

Henrik Wahlström

APPLYING AN ANALYTICS FRAMEWORK
-AN ORGANIZATIONAL CASE STUDY

Master's Thesis
Information Systems
Supervisors: D. Sc. József Mezei
D. Sc. Robin Wikström
Faculty of Social Sciences, Business and
Economics
Åbo Akademi University
May 2020

TABLE OF CONTENTS

TABLE OF CONTENTS	II
LIST OF ABBREVIATIONS	1
ACKNOWLEDGMENTS.....	2
ABSTRACT	3
1 INTRODUCTION	4
1.1 Research Problem.....	5
1.2 Research Questions	6
1.3 Definitions.....	7
1.4 Structure of the thesis.....	8
2 LITERATURE REVIEW	9
2.1 Analytics	9
2.1.1 Big Data	10
2.1.2 Business Intelligence	12
2.1.3 Data Science.....	13
2.1.4 Data Wrangling	14
2.1.5 Internet of Things	15
2.1.6 Industrial Internet and Industrie 4.0	16
2.2 CRISP-DM.....	17
2.2.1 Business Understanding.....	18
2.2.2 Data Understanding	19
2.2.3 Data Preparation	19
2.2.4 Modeling	20
2.2.5 Evaluation	20
2.2.6 Deployment.....	21
2.3 Developing analytics capabilities in organizations	21
2.3.1 Massive Open Online Course (MOOC).....	22
2.4 Literature review conclusion.....	23
3 METHODOLOGY	25
3.1 Action Research Case Study	25
3.1.1 Data collection.....	27
4 CASE STUDY	28
4.1 Background and context	28
4.2 Identified problems	29
4.3 Data collection	31
4.3.1 Project descriptions.....	32
4.3.2 Interview	33

4.4	Adapting CRISP-DM framework to fit case company needs	33
4.4.1	Business understanding.....	34
4.4.2	Data understanding	35
4.4.3	Data preparation	37
4.4.4	Modelling.....	39
4.4.5	Evaluation	40
4.4.6	Deployment.....	40
4.5	IT artifacts to support data analytics	41
4.5.1	Introduction of a data wrangling software	42
4.5.2	Introduction of a visualization tool for data management	43
4.5.3	Introduction of scripting languages	44
4.5.4	Introduction of a source code editor	45
4.5.5	Introduction of data warehouse approach to DM projects.....	46
4.6	Knowledge development at the case company.....	47
4.6.1	Learning from an organizational perspective.....	49
4.6.2	Individual learning.....	50
4.7	Lessons learned from case study	52
5	DISCUSSION.....	54
5.1	Additional findings	57
5.2	Further research.....	58
5.3	Limitation of research	59
5.4	Answers to research questions	60
5.4.1	RQ1: How can companies improve their data analytics delivery model?..	60
5.4.2	RQ2: Which IT artifacts can support the organizations delivering data analytics services more efficiently?	60
5.4.3	RQ3: What capabilities are the most important aspects for developing data analytics?	61
5.4.4	RQ4: How can learning analytics capabilities be supported from an organizational perspective?	61
6	CONCLUSIONS.....	62
7	SWEDISH SUMMARY - SVENSK SAMMANFATTNING	64
	Skapandet av dataanalysramverk: från ett organisationsperspektiv	64
8	REFERENCES	67
9	APPENDIX.....	71
	Appendix A: Questions included in the interview	71
	Appendix B: Table 8: Template for raw data documentation	72
	Appendix C: Table 9: List of features for Trifacta software (Trifacta 2020)	72
	Appendix D: Figure 12: Case company adaptation of CRISP-DM.....	73

List of Figures

Figure 1: Adopted from (Gandomi and Haider 2015), processes for extracting insights from big data, first introduced by (Agrawal et al. 2011).	12
Figure 2: Profile of skills needed as a data scientist (Aalst 2014).	14
Figure 3: CRISP-DM framework adapted from (IBM 2012)	18
Figure 4: The 4x4 Analytics framework adapted by George & Glady (2014).	22
Figure 5: Conceptual approach to the thesis problem.	23
Figure 6: Design of case study. Figure adapted from COSMOS Corporation.	26
Figure 7: Delivery process in the company's standard project.	29
Figure 8: Company work hours in project A and B. Write-off hours signify budget cost overruns.	30
Figure 9: Case company standard delivery model in comparison to CRISP-DM model and roughly sketched which CRISP-DM activities fall into which case company delivery phases	34
Figure 10: Phases of Analytics learning sessions	47
Figure 11: The forgetting curve (Sydänmaanlakka 2007), originally Ebbinghaus' theory (Murre and Dros 2015)	52
Appendix D: Figure 12: Case company adaptation of CRISP-DM.	73

List of Tables

Table 1: Adapted from (T. H. and J. G. H. Davenport 2017) four eras of analytics	4
Table 2: Description of the projects, units of data collection.	31
Table 3: Roles and responsibilities of company and client in the business understanding phase of DM projects. Adapted from the CRISP-DM model (Wirth 2000)	35
Table 4: Data understanding as-is compared to new proposed way of working	36
Table 5: Software and tools utilized for data preparation.	38
Table 6: Main modelling methods applied in projects	39
Table 7: Software, tools and templates introduced.	42
Appendix B: Table 8: Template for raw data documentation	72
Appendix C: Table 9: List of features for Trifacta software (Trifacta 2020)	72

LIST OF ABBREVIATIONS

AR	Action Research
B2B	Business to Business
B2C	Business to Consumer
BI	Business Intelligence
CBM	Condition Based Maintenance
CRISP-DM	Cross-Industry Standard for Data Mining
CRM	Customer Relationship Management
CPS	Cyber-Physical Systems
DataViz	Data Visualization
DM	Data Mining
ERD	Entity Relation Diagrams
ERP	Enterprise Resource Planning
IoT	Internet of Things
IoE	Internet of Everything
IIoT	Industrial Internet of Things
IT	Information Technology
KPI	Key Performance Indicator
MOOC	Massive Online Open Course
ML	Machine Learning
MS	Microsoft
OLTP	Online Transactional Processing
OLAP	Online Analytical Processing
R&D	Research and Development
SQL	Structured Query Language

ACKNOWLEDGMENTS

I want to thank my supervisors, D. Sc. József Mezei and D. Sc. Robin Wikström for supporting me throughout my research journey. I want to thank my family and friends for trusting in me and motivating me to finish.

I thank everyone at the case company for all the growth opportunities and shared experiences, Mikaela, for making many great memorable experiences possible. Thank you, Viktor, for your valuable time supporting my work while being a fun person to be around. As well, to all my supervisors Eeva, Suvi, Anders and Filip. Thank you, Fredrik, for sharing your knowledge, being a great colleague and friend.

And of course, my gratitude goes towards all wonderful people at information systems, Prof D. Sc. Christer Carlsson, D.Sc. Pirkko Waldén, D. Sc. Markku Heikkilä, D.Sc. Anna Sell, D. Sc. Shahrokh Niko and Eija Karsten at Åbo Akademi University who have sparked my interest towards the subject.

Henrik Wahlström

Turku, May 2020

ABSTRACT

Analytics provides decision makers a data driven approach towards businesses operations, frequently outperforming traditional business practices. Data scientists are the persons creating these analytics solutions and requires a holistic knowledge about the business and operational data. The practical task of the data scientist is to transform the data to correct format, enabling for visualization or algorithmic calculations. Thus, the processing of data in sequential, iterative order is described as data wrangling.

This thesis provides insight into a case study of analytics projects performed in a business organization. A literature review was conducted to find existing frameworks and methods and provided directions of the research towards improving data wrangling, with the CRISP-DM framework and supporting creations of IT artefacts. Action research methodology was utilized to generate quick improvements with both quantitative and qualitative methods. During the research period of four years and completion of six analytics projects, an all-round improvement in analytics capabilities was achieved.

Based on literature review and empirical research new concrete ways of working were introduced allowing for more efficient business operations and project management. Analytics knowledge development strategies are presented that can be adapted into the existing knowledge management program. The findings also function as knowledge mobilization, drawing attention to the issue of data wrangling within the case organization.

1 INTRODUCTION

Analytics is described as the use of statistical, mathematical and computational methods to extract valuable information from data and assist in decision making (Cooper 2012). The evolution of analytics since the mid-1950s has been discussed by Tom Davenport. During the last decades much has happened, traditional spread sheets analytics have evolved to enterprise wide information systems with analytics capabilities. The first analytics (1.0) era was dominated by simple statistical descriptive methods used on structured data. Analytics has been used in many companies very successfully and combined with a successful business wide implementation can lead to competitive advantage (T. H. and J. G. H. Davenport 2017). Davenport describe in his book the recent analytics eras, shown in Table 1, distinguishing the evolution of analytics. As the level of sophistication of intelligence for each analytics era increases, the competitive advantage grows. (T. H. and J. G. H. Davenport 2017)

Table 1: Adapted from (T. H. and J. G. H. Davenport 2017) four eras of analytics

Analytics Era	Analytics 1.0	Analytics 2.0	Analytics 3.0	Analytics 4.0
Type of Analytics	Descriptive analytics	Predictive analytics	Prescriptive analytics	Autonomous analytics

Big data and the emergence of business ecosystems have created a state where vast amounts of data are gathered and accessed by various stakeholders. To approach a higher level of analytics, researchers and practitioners have presented business specific concepts to increase analytics capabilities in respective fields. For manufacturing and industrial context, the concepts “industrial internet” or *Industrie 4.0* have recently been envisioned and promoted both in academia, business and political settings to disrupt the industry and support these businesses in staying competitive in the market. (Pfeiffer 2017)

Technology developments and expanding organizational co-operations have led to a position where data is used on an inter-organizational level and data being created at an immense rate. Organizations have adopted certain kind of analytics methods to process the data, Yet, in more and more instances, data quality and data accuracy has led to poor

decision making (Redman 1998). Research show that, top management understand the importance of data-driven decision making, however the top of the line, market disrupting analytics is mostly embraced by big companies. (T. H. Davenport and Bean 2018). These results indicate that there is room for improvement and that change requires time and commitment from companies.

Data Mining is referred to as the extracting of information and knowledge from small or large databases. (M. Chen and Han, Jiawei; Yu 1996). Data mining as can be understood as an application domain instead of technology. (Coenen 2014). Analytics and Data Mining are in a way related as both have the end goal of finding knowledge from data. Analytics is perhaps a more generic term that is, non- domain specific. This leads to the questions; how should companies improve their analytics capabilities and what are some concrete ways this can be achieved?

Now zooming in on individual cases, it is the analysts, data scientists and other data crunching roles, that require the proper knowledge, skills and technological succeed in their respective positions. Individuals in business and academia agree that majority of the time for analysis is devoted to data preparation phase (Hellerstein, Heer, and Kandel 2018). Therefore, creating more effective data preparation will overall make analytics more effective (Stodder 2016).

1.1 Research Problem

Organizations around the world are phasing the same situations, how to extract value out of data? The challenge becomes even more daunting with Big data. What was once only possible in large organizations with large investments in IT infrastructure, is now possible by individuals or small team of data scientists. Technological advancements have made it possible for persons or small teams of data scientist to process and extract big data. Providing the necessary frameworks, tools and knowledge will enable organizations to stay competitive. In practice, organizations need to adapt to this new norm and transform business practices with the help of analytics. The aim of this thesis is to provide an understanding of the conducted data analytics projects and recognize which framework would support projects in the future. At the same time there was a need to evaluate software that could support the data analytics projects. This thesis was ordered because

there was lack of systematic way to perform these projects, the analysts were working overtime and mistakes were made that could have been avoided.

The research problems can be summarized into:

- Big data trends such as ever largening volumes are hindering the delivery of analytics projects.
- Data cleaning and transformation (or data wrangling) is a manual labor-intensive part of the Data Mining project.
- Limited knowledge about capabilities required to be able to perform big data analytics.

These are practical problems and could be observed at the case organization,

1.2 Research Questions

The previously identified research problems strive to be resolved by answering the following research questions. These questions were identified by thorough examination of the current situation in the case study organization. The problems may also be multidisciplinary as it can be seen and researched from multiple perspectives. Research questions one to four are contributory questions that help answer the main research question.

- Main RQ: How can data analytics capabilities be improved in organizations that deliver data analytics services?
- RQ 1: How can companies improve their data analytics delivery model?
- RQ 2: Which IT artefacts (tools, templates or software solutions) can support the organizations delivering data analytics services more efficiently?
- RQ 3: What capabilities are the most important aspects for data analytics projects?
- RQ 4: How can learning analytics capabilities be supported from an organizational perspective?

The objective of the research is to provide a holistic answer to the main research question. It should be noted that the research questions are in no order. Still, the first and second questions are closely related due to socio-technical interdependencies. The literature review section by itself certainly can provide answers to the research questions. However, findings would be validated with the employees at the case company. Benefits of new ways of working was presented in comparison to previous ways and its limitations. The third and fourth research questions are related as they are about knowledge development. Aspects of learning in this thesis, especially mentioned in chapter 4.6, are strictly presented from an organizational perspective, pedagogic considerations are limited in scope.

In short, the objectives of the thesis can be condensed to:

- Finding a systematic method of performing Data Mining projects.
- Finding IT artifacts to support data analytics, specifically for data analysts or small teams of data scientists.
- Finding the most important capabilities to data analytics

1.3 Definitions

Here I will define the most important terminologies used in the thesis. The primary intention of these definitions is to introduce readers to the topic and the context will be clarified in the research.

Data Mining: Discovery of valuable information from structured or unstructured data, achieved by data visualization, statistical methods, algorithms, organizing or restructuring data.

Data Wrangling: The process of transforming data from raw format into a format valuable for analytics or other purposes.

IT Artifacts: Information technology tools, template, hardware infrastructure, software or any other resources supporting personal or organizational goals.

1.4 Structure of the thesis

Chapter 1, **Introduction**, introduces the motives of this thesis, the research background, problem, and aim.

Chapter 2, **Literature review**, contains literature review of relevant topics of prior research in the field of analytics, major trends that affect the business environment.

Chapter 3, **Methodology**, covers the methodology used in the research.

Chapter 4, **Case study**, presents the case study research of the topic divided into sub-sections.

Chapter 5, **Discussion**, brings forth research answers to the research questions, research related considerations, limitations and research possibilities.

Chapter 6, **Conclusion**, summarizes the research findings.

Chapter 7, **Swedish summary**, contains a Swedish summary of the thesis. Thereafter the references and appendix are presented.

2 LITERATURE REVIEW

This chapter will introduce key concepts that are relevant for this thesis. Topics presented in the literature review are derived from the research questions and these chapters introduces major developments in the field of information technology. More focus is set on topics that are more likely to answer the research questions. In the end of this chapter, the conclusion an assessment is presented if the topics assist in finding solutions to the research problem. Most focus is placed into presenting the methodologies and frameworks available for analytics projects (Chapter 2.2). Concepts and topics presented may relate to one other as is the case of Data Mining and data analytics (or just analytics). The topic presented are influenced by trends that the case company employees encounter with client discussions, therefore the topics presented ought to be relevant from an business to business (B2B) industrial perspective. As an introduction to the literature review, the broad concept of analytics is presented with its wide-ranging definitions and ways of approaching it.

2.1 Analytics

The term analytics has been widely used by professionals and academics with different variations (Barneveld, Arnold, and Campbell 2012). However, the overwhelming similarities among all definitions is the data-driven approach to decision making (Cooper 2012). The following definition is suggested by the EDUCASE Learning initiative:

“Analytics is the process of developing actionable insights through problem definition and the application of statistical models and analysis against existing and/or simulated future data.” (Cooper 2012)

In the analytics definition “actionable insights” point to the idea of a person receiving information that leads to conclusions which are valid and proven by e.g. statistical methods. Analytics can be further divided into (1) descriptive analytics, that uses Business Intelligence tools or Data Mining to understand what has happened, (2) predictive analytics, that uses statistical methods, algorithms to forecast what could happen and (3) prescriptive analytics, that uses optimization and simulation to understand what should be done (Cooper 2012). From this definition Data Mining can be described as finding patterns from data and can therefore be interpreted as a subfield in analytics

By Cooper's definition "standard" business reporting, monthly, quarterly, and yearly reporting are descriptive in nature and are the basis for Business Intelligence. IBM describe that the value created by analytics is growing when we are moving towards predictive- and prescriptive and analytics. The noteworthy difference between predictive and prescriptive is that predictive recognizes patterns, predicts events, however prescriptive analytics recommends actions.

Analytics has been used in many contexts e.g. "web analytics" and "consumer analytics"; in this thesis the most appropriate term would be to use business analytics. Analytics as such does not require (although it might mean) having expensive IT software or data warehouses. The methods or systems used are not in the focus, rather how data-driven decisions are done (Holsapple, Lee-Post, and Pakath 2014). Studies show that firms with business analytics (Data driven decision making) are performing better on multiple domains, mainly in productivity and market value (Brynjolfsson, Hitt, and Kim 2011) (Fosso Wamba et al. 2015).

2.1.1 Big Data

During the last decades, the possibility for storing, processing, and exchanging data has increased significantly. Moore's law has year after year proven correctly as the number of components on electric circuits has doubled every year. A similar observation can be linked to the ever expanding creation of data (Rogers 2011). This evolution has led to the creation of terms such as big data and big data analytics, which have been used in the context to describe storage and analyze of massive and or complex data sets (Ward and Barker 2013). Therefore, special software and hardware is required in order to gain meaningful insights (H. Chen, Chiang, and Storey 2012). Some researchers and practitioners have added some indicators of the data amount analyzed being in Terabytes (ten to the power of 12) to Exabytes (ten to the power of 18), ranging from sensor data to social media data. However, these should only be indicative as when technology evolves the definition will also change. Overall defining it quantitatively is complicated and comes with various epistemological problems (Floridi 2012).

