LAPPEENRANTA UNIVERSITY OF TECHNOLOGY 14.11.2018
Business Administration
Master's Programme Supply Management
Examiner Professor Jukka Hallikas

Master's Thesis

SITUATION STATUS BETWEEN DESCRIPTIVE AND PREDICTIVE ANALYTICS IN DECISION MAKING

Ville Lainio 2018

ABSTRACT

Author:	Ville Lainio
Title:	Situation status between descriptive and predictive analytics in decision making
Faculty:	School of Business and Management
Master's Programme:	Master’s Programme in Supply Management
Year:	2018
Master's Thesis:	Lappeenranta University of Technology
	72 pages, 7 figures, 4 tables, 1 appendix
Examiners:	Professor Jukka Hallikas
Key Words:	Data management, Data Mining, Analytics, Descriptive methods, Predictive methods, Data science

Statistical mathematics has received new hype words, like machine learning, advanced analytics and predictive analytics. Descriptive analytics has decades been in help and support of decision making in business environment. Nowadays predictive analytics is strongly discussed and linked to decision making. This master's thesis research, what is the situation status between descriptive and predictive analytics in decision making. The purpose is to examine, what is the gap between descriptive and predictive analytics in terms of data management and business users. The empirical research is conducted as multiple case study by interviewing consultants that have several years’ experience of data managements in business cases. The findings from the empirical research explained, that the technical aspect of descriptive and predictive itself do not have or create gap. But instead, managerial decision and understanding the holistic data management process are the reason for gap between these two methods. Additionally, it came to notice, that there is enough competence in available workforce, but the amount of competence is far too less to correspond the current need. Thirdly, budget is barrier factor which cause gap between the methods. The study contributes on the existing literature on data management by giving general view of data process, data mining and analysis methods that are known.

Tiivistelmä

Tekijä:	Ville Lainio
Otsikko:	Tilannekatsaus raportoinnin ja ennustavan analytiikan välillä
Tiedekunta:	School of Business and Management
Maisteriohjelma:	Supply Management
Vuosi:	2018
Pro-Gradu tutkielma:	Lappeenrannan teknillinen yliopisto
	72 sivua, 7 kuviota, 4 taulukkoa, 1 liite
Tarkastaja:	Professori Jukka Hallikas
Hakusanat:	Tiedonhallinta, Tiedonlouhinta, Analytiikka, Raportointi Ennustava-analytiikka.

Tilastomatematiikka on saanut rinnalleen uusia supersanoja, kuten koneoppiminen, edistynyt data-analytiikka ja ennustava analytiikka. Kuitenkin, raportointi on vuosikaudet ollut johtamisen ja päätöksenteon apuväline liiketoiminnassa. Tänä päivänä, ennustava analytiikka on nostanut päätään ja se on vahvasti linkitetty osaksi päätöksenteon apuvälineeksi. Tämän pro-gradu tutkimuksen tarkoituksena on tutkia, millainen kuilu perinteisellä raportoinnilla ja ennustavalla analytiikalla on datahallinnan ja liiketoiminnan näkökulmasta. Empiirinen osuus on tehty monitapaustutkimuksena, jossa haastatellaan alalla olevia konsultteja, joilla on useamman vuoden kokemus liittyen tietojohtamiseen ja analytiikkaan. Tutkimuksen tuloksista selvisi, että tekninen osaaminen niin raportoinnissa kuin ennustavassa analytiikassa ei ollut havaittavaa kuilua. Sen sijaan, liiketoiminnan päätökset ja puutteellinen kokonaisuuden ymmärtäminen datahallintaan loivat kuilua. Tämän lisäksi tutkimuksessa ilmeni alan työvoimasta, että osaaminen on riittävä mutta määrällisesti tätä osaamista on niukassa. Kolmantena, liiketoiminnan budjettirajoitukset luovat kuilua raportoinnin ja ennustavan analytiikan välillä.

Acknowledgements

Past two years has been really busy and event rich period. Still, I am now reaching to one of my top goals. It is obvious, that I could not get there without support. Thank you,

Professor Jukka Hallikas, for your help and guidance, but the most, your support towards my decisions.

Professors and their assistances for high quality education in both Supply management as in Business Analytics. With combination of these, I have managed to get excellent portfolio for my career.

My colleagues and dear friends. With your commitment and effort, we managed to accomplish our goals.

Lappeenranta 30.10.2018
Ville Lainio

Table of Contents

List abbreviations

BI = Business Intelligence

CRISP-DM = CRoss-Industry Standard Process for data mining

DB = Database

DM = Data Mining

DW = Data warehouse

KD = Knowledge Discovery

KDD = Knowledge discovery in database

POC = Proof of concept

PCA = Principle component analysis

List of Figures

List of Tables

1. INTRODUCTION

Business analytics is showing its place in managerial implications in today's business environment. It is the fastest popularity gaining subject than any other managerial paradigms have witnessed in recent years. Main reason for this is, that it potentials provide managers to take advantages of data and use it for better decision making. Effectiveness of business analytics systems lies on volume and quality of data, accuracy, integrity and timeliness. This all come together with suitable, efficient tools and processes that is needed when wrangling with data. (Delen, Demirkan 2013)

Acito and Khatari (2014) describe business analytics' core being about extract value from data. They address, that data should not be referred as the "sludge of the information age" but more as "the new oil". It is not easy task to extract value from data, especially when volume and even velocity is high. It surely offers opportunities and data can also be used to identifying market niches, discovering new ways to develop new products and services.

To dig deeper and providing solid foundation for the thesis. Davenport and Harris (2007) described business analytics being concerned with "the extensive use of data, statistical and quantitative analysis, explanatory and predictive models, and fact-based management to drive decisions and actions". Main idea or more suitable approach to these three definitions is that business analytics affected with decision-making.

In the same token, Vidgen, Shaw and Grant (2017) writes about the popularity of business analytics and how it is increased tremendously in the last decade. They rise a view that world is now in state where there is an enormous amount of data. Referring to digital trace, which means that where ever people go or do, there will be some kind of digital mark and it is recorded and stored. On the other hand,

there is much potential but also chance for pitfalls. Around business analytics organizations are trying to figure out ways to explore massive amount of data and how to extract it to create value. Data analytic methods are being used in many and varied ways. For example, different ways of predictions based on history data. Classical example would be prediction of consumer choices or predict the likelihood of medical condition. Today's popular way to wrangle with data is related to social networks and social media. (Vidgen, Shaw & Grant 2017)

It has been hard not miss discussions and growth of analytics in last decade. Articles in media and literature of business and technology related books have introduced different of ways to interpret and use of analytics. It rises subjects of collection, storage and analysis of massive amounts of data. Data is collected virtually every aspect of human activities. These collected data has been used carefully in designed experiments and investigations. Additionally, data has been collected also from operation of vehicles, factories and natural phenomena's. The term data mining was long time ago reflected to sort of negative action which was used to describe unguided sifting through numbers in hopes of discover insights. These turned to be too fragile or illusory. Nowadays, data mining and related techniques have become accepted and often lead to useful discoveries. (Acito, Khatri 2014)

This leads to word called business intelligence. Thomas and Charles (2006) defines it as the extraction of insights from structured data that has a long history. It reflects with previously introduced concepts as decision support, data warehousing and data mining. The business intelligence literature is full of discussions of technologies that extracts, transforms and loads (ETL) data for statistical analysis and descriptive reporting.

Hindle and Vidgen (2017) address in their research paper about methodologies in information systems development. They can range from the software-focused to organizational. They appear to be less common in business analytics and data science. As Hindle and Vidgen points out the lack of methodologies in literature, they found one exception from these claims and that yields in field of data mining.

The researcher, while writing this master's thesis is working among business intelligence and has occur these previously mentioned in daily basis. Still, there was no clear understanding of the situation or status among descriptive and predictive methods as part of decision making when data is involved. Therefore, next chapter presents the research problem and the questions related to it.

1.1. Research problem, objectives and delimitation

It is possible to get caught and lost with all hype-terms related to data management, especially when issuing with analytics and reports. This paper is trying to get understanding of the current situation what comes to usage of descriptive and predictive methods as part of decision making. Main point is to figure out the real situation behind the curtains. Therefore, the main research question is:

1. *What is the gap between descriptive analytics and predictive analytics in Finnish companies that uses business intelligence in decision making?*

The idea of this question is to search possible gaps between two methods. Especially because descriptive methods are rather long been in the game and predictive methods has just gained more headlines in recent years.

Additionally, there are two sub-questions which are to support the main question. To research gaps, it is quite naturally related to competence and resources. Therefore, two sub-questions are:

1. *What kind of capabilities is needed for companies to take advantage of predictive methods?*
2. *What kind of resources companies need to have for implementing predictive methods?*

Aim for this qualitative research is to obtain new insight and knowledge by interviewing experienced consultants that have implemented and guided data manage-

ment processes at different industries in Finland. As this thesis is based on personal interest of the writer in field of data science, not in supply management. The structure and issues related to research are written in underlying level so that reader with no touch of data science would have still solid understanding of the research after reading it.

The research is interesting not only by personal interest, but also, because World Wide Web platforms, journals, social media posts are putting emphasises to write about advanced analytics in part of business. It would be interesting to see, if it is still in talk level or are companies really implemented predictive methods along descriptive methods as part of decision making.

Therefore, the results of the research could be interesting for junior consultants or junior position data scientist that enters to business world. Even more, this study can be beneficial to university students that are studying data science.

1.2. Conceptual framework

The literature review is based on three concepts that are presented in logical structure, so that together they form big picture as paper goes towards empirical part. Figure 1 illustrates these three subjects: data mining, data management process and advanced analytics. These topics creates the border for this thesis. Subjects builds on top of each other, so that it is more understandable for reader (also for the writer) to continue empirical part of the thesis. Notice, word data mining can have different meaning in different environment, but here, it represents the path from data management towards analytic methods.

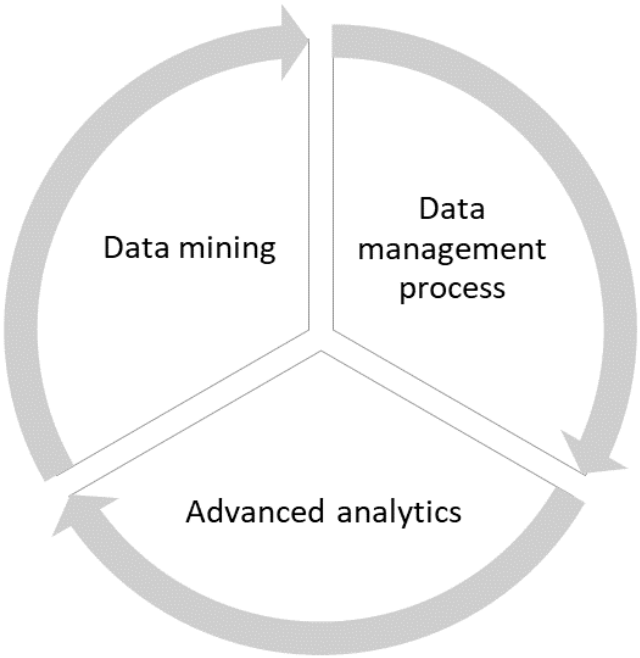

Figure 1 Conceptual framework of the master's thesis

Data mining introduces concept of data mining and presents fundamentals of three different analytical categories. Additionally, data mining results and benefits are explained using existing literature.

Data management is fundamental base for the paper and it presents two major scholars, what comes to data management literature. As the paper introduces also about Knowledge Discovery in Database (KDD), the interviews and the discussions are based on cross-industry standard process (CRISP-DM) method.

The last layer in conceptual framework is advanced analytics, which is the climax of the literature review. By understanding through previously explained subjects, this part of the layer concludes and gives a reasonable understanding towards the main research question (1) “What is the gap between descriptive analytics and predictive analytics in Finnish companies that uses business intelligence in decision making?” Additionally, two sub-questions are all related to these three layers

(2) “What kind of capabilities is needed for companies to take advantage of predictive methods?” and (3) “What kind of resources companies need to have for implementing predictive methods? “.

