LUND UNIVERSITY
School of Economics and Management
Department of Informatics

Data Science

Influences on the Expected Outcome of Data Science

Master thesis 15 HEC, course INFM10 in Information Systems

Authors: Sanna Briskog
Anna Charlotta Riley

Supervisor: Odd Steen

Correcting Teachers: Osama Mansour
Miranda Kajtazi

Data Science: Influences on the Expected Outcome of Data Science

AUTHORS: Sanna Briskog and Anna Charlotta Riley

PUBLISHER: Department of Informatics, Lund School of Economics and Management, Lund University

PRESENTED: June, 2020

DOCUMENT TYPE: Master Thesis

FORMAL EXAMINER: Christina Keller, Professor

NUMBER OF PAGES: 131

KEY WORDS: Data Science, Data Science Challenges, Data Science Success

ABSTRACT (MAX. 200 WORDS):

By using data to generate insights, Data Science has increased in popularity and has become a substantial part of many companies. However, Data Science efforts often result in failure, leading to companies not having the expected outcome of their Data Science investments. There seems to be knowledge lacking to why, therefore, this thesis aims to describe what may influence the expected outcome of Data Science with the research question: *What influences companies' expected outcome of Data Science?* To conduct this research, a qualitative method was chosen where six interviewees from various industries working with Data Science were interviewed. Conclusively, there are both negative and positive influences on companies' expected outcome of Data Science. The negative influences identified mainly concern data quality, ethics, knowledge and organizational support. The positive influences are related to creating value from data used in Data Science, the capabilities of companies to adapt, seeking help externally to solve problems and having clear goals, clear problem formulation and a clear division of responsibilities. However, since Data Science is a broad concept with many areas of application, the expected outcome of Data Science is subjective, and depends on the organizational context and maturity.

Content

Tables

1 Introduction

In 2017, The Economist published the article "The world's most valuable resource is no longer oil but data" where data is described as a new commodity in a rapidly growing industry, a commodity that one century ago was named oil (The Economist, 2017). Data is still referred to as a vital asset for companies (Treder, 2019) and according to Raviv, Jain and Bruck (2020) data is the most important asset of the information age with a significant impact on society. Data and its technologies open new opportunities (Raguseo, 2018) and has become a way for companies to gain competitive advantage (Treder, 2019). Additionally, according to Hoffman (2017), data enables more knowledgeable insights and increased economic and social winnings, if managed correctly. Hence, the data availability makes areas such as product development, marketing and business services entirely based on data (Pagán, 2018). Due to the increased volume of data available today, data has become increasingly valuable, and is an imperative asset for organizations in the decision-making process used to generate value, draw insights from data and base decisions on these insights (Atwal, 2020; Syed Fiaz, Asha, Sumathi & Syed Navaz, 2016). This has led to companies investing in Data Science as a way to differentiate towards their peers (Provost & Fawcett, 2013b).

The concept of Data Science became publicly known during the first half of the decade (Braschler, Stadelmann & Stockinger, 2019) and has in the past years emerged to be an important part of businesses (Kotu & Deshpande, 2019; van der Aalst, 2016). Data Science assists companies in scanning through billions of rows of data to analyze patterns and create insights about their businesses (Treder, 2019). Data Science disrupts the way businesses work by offering advantages of automation and better support in decision processes that usually are managed manually by employees (Braschler, Stadelmann, & Stockinger, 2019) which has resulted in its increased popularity (Kotu & Deshpande, 2019). However, Data Science is a buzzword and there are still varying definitions (Cao, 2018). For instance, the concept might both be associated with Data Engineering, Business Intelligence, Data Analysis (Kotu & Deshpande, 2019), Computer Science, Statistics (Braschler, Stadelmann & Stockinger, 2019) and machine learning (Donoho, 2017). Moreover, van der Aalst (2016) claims Data Science to be a combination of many different fields and might include several aspects and that many approaches need to be combined in order to make the data valuable for both individuals and organizations.

However, managing large quantities of data is no easy task and managers must realize the complexity as well as the benefits and challenges that are related to such a task (Raguseo, 2018). To reach success in Data Science, Treder (2019) emphasizes the importance of companies being able to change and adjust their business. Raguseo (2018) also explains that management must adapt due to the changes this brings regarding roles, customer relationships and business models. Hence, the responsibility of making data valuable lies upon an organization's capability of transforming the data into something that is meaningful and actionable (Hoffmann, 2017).