Originally three V's has been used to define big data; Volume, Velocity and Variety. Later Veracity was added to the definition (Ward and Barker 2013).

- Volume - the huge amounts of data gathered every second as transaction-based services are logging activity of users such as Facebook.
- Velocity - the fast phase of data gathering process data, batch data, stream data and real time.
- Variety – Structured, semi-structured and unstructured data
- Veracity – the uncertainty that is added to the data

The definitions have been further broadened to include seven V's (Sivarajah et al. 2017).

- Visibility (or visualization) – Explains the requirements for complete data to gain a holistic information to perform correct decisions.
- Variability - refers to the complexity within the data, as data itself is always changing. Differently expressed, the meaning of the data evolves over time.
- Value – Extracting knowledge/value from the data (events, correlations, hypothetical, statistical etc.)

Authors such as (Baraniuk 2010; Agrawal et al. 2011) describe the phenomenon of data deluge or data tsunami as data being generated at an overwhelming speed leading to problems at organizations and researchers. This can lead to problems such as privacy, network infrastructure limits, algorithms etc. (Provost and Fawcett 2013) reason that, once firms are capable of processing large amounts of data in a reasonable manner, they will come up with new ways of leveraging data and in better ways than today. The authors estimate that, this will eventually lead to a golden era of data science.

(Acito and Khatri 2014) describe the business analytics as a revolution, wherein value is being extracted from data. A perhaps populist view of this same phenomena is obtaining signals from the noise (Silver 2012), while authors point out the problem in big data is the comparatively small signal to noise ratios. In small-scale datasets the noise, such as outliers and errors are easily observable. The opposite is also true, predictive models using large amount of data can be affected by the increased signal to noise ratio. Reducing the signal to noise ratio has proven in multiple cases to improve predictions based on big data. (Hassani and Silva 2015). What could be emphasized here is, to understand the increasing data sizes might hide noise that is simply apparent.

Data by itself is useless, however the potential is unlocked once applied to decision making. (Gandomi and Haider 2015). The same phenomena was noticed by the mathematician Clive Humby (Humby 2006) and later journalists noted that the most valuable resource of the world is no longer oil, rather it is data (Economist 2017). The crude oil analogy works with data as well, oil by itself is not that valuable, however when refined into products they become valuable. The process of gaining insights from big data can be summarized in Figure 1, in which the distinction of data management and analytics should be noted.

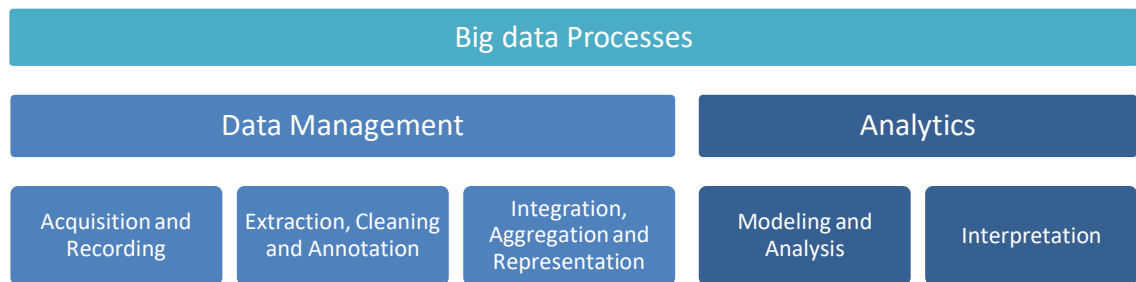


Figure 1: Adopted from (Gandomi and Haider 2015), processes for extracting insights from big data, first introduced by (Agrawal et al. 2011).

Due to the extensive technological development in the big data era and demand for business domain expertise, particularly in decision making, leads to new demands from the practitioners.

2.1.2 Business Intelligence

The term Business Intelligence (BI) was first mentioned by Hans Peter Luhn in 1958 defined by the two components Business and Intelligence (Luhn 1958). BI became popular in the 1990s and 2000s when it became a umbrella term to describe concepts and methods to support business decision making (Tutunea and Rus 2012).

Databases have become standard tools for professionals in almost every field. A fundamental building block of information systems is relational databases and transactional processing, of which the most common language is SQL (Structured Query Language). Traditionally OLTP (Online Transactional Processing) databases are in use for operational applications, which consist of everyday transactions. Data warehousing is a collection of decision support systems and has pushed the focus more on data driven decision-making. OLAP (Online Analytical Processing) operations are query searches

that consolidate data. Decision makers and knowledge workers are using OLAP data to create a summarized view of an operation, business, industry or alike. OLAP queries have become a cornerstone in Business Intelligence (Vassiliadis and Marcel 2018) (H. Chen, Chiang, and Storey 2012).

A newer concept for data storage is data lakes. These can be described to be more agile approach into storing heterogeneous data. The advantage is in having a curated and protected pool of data for users to access that does not necessary require the IT resources than a traditional data warehouse. Data lakes can have both enterprise internal structured- and external unstructured “raw” data (Terrizzano et al. 2015). Data lakes are also a way for enterprises to combine siloed data into one central data lake. Silos of data are a natural outcome after mergers and acquisitions or retired business data architectures (Chessell 2014). Certain data lakes architecture support low latency processing, needed for real-time or near real-time data processing, supporting big data analytics of time-sensitive solutions (Miloslavskaya and Tolstoy 2016).

2.1.3 Data Science

The recent development in the information society has led to a rapid growth in demand for data scientists. Receiving publicity in the press by being coined the “*The sexiest job of the 21st Century*” (T. H. Davenport and Patil 2011). The definition of data scientist has emerged as disciplines such as mathematics and computer science have combined into what is today known as data science. As seen in Figure 2, data science is a multi-disciplinary field and many skills are required to perform successfully (Aalst 2014).

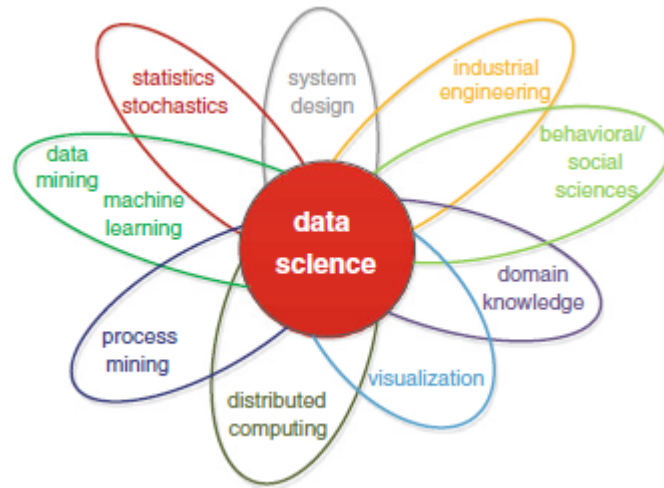


Figure 2: Profile of skills needed as a data scientist (Aalst 2014).

The central message in this profile is to grasp the extensive knowledge required. Employers and academics are now realizing that these people also require the skillset to communicate their ideas with others and apprehend complete solutions. As technology is changing at fast phase, so are the algorithms as well as demands for computer science skills. The role of distributed computing gains importance as datasets sizes increase and processing of data transitions into real-time where executions need to occur in split seconds. Process mining meaning being able to analyze past process data to find patterns. In 2013 the data science center Eindhoven was created and has focused their research around the following themes.

- Internet of Things: Gathering data
- Data analytics: Turning data into information
- Understanding and influencing human behavior

Another fundamental concept of data science is the intention of creating analysis that in that consequently can be sold as products or services.

2.1.4 Data Wrangling

In the big data age, the processing of large data sets has become increasingly more time- and resource consuming. (Endel & Piringer, 2015) In order to perform an analysis the data must be gathered from various sources, aggregated, derived and cleaned (Sean Kandel et al. 2011). These researchers have noticed the amount of time data preparation consumes and therefore researched ways to improve the data transformation process.

Introduction of a predictive interaction framework for data wrangling has even helped non-technical businesspersons in preparing data for analysis.

Data wrangling is roughly described as the transformation step in the data preparation phase in the CRISP-DM model (see chapter 2.2). (Sean Kandel et al. 2011) tested the usefulness of the Wrangler software in comparison to Excel. In three tests performed by 12 data analysts', the wrangler proved to be more efficient, although the persons had no previous knowledge of the wrangling software. It is further pointed out that data wrangling is iterative and a multidisciplinary process. (Endel and Piringer 2015) The authors mention that there are multiple solutions available to address wrangling tasks, however there is no one solution providing an end-to-end wrangling process.

Desktop studies showed that script-based programming languages such as *R* and *Python* provide data wrangling tools. Out of these two, *Python* appears to be all around great solution as it is easy to learn and there are add-on packages that extend the capabilities of the language. (Ayer, Miguez, and Toby 2014) One of these packages is *Pandas*, an open source data analysis project that aims to be “*the most powerful and flexible open source data analysis / manipulation tool available in any language*”. (Wes McKinney 2016) *Pandas* achieves this by utilizing data structures, called DataFrames. These can overcome the problem of memory errors. *Pandas* offer ways of manipulating the contents of tables in a flexible way. For example, “dropna” function is used to drop all rows in column which have missing values and the missing values can be filled with a specified value. The *R* language also offers other data wrangling packages such as *Dplyr*; for deriving new columns from previous columns, filtering rows, arranging rows, adding new columns. (Wickham et al. 2015) *Tidyr* is another package of tools and methods to create tidy data. Among these are the ability of pivoting, nesting, unnesting of data and turning data into rectangular data frames. (Hadley and Henry 2018)

2.1.5 Internet of Things

Global megatrends such as globalization, digitalization, development of Moore's law has led to significant cost reduction of electrical components, which in turn has led to products and services for the consumers incorporating internet connection to electric devices. Smart devices cover three elements: (1) physical device, (2) smart components and (3) the ability to connect. This has led to consumers receiving a great deal of added value to

their products (Xu, He, and Li 2014). Different industries have noticed the potential for this process of information creation. Companies have three main ways to incorporate the possibilities of this technology: (1) increase the performance of the current business (evolution), (2) create totally new business operations (revolution), or (3) create added value to their products (Juhanko et al. 2015).

The majority of industry experts still agree that the Internet of Things (IoT) technologies and applications are in need for improvements, more specifically in the area of technology, standardization, security and privacy (Xu, He, and Li 2014). Despite the complications, visionaries are looking towards the future. The networking hardware company CISCO and semi-conductor company Qualcomm is promoting the term Internet of Everything (IoE), not just include things, but to include people, process and data (Miraz et al. 2015). Analytics of IoE data becomes more effective as new analytics frameworks, new data models and methods are being applied (Mishra 2018).

2.1.6 Industrial Internet and Industrie 4.0

Industrial Internet of Things can be defined as “*...the use of IoT technology in manufacturing*” and Industrial internet can be defined as “*...industrial applications of IoT, or IIoT*” (Boyes et al. 2018). The industrial internet as a concept is speculated to combine the benefits of both the industrial revolution and the internet revolution (Qin, Chen, and Peng 2020). The research institute of the Finnish economy has conducted a report of the industrial internet and its effects on the future (Juhanko et al. 2015). The Industrial internet is seen as a opportunity for new markets and innovations on a global scale. Parallels can be derived from Internet of Things in the consumer markets, where this has lead to innovations that have revolutionized our daily lives.

Traditionally, organizations have pursued creating siloed systems with technologies such as firewalls in order to retain secure communication channels (Boyes et al. 2018). Now, technological advancements in computing power, cloud technologies, IT platforms etc. has lead and will lead to new innovations in industrial settings. Technologies are disrupting how communication flow within manufacturing operations occurs, for instance blockchain technology could provide a solution for ever increasing amount of IoT connected devices within a industrial setting (Wan et al. 2019).

New solutions are assumed to disrupt traditional business models and there are examples of these already in use, such as Condition Based Maintenance (CBM) (Muhonen, Ailisto, and Kess 2015) and fleet management (instead of monitoring one equipment a whole fleet of equipment is monitored. (Backman et al. 2016). Different The industrial internet concepts contains technological solutions such as cloud computing, wireless factory networks, sensor technology and in-house network security systems.

In 2011, the German government coined the term *Industrie 4.0* and set a top priority research agenda which is set to push the fourth industrial revolution, refering to convergence of Smart technology, IoT technology and Cyber-Physical Systems (CPS). (Bhuskade 2015) Strong focus lies in monitoring industrial systems from a top down perspective in order to optimize production and automate interoperability between systems. Machine-to-Machine (M2M) platforms play a substansial role in this aspect, as these provide a communication network for systems, that are in the manufacturing context assembly line robots, material transport systems, products themselves etc. The trend in production monitoring is to provide real time analytics to react as fast as possible to any disturbances in the manufacturing processes (Pfeiffer 2017). Vendors such as *Microsoft*, *Oracle* and *Thingworks* are providing these types of platforms and due to different requiremets of the platforms different end purpose is given. It is estimated that in the future the winning platforms will create ecosystems that will dominate the market. Additinally, plug and play principle to modularity is signified as important when designing flexible systems, to provide the means for manufacturing products according to changing demand.

2.2 CRISP-DM

The standard business process model for applying Data Mining has become the CRISP-DM, abbreviation for Cross-Industry Standard Process for Data Mining (Wirth & Hipp, 2000). The model was especially created to be an application-neutral model open for everyone from academics to practitioners. The purpose for this manual is to guide data miners through the whole data project lifecycle. The model includes six phases, (1) business understanding, (2) data understanding, (3) data preparation, (4) modeling, (5) evaluation and (6) deployment. These phases are further hashed into specific tasks, and deliverables (or outputs) for each task are specified.

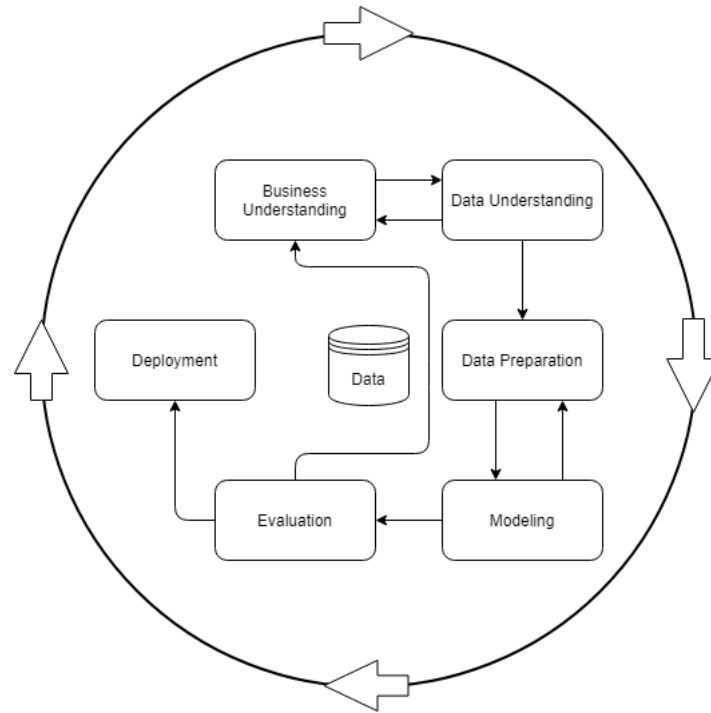


Figure 3: CRISP-DM framework adapted from (IBM 2012)

CRISP-DM methodology has its roots in KDD (Knowledge Discovery in Databases) which includes five stages; Selection, Pre-processing, Transformation, Data Mining and Interpretation/Evaluation. SAS Institute has also developed a framework for DM: SEMMA (Sample, Explore, Modify, Model, Assess). In this thesis the focus will be on CRISP-DM as it appears to be the most useful in guiding users to apply DM in real world projects (Azevedo and Santos 2008).

2.2.1 Business Understanding

Business background should describe the business organizational structure, to identify the key individuals in the organization. Clearly defining internal sponsors will ensure the required support will be received, both in terms of knowledge and financial support. The second phase of business understanding is to identify the business objective. Describe the goals of the project, which can be simple objective criteria e.g. creating a model to predict next year sales. More subjective goals would be discovering reoccurring patterns in data. These critical success factors should be clearly defined to prove all stakeholders the success of the project. Assessing the situation includes examining which resources are available, resources meaning available personnel available (business experts, data experts, Data Mining experts etc.), computing resources (hardware and software) and data

(access to data, data extracts or operational data). Additionally, some information needs to be presented if any current solutions are in use for the current business problem. All this information should then be reported in a resource inventory report. (IBM 2012)

2.2.2 Data Understanding

Throughout this phase the focus is on exploring the data, describing it and finally writing it in a report format to the project documentation. First the data needs to be gathered from various sources external or internal, also referred to ‘collect initial data’. The authors point out the importance in this step to understand which attributes in the data seem promising for analysis. Moreover, the data should be evaluated on missing values and how these affect the results. Value types should be reported that explains how the data is represented, in which format, numeric, categorical, or Boolean. Output for this subtask is a data description report. In the SPSS data guide this information should be included in a data collection report. Data understanding and business understanding are closely related and therefore working with these two together is recommended, if not necessary. (IBM 2012)

2.2.3 Data Preparation

According to IBM CRISP-DM guide the data preparation step consumes between 50-70% of the project time. (IBM 2012) In order to make this step as efficient as possible, enough time and effort should be devoted to the business understanding and data understanding part. Steps in the data preparation step include merging and combining data, deriving new attributes, removing, or replacing blank or missing values. The goal of this is to create the dataset in its final form to be used in the modeling phase. Selecting data involves assessing which items (rows) and attributes (columns) are of value.

