research process.

## 8.1 Research outcomes

The most important findings on obtaining the research outcomes is that organizational data is hard to acquire, understand and analyze. This all has its effect on the outcome of the research. I will discuss this in the following.

Data is one of the most precarious assets of an organization. Especially when this data concerns customers of that organization. This is why only a limited group of people has access to the data. While in this research data from different data warehouses had to be combined and merged, only people with the highest authorization level of data access could perform this task. For the progress of my research, these people were critical actors. However, they were also very scarce, and they handle requests according to prioritization, not through first come, first serve. At the time, it was no option to acquire the authorization myself because of my role as an intern. This made it hard for me to acquire the data that was needed for the analysis. Initially, the goal was to perform a second case study to validate the framework. However, because of the time it costs to acquire data, the decision has been made to perform thought experiments instead of case studies. If future graduation projects appear to have the same situation, I would recommend the following: acquire the authorization needed to remove dependencies [redacted] currently offers more possibilities to acquire an authorization), or request data really early, in a very detailed scope and plan for contingency.

While in academic research projects, data is often complete and comprehensive, organizational data often is not. The data that you are going to use, is not collected for the sole purpose that you want to use it. [redacted] Nevertheless, it is of utmost importance to really get acquainted with the data you are going to use, and to get acquainted with the way in which the data is filled. It is advisable to quickly build relations with experts on the contents on databases, that can elaborate on what a variable actually means. This understanding will eventually lead to better project outcomes.

The fact that an [redacted] customer has a lot of possibilities on the [redacted] web site, is both a blessing and a curse for the project. [redacted]. Moreover, some

extra possibilities of the algorithm, such as pattern clustering, was not available after the complexity reduction. If this pattern clustering could have been done, an outranking method (as described in 5.2.2) would become more feasible. A lesson learned from this is that it is good to be very ambitious on the case study you want to perform, but also be realistic and weigh your ambition to the time window and the organizational complexity. Critically examine the literature not only for theoretical notions of the method, but also for successful applications.

I am very pleased that the case study did not only lead to the desired insights, but also delivered Call and Internet possibilities [redacted]. The case study was focused on the improvement of the web site and not o [redacted] That this case study delivered knowledge that goes beyond the boundaries of the case study is very pleasing. This should motivate [redacted] employees and data scientist to explore more possibilities of data mining and to look beyond the boundaries of a project scope. Moreover, I would recommend every graduate or intern to actively promote your research deliverables throughout the organization, you may never know what departments can benefit from them.

## 8.2 Research process

To evaluate the research process, I will reflect the three main parts of this research: research methodology, literature research and the case study.

### 8.2.1 Research methodology

For this research, I used the design science research methodology developed by Hevner and Hevner et al. (Hevner, 2007; 2004). This methodology forces you to explicitly take into account the goal of the literature research and the case study. Since both the case study and the literature research has been carried out quite detailed, the methodology worked well to keep once in a while take a step back and reflect on the goal of certain efforts. For a research that was as time consuming as this, it was very helpful to structure.

By taking the goal of the research explicitly into account, the researcher is helped to the keep the long term planning in mind. Graduation projects are often the largest project a student has done so far. Therefore, the end of a research may often not seem near. The design science framework forces a student to think about the end-state of the research, a validated artefact. Having the final state in mind helped me logically consider what research steps were needed. This makes a graduation project manageable. Besides that, it also serves as a stop signal for the ambitious student. My ambitions regarding the quality of my deliverable have been very high, and I am not easily satisfied. Because I had a clear view of the end state of my project right along, it helped me to know when to stop (although the tendency to do more still exists).

Finally, the combination of the design science framework with requirements analysis helped me to logically structure my findings from the case study and literature research. Being an inquisitive student, I have the tendency to dive deeper and deeper into the details and constantly find out more. Going back from that level of detail to the design science framework is a large aggregation step that was not easy for me to make. By the addition of the requirements analysis, I introduced a mid-level aggregation step between the design science framework and the detailed analysis. This helped me to relate my insights gained to the eventual framework.

### 8.2.2 Literature research

As already mentioned, I am an inquisitive student who likes to know all there is to know about certain subjects. As a consequence of that, I have performed an extensive literature review on data mining. This was also necessary, since this was my first encounter with data mining.

Although SEPAM is about analyzing a certain topic from multiple perspectives, a SEPAM graduate student should not try to incorporate all scientific worlds in his or her analysis. Being enthusiastic about all the discoveries I did in the field of data mining, I dove in all kinds of literature. At some point, I found myself reading psychological literature instead to substantiate problems between deduction and induction. This for me was a realization moment that I had to canalize my enthusiasm and should critically consider the added value of everything that I was doing.

The extensiveness of the case study made it quite hard to combine all the information in one chapter. This affects the readability of the chapter, especially if the research topic itself is already complex. Eventually, you as an academic should make a trade-off between the amount of information you want to communicate and the clarity of your line of argumentation. I would advise every graduate student to keep the reader of your report in mind.

### 8.2.3 Case study

The case study has been conducted following the rules of participatory action research. This fits the design science approach really well. PAR actually takes place in the real world, with real stakeholders on real problems. This secures the relevance of the project. Moreover, the threshold to come up with new ideas is lowered in PAR, so new ideas can be tested quickly. This is a very suitable way to stimulate innovations in the data mining process. The fact that [redacted] have indicated benefits on all levels of research is a huge boost for the quality of the research and hopefully the further developments of the framework. This strengthens me in the opinion that artifacts that have been co-developed by the academic world and the corporate world are robust and more likely to have an impact in the real world. And when things have an impact in the real world, this attracts academic attention.

Besides that it is a blessing, the freedom PAR offers is also a curse. PAR is a relatively unstructured method of performing a case study. Unlike interviews, insights gained in PAR are not well documented. In this way, it sometimes makes it hard to attribute an insight to a certain party or person. For the insights to be really supported, they must be more formalized in other studies.

My personal opinion is that I did not choose the easy way: solving a complex problem with a complex technology, and analyzing that technology from a totally different perspective than currently has been done. One of the main motivations for me was that I got to perform a case study with state-of-the-art technology to solve a difficult problem with little initial knowledge. I have really experienced it as a discovery process, and every insight motivated me to go deeper and find more.

### 8.2.4 Planning

Although I fancy a challenge and have a high level of ambition, the combination of a complex problem and new technology with a new perspective did also cause problems and delays. This research started in October 2014 and ended in August 2015. It is clear that the duration of the project exceeded the planning (although I have taken some time off, especially in summer). This is partly because of me, since I have clearly communicated that I rate quality over planning. However, the dependency on a

data scientist's access to the datawarehouse and his partial involvement in the project, also caused me some delays. Moreover the computational problems I encountered halfway in the project was an unforeseen delay, and is inherent to the complexity of the project. . Since no one at [redacted] or TU Delft had experience with this data mining setup, it sometimes was difficult to request help. Luckily, this all worked out well in the end, since the quality of the research was a higher priority to me than the duration of the project. For future graduates, I would advise to explicitly state your ambition level in terms of planning and quality, and adjust the complexity of your project to that.

# 9 Conclusions and recommendations

## 9.1 Conclusions

When embedding data mining in organizations, two complex worlds are integrated. Data mining on itself is a complex method, because of the magnitude of the datasets and the combination of several advanced analytics. This leads to two problems: data mining delivers many results, and not all results are relevant.

Organizations are complex because of the amount of stakeholders that are involved in decision making, that can have different views, opinions and competencies. When data mining was applied in business, it often did not deliver the desired results. This is because data mining and decision making is badly aligned. Data mining, being an inductive approach that searches for objective insights, conflicts with decision making, a deductive process with subjective elements.

Both data mining and organizational behavior have been extensively studied, but if data mining is to be applied in an organization, other aspects for data mining patterns become of importance. Not only extracting patterns is then relevant, but also decision making on basis of these patterns. The existing body of knowledge has focused too little on improving the interpretation of data mining patterns. Furthermore, the position of data mining in the organization has largely been neglected.

This research focused on embedding data mining and the process of data mining in organizations. To this end, the following research question was at the heart of the report:

***How can data mining be integrated in organizations in such a way, that the results can help decision making in the organization?***

### 9.1.1 Requirements for data mining results in organizations

After performing an extensive literature study and carrying out a case study at [redacted] bank, the research question can be answered. After evaluating literature on data mining, it come to light that the current set evaluation measures has been ill-defined. A lot of types and specific measures exist, but they are not in line with the requirements for a data mining result. While in practice only the validity was measured in the evaluation, a true data mining result unveils new knowledge and has a certain value to the decision maker. This caused situations where the decision maker is confronted with a large amount of patterns, which only makes it harder to interpret the data. Therefore, the criteria for a data mining result in an organization have been defined as *validity*, *utility* and *novelty*.

The validity of a data mining result determines that the pattern extracted from the dataset is very likely to exist in the real world. This requirement measures the objective value of a result. The value or *utility* of a result measures the subjective value of the result for the decision maker. This value is focused on improving the interpretability and the actionability of outcomes. The novelty of a result ensures that this result does not represent knowledge that has been known before.

These three requirements have been taken into account for the first time in pattern evaluation. Since utility (converging to what you know) and novelty (diverging to what you do not know yet) may conflict with each other, their role in the process has been defined in the CDE-framework. Instead of one integrated evaluation, the evaluation steps are decoupled (see Figure 42). The validity of a pattern

serves as a filter to leave only statistical relevant patterns. These patterns are then awarded a utility value, a value that is built up out of predefined case-specific criteria that determine the value for the organization. Utility is used to converge the scope of the decision makers to patterns that are relevant to him. Novelty evaluation can then be used to broaden the scope again for new information. In this report novelty was used to reflect on the insights of the utility evaluation. This can be done in the form of an additive learning step, or as a meta-evaluation to discover a pattern within patterns.This search for novel information is now more focused: it aims to find additional, underlying factors that can explain a difference between high-utility patterns and low-utility patterns. With this information, a general rule can be hypothesized out of a pattern. Moreover, novelty evaluation offers multiple opportunities. For instance, it may be used to reflect on the validity of a model. For this, future research is needed.

For the sake of adequacy, simplicity and to prevent friction between utility and validity, the three evaluation steps are decoupled in the evaluation phase.

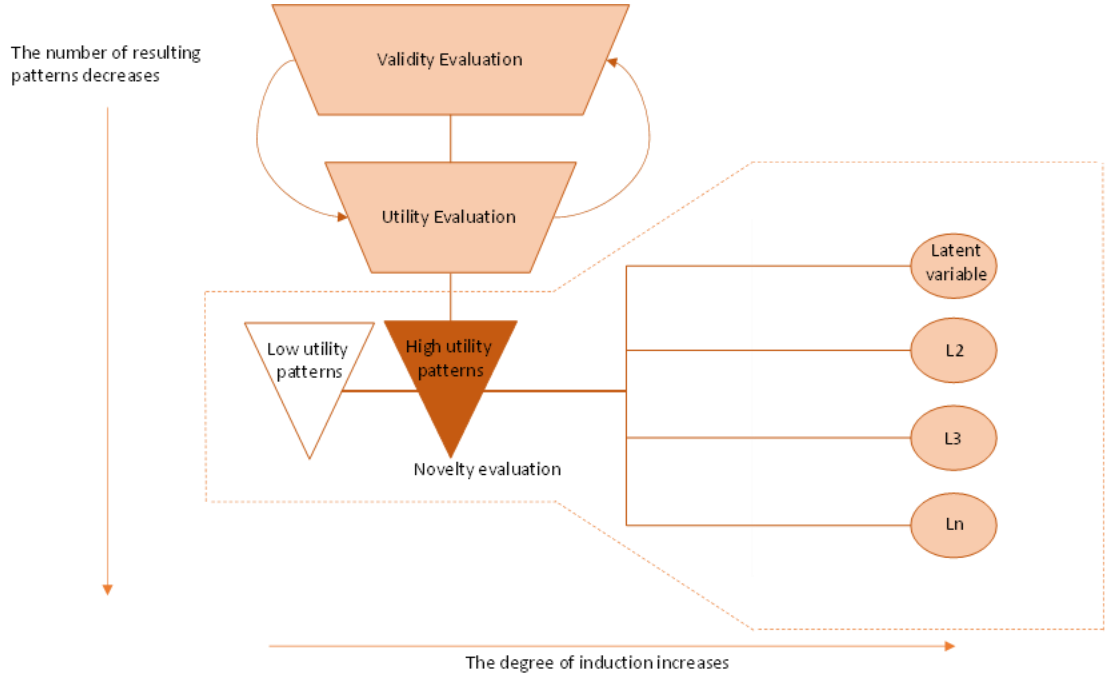



**Figure 42: Defining the roles of the three types of requirements in pattern evaluation**

### 9.1.2 Embedding data mining in an organization

To successfully apply data mining in organizations, the process must bridge the gap between data science and decision making. In order to do so, the process of data mining needs to become more interactive. Since data mining and decision making are so different, one should not expect one party to successfully take account of the other's goals, norms and values. The CDE-framework actively involves decision makers early on in the data mining process. This should create a common perspective on the problem situation, the goal and the data mining method to be used. This also results in a joint determination of the validity measures, the utility measures and the variables to take into account for the novelty evaluation. If agreements have been made on these aspects, data scientists and decision makers agree on the process rather than the outcome of the process. In this way, subjective knowledge can be obtained in an objective way. All these agreements create a sense of ownership on the project itself, leading to a higher acceptance of the results.