Understanding these three subjects through literature is essential and it build up solid foundation for empirical part of the study. The framework will be used in a way, that research questions reflects to them and it also gives correct direction when analysing the results and making conclusion.

1.3. Methodology

The study is conducted by qualitative method using semi-structed interviews. For data collection structured list of questions were made and asked from the case. After collecting the data. The writer used with-in analysis and cross-case analysis to find similarities and differences that may or may not explain the gap between descriptive and predicative analytics in decision making.

1.4. Definitions of key concepts

1.4.1. Data management process

Data management presents the structure or method how unstructured data can be handled in a way that is usable. In another words, data management is a process of organizing data which can give leverage in terms of achieving sustainability, improving innovativeness and being able to reply environmental changes. (Argote & Ingram, 2000; Davenport & Prusak, 2000; David J. Teece, 2007; Thrassou & Vrontis, 2008) Garcia, Herrera and Luengo (2015) addresses it as ”the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data”. It could additionally be said, that data management conduct an automatic exploratory data analysis of large databases.

1.4.2. Data mining

Data Mining is subject, which try to solve problems by analysing data in real databases. Nowadays, it is qualified as science and technology for exploring data to discover already present unknown patterns. (García, Herrera & Luengo 2015) When data management collects pieces of data from different sources, even they are irrelevant to each other. Data mining process gives opportunity to investigate these data as whole new and useful information may emerge. (Wang et al., 2018)

1.4.3. Advanced analytics

Advanced analytics in simple term, predicts what's ahead. Practical example would be price optimization for big store chain or just for a local store, using existing data of product prices versus purchase prices by applying statistical tools. (Bradlow, et al., 2017; Hashimzade, et al., 2016) Lorenzo et al. (2018) summarises: "*predictive analytics has been exploited for several years by many lucrative business endeavours to individualize and maximize their reach to potential consumers, monetizing based on the rich profiling generated by these vast amounts of data".*

1.5. Research process

First glance of the research happened in autumn 2017, when the writer and professor sat down and talked about the topic. It was very clear in the beginning, that the personal interest for research lies somewhere among business analytics. After few conversation sessions and email exchange, the frame for the master's thesis started to appeal. During the writing process, the writer is working full time which created barriers towards timeframe. To be able finish given one-year timeline for the thesis, solid time management structure needs to be conduct.

The researcher divided time scale from September 2017 to October 2018 in four section. Figure 2 illustrates the process. First, the literature review will be written till end of January 2018 and the requests for interview sent. During February the

structure of interview need to be done and verified by professor. In third section, interviews were done and also research plan presented respectively in master's thesis seminar -course. The last section, during the writer's summer holiday, empirical part would be written and the adjustment for the thesis would be done.

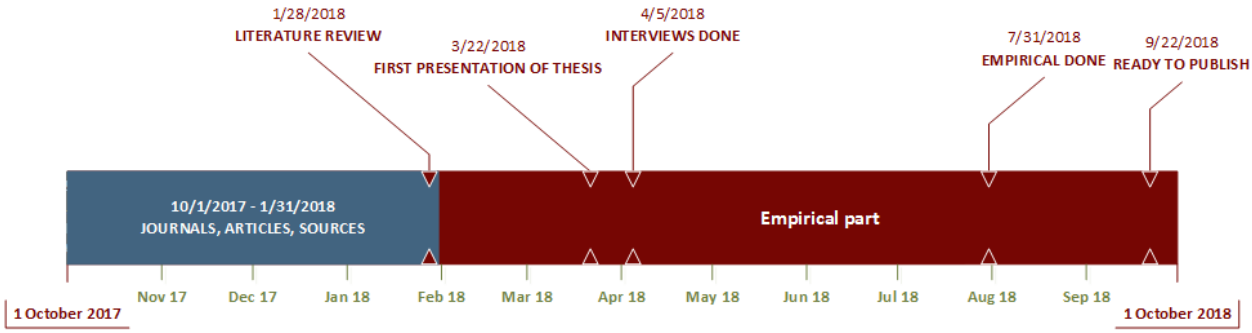

Figure 2 Research process

Because very strict timeline, two out of five interviews were able to conduct. Both of them started end of the second time section, which postponed progress of the thesis. Despite all the time factors, writing continued on July 2018 and thesis was ready to be evaluated in November 2018.

During the whole process, the writer kept professor informed by emails. These emails usually consisted updated versions of the thesis or suggestions on top of previous written version. For time and process management, the writer used office 365-programs to create structured work path.

1.6. Thesis structure

Structure of the research is demonstrated in figure 3. It is built so, that reader could form solid mindset towards what is being investigated. Thesis starts with introduction and moves towards data mining and its categories, presenting basic concept of data mining. Additionally, section introduces to three different categories of analysis: descriptive-, predictive- and perspective analytics. Second section of the thesis is dedicated to data management processes, which focuses on core basis of known

literature of data handling. It introduces to two famous concepts; knowledge discovery in database and cross-industry process standard for data mining.

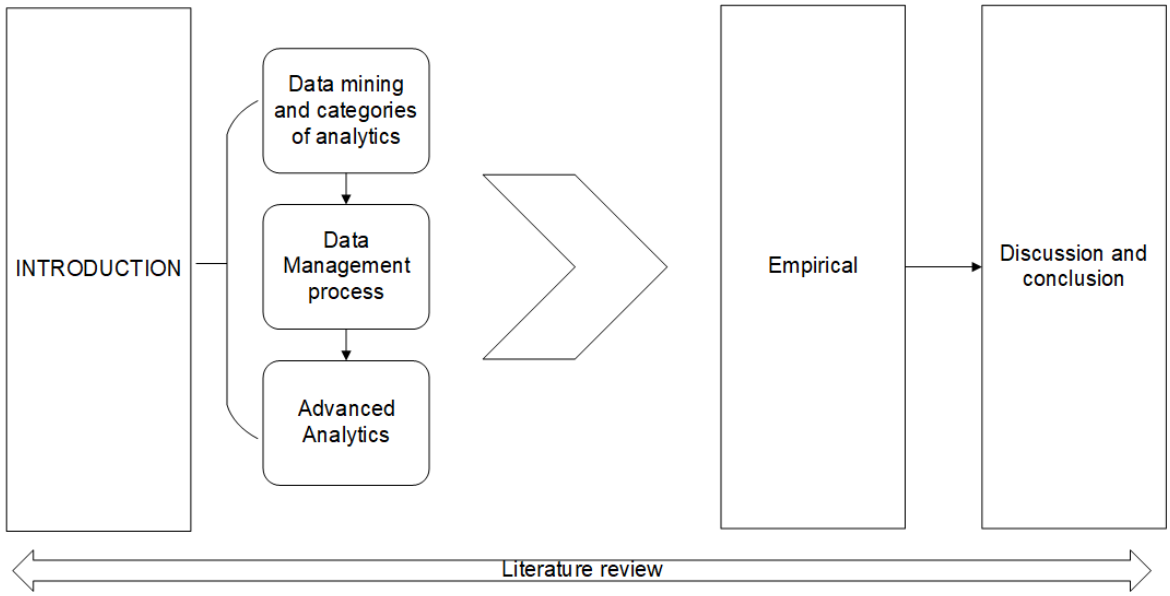

Figure 3 Thesis structure

Third part of the literature focuses underlying level of advanced analytics. It introduces to two categories of statistical methods for prediction: supervised and unsupervised learning. Lastly, the chapter presents two groups of predictive analytics; classification and regression cases. After literature review, methodology and data collection methods are presented. Chapter gives broader understanding of selected qualitative research method and also justifies why semi-structured questions are used. Furthermore, reliability and validity are explained.

Sixth chapter presents the outcome of the interviews as also their comparison between each other. Thesis continues to discussion with conclusion and ends with future research suggestion.

2. DATA MINING AND CATEGORIES OF ANALYTICS

Chapter issues two topics; data mining and three categories of analytics. Data mining will be referred in the research as a process, that is part of data management but without the base of data management, it cannot be used. Therefore, when issuing analytics methods, understanding underlying theory of data mining is important. Categories of analytics, all three methods are presented but only descriptive and predictive methods are essential for the research. Perspective method is introduced as complimentary.

2.1. Data mining

Data is something that we collect and store. Knowledge is the information that we want to get out from the data, to make better decisions. The extraction of knowledge from data is called data mining. It is a method to discover meaningful patterns and rules from large quantities of data by exploration and analysis. (Negnevitsky 2011)

Negnevitsky (2011) address that world is now on phase where data is rapidly expanding. The quantity of data roughly doubles every year and professionals are struggling finding the information in huge amount of data. Negnevitsky rise few examples, NASA has more data than it can analyse, and Human Genome project researchers have to store and process thousands of bytes for each of three billion DNA bases that make up the human genome. Every day huge amount of data passes through internet and there is urgent need to have right methods, to extract useful and meaningful knowledge from that data mass.

In modern, competitive business world data mining is becoming essential. Data mining has referred to gold mining, because large quantities of ore must be processed before the gold can be extracted. Same idea goes with processing data. There are sometimes million or billion rows of information and it needs proper handling, so that the value knowledge can be identified.(Negnevitsky 2011)

Organisation that want to be successful, especially in data driven world, they need quickly respond to changes in market. To accomplish this, accessing to current data that usually are stored in operational databases, by meaning organisations should have some kind data warehouse solution. Furthermore, an organisation must also determine which trends are relevant to their business. (Negnevitsky 2011)

Data warehouse is like a big pool that can have huge capacity. Data warehouse could include million, even billion of data records. Negnevitsky (2011) describes it as “time dependant – linked together by the times of recoding and integrated. All relevant information from the operational databases is combined and structured in the warehouse”.

Data is handled by user-driven techniques where user generates a hypothesis and then test and validates it with available data. Furthermore, human mind at best, can handle three or four attributes when searching correlation. In real world, truth is that in data warehouses can lay records with dozens of variables and there may be hundreds of multifaceted relationships among each other. Human brain would have hard time to process all that. (Negnevitsky 2011)

Negnevitsky (2011) adds, that statistic and regression analysis are powerful way to interpret with data. Statistic collect, organise and utilise numerical data and it gives general information about data. Statistical numbers like average and median values, distribution of values and observed errors. In other hand, regression analysis, one of the most popular technique for data analysis. Statistic is suitable in analysing numerical data, but it does not solve data mining problems. These problems are discovering patterns and rules in large quantities of data.

Provost & Facett (2013) distinguish two things from mining data. It can be categorized in two set: the difference between mining the data to find patterns and build models and using the result of data mining. It is not rare that people confuse these two processes while talking about data science or business analytics. “The use of

data mining result should influence and inform the data mining process itself, but the two should be kept distinct.

2.2. Analytic categories

Data mining which operates as gateway to analytics establish several ways how we can interpret with data. Reporting data to analyse trends, creating predictive models to identify potential challenges and opportunities in near future, providing new ways to optimise business processes to enhance performance. There are three main categories in analytics: descriptive, predictive and prescriptive. (Delen & Demirkan, 2013)

2.2.1. Descriptive analytics

Descriptive analytics which also may refer to business reporting and it is used to answer questions like “what is happened” or what is happening”. It stands for rather basic and simple analytic form for business, example given, ad-hoc or on-demand reporting but as well dynamic and interactive reporting. Main issue for descriptive analytics is recognizing business problems and opportunities. (Delen & Demirkan, 2013) Lestringant et al. (2018) researched, how have conventional descriptive analysis methods really been used. The outcome of the analytics is typical descriptive analytics method. Using summary statics at different levels to get answers.