Cleaning data defines the act of analyzing data and deciding actions for the following faults; missing data errors, should it be left blank or derived new value. Coding inconsistencies exist because of different datasets have different names for some values. Where there are clearly faulty symbols or inputs. Missing or bad metadata will confuse users and might lead to misunderstandings, these should be added or modified if possible. Constructing new data can include creating new data rows or columns, based on information within the data. Integrating data can be done by two processes, merging or appending data. Merging data involves adding more columns by integrating data with same key identifiers, such as customer ID. Appending data describes the action of

integrating data where it adds cases or observations. Sometimes, depending on the analytics case, formatting of data is required or preferred. Certain algorithms in the modeling phase might require this or becomes less demanding computationally. If Machine Learning algorithms are utilized, then the data is split into training and test data sets. After the data preparation phase the data should be in a tidy format ready to the next phase. (IBM 2012)

2.2.4 Modeling

At this point is where the previous work begins to show results, this is where the analytics methods are chosen. These could be mathematical models such as regression models, Machine Learning algorithms, neural networks or visualizations. Data Mining objectives can usually be answered in a variety of ways, nonetheless it is the data scientist duty to select methods that fits the needs of the objective and the data. Seldom the correct choice of model is found instantly, more likely experimentation of different methods ensues. If the methods selection proves non-functional, the data miner has the option of selecting alternative modeling methods or fine-tuning parameters within the existing model. During this phase, some initial conclusions can be drawn regarding effectiveness of the Data Mining, as well as reveal if returning to tweak the data preparation is needed. Throughout, the modeling technique and modeling assumptions need to be specified and all parameters documented, for anyone replicating the procedures to achieve similar results as the model creator. The end goal of data modeling is to have accurate and relevant results. (IBM 2012)

2.2.5 Evaluation

During evaluation phase the results are evaluated according to the business success criteria. The results are then logged and reported. Attention should be given to the way the results are presented, such as unique findings that need to be highlighted and any new questions being raised from this newly generated information. Finally, a decision should be made to proceed to the next step or return to previous steps and refine or replace model. If the model passes evaluation criteria and satisfies the business goals, it is time to progress into deployment phase. (IBM 2012)

2.2.6 Deployment

The final stage includes deploying the solution in the organization. The extent of the deployment depends on the business use case and the end-users. The solution can be for example enhance a prior data warehouse or rolled-out in a BI platform. Solutions can be stand-alone or integrated into pre-existing or new business processes.

In a full *monitoring and maintenance deployment plan* the changes in circumstances are taken into consideration and decided the appropriate actions in case these occur. Typical changes are changes in database structure, models become outdated because of newer data or the business issue has expired. Producing the final report includes documenting the whole Data Mining process and its connection to the original business problem. While generating the reports, the creators need to take into consideration to whom the documentation is directed, non-technical end users, technical experts or sponsors in the project. Reviewing the DM project should include impressions from business, lessons learned from the CRISP-DM process or from the project in general. (IBM 2012)

2.3 Developing analytics capabilities in organizations

Substantial effort in current research has focused on the technical aspects of big data analytics. Less efforts have been in conceptualizing elements required to add business value to companies. Findings suggest that in order to improve data analytics capabilities, there are four aspects in the human skills and knowledge resource types, namely Technical-, business-, and relational knowledge and business analytics (Mikalef et al. 2018). These coincide with the theoretical framework by (Martine, George & Nicolas 2014) in order to understand the skillset required in a dynamic business environment. The four skillsets are depicted in a circle in Figure 4: Analytics, IT, Business and Communication. The analytics skills include the ability to perform quantitative analysis, understanding performance management etc. IT knowledge covers technical understanding and skills of data management, database design, IT-programming and BI Tools. Business encompasses the understanding of business environment, making advice that is actionable, business standard practices (including urgency, flexibility, relevancy

etc.). Communication includes the skill to efficiently communicate and have appropriate people skills to communicate with decision makers. (Martine, George & Nicolas 2014)

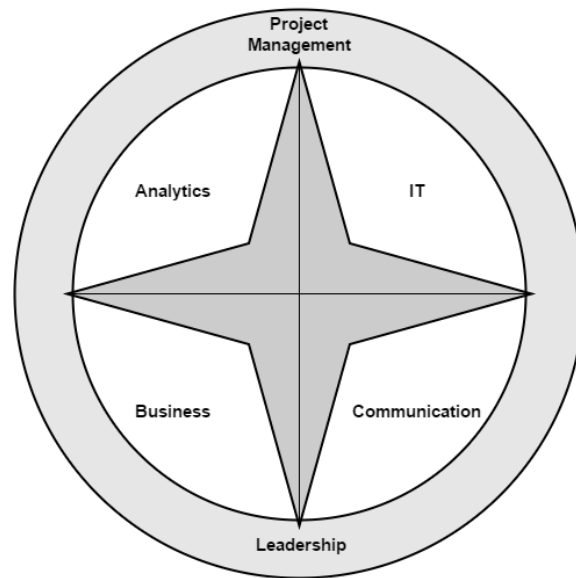


Figure 4: The 4x4 Analytics framework adapted by George & Gladly (2014)

Beside the four skills (shown above) are required as well as four analytics practices within organizations are required; sponsorship (1), recruitment (2), Talent and Knowledge Management (3), and Cross-fertilization (4), hence the name 4x4 analytics framework (Martine, George & Nicolas 2014). Sponsorship is the form that management is convinced by the power of analytics to have an impact on the performance of the organization. In the recruitment practices candidates are often selected by their technical and analytical skills and less emphasis is put on business and communication skills. Thirdly the data scientist position is often occupied by a person with higher education, PhD or similar. The management of talent and knowledge specifically requires keeping the persons intellectually engaged. Career path of a data scientist is regarded as an expert position and the management should consider this when planning career advancements possibilities. Working with diverse teams will lead to cross-fertilization of talents in organizations. This will expand the knowledge of business analytics

2.3.1 Massive Open Online Course (MOOC)

Today it is possible to learn most data science related topics online. E-learning has been possible in the past decades however, during the last years Massive open online courses (MOOC's) have become a disruptors in the education scene (Pappano 2012). These free

courses are offering knowledge in from disciplines such as social science, economics, business, science, healthcare, math or humanities, thereby providing domain specific knowledge. The business model is highly scalable and therefore can offer learning to thousands of people. Providers such as Coursera, edX, Udacity have portals serving the learners courses which can be accessed anywhere every time with a device with internet access.

In a survey of why people enroll in MOOC's 50,1 % of the recipients responded "Curiosity, just for fun" and 43,9% responded "gaining specific skills to do my job better". (Christensen et al. 2013) Therefore it seems that people with jobs are taking these courses to further improve their skillsets, which is also evident in the survey as 50% of the respondents are full time employees. MOOC's are known to have a problem with dropout rate, while thousands enroll in the courses only small proportion of learners complete the course. Another recognized problem is cheating such as CAMEO, the act of Copying Answers using Multiple Existences Online, which is problematic for the integrity of the MOOC providers certification program. (Northcutt, Ho, and Chuang 2016)

2.4 Literature review conclusion

Because of the literature review and discussions at the case company the thesis problem can be approached by three conceptual ways described in Figure 5, to overall improve the company analytics capabilities. Described in detail in Chapters 4.4, 4.5 and 4.6.

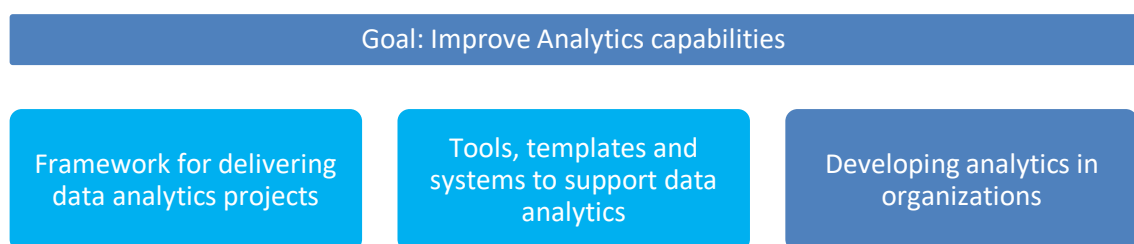


Figure 5: Conceptual approach to the thesis problem

According to the project team, the most problems emerged in the data understanding and data transformation process. This was in alignment with the literature review findings. Due to the importance of data preparation, considerable empiric resources were focused on that specific topic. The areas presented in the figure above is suspected to affect each

other, researching and improving any of the three areas will in improve the analytics capabilities. Finding suitable frameworks, one stood above the rest CRISP-DM seems to be an industry accepted model and has large numbers of citations.

Another interesting finding are the diverse word choices for the same data analytics tasks. In principal, Data preparation, data transformation and data wrangling describe the same process, although the scopes of the phases might slightly vary. Perhaps the different wording depends on the target group of the research authors.

There are multiple ongoing major trends affecting business', Big Data, IoT and data science, among others. The trends introduced are the ones commonly referred to in information systems research. It is unclear which of the trends will have largest impact on the case company business. The relevance of each topic will also shift depending on the strategic direction of the case company's clients.

3 METHODOLOGY

This chapter outlines which research methods are used in this thesis. Overall, this thesis work will be of mixed type, both quantitative and qualitative research methods are applied. A more detailed breakdown is presented below in sub-chapter 3.1.

3.1 Action Research Case Study

Information systems research is an interdisciplinary field, combining information system component with people. Case study research method has proven particularly valid for information system research, as focus has shifted from organizational perspective opposed to technical issues. (Benbasat, David, and Mead 1987) (Klein and Myers 1999). One of the most cited criteria for case study research is presented in the book *Case Study Research: Design and Methods* by Yin (Yin 2003) and its multiple editions. Case study research methodology is suggested if (a) research is meant to answer research questions such as “how” and “why” (b) one cannot manipulate the behavior of those involved (c) contextual matters are relevant to subject of study (d) boundaries between phenomenon and context are not clear. (Baxter and Jack 2008). Case study research certainly has its disadvantages, such as scientific generalizations is difficult (Yin 2003).

The circumstances for the study justify a single-case design. The context of the research is both rare and unique. The case company is a research-based management consulting specialized in B2B industrial business. Client projects are the objects of research and the ones described in this thesis can be regarded as representative of Data Mining projects. Although the projects described in this thesis are new to the case organization, it is apparent that similar projects are done in many organizations. The design of the case study is visualized in Figure 6, presenting an embedded research design where the context is the case company, case is data analytics services and embedded unit of analysis are the data analytics projects. The study contains both qualitative and quantitative research methods, both of which are valid research methods in information systems. (Kaplan and Duchon 1988). Information systems research known for its interplay between positivistic and interpretivism approach towards research (Pather and Remenyi 2004). This certainly affect this thesis, as well as the research methodologies, approaches to conclusions and choices of references.

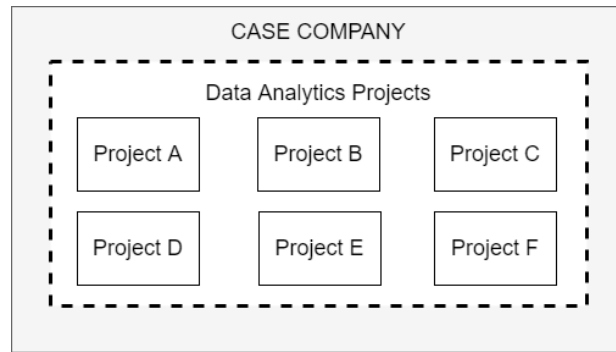


Figure 6: Design of case study. Figure adapted from COSMOS Corporation

The research project was conducted as an Action Research (AR) where the author was working first as a researcher and then hired to work in the case organization. The role of the Action Researcher is being an objective active participant, while being self-reflective, critical, and having a systematic approach to the research. The Action Research model consists of phases (1) *plan*, (2) *action*, (3) *observe* and (4) *reflect*, within continuing cycles of action. These align with analytics project lifecycles where *plan* is considered sales phase of scoping the project, *action* equals delivering the projects, *observe* seeing how actions affect the client and *reflect* being the lessons learned (Coghlan 2019). The Action Research method allows for obtained knowledge to be implemented quickly, while providing a iterative process of connecting theory and practice (Baskerville and Wood-Harper 1996). Previously presented case study research can be categorized as explanatory, where answers to questions are presumed to be too complex for surveys or experimental strategies (Baxter and Jack 2008). In summary, Action Research from an insider perspective is valuable as it generates useful knowledge about how change and is perceived and how change is managed (Coghlan 2003).

Trust is an important factor in organizational success, especially encompassing process change (Morgan and Zeffane 2003). The research for this thesis began with observation of the current practice of performing analytics, hereafter findings were implemented incrementally to build up trust with the employees at the case company. Perhaps, the longer into research the more social factors benefitted the overall goal due to extended interpersonal connections with the employees. In contrast the objectiveness of the research is in scrutiny as biases from the organization might affect the researcher (Norris 1997). Judgements on biases in assumptions and views are challenged by supervisors of the work, self-criticism and introspection is practiced by the researcher.

3.1.1 Data collection

Most of the research was of observational research, by working closely with employees at the case company as the author was involved in the end phase of both the projects. The observational setting involves often working in pairs or with three persons together to solve mostly data related problems. Included are projects completed between 2015 and 2019 and the unit of analysis in this thesis are the data analytics projects. Included in are projects that fulfil the definition of Data Mining or analytics, in other words data was crucial for the project and required specific knowledge to deliver the project. During the time at the case company the author did observations, wrote down meetings memos and held discussions to use them as reference in this thesis. Most of the communication was held face to face with employees or via internal e-mails. Project management documents before, during and after projects delivery where utilized. Sales material with references where created after projects where delivered and proposals where created of analytics services as they were requested by the clients. The documents referenced in the thesis are referenced to further prove a point and using only written knowledge would limit observations at the company. Statements that directly originate from employee at the case company are mentioned by citation.

Additionally, a one-hour semi-structured interview was also held with one of the senior employees at the company. This interview was held in Swedish and was recorded and transcribed and shared with the company CEO. Findings of a company personnel climate survey was included. To add more context to the research the company background is briefly presented in chapter 4.1. In addition, company internal documents were used when appropriate, especially in chapter 4.6, The work hour analysis presented in *Figure 8: Company work hours in project A and B*. Write-off hours signify budget cost overruns and was exported from the organization's ERP system.

4 CASE STUDY

This chapter describes the research done for this thesis. First the case study company (hereafter company) and projects are presented to add context to the research. Thereafter three different ways to approach the research questions are presented. All three ways seem to support the aim of developing analytics capabilities.

4.1 Background and context

The case company's previous data projects have been limited in size and scope. However, during the last years the projects have become more complex as more data sources were included in the analysis. Two major cases demonstrated the company's ability to work with data projects. The following chapters will describe observations and discussions within the company. It is of interest to both the consulting company and client that the analytics process is as coherent as possible. Structuring the project into a framework ought to increase the likelihood of projects being delivered within resourced time and budget. If the project structure becomes comprehensible, the communication and cooperation in turn will become easier. This in turn leads to more satisfied customers. All these factors make it that projects have better margins and continuation of co-operation is more likely.

The company in which the case study is undertaken is a research-based management consulting firm, providing services to B2B industrial firms, specialized in project based business. The company is divided into sub-organizations: (1) strategic consulting, (2) analytics (previously diagnostics) and (3) Research and Development (R&D), and division of labor in the company is quite imprecise, therefore persons working for one division might also work for the other, or all three. The standard way of operations is ongoing communication with business owners and top management, and the services aim to resolve problems or provide knowledge to clients. The services are sold on a project basis, although certain customers have annual contracts. Thereafter the work can be divided among personnel, with diagnostics mostly responsible for running the operations side of projects. Diagnostics is also in charge of daily, weekly and monthly operations. Later, during the research, the consulting and analytics departments were merged.

The R&D department is partly outsourced to a University department. Close co-operation makes it possible for conducting research that is relevant both to academia and real-world business problems. Findings in the R&D department can be piloted in client projects. The research department having a good reputation has led to cooperation with respected universities around the world.

As there are less than 20 employees at company, the core competences of each employee in a way define the offering of the company. Introducing a new field of knowledge i.e. data analytics to the company will expand the offering and at the same time broaden the perspectives of the knowledge workers. The company focus since its beginning has been to focus on long term partnerships. Benefits of this approach is to develop a real understanding of the client businesses and to see their real needs. Therefore, finding possibilities for improvement in client business is easier. The main business focus is in marine & logistics, energy, and mining industries. To its current date the company has completed more than 500 assignments (Anonymous 2019).



Figure 7: Delivery process in the company's standard project

The company uses a standard project lifecycle view to conduct business. Each project phase has defined milestones and outcomes. Sales phase consists of understanding the clients's main problems, what they want to achieve and ends with approval of project proposal. After sales comes handover to project delivery team and after delivery comes project closing. Sales personnel are usually somehow involved in project delivery phase e.g. in advisor role or account manager purposes. Projects are followed-up on weekly or even daily basis on standard PM metrics; time, cost and quality.

4.2 Identified problems

Arriving at the company I was introduced to the situation with two ongoing projects. The first weeks I took a more observing role to identify the problems from the company context and asses the current situation within the projects. After the initial introduction to

the company and incrementally comprehending the way of working I started actively collaborating with two employees in performing similar data wrangling tasks.

I came to the same conclusion as the two employees that the data wrangling (i.e. data understanding and data preparation) was tedious, leading to complications in the project delivery and quality. Simultaneously, it was also reported by the employees; *“during two to three years the volume of data increased in projects receiving larger data sets from clients”* (Wikström, Sundholm, and Wahlström 2015), coinciding with the big data trends presented in the literature review section. The two employees assigned to the project were self-taught in transforming the data and gave an impression in having a limited knowledge about best practices. The two employees were actively collaborating in running the Data Mining process from and both agree that the most time-consuming part of projects is understanding the data and preparing the data into correct format. The data understanding, and preparation were done mostly in Excel.