The agreements result in an explicit goal and in an experimental setup to solve the problem. The attainability of this setup has to be tested quickly, in order to prevent delays later on in the process. If the setup is not viable, the agreements have to be adjusted or the project can be terminated.

In order to secure the process, the data mining perspective must acknowledge that their minimum level of criteria satisfaction is higher than that of decision makers. In the end, decision makers are looking for results that they can translate into action and value. These patterns do not necessarily be novel. For the practical use of the data mining framework, decision makers must be allowed to design propositions for patterns that are valid and have a high utility. These propositions must lead to a system or artifact. This artifact can be challenged by the stakeholders, in order to create a process of constant refinement of the artifact.

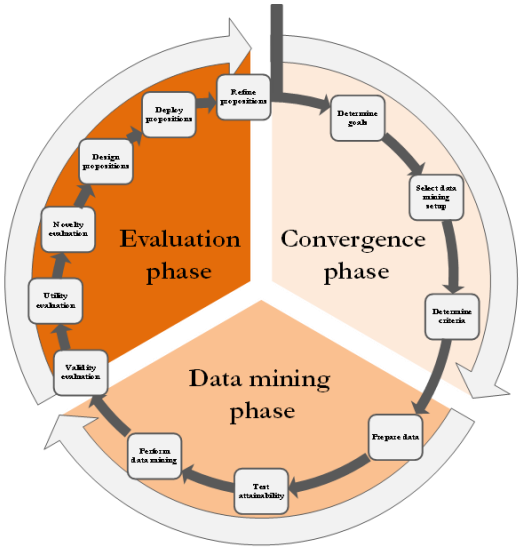



**Figure 43: The Convergence-Data mining-Evaluation-Framework**

## 9.2 Limitations

The CDE-framework has been designed in order to structure the interactions between data scientists and decision makers in data mining projects in organizations. It is designed as a unifiable framework that should be applicable in different contexts and with different data mining methods. However, in some situation and for some goals, the framework has some limitations. Three main limitations are discussed below.

Firstly, the deployment of insights in the organization has only been tested theoretically. Due to time limitations, the deployment of the results could not be tested in the case study. However, the deployment of the results has much influence on the securization of DM in organizations (Sharma & Osei-Bryson, 2009). Possible insights of the relevance cycle on the deployment of results are therefore not included in this research. This is a possible explanation of the friction between a machine learning setup and the utility evaluation step of the CDE-F. Machine learning is a way to deploy insights in the organization, and caused a conflict with the utility evaluation. These issues may have been caused because the deployment phase was out of scope for the case study. Although future research may

shine a light on a solution for the friction, CDE-F is currently less suitable for machine learning projects.

Secondly, the framework does not provide a method for outranking of alternatives (like Electre), since that is labour intensive. On this scale of pattern evaluation, this is becomes an almost impossible approach. The interpretation of patterns that can be turned into action still remains people's work.

Finally, the framework explicitly mentions an attainability test. However, this test cannot yet be defined in a simple formula that gives a clear answer whether the intended data mining method can be carried out. The judgement on the attainability does depend mostly on the professional opinion of the data scientist. The rules of thumb help to structure this decision, but no objective decision can be made without preparing the total model.

## 9.3 Recommendations

On the basis of the conclusions of this research, the contributions of this research to the scientific world can be identified. This will be done via scientific recommendations. Furthermore, some specific recommendations for [redacted] will be made as well.

### 9.3.1 Scientific recommendations

For the academic world, the most important contribution is the consideration that data mining should be analyzed in an organizational context. Previous data mining research was mainly focused on the data scientist and technical improvements of data mining. This research lead to the insight that data mining problems can also be caused by the friction between the data mining perspective and the business perspective. While the business perspective has a focus on results that they can relate to, the data mining perspective aims at patterns that represent new knowledge. This friction causes the misalignment problem, and influences the ability to turn data mining results into action. By incorporating data mining with decision making and the organizational context, the importance of stakeholder participation and –influence is discovered. The CDE-F framework proves the added value of including other perspectives. It formalizes the necessary interactions between data mining and decision making by an interactive approach. The CDE-F also proves that it is possible to incorporate subjective information in an objective process. In this way, the CDE-F can function as a starting point for the development of data mining as a socio-technical system.

Existing data mining frameworks do not go in depth on the evaluation phase of the data mining process. On the technical side of the evaluation, a discrepancy exists between the conceptual criteria for patterns and the types of criteria that are actually used. By formalizing the evaluation criteria *validity*, *utility* and *novelty* in the CDE-F, the data mining process specifies what types of evaluation methods must be included in the process. These criteria are generalizable for a large variety of data mining applications. This research provides insights that further define the concepts of utility and novelty. While utility is so case-specific that no general utility function exists, novelty is a concept that can be measured on a regular level and a meta-level. Furthermore, the three evaluation criteria they have a direct focus on dealing with the rule quantity problem and the rule quality problem. This contribution may be used to develop the data mining process towards a more standardized evaluation process of data mining patterns.

For adopters of the framework, it is important to be aware of the limitations of the framework in its current form. No empirical validation for the deployment of data mining results was collected due to

time constraints. For further development of the CDE-F, these empirical insights are needed. This will eventually lead to a mature, validated framework.

### 9.3.2 Societal recommendations

Considering the societal relevance of this project, this research provides the insight that a data mining project should be fitted to the organization in which it is applied. The research builds understanding on the factors that determine how data mining should be fitted in the organization. Moreover, it defines in what way the organization should participate in data mining and provides practical solutions that deal with frictions that exist between data mining and decision making. It defines topics on which business stakeholders and data scientist need to reach agreements on. All these insights create the possibility that data mining could be successfully applied in various types of organizations.

Moreover, insights from this research can be used in practice to cope with two main problems in data mining: the rule quantity problem and the rule quality problem. It does not only do this by providing evaluation measures, but also shows the importance of the preliminary phases of data mining on the project outcomes.

The CDE-F emphasizes the importance of the preliminary phases for the success of the data mining project. By highlighting the important aspects of the preliminary phases, the complexity of the actual data mining is reduced. If organizations adopt the CDE-F, they can break up the complexity of such a project in this way.

If an organization is to adopt the CDE-F, it is important to know that is toes not provide a holistic view on a data mining project, since the context of every project is unique. In order to keep the framework generalizable for all types of data mining, the framework does not provide a step-by-step guide of all tasks in a data mining project. For instance, it does not specify the content of the utility criterion or takes into account that a project planning has to be made. If the CDE-F is adopted, it has to be translated to the problem situation.

For [redacted] this research was relevant on the micro-, meso- and macro level. On the micro level, [redacted] . On the meso level, [redacted] On the macro level, [redacted] For [redacted] it is important to be aware of the technical limitations of the micro level insight. Due to computational complexity, the analysis of the patterns had to be limited. The way in which web pages are logged required to make some assumptions. If [redacted] wishes to act on the micro level insight, they should support these assumptions and accept the limitations of the analysis.

## 9.4 Future research

As often is the case in science, the deeper you go, the more questions you get. Also in this research, some interesting elements have been identified that could not be solved or discussed in the course of this project. However, they do fit the scope of this project. Therefore, four proposals for future research are discussed below.

This framework for data mining in organization is aimed at combining the competencies of decision makers and data scientists. This secures the interactivity of the process, but could also lead to sluggish decision making. Another emerging theory on data mining in organizations is the concept of the data scientist as a story teller. A data scientist should have sufficient soft skills to take individually take care of the interpretation phase. This is a totally different solution on the rule quality problem. However, it is still a niche term that has received little academic attention. In order to fully compare the pros and cons of each of the two methods, research could be conducted on the implications of the data scientist as a storyteller.

The generalizability of the data mining framework has been discussed multiple times in this report. This is common for DM methods, as their multi-applicability is often a problem (Sharma, 2014). Due to time constraints, the CDE-F has not been tested on different data mining methods and DM suites, they have only been blueprinted on other projects conceptually. Moreover, these projects were all at [redacted] In order to investigate the extent to which this framework is generalizable on other types of projects in other types of organizations, other types of design research evaluation could be applied. Especially a field study would deliver necessary information on the generalizability of the framework (Hevner et al., 2004). For the issue of combining the user perspective with machine learning, the added value of Bayesian learning can be investigated.

With respect to the rule quantity problem, the goal of this research was to cope with the problem instead of solving it. In order to facilitate the possibility for the outranking of alternatives, DM should deliver much fewer patterns than is currently the case. This can currently be done by aggregation of patterns, but then there is a severe chance on information loss. Further research can be conducted in how to further decrease the size of the patterns that exist. In this way, the rule quantity problem can be solved, so that DM patterns can be used for outranking methods, such as ELECTRE (Choi et al., 2005).

The CDE-F is the first data mining framework that specifies the interactions between the data scientists and decision makers. However, the deployment of the results in the organization could not be supported via the case study, subject to the given timeframe of the research. In order to investigate the collaboration of the implementation of the new knowledge in the organization, further research could be conducted in this field. The author advises to again incorporate Mintzberg's Model, as the deployment of data mining is expected to cause friction the data scientists' tendency to standardize, while the backbone of the organization may want to maintain its professional character.

Future research may reflect on extending the possibilities of novelty evaluation. The expert validation shined a light on the possibilities of the added value of the novelty criterion. As in the current research, novelty was used to reflect on the utility evaluation, novelty may also reflect on the validity of a model. Novelty may then be treated as a cost-benefit evaluation of the validity of the model, at the cost of the number of assumptions you put in. A possible solution direction for this is Akaike's information criterion (Akaike, 1981).

# Bibliography

Aggarwal, C. C., & Yu, P. S. (1999). Data Mining Techniques for Associations , Clustering and Classification. In *Methodologies for Knowledge Discovery and Data Mining* (Vol. 1574, pp. 13–23). Beijing: Springer Berlin Heidelberg. doi:10.1007/3-540-48912-6_4

Agrawal, R., Imielinski, T., & Swami, A. (1993). Mining association rules between sets of items in large databases. In *SIGMOD Record (ACM Special Interest Group on Management of Data)* (Vol. 22, pp. 207–216). Publ by ACM. Retrieved from http://www.scopus.com/inward/record.url?eid=2-s2.0-0027621699&partnerID=tZOtx3y1

Agrawal, R., & Srikant, R. (1995). Mining sequential patterns. In *Proceedings - International Conference on Data Engineering* (pp. 3–14). IEEE. Retrieved from http://www.scopus.com/inward/record.url?eid=2-s2.0-0029212693&partnerID=tZOtx3y1

Ahmed, C. F., Tanbeer, S. K., Jeong, B.-S., & Choi, H.-J. (2011). A framework for mining interesting high utility patterns with a strong frequency affinity. *Information Sciences*, *181*(21), 4878–4894. doi:10.1016/j.ins.2011.05.012

Akaike, H. (1981). Likelihood of a model and information criteria. *Journal of Econometrics*, *16*(1), 3–14. doi:10.1016/0304-4076(81)90071-3

Alabdulkarim, L. O. (2013). *Acceptance-by-Design: Elicitation of Social Requirements for Intelligent Infrastructures*. Delft: Delft University of Technology.