2.2.2. Predictive analytics

Predictive analytics uses pre-processed data and mathematics to learn predictive patterns and creates output by interpret the relationship with input and output data. It answers questions “what will happen” or “why will it happen”. To proceed with predictive analytics, it involves steps in data mining, web or media mining and statistical time-series forecasting. The main outcome for predictive analytics is an accurate estimate of possible future outcome. (Delen & Demirkan, 2013)

2.2.3. Perspective analytics

Delen & Demirkan (2013) rise characteristics attributes for perspective analytics by stating that it uses data and mathematical algorithms to regulate alternative courses of actions for decision given a complex set of objectives, requirements and constrains, with aim to refining business performance. These algorithms may lay on data, on expert knowledge or a combination of both. To establish prescriptive analytics, it is essential to think of model optimization, model simulation and multi-criteria decision modelling. The outcome of all these is either the best course of action for a given situation, or rich set of information and expert opinions to conduct best possible action for a decision maker. (Delen, Demirkan 2013)

2.2.4. Brief comparison between descriptive and predictive analytics

Table 1 illustrates the differences between descriptive and predictive analytics. As descriptive method seeks answers for history events "what happened and what is happening" as predictive method emphasises the future scenarios "what will happen and why will it happen".

Table 1 Comparison table between descriptive and predictive analytics

Attribute	*Descriptive*	*Predictive*
What happened and what is happening	X	
what will happen and why will it will happen		X
Uses history data	X	X
Dashboards and score-cards and reports	X	
Data warehouse	X	(X)
Forecasting		X

Methods share same attribute for history data. Both methods require history data, otherwise method is not usable. Characteristic for descriptive analysis are dashboards, scorecard and specific reports. Predictive analysis is result orientated based on forecasting. It could also say that forecasting methods can set in dashboard layer but descriptive analysis it is more common. Forecasting means, using statistical methods to calculate distances using algorithms like classification or regressions. Data warehouse is the typically the base for creating descriptive analytics. Additionally, for predictive methods, it would be good to have database, but it is not necessary, if the input data already exist in correct form.

2.3. Data mining results

Considering the business use of data management in view of supply chain risk management. Fan et. al. (2017) created supply chain risk management concept which held three main categories: risk information and sharing, risk analysis and assessment and risk sharing mechanism. Combination of three steps, the outcome was to apply supply chain risk knowledge to conduct supply chain risk management decision by using data, it was possible to prepare for a risk event before it occurs. Additionally, they suggested to combining supply chain risk management and data mining, to create information sharing platform, which is basis for the risk sharing mechanism among supply chain partners.

As the digitalization of supply chain networks, the vast amount of data will be accessible and that offers faster recognition and responses to potential risks. In supply chain management, simulations based on data can be answer to many risk management problems. Therefore, it can be said, that data mining and its results play critical role in managerial implications when solving complex real-word problems related to supply chains. (Chen et. al. 2013) Govindan et al. (2018) wrote: “Recent studies in the field of big data analytics have come up with tools and techniques to make data-driven supply chain decisions. Analysing and interpreting results in real time can assist enterprises in making better and faster decisions to satisfy customer requirements. It will also help organisations to improve their supply chain design and management by reducing costs and mitigating risks”.

Several studies issues with the benefits of data mining towards business development and decision making. Peral et. al. (2017) Introduced a research which used data mining to discover key performance indicators. To monitor business performance, dashboards are commonly used to show graphical illustration of key performance indicators. Key performance indicators provide accurate information by comparing current performance, but it is sometimes difficult to identify indicators. As data mining techniques are used forecasting trends and correlations, it can also be used to recognizes possible performance indicators. Furthermore, Amani & Fadlalla (2017) wrote a paper which discovers applications of data mining techniques in accounting. Their framework showed that area of accounting did benefit from data mining techniques in segments like fraud detection, business health and forensic accounting.

3. DATA MANAGEMENT PROCESS

The chapter introduces reader to two most fames concepts of data management, Knowledge discovery in database (KDD) and CRoss-Industry Standard Process for Data Mining (CRISP-DM). As Knowledge discovery is popular in science environment, the other is more industry and business environment friendly. Aim of the chapter is to give holistic view, how data should be process and store. Data management creates base for the next two chapters, which are topics of using the data to get information. Emphasises will be on CRoss-Industry Standard Process concept, and Knowledge discovery in database in complementary part but it plays huge role in literature.

3.1. Knowledge Discovery in database

Fayyad et al. (1996) described in mid 90's that digitalisation is taking fast leaps forward. This means larger amount of data is processed and stored which eventually leads data overload. To handle situation, like fast growing data streams and storages, there is need for better computational power and techniques to extract the useful information from large data mass. Data can be gathered from different sources to needed purpose. For example, local store's checkout register, bank's credit card authorization device, records of people doctor office, patterns of telephone calls and much more. These data information can be stored in databases or as nowadays called data warehouses. With all new, fast generating data there are potential to use them in business. The knowledge from data can be used to introduce new targeted marketing campaigns with potential financial returns. Or another example is from field health and well-being where data is extracted and used to detect medical conditions. (Colak, et al., 2015; Fernández-Arteaga, et al., 2016; Liou & Chang, 2015; Yang & Chen, 2015). These techniques and tools are the subject of knowledge discovery in database (KDD) and data mining.

True value for detecting information in data and interpret it successfully lies in people. Ability to extract useful reports, spot attractive trends, support decisions and exploit data to achieve business, operational or scientist goals. Problems arise

when scale of data manipulation, exploration and interpretation grows beyond human capacities. Therefore, people need to rely on computer technology. The problem of knowledge extraction from large databases involves many steps, ranging from data manipulation to fundamental mathematical and statistical inference, search and reasoning. (Fayyad et al., 1996)

There are several names for the operation, which try to find useful patterns in data. Few of them as example are; knowledge extraction, information discovery, information harvesting, data archelogy and data pattern processing. Term “data mining” is used by statistician and business communities. Fayyad et al. (1996) exclusively uses knowledge discovery in database (KDD) to describe overall process of discovering useful knowledge from data. They add, that data mining is a process step in overall process flow. Fayyad et al. (1996) mentioned their view of KDD position in middle of growing data phenomenon. KDD has evolved, and it will continue to evolve, from the intersection of research in such fields as databases, machine learning, pattern recognition, artificial intelligence, data visualization et cetera. That statement supports several different researches that are using knowledge discovery in database method (Chen, et al., 2014; Dehning, et al., 2016; Neto, et al., 2017; Schuh, et al., 2017).

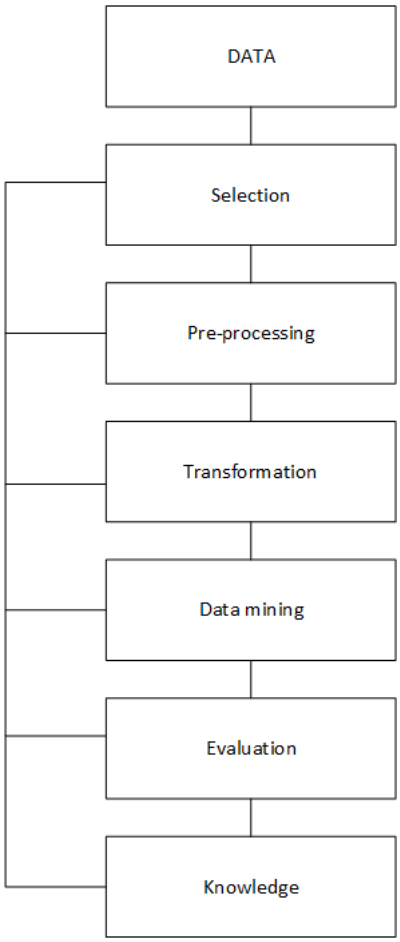

Figure 4 Knowledge discovery in database process flow (Fayyad et al. 1996)

Fayyad et al. (1996) define knowledge discovery process: “The nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data”. Figure 4 presents the knowledge discovery in database process, it is interactive and iterative involving nine steps, described from the practical viewpoint. 1. Learning the application domain: includes relevant prior knowledge and the goals of the application. 2. Creating a target dataset: includes selecting a dataset or focusing on a subset of variables or data samples on which discovery is to be performed. 3. Data cleaning and pre-processing: includes basic operations, such as removing noise or outliers if appropriate, collecting the necessary information to model or account for noise, deciding on strategies for handling missing data fields, and accounting for time sequence information and known changes, as well as deciding issues, such as data types, schema, and mapping of missing and unknown values. 4. Data reduction and projection: includes finding useful features to represent the data, depending on the goal of the task, and using dimensionality

reduction or transformation methods to reduce the effective number of variables under consideration or to find invariant representations for the data. 5. Choosing the function of data mining: includes deciding the purpose of the model derived by the data mining algorithm. 6. Choosing the data mining algorithm(s): includes selecting method to be used for searching for patterns in the data, such as deciding which models and parameters may be appropriate (e.g., models for categorical data are different from models on vectors over reals) and matching a particular data mining method with the overall criteria of the KDD process. 7. Data mining: includes searching for patterns of interest in a particular representational form or a set of such representations, including classification rules or trees, regression, clustering, sequence modelling, dependency, and line analysis. 8. Interpretation: includes interpreting the discovered patterns and possibly returning to any of the previous steps, as well as possible visualization of the extracted patterns, removing redundant or irrelevant patterns, and translating the useful ones into terms understandable by users. 9. Using discovered knowledge: includes incorporating this knowledge into the performance system, taking actions based on the knowledge, or simply documenting it and reporting it to interested parties, as well as checking for and resolving potential conflicts with previously believed (or extracted) knowledge. (Fayyad et al., 1996)

3.2. Cross-industry standard process

In 1996 there was no acceptable approach to data mining method for industries, companies nor organizations. There was call for method, which would adapt academic data mining methodology. The development of a non-proprietary, documented and freely available model would enable organizations to realize better results from data mining. Cross-Industry Standard Process for Data Mining (CRISP-DM) idea started in late 1996 by four companies that interpret with data mining market: Daimler-Benz, Integral Solutions Ltd., NCR and OHRA. Daimler-Benz were that time company which led industrial and commercial organizations to apply data mining in its business operations. Following, ISL was first to provide services based on data mining principles in 1990. NCE was aiming to deliver added

value to its Teradata data warehouse customers. OHRA, Dutch insurance company provided a testing ground for live, large-scale data mining projects. These companies had vested vision for standardised data mining technique for industries and eventually, next several years DM Special Interest Group (SIG) was formed. Idea to develop a standard process model to service the data mining community. (CRISP-DM 2013; Shearer 2000)

CRISP-DM is a data mining methodology which covers whole process from beginning to the end. It is designed to be adaptable different industries to conduct data mining projects. It consists of six phases or a better called, data mining cycle: business understanding, data understanding, data preparation, modelling, evaluation and deployment. Figure 5 Illustrates this cycle. The arrows show the most important and frequent dependencies between phases and circle highlights the cyclical nature of data mining. (Shearer 2000)