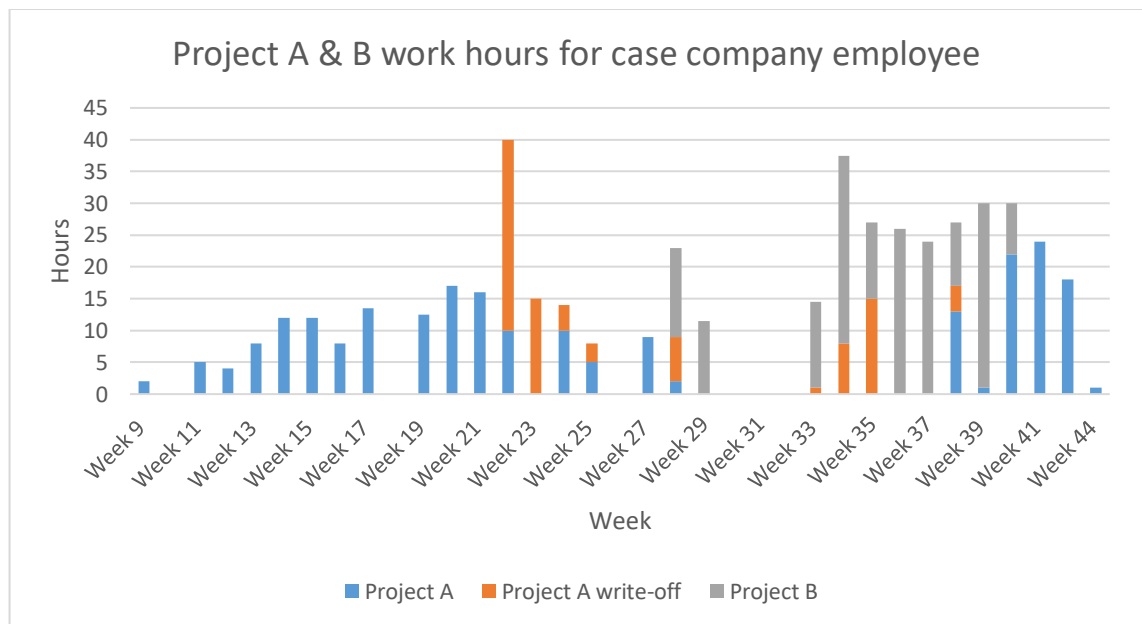


Figure 8: Company work hours in project A and B. Write-off hours signify budget cost overruns.

The Figure 8: Company work hours in project A and B, shows the amount of hours that was billed to the client, the project A showed a write-off of 87 hours due to project cost overruns. However, during discussions with personnel the real number of hours spent is even higher, as other employees participated in the analysis and hours were reported to other projects or internal cost places.

4.3 Data collection

The research-based consulting company (described in Background and context) provided the research data in this thesis. The data originates from client projects can thereby be defined as ‘industrial data’ as it is real world business data and provides insights into problems businesses are facing. As the provided is in raw format, i.e. same format as it would be exported from database systems. The process of gaining insights from the data resembles a “real world” situation companies are facing, where there is a need to find knowledge in data or create Data Mining applications to gain insights. Projects presented in Table 2 are in the case study methodology defined as units of analysis. In the following table the similarities and differences are shortly described. Client names are anonymized and are only mentioned as projects A, B, C, D, E and F. The projects are in chronological order, however certain projects were partly run simultaneously. Data collection lasted between 2015 to 2020.

Table 2: Description of the projects, units of data collection

Project	A	B	C	D	E	F
Type of organization	Global Industrial B2B	Global Industrial B2B	Global Industrial B2B	Global Industrial B2B	Financial services B2C	Global Industrial B2B
Product and service focus	Equipment in marine and energy markets	Cargo handling solutions and services	Solutions and services in offshore and marine industries	Forestry machines and equipment	Banking services	Mining & tunneling
Type of data	Sales/Field Service and CRM	Product level sales	Running hour installed base data	Equipment sales and spare parts sales	Customer data	Spare parts and fleet data
Authors role in Project	Analyst at the end of project delivery	Analyst at end of project delivery	Analyst	Project Manager and Analyst	Project Manager	Analyst

4.3.1 Project descriptions

Project A involved analyzing sales data and what the sales hit rates for different customer segments were for the years 2011-2015. The produced marine segment service dry-docking schedule forecast provided a valuable tool for sales teams to correctly allocating resources and focus sales to specific customers. The analytics showed that the data has a cyclical character. The company has had an extensive collaboration with the client company and had therefore good contacts at the company. The project began as a continuation to a previous analysis done in 2006. The sales process was clear and according to the employees the scope of the project was clearly defined.

Project B began with the client's initiation by contacting the company to request the service for a data analysis. The client for project B has a proactive sales department that required a tool that can provide information about clients that should be contacted in sales purposes. Data was available from four sources throughout the client organization. The data can be divided into product type data and sales data. Product type data includes volumes and specific models of products. As not much about the client data was available, the assessment of data quality was difficult. The main question the client inquired was to analyze the effects of a price decrease of 30 000 items on the business. As well as, what specific parts should be proactively sold.

The focus of project C was to combine financial calculations with industry data to provide evidence of financial profitability. During this project, the main focus became designing the structure of data to be presentable and the outcome coincided that of project A and B, where the data was presented in a BI dashboard.

The client organization for project D has machine and equipment sales globally. They have noticed that in different market areas, customers' spare parts purchasing behaviors and after-sales service needs are different. Due to this reason, this project's aim was to investigate customer spare parts purchasing behavior for specific market areas. The client wanted specifically a data driven approach to the results. The data in the analysis was gathered internally at the client company and then provided to the company.

The customer segment for project E was completely new for the company. The client was strategically important, and it was reckoned important to extend the data analytics capabilities to other business areas. The client work was done on-site at customer facilities

due to strict security policies, and the project was carried out as a thesis project to be delivered in approximately 3 months. The aim of the project was to create a churn prediction model, to predict customers about to terminate company services, allowing for preventive actions to prevent the churn event from happening. The model was created based on a data lake of processed data of customers transaction of two years.

Scope of project F was relatively broad, as the customer needs became clearer during the delivery of the project. In the end, the main goal of this project was to create a tool for the client to estimate the sales potential based on customer fleet composition. Secondary goals included describing customers' spare parts ordering behavior and analyzing spare parts' price change effect on demands of spare parts.

4.3.2 Interview

At the end of January 2016, the company announced an employee change, as one of the managers resigned the company. Employees universally knew that this person carried extensive knowledge in the sales and consulting area. From research perspective, it was interesting as he was responsible for selling the two analytics projects for case A and B and knew great amount of general and practical knowledge. The following main topics and questions needed to be answered:

- How are these analytics services offerings different to previous projects?
- What does the customer value in these services?
- Future of these services impact of change such as big data, industrial internet etc.

The interview was conducted on the 9th of February in 2016 and lasted for one hour and was of semi-structured format. The interview was recorded, and a summary report was created. Interview questions can be found in Appendix A.

4.4 Adapting CRISP-DM framework to fit case company needs

Already working a short time at the company, the problems became apparent. Initial briefings with company personnel showed how their current way of working with analytics projects needed improvement. Due to rapid development in offering at the company the analytics carried out by the company was run *ad-hoc* style with *Microsoft*

Excel as the main tool. The way in which analytics projects were delivered at the company resembled the steps in CRISP-DM. In this chapter the sub-steps of the framework are compared to the case company project phases. The project follow-up was done in Excel and later during research partly moved to a CRM system.

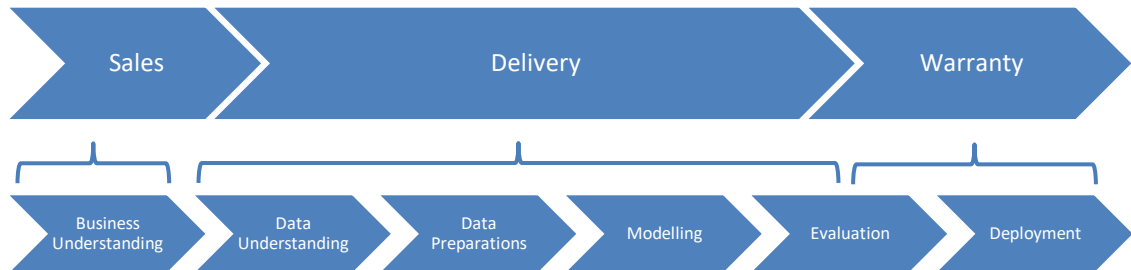


Figure 9: Case company standard delivery model in comparison to CRISP-DM model and roughly sketched which CRISP-DM activities fall into which case company delivery phases

The CRISP-DM methodology has proven to be successful in many organizations. (Wirth & Hipp, 2000). The authors also conclude that the CRISP-DM is especially useful in larger projects with many stakeholders. In internal discussions it was agreed that the complete CRISP-DM model was excessive for the company needs. Therefore, the CRISP-DM was modified to fit within the company business practices.

4.4.1 Business understanding

The data analytics project starts by understanding what the current situation is at the client's company, what are the most pressing problem and strategic goals. According to the company's way of working, this task can be defined as sales. As is evident from Table 3, the company has a large role in this stage. Often the Data Mining project can be linked to specific strategic goals set by the client's management. Therefore, if possible, the business success criteria can be linked to set KPI's (Key Performance Indicators). Even though, this might be difficult in many situations. All these activities can from the company's perspective be part of sales activities in the project lifecycle. Therefore, a successful business understanding phase should in theory lead to won sales, or at least demonstrate a understanding of the current situation and be able to determine a project scope.

Table 3: Roles and responsibilities of company and client in the business understanding phase of DM projects. Adapted from the CRISP-DM model (Wirth 2000)

CRISP-DM task	Company's role	Client's role
Determine Business objective	Part of Sales process situation assessment. Ask Client specific questions. Defined in proposal background section	Describe situation and answer to questions in sales meetings.
Define Business success criteria	Defined in Aims, however usually not quantified value.	Validate Company's presented proposal
Determine DM goals	Specify in technical terms what needs to be achieved. Can provide first draft, however, requires input from client.	Co-create. Provide support e.g. which data could provide best end goals
Produce Project plan	Part of project scope and schedule in proposal. In other words, a sales phase activity	Commit to deadlines e.g. provide data, participate in meetings.

From experience of the project management and sales such as in project F, it was beneficial to start the data understanding phase during project sales phase. This commits the client to gather and deliver relevant data for the project, leading to more likely won sales and the company to provide a more accurate cost estimate for the project. The downside risk is that the data understanding phase is done for free, if the client then declines the proposed analytics project.

For business understanding phase no major new changes were proposed as these activities are seen part of sales tasks. If sales perform well, then all subtasks within the phase should be understood within sales team. Then, the task for sales manager is to organize thorough handover of the project to the delivery team.

4.4.2 Data understanding

This step begins by collecting the data, depending on the client and the type of project, data is already collected, otherwise a data collection procedure is planned. There are usually two methods in which the data is provided, either export provided by client contact or the gathering is outsourced to IT or other department with ownership of the

data. The data received usually contains more columns that is needed for the analysis. At any of the case projects the data volume has not been an issue. However, this is likely to change as clients are collecting larger and larger pools of data. No new way for the collection of data was proposed as new ways can lead to issues with data security when client's sensitive data is transferred.

Table 4: Data understanding as-is compared to new proposed way of working

CRISP-DM task	As-is way of working	Proposed new way of working
Collect initial data	Client collects data and sends to company using appropriate client file sharing software.	
Describe data	No structured way	Structured way using template in Appendix A.
Explore data	Using Excel	Trifacta, R, Python or Excel. Visualization of data sources using Draw.io software.
Verify data quality	Manually check using small samples of data	Tools to support comprehensive analysis of data quality

An overview of the proposed new ways for the company are presented in Table 4. In accordance to the CRISP-DM, the data needs to be reviewed and columns checked for their importance for the dataset. This will work as guidelines for users to focus on data that is relevant for the analysis. The goal is to focus on preparing the data that is required in the analysis. The documentation process also forces the employee to understand the data in that what each column means. This became apparent when in one of the projects the serial numbers for products did at first not seem to have any structure, only when the client contact person sent an email explaining the serial number structure. Thereafter it was possible to label some data.

For project A, the understanding of the data was good as the company had worked with their data before. On the other hand, in project B the data was completely new and there was uncertainty about the contents of the data. During the project it became apparent that

the data quality was lacking, however the project continued, and the analysis was completed with the best precision possible.

The initial project plan should already provide details of what data is required from the client. As was the case in the two projects A and B the data was scattered around the client organization. For example, sales data can be accessed by financial controllers and service data can be accessed by maintenance department managers. Starting with the documentation of the data as soon as possible, proved beneficial, as this document could be reviewed with the data owner from the client side and acted as a cautionary step in avoiding misunderstandings in the data. Reality is that, clients have different ways in which data is entered into systems, highlighting the need for interpreting the data correctly.

The data understanding template also provided a way for the employee to understand the required processing of data. The template for data understanding was used to document the required data cleaning and transformation. This template while exploring the data or together with the client to fill in on-the-go while having a meeting with the client. The IBM CRISP-DM guide (IBM 2012) mentions to write a data exploration report. In the company this step would be made with the Trifacta software, described in further detail in chapter 4.5.1. During or testing of work procedures we did not find it necessary to write a separate report. However, to run the software and include these findings in the raw data documentation file.

4.4.3 Data preparation

How the data preparation phase turns into depends much on the data management practices in place. If the client has a data warehouse where all data is already combined in a structured way, the preparation phase is less time-consuming. However, as noticed in project A and B, no data warehouse implementation had been done for these specific datasets, at least for the data that was used in the analysis. For project C, the data was managed in a large data warehouse which reduced significant time in the data preparation phase. For all the projects, in which results are represented in a Business Intelligence dashboard, the data needed to be in specific structured database format. The as-is way of processing the data was manually done with *Microsoft Excel*. Data was structured in a high-level grouping in manner that visual representation was possible, allowing for

grouping of data into similar values e.g. product type or product group. Discussing with the employees working in projects A and B, an estimated that 90 % of the project time went into preparing the data into correct database structure. This is even a higher estimate than the number provided by (Endel and Piringer 2015) in comparison 50% to 80% of time. As projects were completed and new knowledge about best practices were identified, they were applied to client projects. The development is presented in Table 5.

After reasoning with employees at the case company and by performing this type of practical work at the case company, it became apparent that in smaller data sets the use of *Microsoft Excel* is acceptable to use, however if the data exceeds over 100 000 rows (may vary depending on number of columns) a database software is needed. As the company is not specialized in IT the software should be as intuitive to use as possible. *Microsoft Access* became the choice of software as it is part of the *Microsoft Office* software package and has many similarities with *Microsoft Excel*. As part of the thesis work an internal guide was created to guide users in *Microsoft Access* applied to client data.

Table 5: Software and tools utilized for data preparation.

Project	A	B	C	D	E	F
Modelling method	Excel	Excel	Excel and MS Access	Python code	Python code	Excel and Python code

From experience in projects, if transformations are necessary to the data, then the preferred way of working is to create code or a query that runs the transformation. This technique allows for data preparation performed in transparent structured way and minimizes the risk of having to re-do transformation steps due to complications later in the workflow. To visualize the data preparation phase, the *draw.io* software was introduced, enabling data structuring drawings to be created and presented to clients and company employees. This provided a good view of the overall needed steps for data preparation and supported communication with the clients, as specific parts in this drawing could be referred to in the emails or client meetings.

4.4.4 Modelling

At the company in the modelling stage a mathematical model or visualization structure is established. In project A, B and C the modelling consisted of summarizing data into specific ways, then presented in a dashboard software. This enabled to visualize findings in way for actionable decision making. Most Business Intelligence software's have functionality of importing CSV files. This proved to be a great way for proof of concept solutions to be demonstrated to clients. In the end, the chosen algorithm will state how to prepare the data. In some of the company projects the modelling part can might include creating dummy data points to demonstrate a certain point or logic. This was a way for the company to perform Research and Development, first assumptions about the data were made and later more precise with data is provided by clients and thereafter hypothesis can be proven or rejected. An overview of the employed modelling methods for each project is shown in Table 6.

Table 6: Main modelling methods applied in projects

Project	A	B	C	D	E	F
Modelling Method	Data Viz	Data Viz	Data Viz	Data Viz	Machine Learning model	Data Viz & Interactive calculation tool

Project E provided a more complex modelling phase than any of the other projects. Different Machine Learning models where tested to evaluate which model would provide best customers churn predictions. A desktop study was first conducted to figure out what type of algorithms are used in similar situations. In the end a random forest model provided best prediction scores compared to alternatives tested.

In Project F an interactive tool was created for estimating installed base equipment usage. The client wanted to estimate the spare parts business potential in a said market. Additionally, the spare parts orders where analysed, specifically of interest where the orders with one line of order due to the logistics costs associated and signal a sporadic or reactive ordering of spare parts.

4.4.5 Evaluation

First and foremost, it is evaluated if the business objectives set in business understanding phase are met, as well as verifying data for any inconsistencies. The verification step proved highly important, especially from the client's perspective, as this evidenced to clients the correctness of the numbers, thereby boosting trust in making business decisions with the analytics results. As a matter of fact, the data transformations and calculations might be complicated, and managerial acceptance of these analytics solution or decision support systems, is highly individualized.

The calculations can be summarizations of numbers or statistical calculations, to check and calculate the results different steps can be made: (1) Numbers to compare to e.g. on overall level, (2) compare to small sample or (3) evaluate with domain experts. Choosing which approach to take might be difficult, however these are the ways in which the company is working. Important however is to initiate a verification discussion as soon as possible during the project and avoid leaving the verification to the final steps of the project. This assures that any inconsistencies in data quality, transformations or calculations are noticed in advance.

In projects A and B, the outcome of the calculations is used for budgeting in the upcoming years. Therefore, next year the results will show if the calculated predictions are correct. Here once again long-term partnership will provide further support for the calculations.