Ali, S. (1998). Research Methodology: Back to Basics. *ABAC Journal*, *18*(1), 75–98. Retrieved from http://www.journal.au.edu/abac_journal/jan98/article5.html

Ankerst, M., Elsen, C., Ester, M., & Kriegel, H.-P. (1999). Visual classification: an interactive approach to decision tree construction. *Training*, *1*, 392–396. doi:10.1145/312129.312298

Apté, C., & Weiss, S. (1997). Data mining with decision trees and decision rules. *Future Generation Computer Systems*, *13*(97), 197–210. doi:10.1016/S0167-739X(97)00021-6

Arentze, T., & Timmermans, H. (2003). Measuring the goodness-of-fit of decision-tree models of discrete and continuous activity-travel choice: Methods and empirical illustration. *Journal of Geographical Systems*, *5*(2), 185–206. doi:10.1007/s10109-003-0097-9

Bâra, A., & Lungu, I. (2012). Improving Decision Support Systems with Data Mining Techniques. *Advances in Data Mining Knowledge Discovery and Applications*, 397–417. doi:10.5772/47788

Baxter, G., & Sommerville, I. (2011). Socio-technical systems: From design methods to systems engineering. *Interacting with Computers*, *23*(1), 4–17. doi:10.1016/j.intcom.2010.07.003

Bayardo Jr, R. J., & Agrawal, R. (1999). Mining the Most Interesting rules. In *Proceedings of the Fifth International Conference on Knowledge Discovery and Data Mining* (pp. 145–154). San Diego, CA: ACM New York.

Bell, B., Morgan, G., Kromrey, J., & Ferron, J. (2010). The impact of small cluster size on multilevel models: a Monte Carlo examination of two-level models with binary and continuous predictors.

*JSM Proceedings, Section on Survey Research Methods*, 4057–4067. Retrieved from http://www.amstat.org/sections/srms/Proceedings/y2010/Files/308112_60089.pdf

Bendasolli, P. F. (2013). Theory Building in Qualitative Research: Reconsidering the Problem of Induction. *Qualitative Social Research*, *14*(1), Article 25. Retrieved from http://www.qualitative-research.net/index.php/fqs/article/view/1851/3497#g2

Berkhin, P. (2006). A Survey of Clustering Data Mining. In J. Kogan, C. Nicholas, & M. Teboulle (Eds.), *Grouping Multidimensional data* (pp. 25–71). Berlin: Springer Berlin Heidelberg. doi:10.1007/3-540-28349-8_2

Berson, A., Smith, S., & Thearling, K. (1999). *Building Data Mining Applications for CRM* (First). New York, New York, USA: McGraw-Hill Professional. Retrieved from http://dl.acm.org/citation.cfm?id=555454

Bhatt, G. D. (2001). Knowledge management in organizations. *Journal of Knowledge Management*, *5*(1), 68–75.

Bie, T. De. (2013a). Subjective Interestingness in Exploratory Data Mining, 19–31.

Bie, T. De. (2013b). Subjective interestingness in exploratory data mining.

Blumenstock, A., Mueller, M., Lanquillon, C., Kempe, S., Hipp, J., & Wirth, R. (2010). Interactivity closes the gap: Lessons learned in an automotive industry application. *Frontiers in Artificial Intelligence and Applications*, *218*, 17–34. doi:10.3233/978-1-60750-633-1-17

Boisot, M. (2004). Data , information and knowledge : have we got it right ?, (February).

Bouyssou, D., Marchant, T., Pirlot, M., Tsoukias, A., & Vincke, P. (2006). Evaluation and decision models with multiple criteria: stepping stones for the analyst. *NewYork: Springer*, *86*. doi:10.1007/0-387-31099-1

Boyd, D., & Crawford, K. (2012). Critical Questions for Big Data. *Information, Communication & Society*, *15*(5), 662–679. doi:10.1080/1369118X.2012.678878

Braynova, E., & Pendharkar, H. (2005). Frequent Pattern Mining with Preferences – Utility Functions Approach. In M.-S. Hacid, Z. W. Ras, & S. Tsumoto (Eds.), *Foundations of Intelligent Systems: 15th International Symposium Ismis 2005* (pp. 364–372). Saratoga Springs, New York, USA: Springer.

Brockwell, P. J., & Davis, R. A. (2002). *Introduction to Time Series Forecasting*. (G. Casella, S. Fienberg, & I. Olkin, Eds.) (Second). Fort Collins: Springer.

Brown, B., Chui, M., & Manyika, J. (2011). Are you ready for the era of "big data" ? *Library*, *October*(October), 1–12. doi:00475394

Brown, J. D. (2009). Choosing the Right Type of Rotation in PCA and EFA. *Shiken: JALT Testing & Evaluation SIG Newsletter*, *13*(November), 20–25.

Brynjolfsson, E., Hitt, L. M., & Kim, H. H. (2011). *Strength in Numbers: How Does Data-Driven Decisionmaking Affect Firm Performance? SSRN Electronic Journal*. Cambridge, Massachusetts. doi:10.2139/ssrn.1819486

Brzinsky-fay, C., & Kohler, U. (2006). Sequence analysis with Stata. *The Stata Journal*, *6*(4), 435–460.

Cabena, P., Choi, H. H., Kim, I. S., Otsuka, S., Reinschmidt, J., & Saarenvirta, G. (1999). *Intelligent Miner for Data Applications Guide*. San Jose, California.

Carvalho, D. R., Freitas, A. A., & Ebecken, N. (2012). Evaluating the Correlation Between Objective Rule Interestingness Measures and Real Human Interest. In A. M. Jorge, L. Torgo, P. Brazdil, R. Camacho, & J. Gama (Eds.), *Knowledge Discovery in Databases: PKDD 2005. 9th European Conference on Principles and Practice of Knowledge Discovery in Databases, Porto, Portugal, October 3-7, 2005. Proceedings* (pp. 453–461). Porto: Springer. doi:10.1007/978-3-642-32645-5_19

Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection. *ACM Computing Surveys*, *41*(3), 1–58. doi:10.1145/1541880.1541882

Chandola, V., & Kumar, V. (2005). Summarization - Compressing Data into an Informative Representation. In *Data mining, Fifth IEEE International Conference on Data Mining* (p. no pagenumbering). Minneapolis: IEEE.

Chaudhari, S. C., & Verma, V. K. (2013). Mining High Utility Item Set From Large Database: - A Recent Survey. *International Journal of Emerging Technology and Advanced Engineering*, *3*(5), 685–688. Retrieved from http://www.ijetae.com/files/Volume3Issue5/IJETAE_0513_115.pdf

Chen, H., Chiang, R. H. L., & Storey, V. C. (2012). Business Intelligence and Analytics: From Big Data To Big Impact. *Mis Quarterly*, *36*(4), 1165–1188.

Chen, M.-S. (National T. U., Han, J. (Simon F. U., & Yu, P. S. (IBM). (1996). Data Mining: An Overview from Database Perspective. *Knowledge and Data Engineering*, *8*(6), 866–883.

Choi, D. H., Ahn, B. S., & Kim, S. H. (2005). Prioritization of association rules in data mining: Multiple criteria decision approach. *Expert Systems with Applications*, *29*(4), 867–878. doi:10.1016/j.eswa.2005.06.006

Clegg, S. R., Hardy, C., & Nord, W. (1999). *Managing Organizations: current Issues*. London, UK: SAGE.

Cohagan, C., Grzymala-Busse, J. W., & Hippe, Z. S. (2010). Mining Inconsistent Data with the Bagged MLEM2 Rule Induction Algorithm. In *2010 IEEE International Conference on Granular Computing* (pp. 115–120). San Jose: IEEE. doi:10.1109/GrC.2010.73

Cormen, T., Leiserson, C., Rivest, R., & Stein, C. (2009). *Introduction to Algorithms* (Third). London, UK: The MIT press. doi:10.1163/9789004256064_hao_introduction

Cornwall, A., & Jewkes, R. (1995). What is participatory research? *Social Science & Medicine*, *41*(12), 1667–1676. doi:10.1016/0277-9536(95)00127-S

Costello, A. B., & Osborne, J. W. (1994). Best Practices in Exploratory Factor Analysis: Four Recommendations for Getting the Most From Your Analysis. *Practical Assessment, Research & Evaluation*, *10*(7), 27–29. doi:1531-7714

Courtney, J. F. (2001). Decision making and knowledge management in inquiring organizations: Toward a new decision-making paradigm for DSS. *Decision Support Systems*, *31*, 17–38. doi:10.1016/S0167-9236(00)00117-2

De Bie, T., & Spyropoulou, E. (2013). Subjective Interestingness in Exploratory Data Mining: recent insights and challenges ahead. In *An Information Theoretic Framework for Data Mining Categories and Subject Descriptors*.

De Bruijn, H., & Ten Heuvelhof, E. (2008). *Management in Networks* (First). Abingdon: Routledge.

De Bruijn, H., Ten Heuvelhof, E., & In 't Veld, R. (2010). *Process Management; Why Project Management Fails in Complex Decision Making Processes* (Second). Delft: Springer-Verlag Berlin Heidelberg. doi:10.1007/978-3-642-13941-3

Devlin, B., & Cote, L. D. (1996). *Data Warehouse: From Architecture to Implementation*. Boston: Addison-Wesley Longman Publishing Co.

Drucker, P. F. (1989). *The New Realities*. London: Heinemann Professional Pub. Retrieved from http://books.google.nl/books?hl=nl&lr=&id=WmAsWS9-fFsC&oi=fnd&pg=PA8&dq=Drucker+1989+The+New+Realities:+In+Government+and+Politics/In+Economics+and+Business/In+Society+and+World+View&ots=cPf5l8Kiog&sig=ubYf6CSZF7ARcEylKGa1yVhA-jg#v=onepage&q&f=false

Duin, R., & Pękalska, E. (2006). Object Representation, Sample Size, and Data Set Complexity. *Data Complexity in Pattern Recognition*, 26–58. doi:10.1007/978-1-84628-172-3_2

Eisenhardt, K. M. (1989). Building Theories from Case Study Research. *Academy of Management Review*, *14*(4), 532–550. doi:10.5465/AMR.1989.4308385

Elzinga, C. H., & Liefbroer, A. C. (2007). De-standardization of Family-Life Trajectories of Young Adults: A Cross-National Comparison Using Sequence Analysis. *European Journal of Population / Revue Européenne de Démographie*, *23*(3-4), 225–250. doi:10.1007/s10680-007-9133-7

Emmett, R. B. (2010). *The Constitution of Innovation, Lecture 1: Why "Picking Winners" Robs Us All of a Better Future*. USA: Michigan State University. Retrieved from https://vimeo.com/9235143

European Commission. (2014). European Commission and data industry launch € 2 . 5 billion partnership to master big data Organisation of the PPP. Retrieved from http://europa.eu/rapid/press-release_IP-14-1129_en.htm

Fahrnkopf, K. (2015). Two Frequent Challenges in Data Preparation. Retrieved from http://info.salford-systems.com/two-frequent-challenges-in-data-preparation

Fayyad, U., Piatetsky-shapiro, G., & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. *AI Magazine*, *17*(3), 37–54.