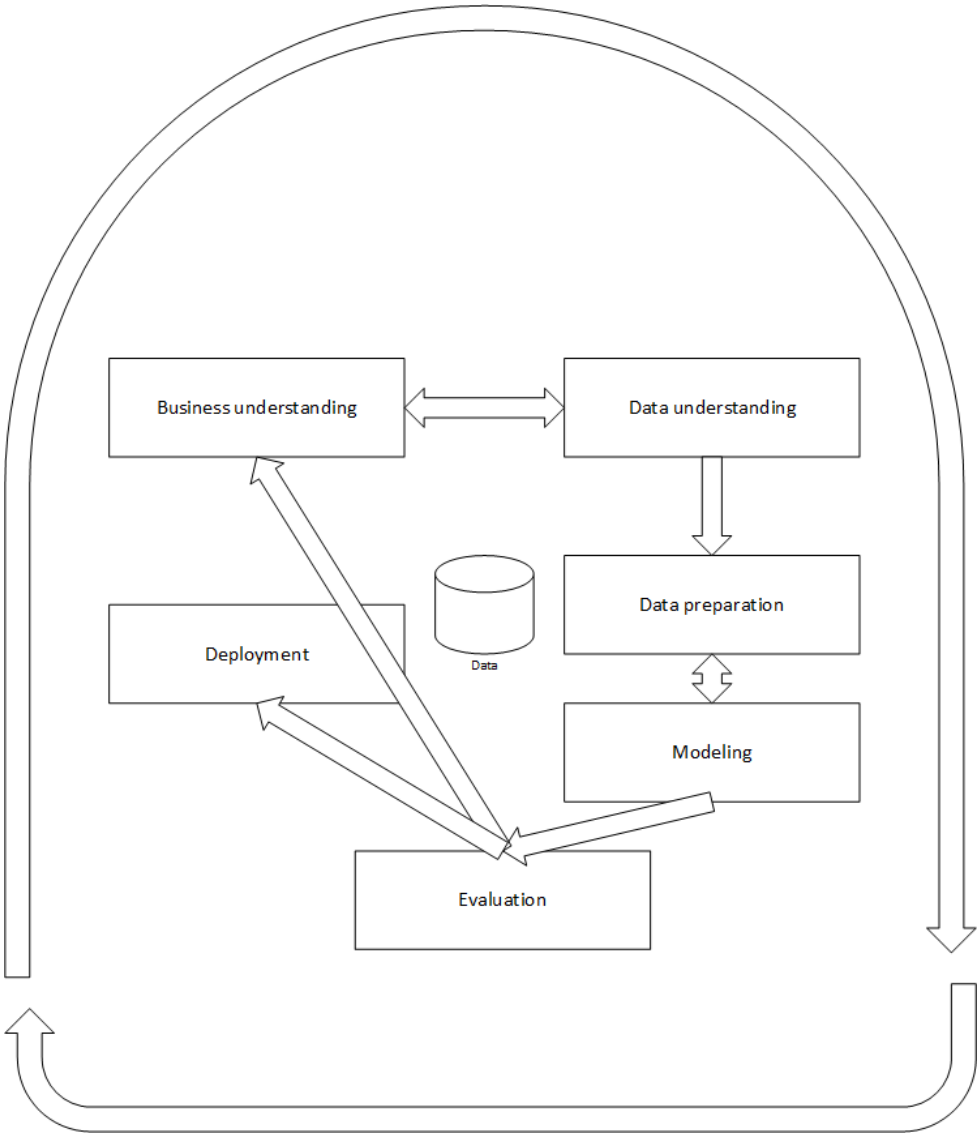

Figure 5 CRISP-DM process flow (CRISP-DM 2015)

There are many researchers that used cross-industry standard process for data mining and it has been adoptable for different industries. Groggert et al (2018) Used CRISP-DM to create scenario-based manufacturing, Poh et al. (2018) used same framework when they investigated safety indicators for construction sites and Morais et al. (2017) when predicting new-borns that need assistance for breathing at birth.

3.2.1. Business understanding

To start data mining project, it is essential to understand the business perspective and collect enough information of it. Ideally, for successful data mining project is to reflect business case with all possible data that might be used during the process. It is crucial to understand which data should be analysed in beginning and which later. Business understanding phase includes several steps: determining business objectives, assessing the situation, determine the data mining goals and producing project plan. (CRISP-DM 2015; Shearer 2000)

Determining business objectives and understanding them is important for data analyst. Providing right answer to wrong question or other way around is a situation that should be avoided. Additionally, when figuring out the business objectives, there is also definitions of measures. To adjust right measures which have rational outcomes instead of measures which are rather absurd from the beginning. Therefore, every settled measure should be connected to business objective.(Shearer 2000; CRISP-DM 2015)

Data analyst should assess the company situation. What kind of personnel is involving to the data mining project, which software is used or needed? Is there existing data that can be used or is there potential to find new data which help to resolve data related questions towards business problem? Risk identification and how to tackle these risks are part of the assessment process. Furthermore, providing cost-benefit analysis for the project involves assessment category. (CRISP-DM 2015; Shearer 2000)

Setting the data mining goals states project objectives in business terms. By this, there should be a realistic goal. Example given, to make a prediction, there should be a certain goal or a measure towards the prediction. Measure can be prediction accuracy or exceeding threshold value. (Shearerm, 2000)

Project planning refers how to execute data mining project and achieve settled goals. This includes outlining specific steps and timeline, assessment of possible

risks and assessment for needed tools and techniques to finish the project. There are general timeline standards: 50 to 70 percent of the time is resourced for data preparation. 20 to 30 percent for data understanding, only 10 to 20 percent is spent in modelling, evaluation and business understanding. 5 to 10 percent is spent in the deployment planning. (Shearer 2000)

3.2.2. Data Understanding

After collecting data, next several steps determine how data should be interpreted. This need wider understanding of what kind of data is available. To identify possible quality problems, to discover potential insights into the data or detect interesting subsets to form hypotheses about hidden information. There are four essentials steps to in data understanding phase: collecting the initial data. data description, data exploration and verification of data quality. (CRISP-DM 2015; Shearer 2000)

Collecting initial data, the data analyst gathers data from one or many sources, loads and potentially adjusts it if necessary. The data analyst reports all possible problems which he or she encountered, so the next person would have information about it if the process is needed to repeat. Gathering and combining source information are typical workflows in this part.(CRISP-DM 2015; Shearer 2000)

In data description part, data is examined to identify data structure. This means, identifying what is the data format, example given; date, character, numeric. Furthermore, investigating the quantity of data, whether all data is useful or is there potential dropouts already in this part. Data analyst should have eye for upcoming data handling phases; perhaps some of information is not needed now but will later on. Outcome is to understand dataset and how it can be used. (CRISP-DM 2015; Shearer 2000)

Data exploration is process where the data gets mingled. There are data related questions which answers may found when scraping the data. This means, using methods like quarrying, visualization and reporting. Querying can be for example,

discover what kind of product certain income group buys. Visualization is powerful when searching patterns in data, like fraud cases. In reporting, data analyst should provide outlines for first findings or initial hypothesis and the potential impact on the remainder of the project. (CRISP-DM 2015; Shearer 2000)

Last part of data understanding process is assessing the data that you have. When working with data, there are chances to have missing data in rows which may occur when data is pulled from different sources et cetera. There may be data type which is in wrong format, length or other issue that may cause the data not working as wanted. Also, it is necessary to check, that attributes are unique, in another words, checking if attributes with different values have similar meanings. Lastly, verifying that any attributes that may give answers that conflicts with common sense. (CRISP-DM 2015; Shearer 2000)

3.2.3. Data Preparation

Data preparation is stage where collected raw data will be processes to fit models that will be used later on. Here data analyst selects needed table, record and attribute to construct dataset. Cleaning and transformation are typical actions in this step. There are five steps in data preparation; selection of data, the cleansing of data, the construction of data, the integration of data and the formatting of data. (Shearer 2000; Brandão et al., 2014; CRISP-DM 2015)

Selecting data for data table that will be used in analysis is based on several criteria's. Data analyst need to reflect selected data with data mining goals which were determined based on business problem. Quality and technical constrains sets boundaries for data table and lastly, there should be documentation why certain datatypes were selected and why certain were left out. Furthermore, attribute comparison, which means determine if some attributes are more important than other is relevant when selecting data. (Shearer 2000; CRISP-DM 2015)

Unclean data will affect negatively to all possible data mining analysis. Therefore, creating clean subset of data which will use in data mining model is important.

As dataset is selected with feature attributes and it is cleaned, data construction can begin. This phase offers opportunity to use feature engineering, meaning that creating new attributes from existing ones. Put it simple terms, here analyst can determine whether combining attributes to a new value is needed. This can lead for better insight of data and may also help modeling algorithms work better. Practical example would be age attribute. If dataset hold several rows of data, like age range, perhaps it is more convenient to transform single age to an age group. Example given, new attribute can be grouped as age groups 18-25, 26-30, 31-50. Modeling tools or algorithms often requires these transformations. (Shearer 2000)

Data integration combines one table to another table. The tables can hold different records with different attributes, but they are connected with same object. In another word, there can have several tables holding information of one store and a table which connects all stores in one table, including all sales information of the store. Idea with data integration is, that analyst can build several tables which hold certain information and later on he or she can use all tables or same attributes of table in a one table. There are several ways to make these joins. Additionally, aggregations to refer operations where new values are computed by summarizing information from multiple records is common technique. (Shearer 2000)

There may be situation when data analyst needs change data structure. These can be simple procedures, like removing unwanted characters from a string or changing data type because of model requires specific adjustments. (Shearer 2000)

3.2.4. Modeling

Modeling phase involves selecting suitable techniques parameter adjustment to achieve best possible model towards built data table. Typically, several techniques exist for the same data mining problem type. Depending on data, there may always chance to need to tweak data so that acceptable model will perform as it should. This phase consists following steps; selection of modeling technique, the

generation of test design, the creation of models and assessment of the models. (Shearer 2000; Brandão et al., 2014; CRISP-DM 2015)

Modeling phase starts by selecting suitable technique which varies from decision trees to neural networks, data analyst can pick one or several methods to find best result. After selecting suitable method and building the model, analyst need to test quality and validity. In supervised data mining checking for errors towards actual data is one way to valid quality. Data will be split into two set, training and test. Model is built on training set and tested towards test set, which is the true right values from history. This allows to measure how well the model can predict from history, before using it to predict the future. (Shearer 2000; CRISP-DM 2015)

Data scientist assess the created models and selects the most suitable. When assessing just the model, it will be reflected to business goals with business people. There can be several technical executions but none of them are useful if it does not match with given business purpose or goal. During this phase, data scientists applies single technique or many techniques and makes result comparison according to evaluation criteria. (CRISP-DM 2015)

3.2.5. Evaluation and deployment

It is important to more thoroughly evaluate the model. Evaluation phase tackles issues of model, before going to deployment phase. Built model need to evaluate and review the steps executed to construct the model, to make sure it has reflected to business objectives. By this step, selected issues have been selected to the model but there could be situation that already dropped information need to be in model. Evaluation that is important before deployment. (Wang, 2011; Brandão et al., 2014; CRISP-DM 2015)

The knowledge which were found during the data mining process, need to transform presentation form, so that business users are able to use and interpret with it. Depending on the requirements, the deployment phase can be as simple as generating a report or as complex as implementing a repeatable data mining process.

Usually, it is not the data scientist who carries out the deployment step, it's the customer. It is important, that the customer understands what actions must be taken in order to actually make use of the created models. (Shearer 2000; Juan Wang 2011; CRISP-DM 2015)

Figure 6 illustrate a poll which was made in 2014. KDNnugets website asked from 200 users: "What main methodology are you using for you analytic, data mining, or data science projects". Outcome of this survey reported that 43 per cent use CRISP-DM, 27,5 per cent use their own methodology, 8,5 per cent use SAS's SEMMA and 7,5 per cent uses Knowledge discovery in database (KDD).

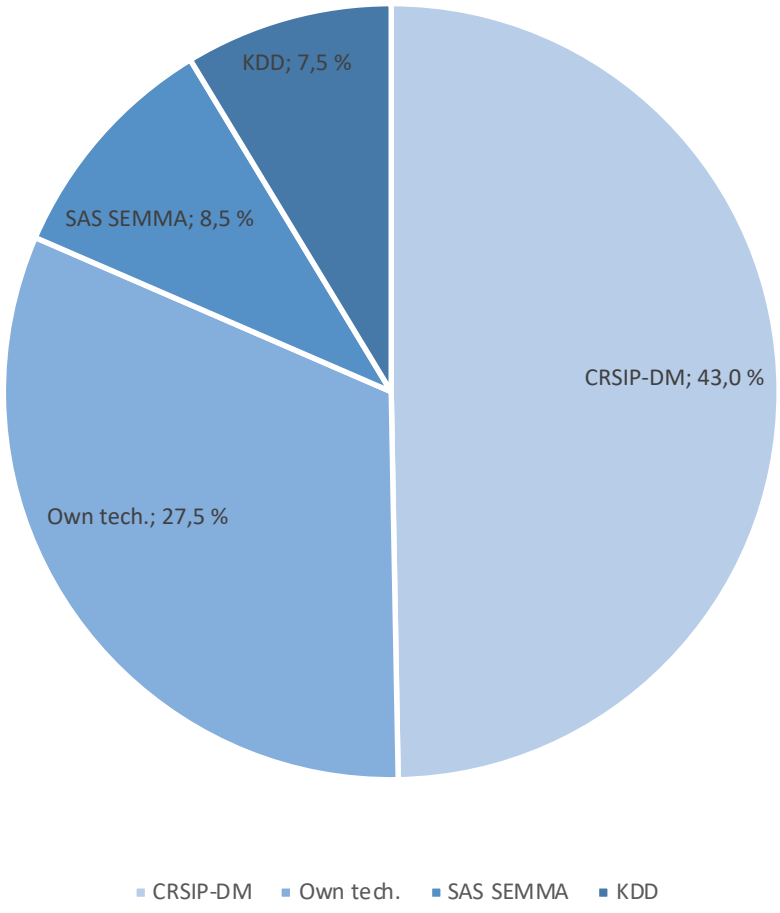

Figure 6 Pie chart of KD nuggets research on methodologies

This chapter collected information from two well-known and successful data management methods, knowledge discovery and cross-industry standard process for

data mining. These two methods are used in different environment but both of them offers powerful process framework for data management.