The most extensive evaluation phase was done for project E where the random forest Machine Learning model had to be tested. Variability in the predication score seemed to be caused by how the randomized training data was chosen. The prediction accuracy was evaluated with AUC (Area Under Curve) scores. It was also part of client's responsibility to evaluate if the produced solution is working as intended. If the analytics solution is not alignment with the contract, the company is required to modify the solution accordingly. To reduce the risk of warranty claims the analytics tasks should support what is agreed upon in the sales phase.

4.4.6 Deployment

The results of the analysis in the projects A and B was visualized in a business analytics dashboard. Training sessions are held for the clients where the dashboard functionality is

shown and educated how to use the dashboard. The follow-up consists mainly of what maintenance is required to keep the databases up to date or algorithms consistent in performance. The dashboard software used in these projects was used for the first time and therefore it required some training by the employees. The software is designed for non-technical users in mind and their website mentions that no SQL knowledge is required to create these dashboards.

The type of projects conducted on by the company could be described as pilot type projects where the analytics solution is demonstrated as a proof of concept. This allows for an agile development and highlights the work previous steps in the CRISP-DM. In coming projects, it is likely that these types of implementation occur on the client's organization Business Intelligence software. The company role would then work as specialist or advisor in implementing the solution (Anonymous 2016b). Company practice is to deliver the deployment as a separate service, either as continuous support or as a project with set time schedule. As the data analytics models become more complicated and intra organizational, the amount of people involved in the projects increases.

For Project E the deployment was done completely on client's side where the model was firstly tested in an A/B- testing setup to evaluate how effective the predictions are, as well as how effective the solution itself is. Project F, the analytics solution was created as an proof-of-concept, and a *Excel* tool and was deployed on client's personal computer, although a server or cloud deployment would be possible.

4.5 IT artifacts to support data analytics

This chapter is divided into different subsections that introduce tools that make the DM processes easier and more efficient with a focus on the Data Preparation in the CRISP-DM phase i.e. data wrangling. The tools and software presented in this chapter are specifically open source or freemium to minimize any additional expenses for the organization. It would be a risk from the company to invest in expensive software if further DM projects are not delivered. Easiness to learn has also been taken into consideration when choosing software. No one solution was found to be able to solve the company needs, therefore, multiple software for specific tasks were introduced, presented in Table 7.

Table 7: Software, tools and templates introduced

Phase	As-is methods, software & tools	Proposed software, tools and templates
Business Understanding	Meetings with client during sales stage	Created templates
Data Understanding	Meetings with client	Trifacta, Python, Draw.io
Data Preparation	MS Excel	Python (Pandas package)
Modeling	BI-platform	BI-platform, Python
Evaluation	Numerical verifications	As-is and ML algorithms
Deployment	BI Platform	BI Platform or custom analytics solution

4.5.1 Introduction of a data wrangling software

As stated in the literature the data preparation phase or data wrangling is a work intensive phase in Data Mining projects, this was apparent in multiple projects at the client company. To solve this problem of slow data preparation the data wrangling framework (Sean Kandel et al. 2011) was developed. In October 2015, a free desktop version of the framework was introduced. (Bartur 2015) The software seemed promising as it has many features available and the software was quick to learn. The developers have divided the data wrangling process into five steps; (1) discovering, (2) structuring, (3) cleaning (4) enriching, (5) validating and (6) publishing. The predictive suggestion shows different actions for the data and shows a preview of how the data will look in the end. The wrangler software is run by a script-based language which is intuitive learn. Large data sets are troublesome as processing power required process the data increases substantially. The software vendor has overcome this problem by taking a small sample of the data so the wrangling can be done on that data. After the wrangling is finished steps performed in the smaller sample of data can be done to the complete data set. (“Trifacta Wrangler” 2020)

Overall time and effort were put in reviewing the software, especially the possibility of using this software instead of *Microsoft Excel*. However, soon it was apparent that the software works best as a complementary tool. Simple transformations such as splits, merges, deleting and renaming were easily performed by the software. The software provided also a great general representation of the data set. On the upper side of the software, bar graphs show the distribution of data for each column. The software can automatically detect the data type for each column, which supports the data understanding phase.

Before introducing the software (“Trifacta Wrangler” 2020) in the company organization some benchmark testing was done. Employees were asked to install the software and to test the software and were given a guide for the most important features. The software vendor promotes their software as having an intuitive user interface; therefore it was decided to verify this with analysts from the company.

As a final thought, the enterprise version includes many features that the desktop version is missing. The evaluation done, included only the desktop version of the software. Implementation of an Enterprise edition requires an on premise or cloud server with a Hadoop filesystem. A more comparative overview of the features for both wrangler, wrangler pro and enterprise version can be seen in

Appendix C: Table 9.

4.5.2 Introduction of a visualization tool for data management

Due to the dynamic nature of the analytics projects more human resources might be needed on a short notice, therefore methods supporting collaborative working are valued in delivering projects. Visualization of the data architecture and data processing needed, allows for more efficient understanding of overall situation. Entity relation diagrams (ERD) are conceptual models of data used in business environment. The projects add in complexity, especially as data is combined from various sources. Typically, in Data Mining projects changes occur during the project. Therefore, quickly designing the requirements for the DM supports the overall projects. The tool was for it to be intuitive to use and easy to modify. The designing of the data requirements scheme is completed in the data understanding and data preparation phase.

The software was chosen, as it has all the features required. (“Diagrams.Net” 2020) The software is run on the web-browser and the diagrams can be shared with anyone with cloud storage, such as Google Drive, OneDrive and Dropbox. Constructing the diagram proved to be easy and made it possible for employees even on meetings to sketch the diagram. These diagrams can be shared with project members and they can also modify the diagrams. Integrations can also be made with project management software and with version-control systems. The software was introduced to project members during the project and to other employees briefly at the internal learning session and in more detail at the master’s thesis presentation.

4.5.3 Introduction of scripting languages

The R statistical programming language is widely used around the world by various organizations. R is especially useful for visual analytics, generating plots charts and graphs out of complex data. RStudio is an open source graphical user interface environment for R programming language. R has become a popular alternative among commercial alternatives and data miner’s statisticians are using it. (Muenchen, Robert A. 2014) The R community creates “packages” for different statistical methods, choosing the most appropriate method depends on what kind of data is available (Smith, 2010).

The software was introduced to the a few employees at the company. The goal is to introduce them to simple scripts by making standard reports that are normally made manually. This would lead to them noticing that the R scripts are simple and could be used to add more advanced analysis in client reports. In charge of this projects was a senior researcher that also seemed to have the most knowledge in data science field. The introduction was sent as an email to employees and the reception was rather mixed and the reason was package dependencies making scripts inexecutable. On the other hand, a few participants become interested in the statistical methods used in the analyses and how the results should be interpreted.

Another approach would be to introduce a software that has a graphical user interface. This would make it easier for people who have limited experience in programming to start analyzing data. The main requirements of the software are to be intuitive to use and supports data analytics. A possible solution is “*Rattle: A Graphical User Interface for Data Mining using R*”. (Williams 2009) Another potential software platform is *KNIME*

(Konstanz Information Miner), offering an environment for working with data to input, manipulate and analyze data. The platform offers a modular workflow environment and can work together with the *R* software packages as well as *Python* programming. Turning analysis into products, the processes can be turned into a pipeline approach where all available data is inputted through right channels and all steps are processed automatically. *KNIME* software is a promising has created an open source solution for an analytics platform. The developers have made it modular and highly scalable platform for inputting, transforming, analysis and visual exploration of data. (Berthold, Cebron, and Dill 2008)

Another alternative for *R* is the *Python* language and provides similarly to *R* a programming language suitable for data scientist. The software *Anaconda* (Analytics and others 2016) is a solution to combine both *R* and *Python* code. Especially the *Pandas* package has extensive features to support data wrangling. *Shiny* is a platform for running *R* as a web application. Running its server allows for multiple users to run *R* applications, Therefore the company deployed a *Shiny* server on one of its development servers. Further investigation is required to assess if data analytics tasks could be implemented to a web interface the software provides. The improvement of this would be to minimize the fault of error due to missing installed packages. *Python* has an equivalent software called *Dash*. (Plotly 2020)

4.5.4 Introduction of a source code editor

Even when working with *Microsoft Excel* one might need to code a bit in order manipulate the data. *Notepad++* is a free editor and has many similarities with *Microsoft Notepad*, therefor the usage is intuitive and requires very little learning from new users. One of the most useful features in the editor makes it easier for the user to code and by visually highlighting errors in the code. The following features are also useful. (Hamid and Scarso 2018)

- Highlighting
- Auto-completion
- Column mode editing
- Multi-editing
- Multi-document (Tab interface)
- Multi-view
- Zoom in and out

The program is also customizable to fit the user's needs. Introduction of the software with a demonstration was arranged at the master's thesis presentation.

A common problem among data analysts, emerges from the way they are working with wrangling the data, that often lead to many dead-ends which requires the analyst to perform different multiple processes before finding the best solution. Therefore, these persons are resilient on documenting their steps. (S Kandel et al. 2012) This same behavior could be observed in the company. In traditional spreadsheet software, meta data is not recorded about processing done to the data. Programming the data wrangling phase with *Python* or *R* would make the process of wrangling more reproducibility and provides a platform for documenting each step. The importance of documentation is highlighted in consulting companies, where knowledge is the main resource the company is selling.

Another similar software proved to be beneficial, *Jupyter notebook*, is an open-source text editor that supports scientific computing of 40 different programming languages. The documents can then be shared across many platforms. The software allows text as well as embedded visualization in documents. (Perez and Granger 2015)

4.5.5 Introduction of data warehouse approach to DM projects

The clients in the project companies sent excel worksheets of the data, often extracts from the client ERP- or CRM systems. It was important not to alter the data in any incorrect form. Data warehouses primarily provide the infrastructure for storing structured data. The Data Mining process can then be streamlined, as the data is imported into database format and queries can be created to receive the data in desired format. Inputting data into a database can be done in multiple ways however, at the case company, when data is received from clients in excel format it can be easily converted to CSV (comma-separated values) file format. The wrangling software also supports this format and can be utilized to clean and transform the data before uploading it into a data warehouse. This structured approach to organizing data enables visualization of the data in a Business Intelligence platform. The company already had an employee working on SQL-databases, however on completely different tasks. However, as the potential and benefits of SQL-databases was acknowledged and was included in the process of improving analytics competence. Additionally, an internal guide was created to support new users with SQL use.

In project F, the client had a separated data lake infrastructure to store preprocessed data. The *cx_Oracle* module was utilized to import the data directly into *Python* by creating a connection between the SQL server and the *Python* notebook (Tuininga 2020). This allowed for easy modification of the data import if any inconsistencies were found. Thereafter, the data could be transformed according to any data modeling needs.

4.6 Knowledge development at the case company

This chapter is divided into sub-chapters that direct the company in both individual competence development and strategical focus from an organizational perspective. The 4x4 Analytics competence and organizational framework (Martine, George & Nicolas 2014) is utilized to support the analysis of the current situation at the company.

In the personnel meeting arranged in the beginning of 2016, the company expressed a strategic focus on improving analytics capabilities. The reason, management has been impressed by the possibilities of analytics from finished projects. The company has a strong focus on knowledge management, and states that in their values that employees should “*continuously develop their competence and learn new things*”. The company has already implemented a competence development framework which is reviewed annually at the board level. On an individual level, employees have annually performance discussions with their superior about their learning goals.

In the context of this thesis work, the knowledge development was divided into three phases (Figure 10) and was done to “keep the momentum”. Analytics is a broad subject and cannot be taught just once, this will also signal everybody at the company the importance of the topic.

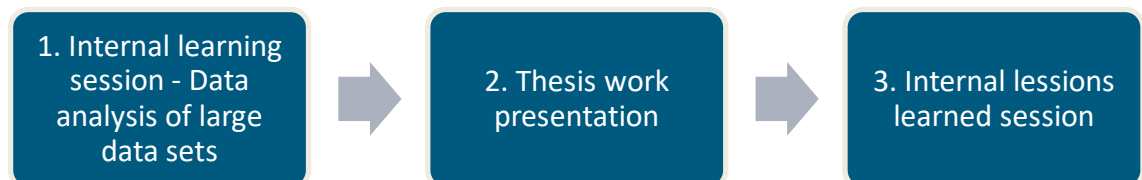


Figure 10: Phases of Analytics learning sessions

In the internal work satisfaction survey employees wrote that within the company internal communication between units were lacking. (*Climate Survey, Autumn (Internal Report)*,

2015) As a conclusion to the first internal learning session this became apparent as some employees had knowledge that others were lacking (Wahlström, Wikström, and Sundholm 2015). This was implicit knowledge that had not been turned into explicit knowledge. The management noticed the significance of this problem and included it in the agenda for the year 2016. The presentation and discussions at the first learning session set out the direction for knowledge development. Everyone in the session agreed to the fact that part of the problem for over allocated project time was because of the “wrong tools” for wrangling data. Therefore, it was agreed upon that the organization would focus on improving individual IT skills and analytics skills and that management should support this work.

After the Phase 1 Internal learning session, a survey was sent to all participants for feedback, which suggested an overall positive response about the topic and session. Many expressed that it was more of an awareness building session and needed and would have required a more training perspective. The real concern became apparent, that the datasets from clients are becoming larger and therefore more knowledge and ability to interact with the data is required. However, communicating the knowledge from these project signals others in the organization the skills and knowledge the company holds, potentially leading to new sales.

The phase 2 thesis work presentation contained both an academic perspective on the problem and a more practical hands-on case where a dataset from a client was introduced and explained what was done and why. Because the session was only one hour, the topics were only briefly discussed. The feedback from the session was positive and again showed that there was a need for this type of work. The session also sparked some discussions about future projects. What became evident to participants was the role of data preparation in data analytics projects. If any predictive modelling is wished to be done, then the prior steps are essential, and the number of workhours required ought not to be underestimated.

The phase 3, consists of project lessons learned sessions and are planned to be had after completed analytics projects. Therefore, the number of sessions depend largely on the number of completed projects. An emerging discussions topic in these sessions is the business impact, however, it could be advised to include data wrangling as one topic within these sessions, exemplifying requirements in these projects and demystifies the

actions behind-the-scenes. It would be worthwhile to reserve more training about the introduced software and elaborate on these tools and how to make practical use of them. It would be advisable to hold this session for a longer duration and have some pre-work required before attending. Especially, as data preparation requires both general understanding of businesses and a technical and analytical perspective on data.

Creating successful analytics-oriented organizations requires substantive effort in terms of management support and organizational culture. The development of data scientist in organizations is in the sake of simplicity divided into development of (1) personal data science skills and to (2) managing data scientists in an organizational perspective.

4.6.1 Learning from an organizational perspective

Coaching or mentoring can be a very powerful way of transferring knowledge inside an organization (Swap et al. 2001). The company has a professional structure in which the employee can choose the researcher path or the analyst path. Managers oversee teaching and mentoring juniors, and have a strong focus on learning by doing, therefore less experienced employees are early on given responsibility in projects. For example, in the projects A and B a manager collaborated on a weekly basis with an analyst to discover the most effective ways of process and analyze the data. In this case, the manager with over 10 years of experience in the business could provide useful insights about the business and specific project situations

The analytics capabilities required for the analysts is to have the ability to carry out quantitative and qualitative analyses, combine and present relevant information, present at internal seminars, ability to automate own work (BI and excel). This clearly, falls into the data science domain, however these definitions are somewhat vague, and a more detailed list could point the organization into improving the correct skillset. As mentioned in chapter Background and context, the company has also a separate R&D department and their employee career path leads to the senior researcher. This position has many similarities with the data scientist; The senior researcher requires a doctoral degree, be able to manage research projects, lead teams and participate in strategic commissions with own area of expertise. Later the researcher path was removed in official documents and merged to the analyst path of professional structure.

In the first meeting of the year 2016 the partner and founder of the company mentioned that a focus of the year will be to go more into predictive analytics, and it is something that need to be developed. Therefore, there seems to be at least verbally a sponsorship to support analytics from the highest management position. During the authors time at the company new sales material was created based on data analytics projects, first reviewed by one member of the management, and then shared with the company. The feedback about the material was positive, however some sales personnel think the concepts are still a little unclear and require some clarification.

Regarding recruitment, in 2018 a student was hired to conduct an analytics project, namely project E described in chapter 4.3.1, for a client organization as a master thesis assignment. The project was split into three months conducted on-site at the client company and three months at the case company writing the thesis. Thus, having an intensive learning experience at the client and could transfer knowledge from learned to the case company.

4.6.2 Individual learning

The proposed methods in this chapter supports the 4x4 analytics framework (Figure 4), the presented solutions are mostly related to the Analytics and IT skills. While the other competencies; business, communication, project management and leadership are important, they are outside of the scope for this thesis, and it could be argued that during normal operation the employees are learning these skills in projects.

Feedback and discussions on the first training session about “Analyzing large datasets” suggested that it is not possible to create IT-, software or programming knowledge in one training session. MOOC’s can offer a solution for this problem, as the courses are ongoing for multiple weeks, leading to a continued learning experience. Study groups could be created to motivate each other or help each other with problems, weekly or bi-weekly meetings would let employees have a discussion of lessons learned and share their progress to support collaborative learning.

Learning goals for online courses could be added in the internal performance measurement system. Internal learning sessions are already measured, and similar key performance indicators could also be set for MOOC’s. Having a list of accepted courses would let employees choose what they want to learn, alternatively there could be different

tracks so to learn the necessary for different titles in the company. This would also provide a way for the employees to receive career guidance of which skills and knowledge is in demand by the case company. Due to certification possibilities the achievements could be highlighted and thereby showcasing the knowledgeability of the employees in social media such as LinkedIn.