1.1 Problem Area

A media buzz has been created concerning Data Science due to its advances, breaking the ground in many ways with the vision the concept anticipates (Braschler, Stadelmann & Stockinger, 2019). Companies such as Netflix and Alphabet have generated invaluable success stories using Data Science (Atwal, 2020) and trending companies as Google, Amazon and Facebook rely heavily on using Data Science (Braschler, Stadelmann & Stockinger, 2019). Hence, companies have started to realize the potential of Data Science and are focusing on investing heavily in Data Science technology and employing data scientists (Atwal, 2020). The increased interest in Data Science is also indicated in statistics (LinkedIn Economic Graph, 2017; Atwal, 2020). For instance, one of the four top jobs that are increasing in demand shown on LinkedIn are the data scientists and this role has increased with a staggering 650% between the years of 2012-2017 (LinkedIn Economic Graph, 2017).

However, the excitement about Data Science and its success and potential may be misleading signs, according to (Braschler, Stadelmann & Stockinger, 2019). According to Atwal (2020) only a small percentage of organizations investing in Data Science have successful and meaningful outcomes. As shown by Forrester (2016), a mere 22% of organizations generate profits from Data Science investments. This correlates with the statements of technology analysts back in 2010 to 2012, that the percentage of companies reaching failure in Data Science amounted to 80% (Braschler, Stadelmann & Stockinger, 2019), a percentage that has not changed due to Veeramachaneni's (2016) predictions of failure rates of Data Science still being at 80% in 2017. It is clear that many companies still do not have the capabilities to facilitate data correctly (Berntsson Svensson, Feldt & Torkar, 2019) and both Raguseo (2018) and van der Aalst (2016) argue that companies are meeting challenges in generating the expected value of their data. Additionally, insights generated from algorithms are often undervalued and not adopted correctly by users (Veeramachaneni, 2016) and creates a gap between the expected results compared to its real value creation, that needs to be decreased to reach the intended potential of Data Science (Braschler, Stadelmann & Stockinger, 2019). It is essential to gain extensive knowledge about challenges of Data Science to be able to improve its performance (Atwal, 2020). Also, to lack awareness of the risks and the challenges, not actively taking actions to avoid them and not doing the right things is costly for the company in many regards (Treder, 2019).

1.2 Research Question

Based on the above stated problem area the following research question is formulated:

What influences companies' expected outcome of Data Science?

1.3 Purpose

The fact that many companies fail to reach their expectations in Data Science indicates a lack of knowledge in what may influence the expected outcome within this field. Therefore, in order to decrease this gap in knowledge, this thesis aims to describe what may influence the expected

outcome of Data Science, by comparing literature and empirical results. This thesis further aims to contribute to the field of Information Systems by determining what may influence the expected outcome and thus assist to lay the foundation for future research.

1.4 Delimitation

The thesis first and foremost investigates what influences companies to have their expected outcomes by taking the perspective of the data scientist. Since data scientists are the employees within a company working closely with Data Science, an assumption is they have many insights due to this and hence, this perspective was considered to be the best fit for this research question and aim. Thus, the thesis will not focus on collecting data from other roles that might happen to be involved in Data Science, such as executives and managers.

A selection of themes and concepts will be derived from central aspects in the literature review which will be used to guide this research. This is since it is not possible within the scope of this thesis to consider all possible concepts that influence companies' expected outcome of Data Science.

2 Literature Review

The literature review begins with an explanation of the main topic, Data Science, including value creation, use areas, the data scientist and the expectations. Thereafter, now that the concept of Data Science has been explained, the main challenges with Data Science that are identified in literature are stated; data quality, ethics, knowledge and organizational support. When the challenges in Data Science have been described, how to succeed in Data Science can be explained. Hence, the following part of the chapter mentions actions impacting success of Data Science that have been identified in the literature and what responsibility the organization and the data scientists have in this matter. The literature review is concluded with a literature summary that includes a short summary of central themes and concepts, and a Thematic Overview showing the connection that each piece of literature has to each theme and concept.