4. ADVANCED ANALYTICS

Provost & Facett (2013) state that, massive amount of data which is available, companies across industries are exploring ways to benefit from data to achieve competitive advantage. Back in the old days, companies could hire statisticians, modellers and analysts to work with data manually, but nowadays the volume, variety and velocity of data have outperformed the capacity of manual analysis. While data is evolving and growing, in the other hand, computers have become more powerful than ever at the same time. Networking is in state that it can be found everywhere, and algorithms have evolved so powerful, that it can give deeper and wider analysis than ever before. To sum all of these, there is a rise to the increasingly widespread business application of data science principles and data mining techniques. (Provost & Facett, 2013)

When issuing with predictive analytics, term machine learning is rather strong in literature. Machine learning (Yue Liu et al. 2017) is method for automating analytical model building to extract usable information from data and use it to make predictions. Algorithms are performed iteratively going through given data and it allows computers to discovery hidden insights without making assumptions or orders towards given dataset.

Predictive analytics shows good applicability in classification, regression and other tasks which are involving with large dimensional data. Characteristics for predictive analytics which goes along as a synonym with machine learning, is to extract value knowledge from massive databases. Functionality is based on learning method, where algorithm teaches computer from previous computations to produce reliable, repeatable decisions and results. Therefore, it is considered to be a huge game changer in decision making and especially in fields like speech recognition, image recognition, bioformiatics, information security and natural language processing as well in business world. (Yue Liu et al. 2017)

Before jumping to classification and regression, clarification of supervised and unsupervised learning is in place. These two abstracts are from field of machine

learning. Supervised learning can be represented as a teacher, who have answers to the questions and set of examples which leads to the answer. As unsupervised learning, it can use same set of examples, but it would not have the correct answers to present as supervised learning has. So, unsupervised method forms its own conclusion about what the examples have in common. (Provost & Facett, 2013) James et. al. (2013) refers that many problems fall naturally into the supervised or unsupervised learning paradigms. However, sometimes the question of whether an analysis should be considered supervised or unsupervised is not always unambiguous.

To draw more guidelines, quantitative problems are commonly related to regression problems, while situations that are involving a qualitative, response are referred to classification problems. In given dataset, variables can be either quantitative or qualitative. Distinction between these two words are, that quantitative variables take numerical values and qualitative are more like different classes or categories. For example, quantitative variable can describe person's age, height or income, some value of property and very common, it can be a stock price. As qualitative variable, it can include person's gender, the brand of product purchased or simply yes and no options for loan application. (James et al. 2013)

4.1. Supervised learning

Supervised learning (Kotsiantis, 2007; Zhang & Tsai, 2006) happens when algorithms are provided with training data and correct answers and Patel et. al. (2016) stated that learning is performed if all of the data is labelled. Portugal et. al. (2015) wrote that supervised algorithms learn or teach itself based on the training data. After algorithm has been taught, it can be used on test data, which is, in another words, new inputs or real data, which algorithm has not seen yet. Based on new inputs, it will give prediction. As an example, in supervised learning (classification problem) algorithm can be used for classification in a bookstore. Training set can be a dataset relating information about each book to a correct classification. Information about each book may be title, author, or in extreme case every word a

book contains. The algorithm first learns with training set, a set that is given to algorithm to see. When a new book arrives at the bookstore, the algorithm is now getting new information (inputs) and based on what it has learn, algorithm can classify the new book. James et. al. (2013) describe supervised learning that for each observation of the predictor measurements there is an associated response measurement. Algorithm needs to fit a model that relates the response to the predictors, providing accurate predictions for future observations (inputs) or illustrate understanding of relationship between the response and the predictors. The set of methods that uses supervising learning are for example: linear regression, logistic regression and support vector machines.

Kavakiotis et al. (2017) addressed in their latest research paper, that supervised learning as "the system must learn". The objective function is used to predict the value of a variable, this is called dependent variable or easier to understand, output variable. From a set of variables, which are addressed as independent variable or input variables or description of features. The set of possible input values of the function, its domain, are called instances. Each case is described by a set of characteristics. A subset of all cases, for which the output variable value is known, is called training data or examples. By this training data algorithm will be given new input variables which is called test set, a dataset which trained algorithm has not seen yet. The combination of training and test set, supervised algorithm can be used with new, upcoming data.

4.2. Unsupervised Learning

Clear difference with supervised and unsupervised algorithms is, that unsupervised do not use training set to perform predictions. For unsupervised algorithm, the dataset is shown as it is in real world and algorithm function is to come up for a resolution based on that given information. Characteristic for unsupervised learning is, that algorithm tries to find hidden patterns that are in data and use it to conclude synapsis that creates outcomes. Portugal et. al. (2015) put it to an example using demonstration of social network. If algorithm can have access to social media database, it can separate users into personality categories, such as outgoing

and reserved. In another word, algorithm learns by comparing inputs with different possible behaviours types of an outputs. By this information for example companies can do target advertising more directly at specific groups of users.

In comparison, James et al. (2013) describe unsupervised learning to somewhat more challenging situation in which for every observation, there is observation of a vector of measurements but no associated response. This means and it out rules unsupervised to be used in linear regression, because there is no response variable to predict. By other word, this means unsupervised algorithm is in some sense working blindly. Therefore, major characteristic for unsupervised algorithm and learning is to seek and understand the relationships between the variables or between the observations. Kavakiotis et al. (2017) rise that the system tries to discover the hidden structure of data or associations between variables. By given that, training data contains instances without any corresponding labels and also Patel et al. (2016) mentioned same, that unsupervised learning is performed when all of the data is unlabelled.

Schrider and Kern (2017) address, that "unsupervised learning is concerned with uncovering structure within a dataset without prior knowledge of how the data are organized." Practical example of unsupervised algorithm learning is principle component analysis (PCA) which main functionality is to discover unknown relatedness relationships among variables. It works by taking as an input dimensional matrix and from there it produces a lower dimensional summary that can reveal set of clusters or just a cluster based on input data.

4.3. Classification

Classification related problems are qualitative observations which are classifying categories or class, hence they are not presented as numerical observation. Like regression, classification problems act like regression, because usually classification first predicts the probability of each of the categories of qualitative variable, as the basis for making classification. Generally, classification problems are yes or no type of questions. Example given, a classification question can be "is person A attending to continue mobile contract". Classification algorithms calculate the probability of yes or no answer based on attributes that are given as inputs. (James et al. 2013)

4.4. Regression

There different kind of regression models, such as linear regression, logistic regression, polynomial regression and so on. Liner regression represents the simplest and most used method. Its task is to predict quantitative results based on input data. At simplest, linear regression predicts value over a time by predictor variable. James et al. (2013) explains in their book a case example: To examine relationship with sales and TV-advertisement. There are data for the amount of money spent on advertising on the radio and in newspapers. With that data, it is possible to calculate if they have any effect on product sales. If it can be proven, that advertising increases or even decreases the product sale, using liner regression model, it is possible to forecast time forward how much future advertising campaigns can bring more sales.

Garrett (2016) addressed in review article: "Regression models are widely used across a range of scientific applications and provide a very general and versatile approach for describing the dependence of a response variable on a set of explanatory variables".

To summarize chapter. Predictive methods are meant to use statistical methods to forecast future outcome based on history data. There are two types; categorial to

answer yes or no questions. Then there is quantitative to answer with numbers, for example forecasting next summer ice cream sales during holiday season.

5. METHODOLOGY AND DATA COLLECTION

Chapter presents in-detail the research method as well the process for data collection. Explanation for case study selection and also reason for with-in and cross case analysis. Master's thesis reliability and validity and brief description of the cases are presented in the end of the chapter.

5.1. Qualitative method

This research is done by using qualitative method, which Creswell (2013) describes as “situated activity that locates the observe in the world”. It holds many gathering styles of information, example given: field notes, interviews, conversations, photographs, recordings and memos. Characteristic for selected method is that phenome is investigated in their natural settings which also refers to naturalistic approach to the world. Qualitative research is done by face to face basis. Therefore, researcher's ability to interact with interviewees is important. With unprepared session or misinterpret of a conversation could lead for bad data or even get lost from original questions and intentions. Structured and well-prepared interviews are required for proper data collection. The best learning is reached when researcher suspend own judgements while interacting and uses tools of qualitative inquiry to learn and represents case perspectives. (Lapan, et al., 2011)

5.2. Case study

The Research is based on case study where it investigates a contemporary phenomenon in its real-life context Additionally, Saunders et al. (2009) rely Robson’s definition for the case study as “a strategy for doing research which involves an empirical investigation of a particular contemporary phenomenon within its real-life context using multiple sources of evidence”. Its mission is to gain rich understanding of the context and it is obtained by questions like how, why and what. According to Yin (2003) a case study design should be considered when: (a) the focus of the study is to answer “how” and “why” questions; (b) you cannot manipulate the

behaviour of those involved in the study; (c) you want to cover contextual conditions because you believe they are relevant to the phenomenon under study; or (d) the boundaries are not clear between the phenomenon and context. By that, specificity for research is inductive. Gillham (2000) explains inductive that the researcher needs to know what other have done but cannot be sure they are relevant, in another word, making sense of what you find after you have found it. (Järvinen 2001) address, that “in the inductive theoretical research a theory is derived from empirical generalisations or by interpreting old results in a new way.

Additionally, multiple-cases is used in this research paper. Interviewees are working among data management and business intelligence, but their expertise are from different industries; private sector and public sector. Using multiple-cases offers broader view than single case study. Yin (2009) describes, multiple-case studies have higher possibilities for analytical generalization than single case study.

5.3. Data collection

Data, in general, can be divided into two different types: primary and secondary. Primary data refers to new data that is gathered for the specific research which is being conducted, whereas secondary data is already gathered data for another purpose, which can be reanalysed for the current research. (Saunders et. al, 2009). Lapan et al. (2011) address: "The qualitative data collection tool kit is substantial and qualitative researcher have many choices to make in terms of study sire, study sample and the specific tools for data collection".

As data collection method, semi-structured interview is used. Question form consist of pre-determined questions that are reflected to the main research questions and its sub-questions. Characteristics for semi-structured is that the interviewer and respondents meet in a formal interview. The interviewer creates and uses an interview guide which usually is a list of questions and topics that need to be covered during the conversation. As the interviewer follows the guide it is acceptable

to follow the direction in the conversation that may stray from the guide if the researcher feels this is appropriate. (Flick 2010) The questions which were used to conduct interviews are in appendix 1.

For data collection, target audience are professionals that have deployed several business intelligence solutions to a mid or large –size Finnish companies or health and well-being sector. They have solid and long background working in business intelligence field, providing analysis (descriptive or predictive) using data management techniques. Outcome of interviews will be aggregated in analysis part of the thesis.

All the data will be stored in word document that is located in cloud-based service which is owned by the researcher. Every new document file follows strict naming pattern and timestamp. When creating tables or figures from answers, information will move to the excel-software and it uses same strict naming pattern and timestamp as word document.