(Martine, George & Nicolas 2014) mention the importance of cross-fertilization within organization dynamics aspects, due to the small size of the company, the problems risen with large organizations communication is avoided. As the company has an office in another city as well there might be knowledge held in either the company offices. Some employees have more contact to the other office staff, while others have less. Strategies for implementing any change in the company must be planned for successful adaptation. Initial reactions have a strong psychological effect for the reciprocity of employees. Finally, it should be advised that the partaking in these courses could be discussed during the performance review for each year. As people have preferred individual learning styles, some prefer learning by observing while others by doing. The company also uses this tactic while training new employees, learning this by firsthand experience, the employees are placed as soon as possible into tasks in client projects. This applies for data analytics projects as well and develops skills along the project. Although this is disregarding any budgetary restraints for the projects. The company encourages personal development by financially supporting employees in educational trainings, this was evident as the author could attend a week-long intensive university course in Data Mining.

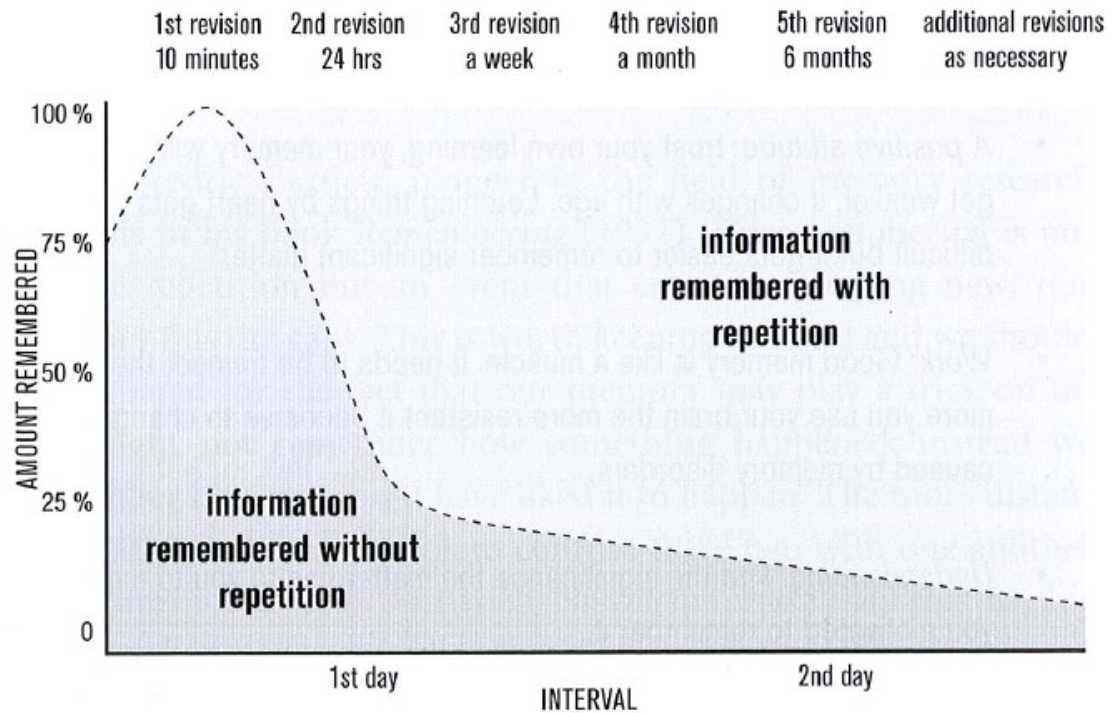


Figure 11: The forgetting curve (Sydänmaanlakka 2007), **originally Ebbinghaus' theory** (Murre and Dros 2015)

To implement lasting change, the process of forgetting information, described in figure 9, must be overcome and by repeating information, the information can be stored in the long-term memory (Murre and Dros 2015). Due to the realization of this mechanism, further justifies the MOOC way of learning, where small amount of effort is repeatedly. This feasibly provides an alternative or addition to the one-time learning session about a specific topic. Therefore, repetition and focus on long term learning is desirable.

4.7 Lessons learned from case study

Although the data preparation phase was cumbersome, the scopes of the projects A and B were underestimated partly due to lack of a structured approach. Project analysts were using traditional spreadsheet software for most Data Mining tasks. Results from assessments, interview and several discussions with the employees resulted in the following conclusion.

- Projects A and B scope was underestimated, and led to project complications, decisions were made in a rush, which led to some misjudgments, especially during the process of data wrangling.
- The software and hardware available for the projects were limited. Simple spreadsheet programs were used for combining multiple data sources, which led to memory overflow problems and many times to crashing of software.
- As time progressed and company focused on analytics, more complex analytics assignments were accomplished such as projects D, E and F.

The results of these two projects proved the need for more planned process for data analysis. There was also a need to better define the amount of time allocated for each task within the projects, the data quality aspect and requirements for cleaning and processing needed issue came apparent. Although these projects went over the allocated resources, it affected mainly the company by allocating more hours to the project than first planned. The positive outcome of the projects was the continuation of partnership with the clients' firms. Therefore, these analyses provided opportunities for further projects.

When similar projects are led, it is advisable, to have knowledge transfer to other employees of what was learned in the projects. The current learning sessions held at the company prove to be good occasions in transferring knowledge to others, however it is unclear whether these are sufficient due to the restricted time. In the projects a great deal of time was put in transforming the data with code and these parts of code were saved in text files. These codes proved to saved considerable time and proved to be valuable for finishing the projects. The code was specifically created for the received client files, however with small modification these they can be made more general. Instructions were added as comments to the code to explain what the code performs; this in turn makes it easier for the analysts to understand each step of the code. As the expertise of the company personnel is diverse and the company is small, some people might be working on projects by themselves. This means that the knowledge transfer is limited during project delivery phase. There is always a risk for the analysts to important if the original project workers become unavailable. In the instance of Data Mining projects, there is a risk that only the creator knows how the data wrangling is performed. It stands further to point out that, both informal and formal communication within organization cultivates the passing along of project best practices.

5 DISCUSSION

Industrial organizations are increasingly recognizing the potential of analytics. This provides an opportunity for the company to re-evaluate strategic focus and possibly invest resources into analytics capabilities. As noted from the case study, the different stages of analytics can be made more efficient. There are multiple research findings indicating improved performance of organizations utilizing analytics in decision making. It could be argued that analytics provides a competitive advantage, if deployed correctly. As of writing, there has been an upward trend towards self-service analytics, meaning software developers are creating solutions where the required steps of analytics are provided within one user-friendly software platform and it remains to be seen if these solutions are versatile enough to provide for data scientists needs.

For the case company, the introduction of CRISP-DM and its modified version, presented in Appendix D: Figure 12, was met with approval due to its clear arrangement and validated research background. The case company's offering is specialized for project business clients therefore similarities of model was appreciated. The CRISP-DM model is now about two decades and still seems to be relevant. Now though, new models have been created which might become practical. (Martinez-Plumed et al. 2019) The clients are to a varying degree aware of the methods used that case company. Nonetheless, client acceptance of the models will surely affect if the model is being utilized. Only one of the clients' main contacts (in project E) mentioned that CRISP-DM would be unsuitable due to its restricted form, specifically its compatibility with agile methods.

Another aspect became apparent in project D, here perhaps the Data Mining steps were rushed. Weekly progress reviews were performed with the client contact person to gain common understanding of findings. Still, at project closure, one of the project deliverables had a mismatch in numbers as the client was presenting the results to other business departments. This further emphasises the importance of thorough the CRISP-DM steps even in cost-restricted projects. To sum up, restricting proper steps might reveal missed data quality issues, misunderstanding of data content or reveal other data issues. The severity of the issues will have increasingly more impact in the project deliverability at later stages of project lifecycle, meaning, if inconsistencies are found in the verification phase, the workload is among the highest.

What became apparent during research is that improving analytics requires a holistic approach. The 4x4 framework (Martine, George & Nicolas 2014) is still relevant to this day. Nevertheless, there lacks a one-size fits all solution, therefore it is the organizations responsibilities to decide on best strategic efforts. What has been encouraging to see, is the development happening during the time of research. In my opinion the single most significant leap in analytics capabilities occurred with the hiring of a part time analyst explicitly to work as a data scientist. This affected greatly on the tools and software being utilized at the case company. Among all the introduced software and tools, the scripting languages, particularly *Python*, was adopted whenever a more complex analytics task was presented. Documentation and reports of the Data Mining steps were included in the script-notebooks, as this was preferred by the analytics team members. The data warehouse approach presented in chapter 4.5.5, found less uses than expected in projects. Perhaps one reason for this, was the fact that the person most knowledgeable in SQL left the company. One major change to the original way of working was the depreciation of a BI software, that previously was utilized to present the analytics results.

Unquestionably, company culture, teamwork in projects have a powerful impact on project success and how issues are viewed and discussed. Many team members noticed an improvement in group dynamic as the departments within the company were merged. Top management support for analytics certainly affect how these projects are perceived. During the time at the case company the CEO changed three times, all having slightly different perspectives of analytics, although all have verbally been supportive of analytics and agree to its importance. All these internal changes accelerate the rate of change, as well as external factors, such as increased demand from customers or industry trends, presented in literature review. As data analytics projects are relatively few in numbers, the importance of rigor reflection internally or with customers is advisable to assure effective organizational development. The company project lessons learned are valuable in promoting continuous improvement.

Certain aspects of analytics change over time, such as software skills in demand. Chapter 4.6 provides ways for the case company to improve their analytics capabilities from individual- and organizational perspective. Notice that the methods and might change as new more improved solutions are available. MOOC's seem to have a mixed reception based on practitioners and academics, partly because of the newness of the technology.

Still, the business objectives were achieved in both projects. In one of the projects, some failures were made in the final steps of the Data Mining phases which led to some redoing. Working in a team of data miners makes it possible to evaluate the data from different perspectives and possibly prevent errors in the process. Especially in the interview with one of the Senior managers, it became apparent that successful execution projects will likely lead to further cooperation with respective clients. This happened for project A as at the same time willing to promoting the case company in new sales cases, spiraling into further possibilities for new assignments.

During the four years working at the case company the maturity of the analytics services has certainly increased. With the available software, hardware, and knowledge it seems the company would have the ability to successfully complete similar projects within time and budget restraints. It will remain to be seen how the company strategically will evolve and whether it will scope projects to deliver proof of concepts or take a director type of role in driving analytics change. One manager company employee, by their own words said, *“we do not want to become an IT consultant company”*, therefore the role of the company analysis services needs to demonstrate a business-oriented approach to analytics. Certainly, this is a balance the case company is contending with, focusing on providing strategic consulting services while attempting to exhibit the skillset of data scientists.

The preferred choice of deployment method of analytics services for clients is still unclear, will the case company implement the analysis services in own infrastructure or will they be implemented into clients' information systems. This opens the discussion to data privacy and security, the analytics solutions often process sensitive data while requiring decent latency and bandwidth access to the data, prompting for on-site or cloud solutions. The infrastructural choices are questions that the case company need to contend and provide solutions that are feasible and secure to all parties.

In addition, the company recognizes digitalization and AI as business disruptions, The installed base of IoT devices increases year by year (Sommarberg 2018). IoT seems to have a substantial impact in maintenance driven business, due to the possibility of monitoring asset condition and thereby forecasting the need for maintenance. Certain businesses are still resiliant to invest in IoT projects as these are still in early stage of development, and managerial, social and behavioural aspects are preventing further

developments. However, if IoT technology becomes mature and accepted technology among industry actors, it will surely provide opportunities for analytics services.

5.1 Additional findings

In the interview and discussions at the company it became apparent that the pricing of analytics services is difficult. The problem lies in estimating how much time is spent for each activity in the Data Mining steps, data understanding and data preparation. These activities were relatively new to the company therefore all prices were estimated. It is jokingly said in the company that 90% of the project budget is spent on just cleaning the data. It could be argued that, projects A and B were out of budget due to limited knowledge of working with Data Mining projects and the project being sold with too little information leading to unrealistic pricing. The projects' outcomes indicated that inconsistencies in projects' sales phase (Anonymous 2016a). With the new analytics framework consultants were able to sell projects with more confidence. The framework showcased to clients a knowledge about analytics and being able to describe what is required in each step, additionally being able to convince clients of having an easier follow-up from project management perspective. In the end it will remain the managerial responsibility to provide realistic pricing while providing enough resources and support for proper Data Mining steps.

Discussions with managers and senior personnel indicate that the way to sell these projects is by making the data understanding as part of the project scope and divide the data understanding steps into smaller tasks and include in the pricing strategy how many data sources are involved. This however is risky as cost aware clients could exclude additional data sources just because of price concerns. The additional step of data assessment in data understanding, at the beginning of the project ought to reduce the risk of falsely estimating the required time for each step. Furthermore, to transfer portions of data understanding tasks into sales activities will have both an effect of understanding project scope and improving odds of won sales due to increased client commitment. Many project ideas have been discarded due to data quality issues, here this can be verified. If, however, the pricing is inadequate, scope changes are possible and are written in all project contracts, however, these are rarely exploited at the company. As a result, unexpected difficulties lead to project extensions, while request for additional

compensation is avoided to avert negative client feedback. Usually, if the projects are successful, the client is satisfied which makes them more reciprocal to larger scope projects. The experienced sellers are talking about making a smaller “teaser” project to increase the scope for the next project.

During my time at the company, development efforts were allocated to packaging project types into services groups, the idea being, to have a somewhat productified service portfolio, saleable products with initial scopes. These projects also gave an indicator of cost and resources needed in each step allowing the sales team to make a better assessment of the cost and resources needed for projects. Another concern for the company was being able to deliver reasonably priced analytics services. While some senior employees have a better knowledge of data preparation, assigning them to the project would easily lead to budget overruns. Therefore, it was found to be important to pass along the knowledge from these persons to more junior employees and thereby contributing towards competitive pricing.

One of the case company’s expertise is value-based pricing, this concept is brought into projects and related discussions arise during meetings with client. The interest would be in creating a competitive offering to the client that adds value, however the dilemma is the Data Mining process is so work intensive especially if many data sources are used. Somehow the added value needs to be proven to the client before the results are achieved. In certain consulting projects, it has been possible to create a quantified value proposition with the help of a business case calculator. This was somewhat possible in an analytics project outside of this thesis scope and allows for business minded persons at the client c to understand the benefit of the analytics solution.

5.2 Further research

Research in this thesis advances the understanding of completed data analytics project and presents improved ways for in a traditional project service business. Recently, however a strategic mandate was defined to (1) improve existing analytics services and (2) restructure current or new analytics services with alternative pricing models (Anonymous 2020). The second point would certainly be of interest as plenty of research is targeted towards organizations and less specifically towards analytics service providers. The CRISP-DM framework is of an iterative continuous character, still the

suitability of the model could be evaluated or alternative models. In addition, this service model could be compared to data analytics self-service models, which solution would benefit customers more. This leads to another contemplation; how should analytics most effectively support data driven decision making in organizational setups. Further broadening the spectrum, one could research how analytics services aid companies within an industry ecosystem.

Desktop research, discussions about the topic and listening to clients suggest a need for further understanding of customers' purchasing behavior, especially for spare parts sales. The case company has experience in analyzing clients' databases consisting of sales and quotation data about spare parts sales. This data could be used to mine frequent item sets, there has been already plenty of research in frequent item sets, therefore these findings could benefit the study. These algorithms have largely been used in B2C sector, however discussions with representatives from industrial B2B companies have led me to believe that there is a need for a more "algorithmic" way of analyzing spare parts sales. The project could be delivered with the framework presented in this thesis. Moreover, data driven decision making could be researched in B2B e-commerce platforms.

5.3 Limitation of research

Obviously, methodological choices affect the research findings and conclusions. As mentioned before, case study research makes it difficult to generalize findings. Due to the character of consulting business, multiple projects often run simultaneously, and project decision making is often done quickly without proper documentation of the taught process. According to project management theory, it is precisely this critical path of decisions and events that lead to desirable or undesirable outcomes and understanding what has happened in detail would provide details of project consequences.

The evaluation criteria of the introduced software and IT artefacts presented in 4.5 led to exclusion of paid commercial software and the evaluation was performed without any proper framework, then again, working in real-world environment, the solutions that provided real value were kept and solutions deemed less appropriate for the use case were discarded. Moreover, the knowledge development aspect of this thesis, mainly in chapter 4.6 was limited in scope, one of many reasons was the postponing and cancelation of analytics project lessons learned sessions, which perhaps lead to knowledge being

withheld from the rest of company. Normally, these sessions are a way of presenting the knowledge in a structured and condensed format resulting in documentation providing useful information for research purposes. Taking into consideration all the limitations of the research, there are still countless companies in same size and field that analyze large amounts of data that benefit from this research.

5.4 Answers to research questions

In the following section I will try to answer the research questions.

5.4.1 RQ1: How can companies improve their data analytics delivery model?

Companies can utilize one of the existing frameworks. Literature review (Chapter 2.2) revealed that CRISP-DM is the de-facto standard and was therefore proposed to the case company. The framework, having an industry wide acceptance, while being relatively easy to understand for business minded individuals, was appreciated by the case organization. Chapter 4.4 describes how the framework is to be utilized in the case company context, both in overall project delivery context and stages within projects. Previous ways and new ways of working are compared in each phase within the CRISP-DM. There are alternative frameworks (SEMMA and KDD, Chapter 2.2). However, twenty years after its birth, CRISP-DM still appears to be the most popular option among data scientists (Martinez-Plumed et al. 2019).