Freitas, A. A. (1997). A Genetic Programming Framework for Two Data Mining Tasks: Classification and Generalized Rule Induction. *Genetic Programming 1997 Proceedings of the Second Annual Conference*, 96–101. Retrieved from http://kar.kent.ac.uk/21483/

Freitas, A. A. (1999). On rule interestingness measures. *Knowledge-Based Systems*, *12*(5-6), 309–315. doi:10.1016/S0950-7051(99)00019-2

Gabadinho, A., Ritschard, G., & Studer, M. (2011). Analyzing and Visualizing State Sequences in R with TraMineR. *Journal of Statistic Software*, *40*(4), 1–37.

Gabadinho, A., Ritschard, G., Studer, M., & Nicolas, S. M. (2011). *Mining sequence data in R with the TraMineR package :* Geneva.

Gabadinho, A., Studer, M., Muller, N., Buergin, R., & Ritschard, G. (2015). *Package "TraMineR ."*

Geels, F. W. (2002). Technological transitions as evolutionary reconfiguration processes: a multi-level perspective and a case-study. *Research Policy*, *31*, 1257–1274. doi:10.1016/S0048-7333(02)00062-8

Geng, L., & Hamilton, H. J. (2006a). Interestingness measures for data mining. *ACM Computing Surveys*, *38*(3), 9–es. doi:10.1145/1132960.1132963

Geng, L., & Hamilton, H. J. (2006b). Interestingness Measures for Data Mining: A Survey. *ACM Computing Surveys*, *38*(3), 9–es. doi:10.1145/1132960.1132963

Guillet, F., & Hamilton, H. J. (2007). *Quality measures in data mining*. Berlin: Springer.

Han, J., Cheng, H., Xin, D., & Yan, X. (2007). Frequent pattern mining: current status and future directions. *Data Mining and Knowledge Discovery*, *15*(1), 55–86. doi:10.1007/s10618-006-0059-1

Han, J., & Kamber, M. (2006). *Data Mining: Concepts and Techniques* (second). San Franciso: Elsevier. doi:10.1016/B978-0-12-381479-1.00001-0

Hand, D. J. (1998). Data Mining: Statistics and More? *The American Statistician*, *52*(2), 112. doi:10.2307/2685468

Harding, J. a., Shahbaz, M., & Kusiak, a. (2006). Data Mining in Manufacturing: A Review. *Journal of Manufacturing Science and Engineering*, *128*(4), 969. doi:10.1115/1.2194554

Hevner, A. R. (2007). A Three Cycle View of Design Science Research A Three Cycle View of Design Science Research. *Scandanavian Journal of Information Systems*, *19*(2), 87–92.

Hevner, A. R., March, S. T., & Park, J. (2004). Design Science in Information Systems Research. *MIS Quarterly*, *28*(1), 75–105. doi:10.2307/25148625

[redacted]

[redacted]

[redacted]

Jagannath, S. (2003). *Utility Guided Pattern Mining*. Raleigh, North Carolina, USA.

Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: a review. *ACM Computing Surveys*, *31*(3), 264–323. doi:10.1145/331499.331504

Kantardzic, M. (2011). *Appendix B: Data-Mining Applications*. *Data Mining: Concepts, Models, Methods, and Algorithms* (Second). Louisville: IEEE & John Wiley & Sons Inc. doi:10.1002/9781118029145.app2

Kasperczyk, N., & Knickel, K. (2004). Analytic hierarchy process (AHP). *IVM Institute*, 1–6. Retrieved from http://www.ivm.vu.nl/en/Images/MCA3_tcm53-161529.pdf

Keen, P. G. W. (1980). *Decision Support Systems: A Research Perspective*. Cambridge: Massachussetts Institute of Technology. Retrieved from http://dspace.mit.edu/bitstream/handle/1721.1/47172/decisionsupports1980keen.pdf

Kimball, R., & Ross, M. (2011). *The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling* (Second). John Wiley & Sons Inc. Retrieved from

http://books.google.nl/books?id=XoS2oy1IcB4C&dq=data+warehouse&lr=&hl=nl&source=gbs_navlinks_s

Köksal, G., Batmaz, İ., & Testik, M. C. (2011). A review of data mining applications for quality improvement in manufacturing industry. *Expert Systems with Applications*, *38*(10), 13448–13467. doi:10.1016/j.eswa.2011.04.063

Kotsiantis, S., & Kanellopoulos, D. (2006). Association Rules Mining : A Recent Overview Basic Concepts & Basic Association Rules Algorithms. *GESTS International Transactions on Computer Science and Engineering*, *32*(1), 71–82. Retrieved from http://www.csis.pace.edu/~ctappert/dps/d861-13/session2-p1.pdf

Kroes, P., Franssen, M., Poel, I. Van De, & Ottens, M. (2006). Treating Socio-technical Systems as Engineering Systems : Some Conceptual Problems y. *Systems Research and Behavioral Science*, *814*(February), 803–814.

Kusiak, A. (2006). Data mining: manufacturing and service applications. *International Journal of Production Research*, *44*(18-19), 4175–4191. doi:10.1080/00207540600632216

Leinweber, D. J. (2007). Stupid Data Miner Tricks: Overfitting the S&P 500. *The Journal of Investing*, *16*(1), 15–22. doi:10.3905/joi.2007.681820

Lenca, P., Meyer, P., Vaillant, B., & Lallich, S. (2008). On selecting interestingness measures for association rules: User oriented description and multiple criteria decision aid. *European Journal of Operational Research*, *184*(2), 610–626. doi:10.1016/j.ejor.2006.10.059

Lewin, A. Y., & Nonaka, I. I. (1994a). A Dynamic Theory of Organizational Knowledge Creation. *Organization Science*, *5*(1), 14–37. doi:10.1287/orsc.5.1.14

Lewin, A. Y., & Nonaka, I. I. (1994b). A Dynamic Theory of Organizational Knowledge Creation. *Organization Science*, *5*(1), 14–37. doi:10.1287/orsc.5.1.14

Lewin, K. (1946). Action Research and Minority Problems. *Journal of Social Issues*. doi:10.1111/j.1540-4560.1946.tb02295.x

Light, D., Wexler, D., & Heinze, J. (2004). *How Practitioners Interpret and Link Data to Instruction : Research Findings on New York City Schools ' Implementation of the Grow Network*. San Diego.

Luj, S., & Trujillo, J. (2005). A Data Warehouse Engineering Process. *Advances in Information Systems*, *3261*, 14–23.

Lunenburg, F. C. (2012). Organizational Structure: Mintzberg's Framework. *INTERNATIONAL JOURNAL OF SCHOLARLY, ACADEMIC, INTELLECTUAL DIVERSITY*, *14*(1), 291–319.

Mandinach, E. B., Honey, M., & Light, D. (2006). A Theoretical Framework for Data-Driven Decision Making. In *Annual meeting of AERA, San Francisco* (pp. 1–18). San Franciso: EDC Center for Children and Technology.

Markus, L. M., Majchrzak, A., & Gasser, L. (2002). A design theory for systems that support emergent knowledge processes. *MIS Quarterly*, *26*, 179–212. Retrieved from http://www.jstor.org/discover/10.2307/4132330?uid=3738736&uid=2&uid=4&sid=21104304024657

Matheus, C. J., Chan, P. K., & Piatetsky-Shapiro, G. (1993). Systems for knowledge discovery in databases. *IEEE Transactions on Knowledge and Data Engineering*, *5*(6), 903–913. doi:10.1109/69.250073

Mcgarry, K. (2005). A survey of interestingness measures for knowledge discovery. *The Knowledge Engineering Review*, *20*(01), 39. doi:10.1017/S0269888905000408

Mckay, J., & Marshall, P. (2001). The dual imperatives of action research. *Information Technology & People*, *14*(1), 46–59.

Mintzberg, H. (1980). Structure in 5's: A Synthesis of the Research on Organizational Design. *Management Science*, *26*(3), 322–341. doi:10.1287/mnsc.26.3.322

Mintzberg, H. (1993). *Structure in Fives: Designing effective organizations*. Englewood Cliffs, New Jersey: Prentice-Hall, Inc.

Mintzberg, H., & McHugh, A. (1985). Strategy Formation in an Adhocracy. *Administrative Science Quarterly*, *30*(2), 160–197. doi:10.2307/2393104

Mitra, A. (2012). Classifying data for successful modelling. Retrieved January 16, 2015, from http://www.dwbiconcepts.com/data-warehousing/12-data-modelling/101-classifying-data-for-successful-modeling.html

Mitroff, I. ., & Linstrone, H. A. (1993). *The Unbounded Mind: Breaking the Chains of Traditional Business Thinking* (8th ed.). New York: Oxford University Press.

Molenaar, J. (2014). *Marketing scrum vs IT scrum - two marketing case studies who now "act first and apologize later."*

Molin, E. (2014). Factoranalyse II. In *Onderzoeksmethoden en dataverwerking 2*. Delft: TU Delft.

Muth, B. (2011). Modern Latent Variable Modeling Methods, (May).

Muthén, L. K., Linda, M., Muthén, B., & Nguyen, T. (2009). Regression Analysis, Exploratory Factor Analysis, Confirmatory Factor Analysis, And Structural Equation Modeling For Categorical, Censored, And Count Outcomes. In *Mplus short courses*. www.statmodel.com.

Natarajan, R., & Shekar, B. (2005). Interestingness of association rules in data mining : Issues relevant to e-commerce. *Sãdhanã*, *30*(June), 291–309.

Natarajan, R., & Shekar, B. (2009). Interesting Knowledge Patterns in Databases. In J. Erickson (Ed.), *Database Technologies: Concepts, Methodologies, Tools, and Applications* (pp. 1652–1666). Hershey, PA: IGI Global. Retrieved from https://books.google.nl/books?id=Nn3jBBZSk9cC&pg=PA1653&lpg=PA1653&dq=solution+to+%22rule+quantity%22+problem&source=bl&ots=lbVs4FZNAU&sig=3vg0c0Ka4VSeoQJtt6NwnFaw1Pw&hl=nl&sa=X&ei=H78vVbnmOo35aJTxgWA&ved=0CCEQ6AEwAA#v=onepage&q=solution to %22rule quanti

Nisbet. (2004). How to Choose a Data Mining Suite. *DM Review*, 1–10.

Nisbet, R., Elder, J., & Miner, G. (2009). *Handbook of Statistical Analysis and Data Mining Applications*. Academic Press. doi:10.1016/B978-0-12-374765-5.00049-8

Nonake, I., & Takeuchi, H. (1995). *The Knowledge Creating Compay: How Japanese Companies Create the Dynamics of Innovation* (First). new york: Oxford University Press. Retrieved from http://books.google.nl/books?hl=nl&lr=&id=B-qxrPaU1-MC&oi=fnd&pg=PA3&dq=nonaka+%26+takeuchi+1995+The+Knowledge+Creating+Company&ots=XhSKlAtfcY&sig=ueUBH0lDmbecZv2q13pB3-bmuhU#v=onepage&q&f=false

Okun, O., & Priisalu, H. (2007). Dataset Complexity and Gene Expression Based. In F. Massulli, S. Mitra, & G. Pasi (Eds.), *Application of Fuzzy Sets Theory: 7th International Workshop on Fuzzy Logic and Applications, WILF 2007, Camogli, Italy, July 7-10, 2007, Proceedings* (First, pp. 484–490). Springer Berlin Heidelberg. doi:10.1007/978-3-540-73400-0

Oppewal, H., Tojib, D. R., & Louvieris, P. (2013). Experimental analysis of consumer channel-mix use. *Journal of Business Research*, *66*(11), 2226–2233. doi:10.1016/j.jbusres.2012.02.002

Papadakis, V. M., Lioukas, S., & Chambers, D. (1998). Strategic Decision-Making Processes: The Role of Management and Context. *Strategic Management Journal*, *19*(2), 115–147.

Parry, J. E. (2010). *Data Mining Overview Introduction to SPSS , an IBM Company*. *Business Analytics software*.