After the interviews, the records were written to a paper and summarised. That means, shrinking the conversations to a smaller piece and leaving not relevant information out from analysis. When both of interviews are issued in their own respectively sub topic. Next step is to make comparison and find similarities and differences between interviews and create conclusion out of them.

Interviews were conducted in late March 2018 and beginning of April 2018. For the interview, the researcher booked a conference room so that interviewee and the interviewer were only one's present. Interviews were recorded using two recorder, second one being the backup device if something happened to the first one. Average time for interviews were approximately 52 minutes. Interview was held in Finnish but afterwards translated to English. Interview followed the semi-structured model by presenting one question at the time, adding follow up questions if needed. After formal part was over, there were small time for free conversations regarding to the topic.

5.4. Reliability and validity

When creating case study, there are together four criteria that should be taken consideration when issuing reliability and validity of the thesis: construct validity, internal validity, external validity and reliability. (Yin, 2009)

Construct validity, the interview questions are carefully selected to seek answers just for the specific situation. The questions that were presented cannot be used to find answer to some another research problem, therefore, it can be said that research questions are constructed so it seeks answers for only this research. As only data collection method is interview, there is always possibility that outside reason makes interviewees answer differently at different time or place. Internal validity is relevant only studies which are researching causal relationship. As this research is not studying casual effect, internal validity does not need more justifying.

The result should not be generalised as its more describes a situation status at that given time when interviews were conducted by the personal professional experiences. The questions are formed so, that they are repeatable as long the persons represent same experience from the industry.

5.5. Brief introduction of interviewees

Interviewee A has been working among business intelligence for over fifteen years. Providing technical solutions regarding data management, descriptive and predictive analytics to different sized companies in Finland. Interviewee B has long career as analyst, especially in public sector at health and wellbeing segment. Working among machine learning algorithms as strong solid knowledge of descriptive analysis. Both interviewees are working as consultants to provide data and information solutions for Finnish mid and large sized companies

6. EMPIRICAL RESULTS AND FINDINGS

Chapter issues both interviews. First as individually then together by cross-case analysis. Both interviews use same main questions but during the interview the supporting questions may vary because different answers were given. Cross-case analysis is written under a research questions, so analysis and the outcome of the answers are presented logical way.

6.1. Case A

There are few angles to overview the situation, which are basically technology and work culture. While latest technology is available to accelerate development, the biggest barrier can be found in managerial implications.

> *“As we live in era where technology is advanced and there are ways to use it. We are still facing barriers from cultural activities in companies. Biggest difference can be found in those leaders, who has data driven mind-set for business intelligence and then there are non-data-driven managers. Typically, there are few managers, even chief executive officer that are recognizing the potential of data, but on contrary managers that holds responsibly of budget and profit margin, they are harder to convince or turn into data driven mind-set*."

While organizations are considering doing something with their data which they possess. There are many issues which are not considered in very base level. Some managers read from articles about these hyper trends of predictive analytics but are still lacking the basics of data warehouse or a place, where data is put in usable form. From that point it is long way use predictive methods.

> *"When we talk about descriptive and predictive analytics or any other trend word (big data, data driven, advanced analytics). The current situation in or-*

ganizations can be that they do not know in real time how their data performs. It could be said that sometimes even the basic mathematics is not used."

Factors that slow progress in field of business intelligence and data science are related to education and managerial performance, but this is not lack of competence. The need and vast growth of data is put our workforce in test.

"I want to say, that technology is in good shape. For coming years, young people that are finishing their university degree in data science, are already built-in this data driven mind-set. They are willing to push the managerial development forward and they are the workforce which future companies needs."

Biggest gap or variety of data-thinking between descriptive and predictive mingles around hype and reality. Usually, when providing predictive solutions to an organization that do not have solid ground for data management, are the ones which creates the gap. Perhaps we would discuss on totally different gap between descriptive and predictive if the managers would have base knowledge of data management.

"Let's take predictive analytics in consideration. There are some managers that assume data can just be taken and use for making prediction and see whether the result can be used or not. This is absolutely contrary thinking in terms of business intelligence and how data is needed to extract transform and load to some valid storage for future usage. By this, it means if a company do not recognize or do not understand which form their data is, the longer distance it is to adapt predictive methods. When companies have clear data handling process, they are more likely ready to implement predictive action on top of their existing layer of data architecture."

Data management and data handling follows quite structured form, like CRISP-DM. Both descriptive and predictive analytics starts from business understanding

to data understanding and leads to data preparation. These three steps are needed before modelling can be take in place.

> *“Descriptive and predictive analytics follows long line same technical flow. Difference is that when they on point x start following different technical structure, then many companies are "in trouble". They are astonished by the fact that data is needed collect, store, transform and put in a logical model - it’s usually something that they haven't fully understand when planning to buy predictive analytics service. I see, that the gap is in business and managerial section, not in technical competence".*

In terms of Strategy, organizations are talking to take big jump forward with their business intelligence solutions, but those leaps are not often contributed immediately business cases which it could to fund itself. They are rather seen as proof of concepts (POC) or as investments which are calculated as part of return of investment, rather as part of business strategy.

> "*With business intelligence framework I have tried to model different kind of competences which are related data driven mind-set. Idea is to handle competence from different point of view; leadership, people, technology and so on. If these base fundamentals are in correct form and in line with company strategy, using data analytics if much easier to adapt".*

There ae increased amount of proof of concept-project or smaller projects which are pending private or governmental funding. For example, Helsingin ja Uudenmaan sairaanhoitopiiri (HUS) established proof of concept where predictive methods tried to forecast dangerous bacteria infection for babies that are born before due date.

> "*Major problem at the moment with proof-of-concepts are, that they are not providing same usage compared to money that has been spent.”*

There is competence to provide descriptive and predictive solutions but usually the problem is related to business itself. Business is an abstract concept which are not typically data classification problems that can be solved using predictive analytics. When asking questions like "how your product is best in year 2025", there are no answer which is based on analytics.

> *"Nokia had best cell phones in the world. When apple first time announced that they will have the best phones in the world - No big data or predictive analytic model forecasted that. Apple just came thru with innovation."*

When all the base foundation for data is in order, predictive analytics can be very useful in terms of decision making. Netflix has proven track record for categorising typical elements of a hit tv-show and use them to create successful tv-series. When data is correctly in order, these kinds of categorical analytics can be done to support decision making.

The future place for data management depends on in which level the decisions are made. High level strategy decisions should still be in hands of business people, taking advantage of data but not give full decision power to analytics. Descriptive and predictive analytics are still just made for support decisions.

> *"Business people are the ones who should in future also make the big decisions and analytics being supportive source".*

On contrary, when predictive analytics is implemented to an operation level, the benefit can be bigger than without prediction. For example, in service desk sits new or not so much experience gained worker. Let's say that there will be an angry phone call from not satisfy customer. With data and combined text analytics - computer could alarm the worker by letting now that customer is not happy, giving the worker leverage to offer a better deal or help, even before the angry customer can even demand it.

Competence is something that will play even higher role. In several cases, the best results are achieved when person with technical responsibility sits in same negotiation table with the business people. Only then the correct questions will be asked, and correct information will be shared.

> "*The future competence is easily recognisable, technical experts in data management need also have really good business understanding - especially in managerial level*"

To reduce descriptive and predictive gap within business could be organisation financial issue. Both descriptive and predictive are valid solutions in decision making, no matter if we speak about business or technical point of view. The general "gap-problem" is when there are more selling points for data the more business intelligence companies has opportunities to sell. This may even more increase the gap because the competence and understanding of a business owner could shatter.

> *"Common problem is that customer wants to have analytics in their business, but they do not realise the importance of data management, especially the role of data warehouse and how much background work it needs to provide predictions based on data".*

Predictive usage over descriptive would not increase, the balance will stay quite the same for some time. Perhaps terminology or new words will increase, but the base level will stay the same. The need of competence in labour market will increase but universities and companies own academy-programs try to answer to this need. The biggest need in labour markets are those skilled people who can manage technical aspect as well the business aspect.

Resources for implementing predictive analytics varies with many things. Main issue is that, first companies should identify their core need for analytics. At the moment many proof of concepts are investments rather than strategy, therefore it is need to overcome the financial barrier and make the proof of concept beneficial

and understand that there cannot be always the "WOW-effect". It is necessary to see even those small beneficial improvements.

6.2. Case B

Descriptive analytics looks time in past. It describes with basic numbers what has been happened and usually these results are visualised. Predictive analytics uses statistical methods. With history data the goal is to predict future outcome, as descriptive only reports what has happened.

> *"Basically, we can use questions "What has happened", "Why is it happened" and "What will happen".*

In health and wellbeing sector descriptive methods are regularised by the government, therefore it can be said, that reporting belongs to that culture. There are several places where reports need to deliver, so descriptive methods have been around for decades. Only by now, analytics (in general) has become part of a toolbox when designing processes or functions. This is, because nowadays budget is really tight, therefore finding new an effective solutions through analytics is essential.

> *"The basic competence for descriptive analytics is there and there is strong historical background".*

When talking about using existing history data to design business improvements or estimate things, cultural aspect is missing. Which means, that there is no business intelligence point of view. By that, there are no enough technical competence or skills to produce beneficial information for the managers out of large data masses.

Medicine research work has been for ages. Forecast models has been done in epidemiology and there is strong culture for that. There are researchers that are

working as a doctor and they have needed competence and understanding of issues. When we talk about the managers in hospitals, their primary goal is to hold on the budget. This means, that resources and time must be in that budget. Only for some time now, past ten years has been realised, that data can be used for something. University hospitals are the first ones to start using data for creating new improvements.

In descriptive analytics the workflow is that from certain information system or source data is picked up and put it to somewhere, and another data is taken from another system or source and put it in same place. This means, that information systems are not co-operating with each other. To establish predictive methods, it is necessary to extract data correctly in a certain place, where it is easy to pick up and use.

> "*It is crucial to identify what information is in data, after that we need to figure out what kind of information is beneficial to business improvements". In another words, predictive analytics need to consider as technical process: first it is needed to understand what we need to lead. Then we can do research to find whether we have data or not. If we have the data, how we can turn it to an information? The most importantly, do we have the financial resources to search that information?"*

Data management, especially in this case, is problematic. Because there are no information systems which could merge data together in one place, all the needed data are collected manually and put it to some software (e.g. excel). When speaking about data management, flow for creating descriptive and predictive analytics are rather same till to the modelling part.

> *"Advanced analytics requires existing data-pool solutions, where all source data are gathered in usable form, locating in easy access place. After that predictive methods are able to use".*

Health and wellbeing sector adapting predictive methods to strategy is complex path. University hospitals are linked to Universities and their research facilities, from there all the newest innovations are coming. It depends which kind of persons there are working and how much is their emphasis to which matter in that time. When moving to central hospital level, it always comes behind compared to University hospitals. There are no innovative researcher community present. Therefore, basically only the valid best practices are used. By other words, this means new innovations moves slowly to central hospital level. Main reason for this lies on people and budget. Local hospital level, staff are just trying to make it through the day.

When talking about strategic management level, there are very strict budget discipline and that is also barrier for creating innovative predictive analytical methods and implement those to a strategy. The most important thing in strategic level is to provide the most quality treatment that exist, within the budget.

> *"In the future university hospitals are those ones that will research and take advantage of advanced analytics. That sector also has the most amount of money in use what comes to research and there are lot of competence regarding to know-how and technical aspects."*

Ideal goal would that the data which is used in descriptive analytics, will be available for both methods. In another words, same data which is used to generate descriptive analytics would also be able to use in predictive analytics. That same data can be used to identify future needs of a health service.

> *"Already existing data can be used to predict ear sickness for upcoming week. This kind of predictions are done in foreign countries but not in Finland. Basically, we use same data for descriptive and predictive methods: we have in year x amount year sickness. With this data it is possible to predict time x forward amount y in that given time."*

There is enthusiastic atmosphere regarding to predictive methods but how to execute the plans, are tied with given budget. Competence is in good level but for every project is needed to apply funding which can take long time.