5.4.2 RQ2: Which IT artifacts can support the organizations delivering data analytics services more efficiently?

Performing analytics often require transforming data into correct format. As the literature review and empirical evidence confirm, the data wrangling or data preparation phase is the most time-consuming part of data analytics projects. Therefore, the most effective way of improving project performance was to improve data preparation step. Thus, Chapter 4.5.1. introduces, a self-service software where users have an easy to use user interface to wrangle the data into correct format (“Trifacta Wrangler” 2020) and 4.5.3 provides an alternative approach, namely, script-based languages such as *Python* and *R*,

which have specific repositories for transforming data such as *Pandas* (Wes McKinney 2016). The second option might be more suitable for employees with computer science or data science background. The employees at case company consists mostly of non- IT-technical employees. Then the first choice, of a more graphical solution might be more relevant. To support the delivery of analytics projects and overall project management a visualization tool introduced to the company, described in chapter 4.5.2. Furthermore, supporting knowledge transfer of complex data structures. Assisting with creating documentation and creating code snippets, a source code editor was introduced, described in chapter 4.5.4, providing visual support for writing code. Data warehouses or data lake structures, presented in chapter 4.5.5, provide organizations with a curated data infrastructure, upon which analytics services can be developed.

5.4.3 RQ3: What capabilities are the most important aspects for developing data analytics?

From literature review it was found that there is a broad spectrum of capabilities required. The 4x4 analytics framework presented in chapter 2.3, Figure 4, provides a simplified map of the capabilities required by data scientists and recommended focus areas by organizations. The framework is divided into four core skills; Analytical, Technical, Business and Communication skills and four aspects of organization dynamics; Sponsorship, Recruitment, Talent and Knowledge Management and Cross-fertilization. An analysis of the organization current situation in regards of the four-component framework is presented in chapter 4.6.

5.4.4 RQ4: How can learning analytics capabilities be supported from an organizational perspective?

Different aspects, are presented both in chapter 4.6.1 and 4.6.2. The options presented are adapted into the case company, specifically to support the existing knowledge management program. Promoting MOOC style learning within the case company organization could provide the required analytics knowledge needed while supporting career development and signaling employee knowledge. The organizational acceptance of these new learning methods remains unclear. In addition, mentoring and coaching are viable options for learning, and do support knowledge transfer between senior and junior personnel.

6 CONCLUSIONS

Digitalization and macrorends such as the rise of big data and artificial intelligence require businesses to adapt to this new environment. Companies able to create data driven solutions in this transformed landscape will perform better than competitors. Designing business analytics solutions requires extensive efforts both from data scientists and domain experts in organizations. Research reveals insights into workflows for creating functional analytics solutions, in these the data preparation unveiled to be the most tedious and time-consuming.

The thesis aims to solve practical problems at the case company, Findings from the company indicated that the data preparation or wrangling phase took considerable time, nearly the same as was found in the literature review. This confirmed, that supporting this stage of the Data Mining projects will overall improve the effectiveness of analytics projects. This was accomplished by both introducing the CRISP-DM analytics framework as well as its modified version and by introducing software and tools, which were specifically chosen to fit the case company needs. Additionally, strategies for knowledge development was proposed aligned with the existing knowledge development program. Overall, this contributed to improving the analytics capabilities, supporting employees in delivering analytics while making customers satisfied and providing revenue to the case company. The research provides an insight into real world cases and reflects how analytics is performed both in from academic research and practitioners. Additionally, managerial perspectives of a small-scale organization are considered, which is unusual in most analytics related research.

The terms analytics and Data Mining are interchangeably used in this work due to their resemblance. The most common framework, CRISP-DM, separates all stages into detail as well as its iterative structure. Actions research methodology allowed for initiatives to be implemented gradually and allowed continued normal business operations. Because the research lasted about four years, it allowed for multiple cases to be included, similarly the extended period obliged to contemplate impact of organizational change throughout the study. The CRISP-DM framework could be adapted to projects at the case company and provided useful information about working with these projects. During the master's thesis, new ways of working with data was introduced to the case company. Software

such as the *Trifacta Wrangler* and scripting languages; *Python* and *R* provided more efficient data wrangling and streamlined the analytics workflow.

The management is aware of the possibilities arising from providing analytics services to clients and strategic effort is placed on creating or modifying a service to support clients. As more of these projects are completed, knowledge increases, and analytics capabilities increases, leading to better offering and data driven services to clients. The development of analytics capabilities requires active managerial action taking in the capability domains, the *4x4 analytics framework* provides a simple overview of how to develop data scientists' capabilities in organizations.

While case study methodology questions the generalizability of the findings, the study provides a holistic overview of real-world challenges organizations are facing. Further research would be needed to determine if data wrangling and analytics services are realized in similar ways, or if the improvement suggestions presented in this thesis would provide any added value.

7 SWEDISH SUMMARY - SVENSK SAMMANFATTNING

Skapandet av dataanalysramverk: från ett organisationsperspektiv

Analytics kan definieras som extrahering av värdefull information från data med hjälp av statistiska-, matematiska- och datorberäkningar för att stöda beslutsfattning (Cooper 2012). Den teknologiska utvecklingen i hårdvara, mjukvara och algoritmer möjliggör för företag att samla in och analysera data och därmed stöda företagsledningens beslutsfattning. Dessa framsteg kräver dock nya system, kompetenser och ramverk för att stöda en datadriven verksamhetsstyrning (Martine, George & Nicolas 2014). De ovannämnda framsteg har också lett till att en stor mängd data (*Big Data*) blir lagrad, och det är svårt för traditionella databasmetoder att bearbeta den. Därför har nya teknologiska lösningar skapats för att förbigå problematiken (H. Chen, Chiang, and Storey 2012). Det har blivit påvisat att företag som är fokuserade på data-analys har en fördel framöver konkurrenterna. (Fosso Wamba et al. 2015)

Avhandlingen genomfördes på uppdrag av ett finskt forskningsbaserat konsultföretag, vars huvudsakliga kunder är industriella aktörer i B2B-sektorn. Företaget var intresserad i att förstå hur dataanalystjänster kan förbättras och levereras effektivare. Forskningsfrågorna för avhandlingen är:

- Fråga ett: Hur kan företag förbättra sina leveransmodeller för data-analys?
- Fråga två: Vilka verktyg, modeller eller mjukvara kunde hjälpa företaget att leverera dataanalystjänster mera effektivt?
- Fråga tre: Vilka kompetenser är mest efterfrågade för en lyckad leverans av data-analysprojekt åt kunder?
- Fråga fyra: Hur kan inläring av data analys kunskaper stödas från organisations perspektiv?

Forskningen påbörjades med en litteraturstudie (kapitel 2) som behandlar koncept och begrepp som ansågs vara nyttiga. Som ramverk för undersökningen användes fallstudiemetodologi där sex kundprojekt, levererade av företaget, fungerade som

forskningsobjekt. Både kvalitativa och kvantitativa metoder har använts. En nyckelperson på företaget intervjuades på semi-strukturerat sätt, intervjun bandades in och transkriberades. Drag av etnologi kan förekomma, då jag under forskningstiden var anställd hos företaget.

Både från litteraturstudien och intervjuer hos företaget framkom att dataförberedelse (*data preparation*) fasen kräver mest tid i projekt. Därför ansågs det nödvändigt att undersöka hur denna fas kunde effektiveras och förbättras. Tidigare hade företaget ingen systematisk metod för att utföra data-analysprojekt. Från litteraturstudien framfördes att CRISP-DM -modellen var applicerbar till företagets behov (Wirth 2000). I kapitel 4.4 anpassades modellen i företagets projektleveransmodell. CRISP-DM -modellen innehåller sex olika iterativa faser där arbetsuppgifterna är beskrivna. Dessa jämfördes med steg som utfördes tidigare i företaget och samtidigt presenteras nya förslag på förbättringar. Då de undersökta projekten var av olika karaktär, gjordes förslagen enligt den kunskap som samlades upp under projektets gång och därefter.

Dataförberedelsen är mer än bara en arbetsprocess, den kräver också lämpliga IT-lösningar. Kapitel 4.5 tar fram mjukvara och olika *Excel*-mallar för att stöda data-analysprojekt i dess olika faser. *Data Wrangling* –mjukvara, som ansågs av analytikerna stöda förståelse för sammanfattningen av data, introducerades till företaget. Visualiseringsverktyg för datahantering togs i bruk för att stöda samarbete mellan analytiker och kunden. Därtill betonades behovet för skript-baserade programmeringsspråk, specifikt *Python* eller *R*, eftersom dessa har en möjlighet att befrämja analysen. Användarvänlighetens nivå i dessa språk är dock tvivelaktig, därför reflekterades ifall alternativa skript-baserade grafiska lösningar kunde användas. Därtill framkom att liknande analyser på företaget kunde också utföras m.h.a. webapplikationer.

Kapitel 4.6 tar fram metoder för företaget att utveckla data-analyskunskaper med hänsyn till företagets position och dess befintliga resurser. Som ramverk användes "*4x4 Analytics Framework*" (Martine, George & Nicolas 2014) där företagets åtgärder analyserades och olika förslag för vidareutveckling av kunskaper togs fram. Begreppet *data science* är sammansmältningen av flera olika domäner av kunskap, som krävs för att möjliggöra en lyckad datadriven organisation. Forskare anser också att en dynamisk businessmiljö kräver konstant utveckling (Martine, George & Nicolas 2014). Därför har jag tagit fram strategier och metoder med hjälp av vilka företaget kan utveckla sina kunskaper, både på

organisations- och individnivå. MOOC-kurser föreslås som ett alternativ, dessa kunde stöda företaget i dess kunskapsutvecklingsprogram.

En ytterligare upptäckt från semi-strukturerade intervjun med managern framkom att prissättningen på data-analysprojekt är svår på grund av osäkerheter i datans kvalitet och projektets omfattning i kontrakten, speciellt gällande dataförberedelsen var oklar, vilket ledde till att extra arbete utfördes utan kostnad för kunden. Alternativa modulerade prissättningsmetoder kunde undersökas.

8 REFERENCES

- Aalst, Wil. 2014. "Data Scientist: The Engineer of the Future." In *Proceedings of the I-ESA Conferences*, 13–26. https://doi.org/10.1007/978-3-319-04948-9_2.
- Acito, Frank, and Vijay Khatri. 2014. "Business Analytics: Why Now and What Next?" *Business Horizons* 57 (5): 565–70. <https://doi.org/10.1016/j.bushor.2014.06.001>.
- Agrawal, Divyakant, Philip Bernstein, Elisa Bertino, Susan Davidson, Umeshwar Dayal, Michael Franklin, Johannes Gehrke, et al. 2011. "Challenges and Opportunities with Big Data 2011-1." *Proceedings of the VLDB Endowment*, 1–16. <http://docs.lib.purdue.edu/cctech%5Cnhttp://docs.lib.purdue.edu/cctech/1%5Cnhttp://dl.acm.org/citation.cfm?id=2367572%5Cnhttp://docs.lib.purdue.edu/cctech/1/>.
- Analytics, Continuum, and others. 2016. "Anaconda Software Distribution." *Computer Software. Vers*, 2.
- Anonymous. 2016a. "Interview Manager Case Company (9.2.2016)." Turku.
- . 2016b. "Personal Communication 10.2.2016."
- . 2019. "Company Presentation."
- . 2020. "Company Internal Strategy Review with Personnel (5.2.2020)."
- Ayer, Vidya M., Sheila Miguez, and Brian H. Toby. 2014. "Why Scientists Should Learn to Program in Python." *Powder Diffraction* 29: S48–64. <https://doi.org/10.1017/S0885715614000931>.
- Azevedo, Ana, and Manuel Filipe Santos. 2008. "KDD, SEMMA and CRISP-DM: A Parallel Overview." *IADIS European Conference Data Mining*, 182–85. <http://recipp.ipp.pt/handle/10400.22/136>.
- Backman, Jere, Janne Vare, Kary Framling, Manik Madhikermi, and Ossi Nykanen. 2016. "IoT-Based Interoperability Framework for Asset and Fleet Management." *IEEE International Conference on Emerging Technologies and Factory Automation, ETFA 2016-Novem*: 0–3. <https://doi.org/10.1109/ETFA.2016.7733680>.
- Baraniuk, Richard G. 2010. "More Is Less : Signal Processing and the Data Deluge" 462.
- Barneveld, Angela Van, Kimberly E Arnold, and John P Campbell. 2012. "Analytics in Higher Education : Establishing a Common Language," no. January: 1–11.
- Bartur, Alon. 2015. "Data Wrangling to the People." Blog Post. 2015. <https://www.trifacta.com/introducing-trifacta-wrangler/>.
- Baskerville, Richard L., and A. Trevor Wood-Harper. 1996. "A Critical Perspective on Action Research as a Method for Information Systems Research." *Journal of Information Technology* 11 (3): 235–46. <https://doi.org/10.1080/026839696345289>.
- Baxter, Pamela, and Susan Jack. 2008. "Qualitative Case Study Methodology : Study Design and Implementation for Novice Researchers Qualitative Case Study Methodology : Study Design and Implementation" 13 (4): 544–59.
- Benbasat, Izak, David, and Melissa Mead. 1987. "The Case Research Strategy in Studies of Information Systems." *MIS Quarterly* 11 (3): 369–386. <https://doi.org/10.2307/248684>.
- Berthold, MR, Nicolas Cebron, and Fabian Dill. 2008. "KNIME: The Konstanz Information Miner." *Data Analysis, Machine ...*, 319–26. http://link.springer.com/chapter/10.1007/978-3-540-78246-9_38.
- Bhuskade, Shrikant. 2015. "Industrie 4.0: An Overview." *International Journal of Advance Engineering and Research Building Information Modeling (BIM)*, no. November: 1089–96.
- Boyes, Hugh, Bil Hallaq, Joe Cunningham, and Tim Watson. 2018. "The Industrial Internet of Things (IIoT): An Analysis Framework." *Computers in Industry* 101 (June): 1–12. <https://doi.org/10.1016/j.compind.2018.04.015>.
- Brynjolfsson, Erik, Lorin Hitt, and Heekyung Kim. 2011. "Strength in Numbers: How Does Data-Driven Decision-Making Affect Firm Performance?" *International Conference on Information Systems 2011, ICIS 2011* 1: 541–58. <https://doi.org/10.2139/ssrn.1819486>.
- Chen, Hsinchun, Roger H. L. Chiang, and Veda C Storey. 2012. "Business Intelligence and Analytics: From Big Data to Big Impact" 36 (4): 1165–88.
- Chen, Ming-syan, and Philip S Han, Jiawei; Yu. 1996. "Data Mining: An Overview from a Database Perspective" 8 (6): 866–83.
- Chessell, Mandy. 2014. "The Journey Continues From Data Lake to Data-Driven Organization."
- Christensen, G, A Steinmetz, B Alcorn, Amy Bennett, D Woods, and E J Emanuel. 2013. "The MOOC Phenomenon : Who Takes Massive Open Online Courses and Why ?" *Social*

- Science Research Network, 1–14. <https://doi.org/10.2139/ssrn.2350964>.
- “Climate Survey Report: 2015 Autumn.” 2015.
- Coenen, Frans. 2014. “Data Mining: Past , Present and Future.” *The Knowledge Engineering Review*, no. March 2011. <https://doi.org/10.1017/S0000000000000000>.
- Coghlan, David. 2003. “Practitioner Research for Organizational Knowledge.” *Management Learning* 34 (4): 451–63. <https://doi.org/10.1177/1350507603039068>.
- . 2019. *Doing Action Research in Your Own Organization*. SAGE Publications Limited.
- Cooper, Adam. 2012. “What Is ‘Analytics’? Definition and Essential Characteristics.” *CETIS Analytics Series* 1 (5): 1–10. <http://publications.cetis.ac.uk/2012/521>.
- Davenport, Thomas H. and Jeanne G. Harris. 2017. *Competing on Analytics: Updated, with a New Introduction: The New Science of Winning*. Boston, Massachusetts: Harvard Business Press.
- Davenport, Thomas H, and Randy Bean. 2018. “Big Companies Are Embracing Analytics, But Most Still Don’t Have a Data-Driven Culture.” *Harvard Business Review Digital Articles*, 2–4. <https://hbr.org/2018/02/big-companies-are-embracing-analytics-but-most-still-dont-have-a-data-driven-culture>.
- Davenport, Thomas H, and D.J. Patil. 2011. “Data Scientist: The Sexiest Job of the 21st Century.” *Harvard Business Review Digital Articles*, no. October 2012: 70–77.
- “Diagrams.Net.” 2020. Seibert Media GmbH. <https://drawio-app.com/>.
- Economist, The. 2017. “The World’s Most Valuable Resource Is No Longer Oil, but Data.” *The Economist: New York, NY, USA*.
- Endel, Florian, and Harald Piringer. 2015. “Data Wrangling: Making Data Useful Again.” *IFAC-PapersOnLine* 48 (1): 111–12. <https://doi.org/10.1016/j.ifacol.2015.05.197>.
- Floridi, Luciano. 2012. “Big Data and Their Epistemological Challenge.” *Philosophy and Technology* 25 (4): 435–37. <https://doi.org/10.1007/s13347-012-0093-4>.
- Fosso Wamba, Samuel, Shahriar Akter, Andrew Edwards, Geoffrey Chopin, and Denis Gnanzou. 2015. “How ‘big Data’ Can Make Big Impact: Findings from a Systematic Review and a Longitudinal Case Study.” *International Journal of Production Economics* 165: 234–46. <https://doi.org/10.1016/j.ijpe.2014.12.031>.
- Gandomi, Amir, and Murtaza Haider. 2015. “Beyond the Hype: Big Data Concepts, Methods, and Analytics.” *International Journal of Information Management* 35 (2): 137–44. <https://doi.org/10.1016/j.ijinfomgt.2014.10.007>.
- Hadley, Author, and Lionel Henry. 2018. “Package ‘Tidyr .’”
- Hamid, Idris Samawi, and Luigi Scarso. 2018. “Notepad ++ For ConTeXt MKIV.” <http://ctan.math.illinois.edu/support/npp-for-context/doc/npp-context-manual.pdf>.
- Hassani, Hossein, and Emmanuel Sirimal Silva. 2015. “Forecasting with Big Data: A Review.” *Annals of Data Science* 2 (1): 5–19. <https://doi.org/10.1007/s40745-015-0029-9>.
- Hellerstein, Joseph M, Jeffrey Heer, and Sean Kandel. 2018. “Self-Service Data Preparation : Research to Practice.” *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering* 41 (2): 23–34.
- Holsapple, Clyde, Anita Lee-Post, and Ram Pakath. 2014. “A Unified Foundation for Business Analytics.” *Decision Support Systems* 64: 130–41. <https://doi.org/10.1016/j.dss.2014.05.013>.
- Humby, Clive. 2006. “Data Is the New Oil.” In *Proc. ANA Sr. Marketer’s Summit*, 2. IL USA: Evanston.
- IBM. 2012. “IBM SPSS Modeler CRISP-DM Handbuch,” 51.
- Juhanko, Jari, Marko Jurvansuu, Toni Ahlqvist, Heikki Ailisto, Petteri Alahuhta, Jari Collin, Marco Halen, et al. 2015. “Suomalainen Teollinen Internet – Haasteesta Mahdollisuudeksi,” no. 42.
- Kandel, S, A Paepcke, J M Hellerstein, and J Heer. 2012. “Enterprise Data Analysis and Visualization: An Interview Study.” *IEEE Transactions on Visualization and Computer Graphics* 18 (12): 2917–26. <https://doi.org/10.1109/TVCG.2012.219>.
- Kandel, Sean, Andreas Paepcke, Joseph Hellerstein, and Jeffrey Heer. 2011. “Wrangler: Interactive Visual Specification of Data Transformation Scripts.” *Human Factors in Computing Systems. ACM*, 3363–72. <https://doi.org/10.1145/1978942.1979444>.
- Kaplan, Bonnie, and Dennis Duchon. 1988. “Combining Qualitative and Quantitative Information Systems: A Case Study.” *MIS Quarterly* 12 (4): 571–86.
- Klein, Heinz K., and Michael D. Myers. 1999. “A Set of Principles for Conducting and Evaluating Interpretive Field Studies in Information Systems.” *MIS Quarterly* 23 (1): 67–94.
- Luhn, Hans Peter. 1958. “A Business Intelligence System.” *IBM Journal of Research and Development* 2 (4): 314–19.