2.1 Data Science

Data Science is a field that connects industry and science (Braschler, Stadelmann & Stockinger, 2019) through the use of data (Cao, 2018). Data Science is a collection of fundamental principles that supports companies to extract knowledge and information from data gathered (Provost & Fawcett, 2013a). Data Science assists human beings by scanning through billions of rows of data records to gather insights (Treder, 2019) in order to interpret customer needs and use cases and to build techniques and tools in a graphically pleasing way (Braschler, Stadelmann & Stockinger, 2019). Data used in Data Science commonly consists of objective facts or signals concerning a specific object (Cao, 2018) and is often associated with technical aspects such as databases, data transmissions and similar technical terms (Treder, 2019). The data sets may vary from sets with only a few numbers, to complex sets of thousands of rows and variables (Kotu & Deshpande, 2019) and are often numerical or from information that are revealed through observations (OECD, 2008). The data may include documents, information gathered from the internet such as chat rooms, or real-world interactions for example interviews and press conferences (Silverman, 2011) or apps, social media and news (Brennan, Chiang & Ohno-Machado, 2018). Moreover, data is a raw representation of information and to use it, it needs to be placed in a context, processed, interpreted and communicated for humans to understand it (OECD, 2008).

Data Science is a relatively new field (Cao, 2018) that became publicly known during the first half of the 2010s (Braschler, Stadelmann & Stockinger, 2019). Data Science has reached increased popularity among many companies and has already become a vital facility for several organizations (Kotu & Deshpande, 2019). Being a buzzword, the definitions and meanings of Data Science are still varying and according to Cao (2018) might sometimes be both conflicting and confusing. Data Science is however often tightly associated or even mixed up with other fields such as Data Engineering, Business Intelligence, Data Analysis (Kotu & Deshpande, 2019), Computer Science and Statistics (Braschler, Stadelmann & Stockinger, 2019). Sammut and Webb (2017) also describe that depending on context, techniques such as Machine Learning, Predictive Analytics, and Data Mining are used in Data Science. According to Müller and Guido (2017) algorithms used in machine learning that reach most success are those that

assist in decision-making, however to succeed, it is important to create a dataset that enables the desired outcome.

2.1.1 Value Creation in Data Science

Data is the main component of Data Science (Cao, 2018), but its value it's often taken for granted (Hoffman, 2017). For data to become meaningful and transformed into valuable information it needs to be assembled intelligently (Treder, 2019). Data is raw material and there are thus many ways to increase its value (Treder, 2019), not least since using data in this sense may generate insights that result in better knowledge (Braschler, Stadelmann & Stockinger, 2019; Kotu & Deshpande, 2019; Hoffman, 2017; Provost & Fawcett, 2013b). Also, Data Science makes organizations consider more than one aspect at a time (Braschler, Stadelmann & Stockinger, 2019) and might help companies to better analyze patterns (Treder, 2019) and address problems (Waller & Fawcett, 2013). However, these insights are useless if they do not have any value (Treder, 2019). Those who can harness this value creation of data may encounter both social and economic gain (Hoffman, 2017) and since many opportunities can be found in digitalization, utilizing data correctly may lead to competitive advantage (Treder, 2019).

Hence, to be able to create value in Data Science, data needs to be managed such as any other valuable business aspect (Treder, 2019). This can be done through classification, cleansing, assembling data sources, pattern identification and continuous data quality management (Treder, 2019) and may also include searching for, acquiring, storing or recovering data (Cao, 2018). Similarly, Philip Chen and Zhang (2014) identify steps to create value in decisions based on data as recording the data, data cleaning, data analysis, data visualization and interpretation and lastly, decision making. However, many companies have difficulties using data correctly and reaching its full value, for example due to data might not be available, the data quantity is often too large, unclear instructions of how to use it, its relevance and how decisions can be based on it (Berntsson Svensson, Feldt & Torkar, 2019). This makes the potential of data often unfulfilled, although businesses are aware of the value of their data (Berntsson Svensson, Feldt & Torkar, 2019).

2.1.2 Data Science Use Areas

Data Science is being used by companies in almost all industries aspiring to use data as a strategic asset, for competitive advantage and to differentiate from others (Provost & Fawcett, 2013b). The concept has been developed from only being used in metrics and investigation purposes, to being an integrated part of all aspects in science and industry (Shreshtha, Singh, Sahdev, Singha & Rajput, 2019). Therefore, the data used is being collected about everything from manufacturing and operations processes to customer behaviours and the performance of advertising campaigns (Provost & Fawcett, 2013b).