Patcha, A., & Park, J.-M. (2007). An overview of anomaly detection techniques: Existing solutions and latest technological trends. *Computer Networks*, *51*(12), 3448–3470. doi:10.1016/j.comnet.2007.02.001

Patidar, V. K., Raghuvanshi, A., & Shrivastava, V. (2013). Literature Survey of Association Rule Based Techniques for Preserving Privacy. *Compusoft: An International Journal of Advanced Computer Technology*, *2*(3), 1–6.

Paul, R., Groza, T., Hunter, J., & Zankl, A. (2014). Semantic interestingness measures for discovering association rules in the skeletal dysplasia domain. *Journal of Biomedical Semantics*, *5*(8), 1–13. doi:10.1186/2041-1480-5-8

Pfeiffer, M., Zhang, H., & Hohmann, A. (2010). A Markov Chain Model of Elite Table Tennis Competition. *Sports Science & Coaching*, *5*(2).

Polanyi, M., & Sen, A. (1967). *The Tacit Dimension*. New York.

Potes Ruiz, P. A., Kamsu-Foguem, B., & Grabot, B. (2014). An Interactive Approach for the Post-processing in a KDD Process. In D. K. Bernard Grabot, Bruno Vallespir, Gomes Samuel, Abdelaziz Bouras (Ed.), *Advances in Production Management Systems: Innovative and Knowledge-Based Production Management in a Global-Local World:* (pp. 92–100). Ajaccio, France: Springer. Retrieved from https://books.google.nl/books?id=6FxbBAAAQBAJ&dq=data+mining+comprehend+decision+maker&hl=nl&source=gbs_navlinks_s

Power, D. J. (2002). *Decision support systems: concepts and resources for managers*. Westport: Greenwood Publishing Group. Retrieved from http://books.google.nl/books?hl=nl&lr=&id=9NA6QMcte3cC&oi=fnd&pg=PR9&dq=Decision+support+systems:+concepts+and+resources+for+managers&ots=DNopsENww9&sig=unjlXMNNnjY5kou45LSXY3qirCE#v=onepage&q=Decision support systems%3A concepts and resources for managers&f=false

Ramsey, F., & Schafer, D. (2012). *The Statistical Sleuth: A Course in Methods of Data Analysis* (Third). Cengage Learning.

Ranjan, J., & Bhatnagar, V. (2008). A Review of Data Mining Tools In Customer Relationship Management. *Journal of Knowledge Management Practice*, *9*(1). Retrieved from http://www.tlainc.com/articl149.htm

Rittel, H. W. J., & Webber, M. M. (1973). Dilemnas in a general theory of planning. *Policy Sciences*, *4*(December 1969), 155–169.

Roman, P. E. (2010). Web Usage Mining. In *Advanced Techniques in Web Intelligence - I* (pp. 143–165). Berlin: Springer Berlin Heidelberg.

Roy, B. (1991). The outranking foundations aproach and the methods. *Theory and Decision*, *31*, 49–73.

Saaty, T. L. (1990). How to make a decision: The analytic hierarchy process. *European Journal of Operational Research*, *48*, 9–26. doi:10.1016/0377-2217(90)90057-I

Saaty, T. L. (1999). *Decision Making for Leaders: The Analytic Hierarchy Process for Decisions in a Complex World*. RWS Publications.

Saaty, T. L. (2008). Decision making with the analytic hierarchy process. *International Journal of Services Sciences*, *1*(1), 83. doi:10.1504/IJSSCI.2008.017590

Sales, P. (1971). The Validity of the Markov Chain Model for a Class of the Civil Service. *Journal of the Royal Statistical Society. Series D (The Statistician)*, *20*(1), 85–110.

Schell, C. (1992). *The Value of the Case Study as a Research Strategy*. Manchester.

Shah, P., & Dua, P. A. K. (2014). cSPADE -UE : Algorithm for Sequence Mining for Unstructured Elements Using Time Gap Constraints. *International Journal of Emerging Technology and Advanced Engineering*, *4*(6), 592–596.

Sharma, M. (2014). Data Mining : A Literature Survey. *International Journal of Emerging Research in Management &Technology*, *9359*(2), 1–4.

Sharma, S., & Osei-Bryson, K.-M. (2009). Framework for formal implementation of the business understanding phase of data mining projects. *Expert Systems with Applications*, *36*(2), 4114–4124. doi:10.1016/j.eswa.2008.03.021

Sharma, S., Osei-Bryson, K.-M., & Kasper, G. M. (2012). Evaluation of an integrated Knowledge Discovery and Data Mining process model. *Expert Systems with Applications*, *39*(13), 11335–11348. doi:10.1016/j.eswa.2012.02.044

Shelokar, P., Quirin, A., & Cordón, Ó. (2013). A multiobjective evolutionary programming framework for graph-based data mining. *Information Sciences*, *237*, 118–136. doi:10.1016/j.ins.2013.02.014

Silberschatz, A. (1995). On Subjective Measures of Interestingness Discovery in Knowledge Bell Laboratories Measures. In *KDD-95 proceeedings* (pp. 275–281). AAAI.

Silberschatz, A., Member, S., & Tuzhilin, A. (1996). What Makes Patterns Interesting in Knowledge Discovery Systems. *IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING*, *8*(6), 970–974.

Simon, H. (1979). Rational Decision Making in Business Organizations. *The American Economic Review*, *69*(4), 493–513.

Stebbins, R. A. (2001). Exploratory Research in the Social Sciences. *Qualitative Research Methods*, *48*. Retrieved from https://books.google.nl/books?hl=nl&lr=&id=hDE13_a_oEsC&oi=fnd&pg=PR5&dq=purpose+of+exploratory+research&ots=NkRGZ_FErI&sig=TMY3iSx4PMUHPKCEwjy-mf48M3k#v=onepage&q&f=false

Suhr, D. (2006). Exploratory or confirmatory factor analysis? *Statistics and Data Analysis*, 1–17. doi:10.1002/da.20406

Sveiby, K. E. (1997). *The New Organizational Wealth* (First). San Francisco: Berrett-Koehler Publishers.

Tan, P.-N., Kumar, V., & Srivastava, J. (2004). Selecting the right objective measure for association analysis. *Information Systems*, *29*(4), 293–313. doi:10.1016/S0306-4379(03)00072-3

Tedmanson, D., & Subhabrata, B. B. (2010). Participatory Action Research : SAGE Research Methods. In *Encyclopedia of Case Study Research*. Retrieved from http://srmo.sagepub.com/view/encyc-of-case-study-research/n246.xml

Vanvoorhis, C. R. W., & Morgan, B. L. (2007). Understanding Power and Rules of Thumb for Determining Sample Sizes. *Tutorials in Quantitative Methods*, *3*(2), 43–50.

Vidulin, V., Bohanec, M., & Gams, M. (2014). Combining human analysis and machine data mining to obtain credible data relations. *Information Sciences*, *288*, 254–278. doi:10.1016/j.ins.2014.08.014

Walter, J., Lechner, C., & Kellermanns, F. W. (2008). Disentangling alliance management processes: Decision making, politicality, and alliance performance. *Journal of Management Studies*, *45*(3), 530–560. doi:10.1111/j.1467-6486.2007.00749.x

Wang, K., Tong, S., Eynard, B., Roucoules, L., & Matta, N. (2007). Review on Application of Data Mining in Product Design and Manufacturing. In *Fourth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2007)* (pp. 613–618). Ieee. doi:10.1109/FSKD.2007.482

Wang, X.-Q., Chen, D.-F., Xu, X. Y., & Xy, X. (2014). Pervasive Computing and the Networked World. In *Joint International Conference, ICPCA/SWS 2013* (pp. 602–609).

Wielenga, D. (2007). Identifying and Overcoming Common Data Mining Mistakes. *Forum American Bar Association*, 1–20.

Witten, I. H., & Frank, E. (2005). *Data mining: Practical Machine Learning Tools and Techniques* (Second edi). San Franciso: Morgan Kaufmann Publishers Inc.

Wold, S., Esbensen, K., & Geladi, P. (1987). Principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, *2*(1-3), 37–52. doi:10.1016/0169-7439(87)80084-9

Yang, Q., & Wu, X. (2006). 10 Challenging Problems in Data Mining Research. *International Journal of Information Technology & Decision Making*, *05*(04), 597–604. doi:10.1142/S0219622006002258

Yen, S., Chen, C., & Lee, Y. (2005). A Fast Algorithm for Mining High Utility Itemsets. In *UBDM '05*. ACM. doi:10.1007/978-1-4471-2969-1

Yi, L., Liu, B., & Street, S. M. (2003). Eliminating Noisy Information in Web Pages for Data Mining. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 296–305). New York: ACM. doi:10.1145/956750.956785

Yin, R. K. (2014). *Case Study Research: Design and Methods* (Fifth). Thousand Oaks, CA: SAGE Publications Inc. Retrieved from https://books.google.nl/books?hl=nl&lr=&id=AjV1AwAAQBAJ&oi=fnd&pg=PP1&dq=yin+case+study+design&ots=gkPaoCNMcR&sig=gJigicPkRNWyL8ikrs3teQgIRgk#v=onepage&q=yin case study design&f=false

Za, O. R., Xin, M., & Va, C. (1998). Discovering Web Access Patterns and Trends by Applying OLAP and Data Mining Technology on Web Logs. In *Research and Technology Advances in Digital Libraries, 1998. ADL 98. Proceedings. IEEE International Forum on* (pp. 19–29). Santa Barbara, CA: IEEE.

Zaki, M. J. (2001). SPADE : An Efficient Algorithm for Mining Frequent Sequences. *Machine Learning*, *42*, 31–60.

Zeleny, M. (1998). Multiple criteria decision making: Eight concepts of optimality. *Human Systems Management*. New York: McGraw-Hill Book Company. Retrieved from http://iospress.metapress.com/index/mh9uvjtkpn8kqtgl.pdf

Zhang, S., Zhang, C., & Yang, Q. (2003). Data preparation for data mining. *Applied Artificial Intelligence*, *17*(5-6), 375–381. doi:10.1080/713827180

# Appendix A: Overview of data mining

## The data mining process

When considering data mining from the technical perspective, the data mining process is divided in five steps. These are shortly mentioned in chapter 1, below is a more extensive description of the different process steps.

### Data selection

In the initial state, there are lots of collected and generated records that are stored in databases. These are often in the magnitude of $10^5$ records in files of terabytes (Cabena et al., 1999; Kantardzic, 2011; Matheus et al., 1993). These are stored in a data warehouse (more on data warehouses in Appendix $). To perform the analysis, the target data that is used in the analysis is selected from the data warehouse. Sometimes data reduction is performed before selection, to select a smaller set of data without significantly sacrificing the validity of the dataset (Han & Kamber, 2006).

### Data preprocessing

In the preprocessing phase, the target data is prepared to be suitable for the analysis that is performed. Typical preprocessing tasks are noise removal, inconsistent data removal, data integration, data reduction, outlier removal and incomplete data recovery (Zhang et al., 2003). Noise removal is the removal of content that can harm the further data mining process, and is described further for Web mining, the type of data mining used in the case study, by Yi et al. (Yi et al., 2003). Inconsistent data removal deals with conflicting cases or discrepancies in codes and names. Removal methods and the value of inconsistent data is described in other literature (Cohagan et al., 2010). If the data is stored in different data warehouses, data integration is needed to merge the desired data. Data reduction is the sampling of data or selecting different instances of data. Outlier removal is the analysis and exemption of extreme records and/or values in the dataset. Incomplete data recovery is done to reduce the ambiguity of some records in the dataset in order to improve the interpretive value of a record. All these processes are very well described by Zhang et al. (2003).

### Data transformation

In the data transformation phase the data is replaced by more interpretable values, if necessary. An example of a transformation in the dataset is the replacement of a url by a substantive name (Za et al., 1998).

### Data mining

In the data mining phase, the data mining setup is selected and is applied on the selected dataset. The combination of a DM method, technique and algorithm is called a data mining suite. Methods, techniques and algorithms are discussed in paragraph 1.1. When deciding on the DM suite, four requirements have to be taken into account (Nisbet, 2004):

- Applicability: is the setup able to give the desired insights?
- Scalability: if successful, can the setup be performed on a larger scale?
- Ease of use: can the setup be performed relatively simple?
- Accuracy: does the setup adequately measure what I want it to measure?