> *“In service and well-being sector if you want to start doing something you need to apply for funding. This can easily take years before the amount of money is collected."*

Future competence relies in two categories; 1) those who can extract the needed data from information systems 2) those who can transform that data in to information, interpret, analyse and create models out of the data.

> "I believe, that information systems will develop, and it will be required that data will be easier to extract for later use in the flow. This means, people who are working with advanced analytics are not so dependent on technical issues regarding to data management processes".

There is no simple solution for reducing gap between descriptive and predictive methods in decision making. Perhaps when business managers are not yet in that point where they fully understand the potential in predictive methods. Generally, gaps can be reduced when understanding of the matter increases. But it can be said, that competence and know how is already excising.

6.3. Cross-case analysis

RQ1: What is the gap between descriptive and predictive analysis in Finnish companies that uses business intelligence in decision making?

Both cases emphasised organization managerial culture towards data and information. Case A illustrated, that many corporate managements are lacking sense of data driven mind-set. As case B, explained the hierarchal of social and wellbeing sector, which also had missing mind-set towards data and its usage, especially when talking about predictive methods.

Case A introduced quite usual problem, which occurs often when companies are wanting something that they do not fully understand. This can be seen in people that are working in management section. Managers that are holding responsibility of company data and are given task to make the best use of it, opposite can be managers that are very budget strict and therefore, cannot see the data driven potential versus the costs. Case B, revealed, that university hospitals are very well-known the potential of data and they are using quite innovative methods. From level down to central hospital and local hospitals are not sharing same method. But end of the day, issues towards data usability sits on managerial level.

Case A emphasises, that if we are talking about gap between descriptive and predictive methods, we should also make in consideration what is the understanding of business managers towards data management. Case B, rises that managers should have the right questions which to ask from their data, without the right questions, how managers can understand what they need or want?

Both cases mentioned the importance of understanding the data management processes, to understand what is needed to produce descriptive or predictive analytics. Case B spoke about using the already existing data and turning it to information, it is crucial to look the data in eyes of business intelligence, which some-

times is not happening. As case B issued, that usually first it is necessary to introduce business intelligence concept, which will help the managers to understand their existing corporate data.

Both cases stated, that descriptive methods are more common in their cases than predictive. Case B stated, that in epidemiology descriptive methods for reporting has been done for ages and it is very regulated by the government. Case A stated, that companies tent to say having analytics in their environment and they may even have some kind of data management structure, but it is usually narrow and short sighted.

Both cases agree, that predictive method is huge trend at the moment. But the technical aspect for using those methods, is not the problem. Main issue is to avoid the shortcuts in data management and understand all the needed steps in technical perspective. Another problem is the managerial level and how they figure out correct questions for their data.

In both cases, budget was a common barrier, in their respective way. Establishing proof of concept or funding innovative method for a research are highly tided to money. There can be one or several proof of concepts which main point is to take advantage of predictive methods. No doubt, these proofs of concepts are expensive. The problem is that these concepts are seen as investment, not a mind-set or part of a strategy and therefore, usually the small but effective improvements are not taking use. Because at that point or some calculated point, it won't give the "needed financial benefit versus to spent money".

Table 2 What is the gap between descriptive and predictive analysis overview

Case A	Case B
Emphasised organization managerial culture towards data and information.	Emphasised organization managerial culture towards data and information.
Lacking sense of data driven mind-set	Hierarchal of social and wellbeing sector

Companies are wanting something that they do not fully understand	Health and wellbeing sector data usability sits only on managerial level.
Business managers competence on data management	Managers should have the right questions to ask from their data
Budget	Budget

RQ2: What kind of capabilities is needed for companies to take advantage of predictive methods?

To take advantage of predictive methods or descriptive methods, it is necessary to understand holistic process of data management. According to cases, competence can be divided in two categories: technical and business competence. Technical competence is at the moment more or less divided to two groups: competence to data management, which in another words means competence to create data warehouse. Second technical competence is way to use descriptive and predictive methods. It could say, that that future need is to handle the both technical aspects.

Business competence means that managers have solid understand between their business goals and data. Cases presented in their respective fields, that it is crucial to get managers to understand how data can be beneficial towards company strategy. It's important that managers can see thru all the hype-words which are clustering around data. To put it simple way: if you want to use descriptive or predictive analytics, you need to have understood of architectural data management framework. Case B, on the other hand issued, that in future data management process can be fully automatic which means, that analysts do not need deeper knowledge towards technical aspect of data preparation.

By these two main competence categories. Cases A and B believe that in future competence it is highly valued that one person can poses both, technical and managerial competence. Case A pointed out, that in client cases the best result has occurred, when technical and business people sat the same negotiation room

and together solve data related need. Case B emphasises, that future managers need to understand to ask right questions towards their data. If you cannot ask what you want your data to show you, it's no point at all.

Table 3 What kind of capabilities is needed overview

Case A	Case B
Technical and business competence	Technical and business competence
Understand architectural data management framework	Automatic data management pre-process

RQ3: What kind of resources companies need to have for implementing predictive methods?

In both respective cases, came across to the same resources need, which they have noticed during their career. To obtain descriptive or predictive, and even more in predictive, biggest resource problems are related to financial issues as also having competent people working within the companies.

Case A and B both gave examples in their respective industry where budget plays the biggest role when adapting predictive methods. Case A addressed, that proof of concepts are still very expensive versus the possible gain. Therefore, there has been many cases which never continued after proof of concepts. In case B, it was mentioned, that in health and wellbeing sector innovative methods are very dependent on budget and funding.

Third research question was seeking answer to competence. The competence is also a resource which companies need to have. Without proper educated workers who understands data management and methods which are used in analytics, it is almost impossible to get the business benefits using data. It is necessary to understand how data management process works, as also having correct questions for the data. Having unstructured data processes or in worst case poor understanding of how to extract correct information from data, that meet business requirements.

Table 4 What kind of resources companies need overview

Case A	Case B
Budget / Financial	Budget / Financial
Know-how	Know-how

7. DISCUSSION AND CONCLUSION

The aim of this research was to find possible gaps between descriptive and predictive analytics. Both methods are different, and it can be tricky for business managers to understand how to handle the data. This research is interesting for the writer because he works among business intelligence, where the boom of hype words like advanced analytics, predictive analytics, artificial intelligence and so on are visible. The goal was find gap between these two methods by interviewing experienced business intelligence consultants, which have strong set of experience in their respective field.

This paper issues three main subject which helps to build enough background, even though reader would not have any experience of data science or data related processes. Topics are written in underlying level and therefore, the thesis did not go in deep, technical level. It would not be in scope of this research.

Introduction started with data mining which may in different environment mean different perspective. For this research, it holds short introduction to three main analyst method, descriptive, predictive and perspective. Second issue is data management which introduces the big machine, the engine of whole data process scenario. Lastly, short introduction of what kind of advanced analytic methods for prediction are existing.

7.1. Conclusion

Before the interviews, the researcher had a mind-set for the gap being in technical issue between descriptive and predictive methods. In another words, assuming that there is gap in technical competence. As it turned out in the cases, this is not the situation. Both methods need very same data management processes to work. It was little bit surprising that the biggest gap in decision making with these two methods are linked to managerial level. Problems are varying of course in different companies, but they share common pitfalls. It is necessary to understand, that

data is needed to modify in many levels and put into data warehouse, from there it is usable for descriptive or predictive analysis, to support decision making.

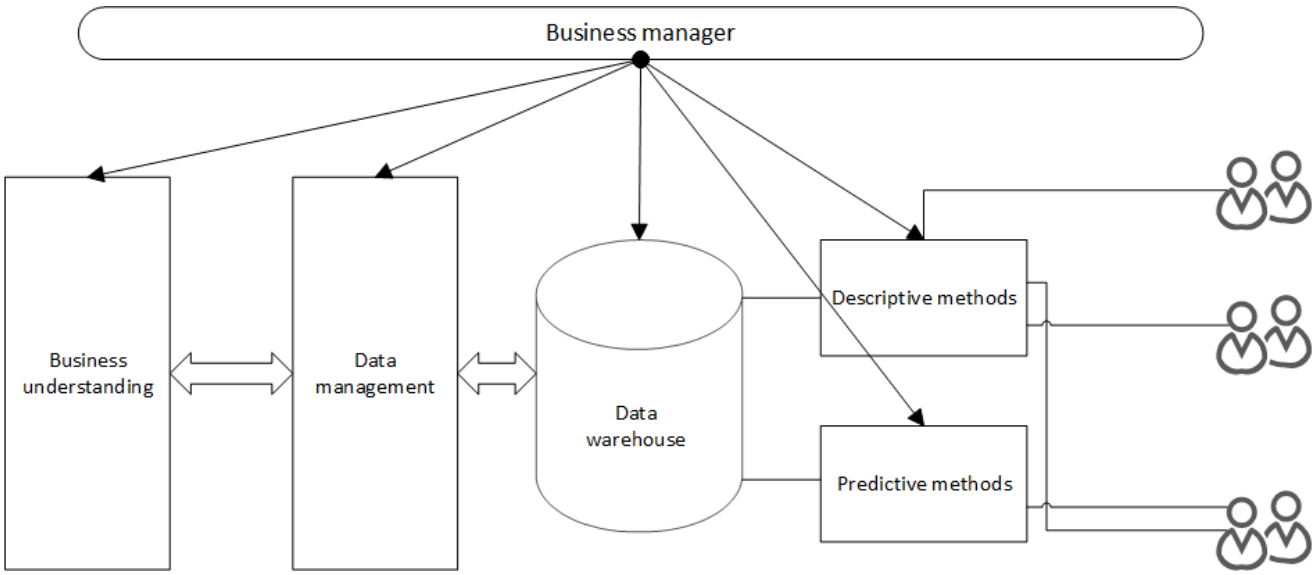

Figure 7 Outcome illustration

Figure 7 illustrates the gap between descriptive and predictive analytics. Business managers should see the whole data management process in order to obtain analytical methods. First issue is to understand relationship between business and data. After that there should be data warehouse solution which is built based on terms of business and data. Next step is to recognise business users and what kind of analytics they need. One user could only need descriptive analytics as other just predictive. Some cases, one user needs both of them. When understanding the holistic process and parts, it will reduce the gap between descriptive and predictive methods. Missing out one or more steps, it will increase the gap between methods. Sheng et. al. (2017) reflects on their research, that foregoing analysis and studies reveals data management and concepts related to have significance affect in business and management improvement. They mention, that data-driven approach does not only touch technical stage, but also promotes changes in management organisation and with additional values can be discovered to accelerate business development.

It seems that all the hype talk about predictive methods has taken the best of us at the moment. Companies (not of course all) are trying by force implement so called

predictive analytics within their everyday life, even though not understanding the full scale of it. Additionally, mix of competence and budget issues affects heavily. Proof of concepts are rather expensive to establish and as these are seen as investment, not as part of business strategy, it will take more time when predictive methods will set its feet better in organization strategy. Hazen et, al. (2017) mentioned, that business analytics is affecting how companies nowadays compete. They point out not only the managerial data-driven aspect, but also dependency on data quality towards analytical methods are crucial.

There are lot of technical competence in existing workforce, but the need for data management is rising rapidly. So, there is need for more competent workforce. Universities in Finland are trying to answer this call. It is also seen in job market, that big consultant companies among other companies are starting their own academies for data management to reduce workforce gap.

7.2. Limitations of the research and future suggestions

The study has limitations as it is conducted by interviewing only consultants but not business users itself. Additionally, because of the time schedule, three interviews could not be done. Having more interviews, it could give wider and deeper results of the given research. The study cannot be generalized but for future research, adding more interviews and participating business users also to it, results could give new insights which was not issued in the research.

Limitations guide towards to future research suggestions. The study is repeatable, and the outcome can vary. This is because, business managers will have more intel about data management processes and it is highly possible, that technical methods will be simplified in future. That can be itself a gap reducing method. Additionally, involving more industry specialised consultants and taking notice business users itself could give new insights to the research.