- Martine, George & Nicolas, Gladly. 2014. "4x4 Analytics Framework: Developing Organization and Their Data Scientists in Business Analytics." *Igarss 2014*, no. 1: 1–5. <https://doi.org/10.1007/s13398-014-0173-7.2>.
- Martinez-Plumed, Fernando, Lidia Contreras-Ochando, Cesar Ferri, Jose Hernandez Orallo, Meelis Kull, Nicolas Lachiche, Maria Jose Ramirez Quintana, and Peter A. Flach. 2019. "CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories." *IEEE Transactions on Knowledge and Data Engineering* 4347 (c): 1–1. <https://doi.org/10.1109/tkde.2019.2962680>.
- Mikalef, Patrick, Ilias O. Pappas, John Krogstie, and Michail Giannakos. 2018. "Big Data Analytics Capabilities: A Systematic Literature Review and Research Agenda." *Information Systems and E-Business Management* 16 (3): 547–78. <https://doi.org/10.1007/s10257-017-0362-y>.
- Miloslavskaya, Natalia, and Alexander Tolstoy. 2016. "Big Data, Fast Data and Data Lake Concepts." *Procedia Computer Science* 88: 300–305. <https://doi.org/10.1016/j.procs.2016.07.439>.
- Miraz, Mahdi H., Maaruf Ali, Peter S. Excell, and Rich Picking. 2015. "A Review on Internet of Things (IoT), Internet of Everything (IoE) and Internet of Nano Things (IoNT)." *2015 Internet Technologies and Applications, ITA 2015 - Proceedings of the 6th International Conference*, 219–24. <https://doi.org/10.1109/ITechA.2015.7317398>.
- Mishra, Nilamadhab. 2018. "Internet of Everything Advancement Study in Data Science and Knowledge Analytic Streams." *International Journal of Scientific Research in Computer Science and Engineering* 6 (1): 30–36. <https://doi.org/10.26438/ijsrcse/v6i1.3036>.
- Morgan, David E., and Rachid Zeffane. 2003. "Employee Involvement, Organizational Change and Trust in Management." *International Journal of Human Resource Management* 14 (1): 55–75. <https://doi.org/10.1080/09585190210158510>.
- Muenchen, Robert A. 2014. "The Popularity of Data Analysis Software," no. April. <http://r4stats.com/articles/popularity/>.
- Muhonen, T, H Ailisto, and P Kess. 2015. "Standardization in Industrial Internet (IoT) and Condition-Based Maintenance." *Proceedings at the Automaatio XXI ...*, no. March. https://www.researchgate.net/profile/Heikki_Ailisto/publication/277668082_Standardization_in_Industrial_internet_IoT_and_Condition-Based_Maintenance/links/5570007508aec226830abb55.pdf.
- Murre, Jaap M.J., and Joeri Dros. 2015. "Replication and Analysis of Ebbinghaus' Forgetting Curve." *PLoS ONE* 10 (7): 1–23. <https://doi.org/10.1371/journal.pone.0120644>.
- Norris, Nigel. 1997. "Error, Bias and Validity in Qualitative Research." *Educational Action Research* 5 (1): 172–76. <https://doi.org/10.1080/09650799700200020>.
- Northcutt, Curtis G., Andrew D. Ho, and Isaac L. Chuang. 2016. "Detecting and Preventing 'Multiple-Account' Cheating in Massive Open Online Courses." *Computers and Education* 100: 71–80. <https://doi.org/10.1016/j.compedu.2016.04.008>.
- Pappano, Laura. 2012. "The Year of the MOOC." *The New York Times*, 1–7. http://www.edinaschools.org/cms/lib07/MN01909547/Centricity/Domain/272/The_Year_of_the_MOOC_NY_Times.pdf.
- Pather, Shaun, and Dan Remenyi. 2004. "Some of the Philosophical Issues Underpinning Research in Information Systems: From Positivism to Critical Realism." *Annual Research Conference of the South African Institute of Computer Scientists and Information Technologists on IT Research in Developing Countries*, 141–46.
- Perez, Fernando, and Brian E Granger. 2015. "Project Jupyter : Computational Narratives as the Engine of Collaborative Data Science." *Retrieved September*, no. April: 1–24. <http://archive.ipython.org/JupyterGrantNarrative-2015.pdf>.
- Pfeiffer, Sabine. 2017. "The Vision of 'Industrie 4.0' in the Making—a Case of Future Told, Tamed, and Traded." *NanoEthics* 11 (1): 107–21. <https://doi.org/10.1007/s11569-016-0280-3>.
- Plotly. 2020. "Dash." Montreal. <https://plotly.com/>.
- Provost, Foster, and Tom Fawcett. 2013. "Data Science and Its Relationship to Big Data and Data-Driven Decision Making." *Big Data* 1 (1): 51–59. <https://doi.org/10.1089/big.2013.1508>.
- Qin, Wei, Siqi Chen, and Mugen Peng. 2020. "Recent Advances in Industrial Internet: Insights and Challenges." *Digital Communications and Networks* 6 (1): 1–13. <https://doi.org/10.1016/j.dcan.2019.07.001>.
- Redman, Thomas C. 1998. "Impact of Poor Data Quality on the Typical Enterprise." *Communications of the ACM* 41 (2): 79–82. <https://doi.org/10.1145/269012.269025>.
- Rogers, Shawn. 2011. "Big Data Is Scaling BI and Analytics." *Information Management* 21 (5):

- 14.
- Silver, Nate. 2012. *The Signal and the Noise: Why So Many Predictions Fail-but Some Don't*. Penguin.
- Sivarajah, Uthayasankar, Muhammad Mustafa Kamal, Zahir Irani, and Vishanth Weerakkody. 2017. "Critical Analysis of Big Data Challenges and Analytical Methods." *Journal of Business Research* 70: 263–86. <https://doi.org/10.1016/j.jbusres.2016.08.001>.
- Sommarberg, Matti. 2018. "Digitalization as a Driver of Industry Transformation," no. April.
- Stodder, David. 2016. "Improving Data Preparation for Business Analytics." *Transforming Data With Intelligence* 1 (1): 41.
- Swap, Walter, Dorothy Leonard, Mimi Shields, and Lisa Abrams. 2001. "Using Mentoring and Storytelling to Transfer Knowledge in the Workplace." *Journal of Management Information Systems* 18 (1): 95–114. <https://doi.org/10.1093/0195165128.003.0012>.
- Sydänmaanlakka, Pentti. 2007. *Intelligent Self-Leadership: Perspectives on Personal Growth*. Espoo: Pertec.
- Terrizzano, Ignacio, Peter Schwarz, Mary Roth, John E Colino, and San Jose. 2015. "Data Wrangling: The Challenging Journey from the Wild to the Lake." *7th Biennial Conference on Innovative Data Systems Research (CIDR '15) January 4-7, 2015, Asilomar, California, USA*.
- Trifacta. 2020. "Pricing." 2020. <https://www.trifacta.com/products/pricing/>.
- "Trifacta Wrangler." 2020. San Fransisco: Trifacta Software, Inc. <https://www.trifacta.com/>.
- Tuininga, Anthony. 2020. "Welcome to Cx_Oracle's Documentation!" 2020. <https://cx-oracle.readthedocs.io/en/latest/>.
- Tutunea, Mihaela Filofteia, and Rozalia Veronica Rus. 2012. "Business Intelligence Solutions for SME's." *Procedia Economics and Finance* 3 (12): 865–70. [https://doi.org/10.1016/s2212-5671\(12\)00242-0](https://doi.org/10.1016/s2212-5671(12)00242-0).
- Vassiliadis, Panos, and Patrick Marcel. 2018. "The Road to Highlights Is Paved with Good Intentions: Envisioning a Paradigm Shift in OLAP Modeling." *CEUR Workshop Proceedings* 2062.
- Wahlström, Henrik, Robin Wikström, and Viktor Sundholm. 2015. "Data Analysis of Large Datasets."
- Wan, Jiafu, Jiapeng Li, Muhammad Imran, and Di Li. 2019. "A Blockchain-Based Solution for Enhancing Security and Privacy in Smart Factory." *IEEE Transactions on Industrial Informatics* 15 (6): 3652–60. <https://doi.org/10.1109/TII.2019.2894573>.
- Ward, Jonathan Stuart, and Adam Barker. 2013. "Undefined By Data: A Survey of Big Data Definitions." *ArXiv.Org*, 2. <http://arxiv.org/abs/1309.5821v5Cnpapers3://publication/uuid/63831F5F-B214-46D5-8A86-671042BE993F>.
- Wes McKinney. 2016. "Pandas: Powerful Python Data Analysis Toolkit." 2016. <https://pandas-docs.github.io/pandas-docs-travis/>.
- Wickham, Hadley, Romain Francois, Lionel Henry, Kirill Müller, and others. 2015. "Dplyr: A Grammar of Data Manipulation." *R Package Version 0.4 3*.
- Wikström, Robin, Viktor Sundholm, and Henrik Wahlström. 2015. "Company Internal Meeting 20.10.2015."
- Williams, Graham J. 2009. "Rattle: A Data Mining GUI for R." *The R Journal* 1 (December): 45–55. http://journal.r-project.org/archive/2009-2/RJournal_2009-2_Williams.pdf.
- Wirth, Rüdiger. 2000. "CRISP-DM: Towards a Standard Process Model for Data Mining." *Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining*, no. 24959: 29–39. <https://doi.org/10.1.1.198.5133>.
- Xu, Li Da, Wu He, and Shancang Li. 2014. "Internet of Things in Industries: A Survey." *IEEE Transactions on Industrial Informatics* 10 (4): 2233–43. <https://doi.org/10.1109/TII.2014.2300753>.
- Yin, Robert K. 2003. *Case Study Research Design and Methods*. Third Edit. Sage Publication, Inc.

9 APPENDIX

Appendix A: Questions included in the interview

Q1: Tell me about Project A?

Q2: Tell me about Project B

Q3: How where these projects sold? Which factors affected the sale?

Q4: How does the analytics projects differ from other projects? Or is it possible to talk about other project, or are all projects unique?

Q5: How has your sales changed after the analytics projects A and B?

Q6: What potential do you see in analytics (or data projects) and how do they affect the business environment?

Q7: Where do you see Analytics services in the case company sales strategy for 2016? (Internal document, 4.6.2015. Knowledge management program “*Sales skills part 1*”, page 8)

Q8: What are your thoughts on case company future analytics projects. Any ideas for smaller or larger projects?

Q9: Regarding analytics competence at case company. What are the biggest challenges?

Q10: How would you create trust with new clients?

Q11: How do you identify and communicate about client needs?

Q12: How do perform pricing of analytics services?

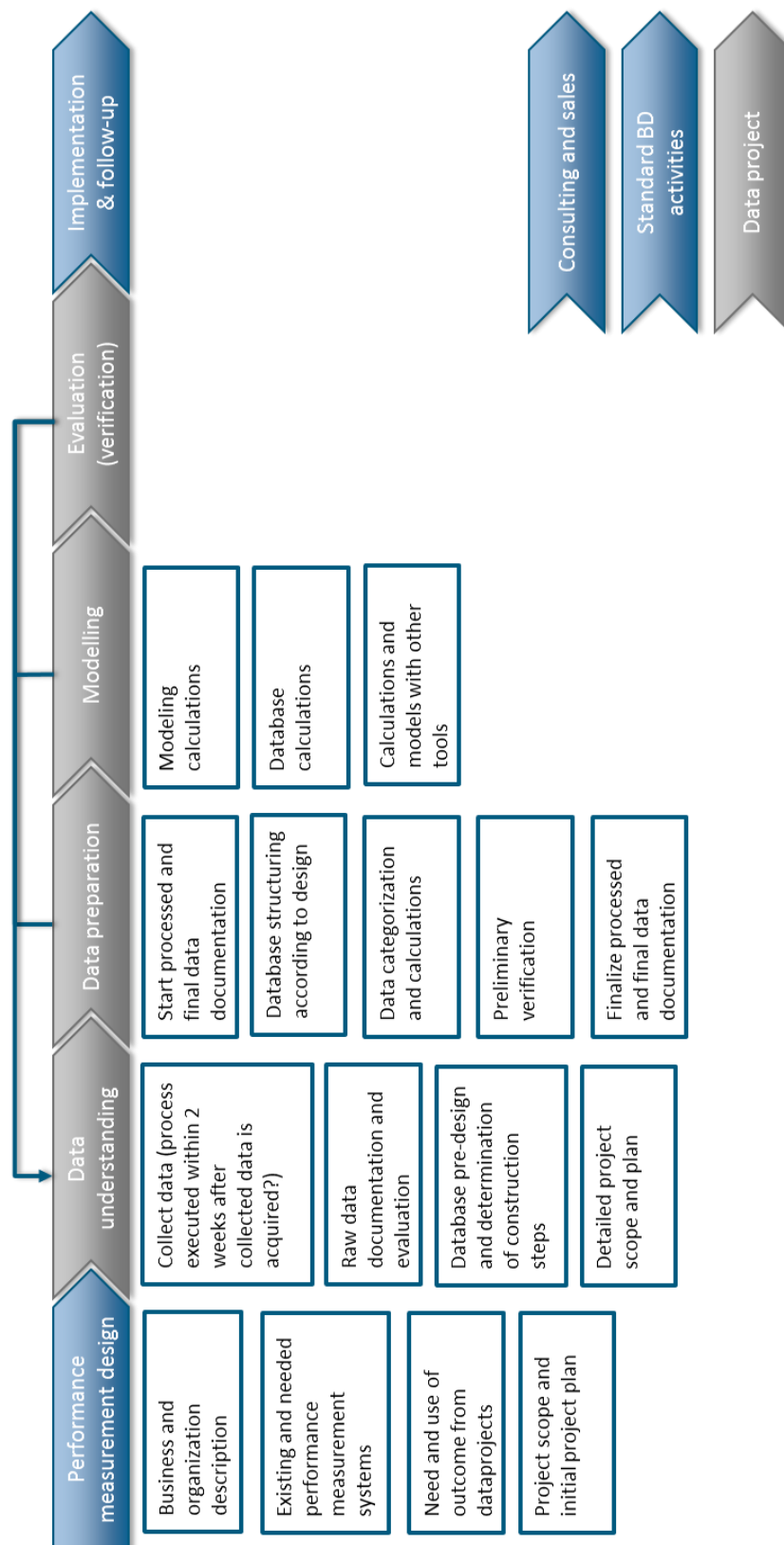
Q13: You held the company internal session on 4.6.2015 with the topic “*Sales skills part one*”. What would part two be about?

Appendix B: Table 8: Template for raw data documentation

Column	Column name	Column description	Relevance	Use	Comment	Format	Cleaning requirements
A							
B							
C							
D							
E							
F							
G							
H							
I							
J							
K							
L							
M							
N							
O							
P							
Q							
R							
S							
T							
U							
V							

Appendix C: Table 9: List of features for Trifacta software (Trifacta 2020)

FEATURES	Trifacta Wrangler	Trifacta Wrangler Pro	Trifacta Wrangler Enterprise
Pricing	Free (up to 100MB)	Starts at \$419/month per user* (billed annually)	Contact Us
Core Wrangling	yes	yes	yes
Collaboration	no	yes	yes
Scheduling	no	yes	yes
DB Connectivity	no	yes	yes
Custom Encryption	no	yes	yes
APIs	no	yes	yes
Run in Customer VPC	no	no	yes
LDAP Integration	no	no	yes
Deployment Manager (SDLC)	no	no	yes
User Defined Functions	no	no	yes
Data Catalog Integration	no	no	yes
Support/Training	Community & Tutorials	Optional 24x5	Optional 24x5, Onsite Training



Appendix D: Figure 12: Case company adaptation of CRISP-DM.