Furthermore, Nisbet argues that it is important that these 4 requirements are in balance. A DM setup that satisfies the four requirements should then be preferred over the most accurate setup that is not easy to use.

### Pattern evaluation

The outcomes of a data mining process are called results or patterns. As data mining executions often generate a large set of patterns, the patterns are either ranked or filtered to structure the outcomes (Carvalho et al., 2012; Shelokar et al., 2013). When this ranking or filtering is done automatically, the pattern evaluation phase can be seen as an integrated part of the data mining phase (Geng & Hamilton, 2006b).

In data mining literature, the concept of interestingness has been introduced to measure rank or filter results (Freitas, 1999; Geng & Hamilton, 2006b; Silberschatz, 1995). A pattern that exceeds a certain threshold of interestingness can be classified as a pattern that can reveal knowledge (Fayyad et al., 1996).

## Data mining methods and techniques

As is already mentioned in Chapter 1, there are many different methods and techniques to perform data mining. During the literature study, these methods and techniques have been investigated in order to gain insight in the differences between the methods. These are presented in Table 16 on the next page.

As can be concluded on basis of this investigation, a data mining method should be picked on the basis of a fit between the goal of the research and the goal of the data mining method.

**Table 16: : overview of different data mining methods and techniques**

| Method of data mining | Goal | Techniques in method | Examples of applications | Further reading |
|---|---|---|---|---|
| Anomaly detection | Detection of unusual cases or outliers<br>*(prediction)* | Statistical anomaly detection<br>Machine learning based anomaly detection<br>Systems call based sequence analysis<br>Bayesian networks<br>Principal component analysis<br>Markov models | Cyber-Intrusion Detection<br>Fraud Detection<br>Medical Anomaly Detection<br>Industrial Damage Detection<br>Image Processing<br>Textual Anomaly Detection<br>Sensor Networks | (Chandola et al., 2009)<br>(Patcha & Park, 2007) |
| Association rule mining | Relationships between variables<br>*(description)* | Frequent pattern mining<br>Frequent sequence mining | Web usage mining<br>Market basket analysis<br>Medical diagnosis<br>Protein sequences<br>Census data | (Agrawal & Srikant, 1995; Agrawal et al., 1993; Patidar et al., 2013) |
| Clustering | Aggregation of cases to undefined classes<br>*(description)* | Hierarchical clustering<br>Partitioning<br>Grid-based clustering<br>Co-occurrence clustering | Mathematical chemistry<br>Crime analysis<br>Social network analysis | (Berkhin, 2006) |
| Classification | Aggregation of cases to predefined classes<br>*(prediction)* | Neural networks<br>Bayesian clustering<br>Decision tree analysis<br>k-nearest neighbor<br>DNF rules | Target marketing<br>Medical diagnosis<br>Library organization | (Aggarwal & Yu, 1999; Jain et al., 1999) |
| Regression | Identify the importance of variables on a target<br>*(prediction)* | Linear regression<br>Logistic regression<br>Mixed models | Economics<br>Transport<br>Consumer analysis | (Brockwell & Davis, 2002; Ramsey & Schafer, 2012) |
| Summarization | Finding a compact description for the data<br>*(Description)* | Statistics tabulation<br>Data characterization<br>Categorical attribution<br>Clustering<br>Frequent itemsets | Text summarization<br>Transaction summarization<br>Online analytical processing | (Chandola & Kumar, 2005; Han & Kamber, 2006) |

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

To determine what can be incorporated, this process is a constant test of business desires along the

[redacted]

[redacted]

**Table 18: classification of three potential algorithms on Nisbet's method requirements (Nisbet, 2004)**

| Scale (– ; 0 ; +) | Easy to use | Applicable | Scalable | Accurate |
|---|---|---|---|---|
| Apriori | + | -- | 0 | + |
| SPADE | 0 | + | + | 0 |
| TraMineR | + | + | + | 0 |

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

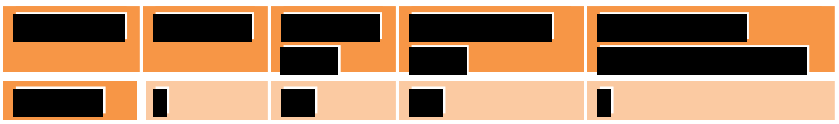
[redacted]

[redacted]

[redacted]

qualitative conclusions, mostly resulting from interaction between the decision makers in team P&S,

[redacted]



### Determine utility function

[redacted]

[redacted]

**Table 20: requirements for a pattern to be actionable**

| | | In the 'successful patterns'- dataset | |
|---|---|---|---|
| | | Exists | Does not exists |
| **In the original dataset** | Exists | Actionable | Not actionable |
| | Does not exist | Not actionable | Not actionable |

If a pattern then is actionable, the value of the potential benefit is determined by the function above. In [redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

- [redacted]
- [redacted]
- Remove loops from page visits
- [redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]



### Test attainability

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

| n-length of sequence | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| **possibilities** | A | AB | ABC | ABCD |
| | B | BC | BCD | |
| | C | CD | | |
| | D | | | |

[redacted]



### Minimum observations check

When the data mining setup with sequence mining did not yet deliver the desired results, the other

### Perform data mining

### Pattern specification

[redacted]

- [redacted]
- [redacted]

[redacted]

- [redacted]
- [redacted]
- [redacted]
- [redacted]



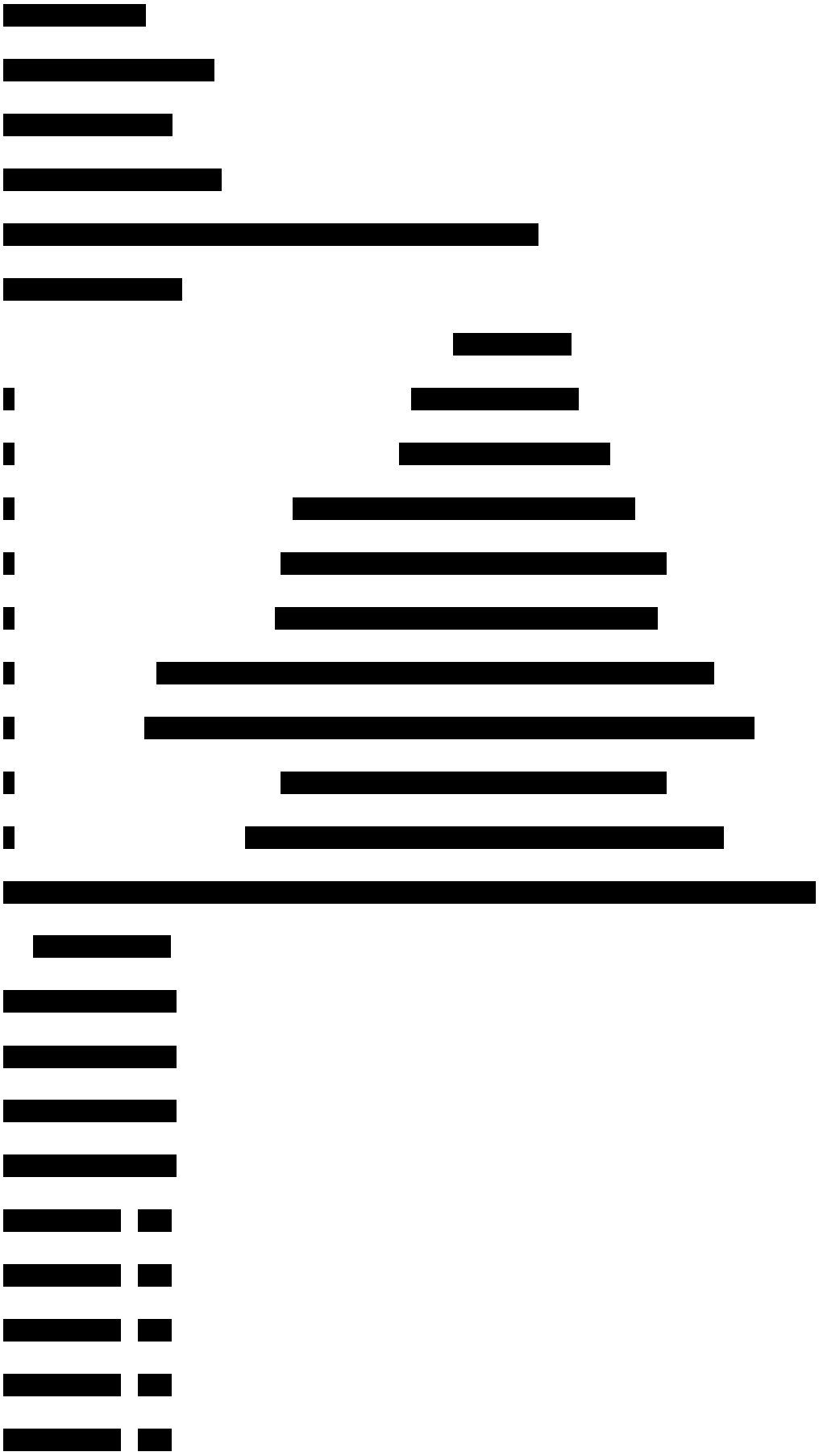
**Figure 47: preview of the transformed data, input data for TraMineR**

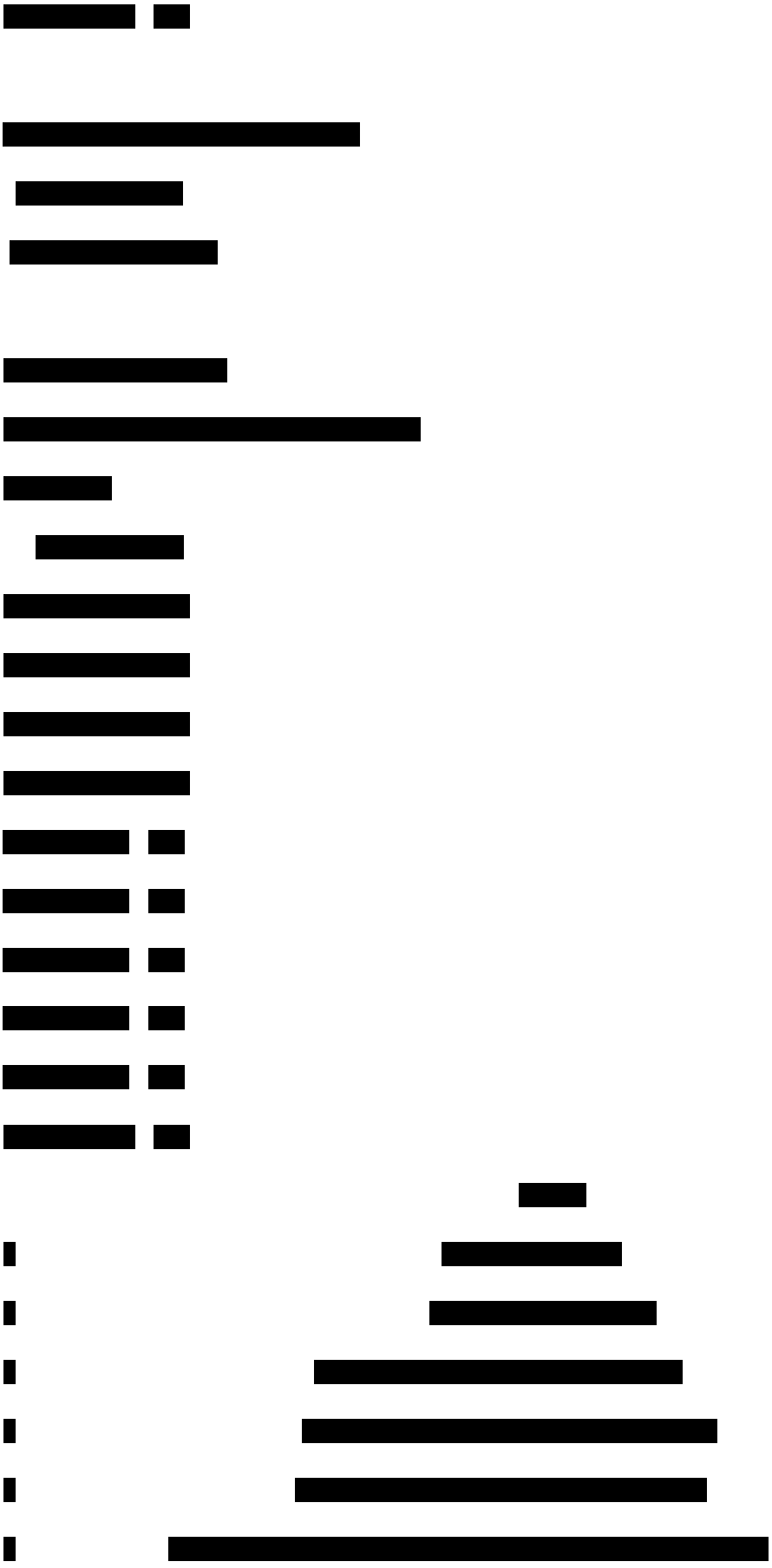
[redacted]