Second research suggestion would be turning the research more technical aspect of view. Creating atmosphere where pure technical competence is measured and

researched. This will allow more specified literature review and use of concepts. In best case scenario, research could lead more efficiencies methods and processes for business units.

Third suggestion would be studying the same phenomena through financial aspect. As now, budget and funding are causing barrier for many predictive methods, studying why this is, could give interesting insight.

Fourth suggestion is related to research the competence around the issue. Now there is competence in available workforce, but studying the possible growth of potential workers with needed competence could be rather interesting in terms of education perspective as business perspective.

REFERENCE:

JOURNALS AND ARTICLES

Acito, F. & Khatri, V. 2014, Business analytics: Why now and what next?.
Amani, F.A. & Fadlalla, A.M. 2017, "Data mining applications in accounting: A re-view of the literature and organizing framework", International Journal of Ac-count-ing Information Systems, vol. 24, pp. 32-58.

Argote, L. & Ingram, P. 2000, "Knowledge Transfer: A Basis for Competitive Ad-vantage in Firms", Organizational Behavior and Human Decision Processes, vol. 82, no. 1, pp. 150-169.

Bradlow, E.T., Gangwar, M., Kopalle, P. & Voleti, S. 2017, "The Role of Big Data and Predictive Analytics in Retailing", Journal of Retailing, vol. 93, no. 1, pp. 79-95.

Brandão, A., Pereira, E., Portela, F., Santos, M.F., Abelha, A. & Machado, J. 2014, "Managing Voluntary Interruption of Pregnancy Using Data Mining", Procedia Technology, vol. 16, pp. 1297-1306.

Chen, C., Khoo, L.P., Chong, Y.T. & Yin, X.F. 2014, "Knowledge discovery using genetic algorithm for maritime situational awareness", Expert Systems With Appli-cations, vol. 41, no. 6, pp. 2742-2753.

Colak, Cemil|Karaman, Esra|Turtay, M. Gokhan 2015, "Application of knowledge discovery process on the prediction of stroke", Computer Methods and Pro-grams in Biomedicine, vol. 119, no. 3, pp. 181-185.

Davenport, T. & Prusak, L. 2000, "Working knowledge", Ubiquity, vol. 2000, no. August, pp. es.

David J. Teece 2007, "Explicating Dynamic Capabilities: The Nature and Micro-foundations of (Sustainable) Enterprise Performance", Strategic Management Journal, vol. 28, no. 13, pp. 1319-1350.

Dehning, P., Lubinetzki, K., Thiede, S. & Herrmann, C. 2016, "Achieving Envi-ronmental Performance Goals - Evaluation of Impact Factors Using a Knowledge Discovery in Databases Approach", Procedia CIRP, vol. 48, pp. 230-235.

Delen, D. & Demirkan, H. 2013, "Data, information and analytics as services", Decision Support System, vol. 55, no. 1, pp. 359-363.

Fan, H., Li, G., Sun, H. & Cheng, T.C.E. 2017, "An information processing perspective on supply chain risk management: Antecedents, mechanism, and consequences", International Journal of Production Economics, vol. 185, pp. 63-75.

Fayaad, U., Piatetsky-Shapiro, G. & Smyth, P. 1996, The KDD process for extracting useful knowledge from volumes of data, Communications of the acm.

Fernández-Arteaga, Verónica|Tovilla-Zárate, Carlos Alfonso|Fresán, Ana|González-Castro, Thelma Beatriz|Juárez-Rojop, Isela E|López-Narváez, Lilia|Hernández-Díaz, Yazmín 2016, "Association between completed suicide and environmental temperature in a mexican population, using the knowledge discovery in database approach", Computer Methods and Pro-grams in Biomedicine, vol. 135, pp. 219-224.

Flick, U. 2010, Anœ introduction to qualitative research, 4. ed., repr. edn, Sage Publ, Los Angeles, Calif. [u.a.].

García, S., Herrera, F. & Luengo, J. 2015, Data Preprocessing in Data Mining, 2015th edn, Springer, Cham.

Garrett, M.F. 2016, "Regression", Diagnostic Histopathology, vol. 22, no. 7, pp. 271-278.

Groggert, S., Elser, H., Ngo, Q.H. & Schmitt, R.H. 2018, "Scenario-based Manufacturing Data Analytics with the Example of Order Tracing through BLE-Beacons", Procedia Manufacturing, vol. 24, pp. 243-249.
Hashimzade, N., Myles, G.D. & Rablen, M.D. 2016, "Predictive analytics and the targeting of audits", Journal of Economic Behavior and Organization, vol. 124, pp. 130-145.

Hazen, B.T., Weigel, F.K., Ezell, J.D., Boehmke, B.C. & Bradley, R.V. 2017, "Toward understanding outcomes associated with data quality improvement", International Journal of Production Economics, vol. 193, pp. 737-747.

Järvinen, P. 2001, On Research Methods, Opinpajan kirja, Tampere, Finland.

Juan Wang 2011, "The Study on Cross-Industry Standard Process for Data Mining in E-Marketing", Applied Mechanics and Materials, vol. 66-68, pp. 2298.

Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I. & Chouvarda, I. 2017, "Machine Learning and Data Mining Methods in Diabetes Research", Computational and Structural Biotechnology Journal, vol. 15, pp. 104-116.

Kotsiantis, S.B. 2007, "Supervised machine learning: a review of classification techniques", Informatica, vol. 31, no. 3, pp. 249.

Lapan, S.D., Quartaroli, M.T. & Riemer, F.J. 2011, Qualitative Research, 1. Aufl. edn, Jossey-Bass.

Lestringant, P., Delarue, J. & Heymann, H. 2018, "2010–2015: How have conventional descriptive analysis methods really been used? A systematic review of publications", Food Quality and Preference, vol. 71, pp. 1-7.

Liou, D. & Chang, W. 2015, "Applying data mining for the analysis of breast cancer data", Methods in molecular biology (Clifton, N.J.), vol. 1246, pp. 175.

Lorenzo, A.J., Rickard, M., Braga, L.H., Guo, Y. & Oliveria, J. 2018, "Predictive An-alytics and Modeling Employing Machine Learning Technology: The Next Step in Data Sharing, Analysis and Individualized Counseling Explored with A Large, Prospective Prenatal Hydronephrosis Database", Urology, .

Morais, A., Peixoto, H., Coimbra, C., Abelha, A. & Machado, J. 2017, "Predicting the need of Neonatal Resuscitation using Data Mining", Procedia Computer Science, vol. 113, pp. 571-576.

Neto, C., Peixoto, H., Abelha, V., Abelha, A. & Machado, J. 2017, "Knowledge Discovery from Surgical Waiting lists", Procedia Computer Science, vol. 121, pp. 1104-1111.

Patel, M.J., Khalaf, A. & Aizenstein, H.J. 2016, "Studying depression using imaging and machine learning methods", NeuroImage. Clinical, vol. 10, pp. 115-123.

Peral, J., Maté, A. & Marco, M. 2017, "Application of Data Mining techniques to identify relevant Key Performance Indicators", Computer Standards & Inter-faces, vol. 54, pp. 76-85.

Poh, C.Q.X., Ubeynarayana, C.U. & Goh, Y.M. 2018, "Safety leading indicators for construction sites: A machine learning approach", Automation in Construc-tion, Portugal, I., Alencar, P. & Cowan, D. 2015, "The Use of Machine Learning Algorithms in Recommender Systems: A Systematic Review", .

Saunders, M., Lewis, P., Thornhill, A. 2009. Research methods for business students. Harlow: Pearson. P.139, 145-146, 256-262, 482-484.

Schrider, D.R. & Kern, A.D. "Supervised Machine Learning for Population Genetics: A New Paradigm", Trends in Genetics, .

Schuh, G., Prote, J., Luckert, M. & Hünnekes, P. 2017, "Knowledge Discovery Approach for Automated Process Planning", Procedia CIRP, vol. 63, pp. 539-544.
Shearer, C. 2000, "The CRISP-DM Model: The New Blueprint for Data Mining", Journal of data warehousing, vol. 5, no. 4, pp. 13-22.

Sheng, J., Amankwah-Amoah, J. & Wang, X. 2017, "A multidisciplinary perspective of big data in management research", International Journal of Production Economics, vol. 191, pp. 97-112.

Thrassou, A. & Vrontis, D. 2008, "Internet marketing by SMEs: towards enhanced competitiveness and internationalisation of professional services", Interna-tional Journal of Internet Marketing and Advertising, vol. 4, no. 2-3, pp. 241-261.

Vidgen, R., Shaw, S. & Grant, D.B. 2017, "Management Challenges in Creating Value from Business Analytics", European Journal of Operational Research, vol. 261, no. 2, pp. 626-639.

Wang, R., Wang, X., Ji, W., Liu, M., Weng, J., Deng, S., Gao, S. & Yuan, C. 2018, "Review on mining data from multiple data sources", Pattern Recognition Let-ters, vol. 109, pp. 120-128.

Xianshun Chen, Yew-Soon Ong, Puay-Siew Tan, NengSheng Zhang & Zhengping Li Oct 2013, "Agent-Based Modeling and Simulation for Supply Chain Risk Management - A Survey of the State-of-the-Art", IEEE, , pp. 1294.

Yang, H. & Chen, Y.P. 2015, "Data mining in lung cancer pathologic staging di-agnosis: Correlation between clinical and pathology information", Expert Sys-tems With Applications, vol. 42, no. 15-16, pp. 6168-6176.

Yin, R. K. (2009). Case study research: Design and methods 4th edition. United States: Library of Congress Cataloguing-in-Publication Data.

Yin, R.K. 1993, Applications of case study research, 1st edn, Sage Publications, California.
Yue Liu, Tianlu Zhao, Wangwei Ju & Siqi Shi 2017, "Materials discovery and design using machine learning", Journal of Materiomics, vol. 3, no. 3, pp. 159-177.

Zhang, D. & Tsai, J.J.P. 2007, Advances in Machine Learning Applications in Software Engineering, Idea Group, Hershey.

BOOKS

James, G., Witten, D., Hastie, T. & Tibshirani, R. 2013, An introduction to statistical learning, Springer, New York.

Negnevitsky, M. 2011, Artificial intelligence, 3. ed. edn, Addison-Wesley, Harlow, England [u.a.].

Provost, F. & Fawcett, T. 2013, Data science for business, 1. ed. edn, O'Reilly, Beijing [u.a.].

WEBPAGES

CRISP-DM 2015, CRoss industry standard process for data mining 1.0: Step by step data mining guide.. Available: http://crisp-dm.eu [2017, 24.12.].

APPENDIX

Main research question: What is the gap between descriptive analytics and predictive analytics in Finnish companies that uses business intelligence in decision making

Sub question:

- What kind of capabilities is needed for companies to take advantage of predictive methods?
- What kind of resources companies need to have for implementing predictive methods?

SECTION ONE. OPEN QUESTIONS

- How you define difference between descriptive and predictive analytics?
- What kind of gap you recognize between descriptive and predictive analytics?
- What methods have been used to take advantage of predictive analytics?
    - Can you illustrate business benefits of these?
- Have you seen corporate strategies changed over the years to adapt more predictive analytics beside descriptive?
- What you reckon, will data mining achieve more important status in the future and what kind of balance there would be among descriptive and predictive analytics?

SECTION TWO: PROBING QUESTIONS

- How would you evaluate the use of CRISP-DM in real life business?
- How would you describe the needed competence that involves in data mining?
- Is there enough competence available to fully take advantage of analytics (both descriptive and predictive)?
- What could be the reasons/factors for Finnish corporations to implement more predictive analytics among descriptive?
- What kind of resources are needed to implement predictive analytics for Finnish companies?
    - Pont of view data mining
    - Point of view descriptive <> predictive
- How would you reduce the gap between descriptive and predictive analytics?

SECTION THREE: SPECIFIC AND CLOSED QUESTIONS

- Will there be increase usage of predictive analytics over descriptive?
- Will be there enough competence available in the labour market to perform predictive analytics?
- Are there enough resources for companies to implement predictive analytics to their business?