[redacted]

[redacted]

[redacted]





[redacted]

### Model summary

After running $summary(data.seq)$ => 6901 sequences, 4721 unique sequences. This means that

[redacted]

[redacted]

[redacted]

[redacted]

| | Support | Count | subseq | | |
|---|---|---|---|---|---|
| 1 | 0,976153 | 5608 | clp:betalen:index | | |
| 2 | 0,968842 | 5566 | clp:betalen:incassos:overzicht | | |
| 3 | 0,828895 | 4762 | clp:betale | clp:betalen:incassos:overzicht | |
| 4 | 0,616362 | 3541 | clp:betale | clp:betalen:index | |
| 5 | 0,574413 | 3300 | clp:betale | clp:betalen:index | |
| 6 | 0,522367 | 3001 | clp:betale | clp:betalen:index | |
| 7 | 0,506701 | 2911 | clp:betale | clp:betalen:incassos:overzicht | |
| 8 | 0,491906 | 2826 | opp:index | | |
| 9 | 0,489991 | 2815 | opp:polls:poll_particulier | | |
| 10 | 0,473629 | 2721 | clp:betale | opp:polls:poll_particulier | |
| 11 | 0,451871 | 2596 | opp:index | clp:betalen:incassos:overzicht | |

[redacted]

[redacted]

[redacted]

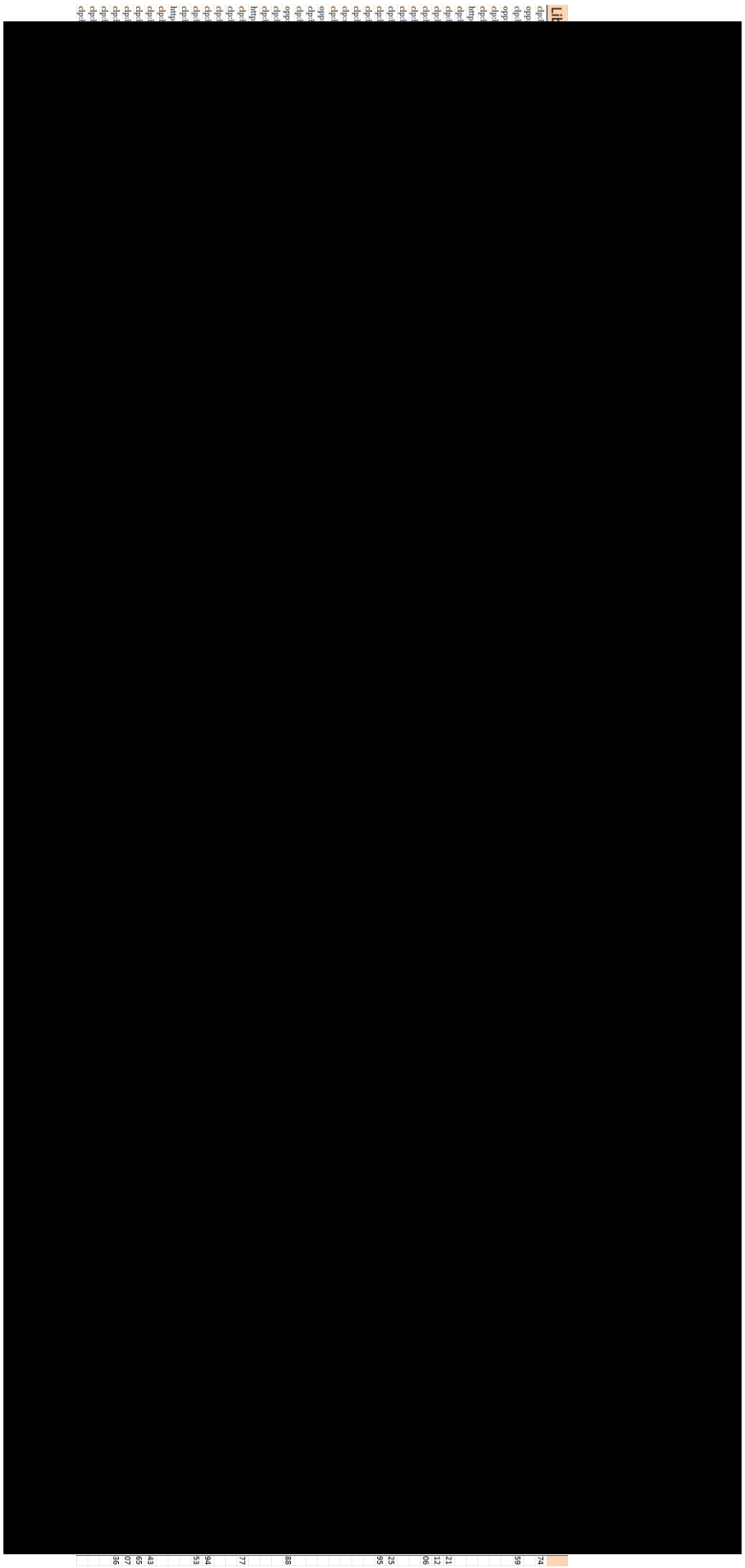


**Figure 49: the library of original and successfull click paths that passed the validity evaluation (snip from large**

## Confidence ratio

[redacted]

[redacted]

[redacted]



[redacted]



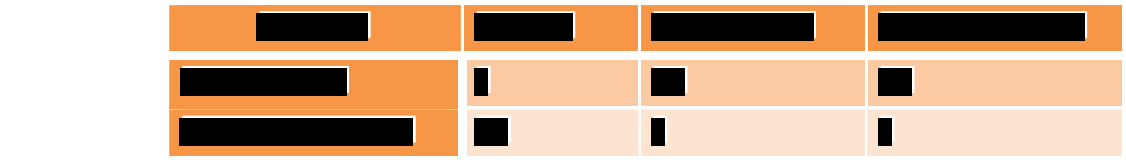
### Utility evaluation dashboard

The



## Sensitivity analysis

[redacted]



[redacted]



[redacted]

[redacted]

[redacted]

[redacted]

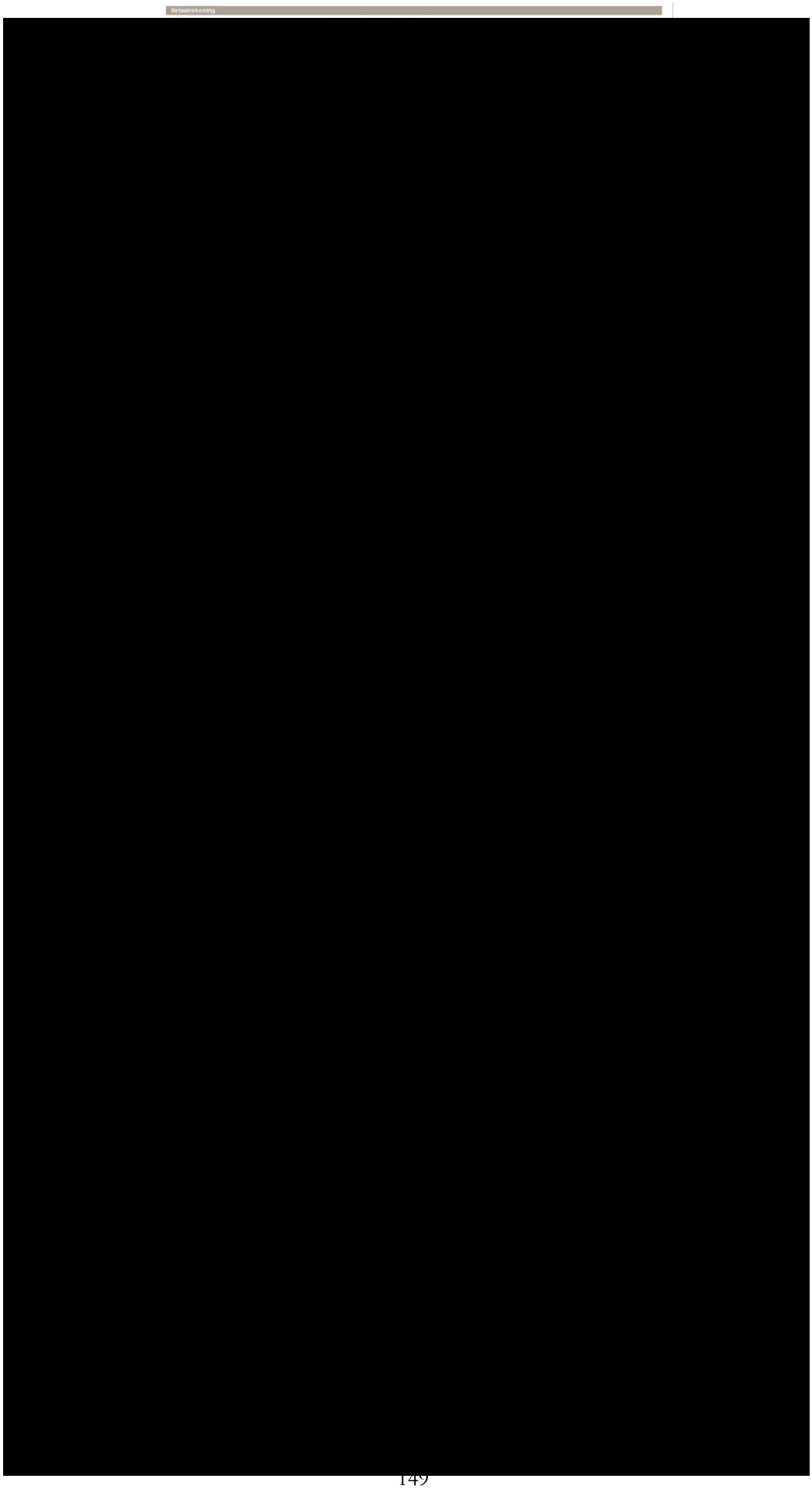
### Interpretation

After these patterns have been specified, the discussion with the decision makers focusses on two [redacted]

- [redacted]
- [redacted]

[redacted]

[redacted]



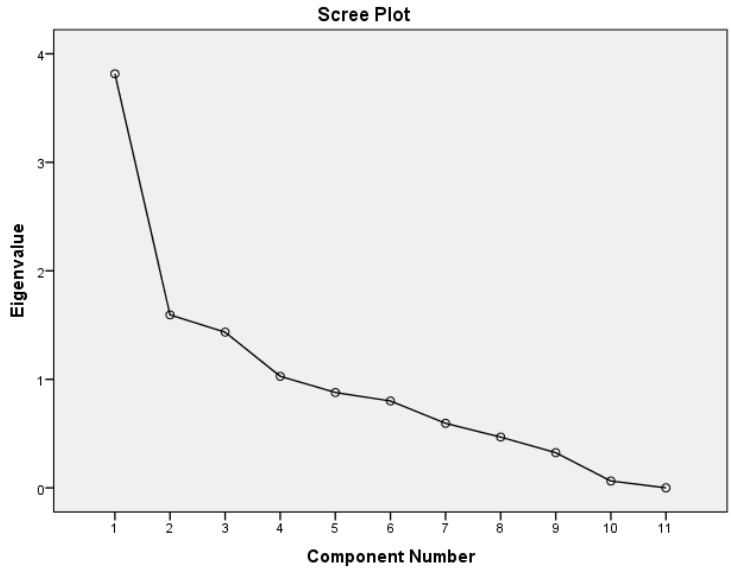
## Novelty evaluation

[redacted]

**KMO and Bartlett's Test[a]**

| | | |
|---|---|---|
| Kaiser-Meyer-Olkin Measure of Sampling Adequacy. | | ,730 |
| Bartlett's Test of Sphericity | Approx. Chi-Square | 2103,284 |
| | df | 55 |
| | Sig. | ,000 |

a. Only cases for which Highutility = 1 are used in the analysis phase.




[redacted]ed Component Matrix[a,b]

| | Component | | | |
|---|---|---|---|---|
| | 1 | 2 | 3 | 4 |
| | ,812 | | | |
| | | | ,529 | |
| | ,966 | | | |
| | ,962 | | | |
| | | | -,659 | |
| | | | | ,822 |
| | ,967 | | | |
| | | ,818 | | |
| | | | | ,702 |
| | | | ,729 | |
| | | ,637 | | |

[redacted] Component Analysis.
[redacted]with Kaiser Normalization.
[redacted]a. Rotation converged in 8 iterations.

b. Only cases for which Highutility = 1 are used in the analysis phase.

**Figure 59: the output of the high-utility factor analysis**

### Low-utility patterns

The set of low utility patterns had a measuring adequacy value of .757, and thus leads to a valid set of factors. The scree plot shows three factors higher than 1, although the rotated factor solution leads to a two factor solution. This two facto solution is more interpretable, and therefore taken into account for further analysis. These two factors are interpreted as follow[redacted]

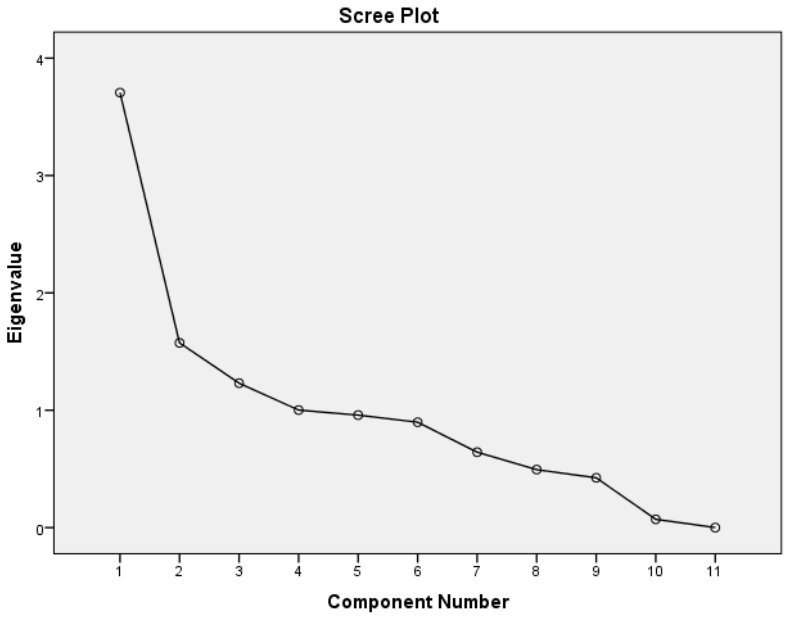
[redacted]

[redacted]

**KMO and Bartlett's Test[a]**

| | | |
|---|---|---|
| Kaiser-Meyer-Olkin Measure of Sampling Adequacy. | | ,757 |
| Bartlett's Test of Sphericity | Approx. Chi-Square | 142485,448 |
| | df | 55 |
| | Sig. | ,000 |

a. Only cases for which Highutililty = 0 are used in the analysis phase.




[redacted]t Matrix[a,b]

| | Component | |
|---|---|---|
| | 1 | 2 |
| | ,777 | |
| | ,964 | |
| | ,959 | |
| | ,964 | |
| | | ,825 |
| | | ,743 |

[redacted]Component Analysis.
[redacted]th Kaiser
[redacted]erations.
[redacted]utility = 0 are used in the analysis phase.

**Figure 60: output of the low-utility factor analysis**

## Interpretation of factors

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

| [redacted] | | [redacted] | |
|---|---|---|---|
| [redacted] | [redacted] | | |
| [redacted] | [redacted] | | |
| | | [redacted] | [redacted] |
| | | [redacted] | [redacted] |

## Design propositions

[redacted]

[redacted]

[redacted]

[redacted]

# Appendix C Validation interviews

The thought experiments focused on the generalizability of the CDE-F. In order to test the generalizability, three project leaders of previous or pending data mining have been interviewed. They all received more or less the same questions, intermitted by an explanation on the CDE-F. :

*What was the goal of the project?*

*Who was involved in the project, and how where they involved?*

*What were the main problems in your project?*

*-(short explanation on the CDE-F)-*

*What elements of the CDE-F do you recognize in your project?*

*What elements of the CDE-F could you have used?*

*What elements of the CDE-F would have caused a conflict in your project?*

The goal of this interview was to discover to what extent the CDE-F could be blueprinted on their projects. Furthermore, special attention has been paid to elements where the CDE-F could not be blueprinted.

The goal of the expert validation is to test the overall acceptability of the framework, and to reflect on the framework given the dynamics of the academic research on data mining.

## Clustering

To validate the CDE-F on a clustering project, the CandY project of data mining at [redacted] savings departments has been selected. Colien Snoep, one of the two project leaders of this project

*What was the goal of the project?*

[redacted]

*Who was involved in the project, and how where they involved?*

[redacted]

[redacted]

*What were the main problems in your project?*

[redacted]

*What elements of the CDE-F do you recognize in your project?*

[redacted]

[redacted] cluster to the other. This could tell us more on the clients' purpose of the savings account.

*What elements of the CDE-F could you have used?*

[redacted]

*What elements of the CDE-F would have caused a conflict in your project?*

[redacted]

## Anomaly detectio

In the fraud detection case, anomaly detection and rule specification has been selected as methods to mine data. Delphine Douchy, a Blackbelt at [redacted] was involved in this project that took place in the Corporate Clients department.

*What was the goal of the project?*

[redacted].

*Who was involved in the project, and how where they involved?*

[redacted]

*What were the main problems in your project?*

[redacted]

*What elements of the CDE-F do you recognize in your project?*

[redacted]

*What elements of the CDE-F could you have used?*

[redacted]

*What elements of the CDE-F would have caused a conflict in your project?*

[redacted]

## Random forest classification

At [redacted] a random forest classification model was made to identify possible new [redacted] Banking customers. For this project, Jasper Boukens was involved with two data scientists and [redacted] Banking employees of [redacted]

*What was the goal of the project?*

[redacted]

*What were the main problems in your project?*

[redacted]

*What elements of the CDE-F do you recognize in your project?*

[redacted]

consensus was reached. Moreover, the data scientist went along with a [redacted] employee to his [redacted]

*What elements of the CDE-F could you have used?*

[redacted]

*What elements of the CDE-F would have caused a conflict in your project?*

[redacted]

## Expert validation

Scott Cunningham, is an associate professor at the policy analysis department at the faculty of Technology, Policy and Management at Delft University of Technology. He published in research on data mining and has made some non-credited efforts in the development of the CRISP-DM.

*What do you think are the most relevant problems for data mining in organizations?*

The main problem is that the potential of data is not fulfilled. This is the indicator of underlying problems on  Apart from the technical problems, the main problem of DM in organization is what I call business discovery. Looking at the CRISP-DM, there should be a feed-forward of business knowledge on data mining and a feedback of data mining discoveries to the to the organization. This concerns the interface between data mining and domain knowledge.

*---(information on convergence)---*

*How do you think that the CDE-F adds to bridge the interface between data mininig and domain knowledge you just described?*

I am really excited that you are thinking about those problems. I recognize the main problems you use in your thesis, and really like the formation of a common perspective on the problem. Although I would like to add two other problems: one is the actionability of results. Even if you converge to a shared understanding of the problem, the patterns that are derived from data mining may not be

directed towards action. A patterns may describe the world to me, but if it does not enable me to make decisions I am not interested in it.

The second problem builds on this example. Often, the desired level of abstraction is not specified for patterns. A pattern that describes the world but is not actionable, may represent an actionable insight on a high-level perspective. Sometimes the decision maker needs an abstracted, high-level view. Not an inaccurate view, but an insight on the systems level. Alexander may talk about this in terms of meta-modelling.

To me, a good model asks for three things: an accurate model, a parsimonious model and you need a theory. An accurate model provides a correct translation of reality. A parsimonious model is efficient in the way it computes and in the results it delivers. A theory is needed, because you need to relate the results to the business problem. You need someone that says: 'Hey, I support this result based on what I experience and believe about the system'. These problems are actually fundamentally unsolved in data science, and is called the model selection problem.

These different worldviews may differ greatly across organizations. I have some background to support this. Imagine they put me in two organizations that deliver consumer goods. These organizations are filled with business experts that have a lot of knowledge on how their world works, but somehow the worldview differs between the organizations and even within the organizations. This is the result of what I think is stickey knowledge, based on experience and observations. These worldviews are not bad, but a method needs to be devised for data mining to comply with these different worldviews.

For me, this is where SEPAM students have added value. Sepam is able to translate knowledge between the domains and make this knowledge explicit. In this way, domain expert and data scientists gain insight in the worldview they need to understand.

*---(information on how to incorporate subjective information in data mining, and introduction of validity, utility and novellty)—*

*How do you view the this explicit triple evaluation approach?*

I like the way you approached this, and I think it is a useful way to structure the different insights that can be gained from a model.

You want to find a sufficient solution that satisfies all requirements, so it should be more regarded as a process of triangulation. You try to converge on something in the middle.

About the novelty part, this is an important expression. There are certain things that you have to put into the model, you have to make some assumptions. In that terms, you want to make sure that what you put into the model, is less than you get out of the model. Think about this as a cost-benefit thing. Akaikes information criterion says something about this. How large is the validity of the model, but at the costs of what parameters is this validity achieved? It is a way to evaluate a model, for instance a rule induction model. Let's say you want 12 rules out of the model. Then you want to get the rule that give you the best results with a relatively little amount of assumptions.

There's something very interesting about novelty that is on people's minds. This novelty might be investigated further.

*Is the CDE-F applicable on all types of data mining?*

In any case, the lessons that are in your framework are valuable for every type of data mining project. It is holistic,

Machine learning in my view is just probabilities with crafts. To me this is insufficient information, I would rather incorporate a degree of conviction, an attempt that describes my worldview and a value that described how to what extent the pattern accommodates this worldview. This might become a little more difficult to apply your framework on, but I don't think you should limit your framework to a given set of applications. It seems generalizable to me. Probably Bayesian learning can add something here.

*How would you position this framework in the current developments of data mining?*

This is an interesting question. You know, developments that contest this, there are some people that state that data scientists should be story tellers. The constant translation between data mining and domain knowledge is then facilitated with soft skills of the data scientist. This idea is not yet crystallized, and is in a very preliminary phase of scientific attention. It is not that I disagree with them, but I am curious to what they mean.

What's nice about this, is that it is a simple solution. You just say: 'okay data mining, this is all on your shoulders', and the data mining should make it happen. If he succeeds, he can be viewed as a Guru, but the Storytelling idea does not yet offer a facilitation of storytelling skills. In the CDE-F you can have a more full-organization discussion about it. This makes the process more complex, but currently more attainable. Besides that, this discussion has some other benefits, such as ownership.