Data Science is primarily used for business purposes and to solve technical problems through its extracting, gathering, visualizing and protecting of data (Shrestha et al., 2019). According to Provost and Fawcett (2013b) the main goal of data collection for many companies and why many invest in Data Science is to become better at marketing and advertising by for example improving targeted marketing or increasing the reach of online advertisement (Provost & Fawcett, 2013b). However, Braschler, Stadelmann and Stockinger (2019) claim that Data

Science can be used to solve an infinite number of problems, ranging from challenges regarding specific phenomena to problems that occur in people's everyday life. Data Science is used to collect information about businesses and later determine if aspects are deemed as valuable to help businesses gain understanding about leadership and to produce predictive models about the future (Shreshtha et al., 2019). For example, Data Science can be used to gain knowledge about past, current and future situations to understand businesses performance (Braschler, Stadelmann & Stockinger, 2019). Additionally, companies collect external data such as its competitors' performance and trends on the market, to manage customer relationships in order to maximize the customer value (Provost & Fawcett 2013b).

Furthermore, Shrestha et al. (2019) identify four current utilizations of Data Science which can be summarized with the terms expectation, security, computer vision and natural language processing. Expectation means using Data Science to make predictive models since it can gather and analyze patterns of large sets of data which is the foundation of Machine Learning (Shreshtha et al., 2019). Expectation and prediction is also a use area stated by Braschler, Stadelmann and Stockinger (2019), and claims that Data Science can be used to forecast patterns in traffic and to forecast how likely it is for customers to positively react to advertising and marketing. The security area entails that Data Science gathers data from analyst logs and assists in identifying fraud which can be useful for banks or other monetary institutions (Shreshtha et al., 2019). Similarly, Provost and Fawcett (2013b) explain that Data Science can be used within finance for credit scoring and within operations to detect frauds. Moreover, computer vision entails that data from videos and pictures are used to provide the vision for PCs and Data Science does this through examples such as autonomous driving and in human-computer cooperation (Shreshtha et al., 2019). Braschler, Stadelmann and Stockinger (2019) also state autonomous cars and other types of technologies where humans are replaced by algorithms as use areas for Data Science. Lastly, Data Science can be used in natural language processing to use data in order to undertake for example, machine translation and parsing (Shreshtha et al., 2019).

2.1.3 The Data Scientist

In order to gain successful results from Data Science, it is vital that data scientists lead Data Science projects (Treder, 2019). Braschler, Stadelmann and Stockinger (2019) describes the data scientist as an engineering, analytical, entrepreneurial and communicative professional. Data scientists also need to have certain skills (Atwal, 2020; Saltz & Stanton, 2018), such as understanding of the business and application domain (Atwal, 2020; Saltz & Stanton, 2018; Waller & Fawcett, 2013; Braschler, Stadelmann & Stockinger, 2019), communication skills with users, knowledge of data visualization and presentation, data transformation and analysis and the ability to value quality and ethics (Saltz & Stanton, 2018). Additionally, Atwal (2020) mentions skills such as the capability of identifying and preparing the right data, applying the correct data algorithms, persuading action from stakeholders, operationalizing the data and using metrics to analyze the results. Having such skills would enable decision-making of higher quality (Atwal, 2020). People who have these many skills in Data Science are referred to as unicorns, however, finding one or becoming one will lead to disappointment since they do not exist (Braschler, Stadelmann & Stockinger, 2019).

Similarly, Provost and Fawcett (2013b) emphasize the importance of having the right data in the role of the data scientist. Managing data is one of the main activities of a data scientist according to Saltz and Stanton (2018) where tasks concern data architecture, acquisition,

analysis and archiving This means that the data scientist needs to have knowledge of organizing the data, how to gather data correctly, how to summarize, sample and visualize data and how to reuse the data for future purposes if needed (Saltz & Stanton, 2018). However, main tasks are often divided into subtasks by the data scientists to easier find a solution on issues within businesses and to provide an overall solution to the main problem (Provost & Fawcett, 2013b). However, if the appropriate data is not available, even the most qualified team of data scientists will not generate any value (Provost & Fawcett, 2013b). Saltz and Stanton (2018) also explain that without effective communication through visualization to the user of the data, the results created will be useless, no matter the quality of the analysis.

2.1.4 Data Science Expectations

The expectations of Data Science vary (Cao, 2018) and where to draw the boundaries of Data Science has been discussed for a long time (Braschler, Stadelmann & Stockinger, 2019). Data Science is sought after in all industries to be used as a strategic asset (Provost & Fawcett, 2013b) and a vital facility (Kotu & Deshpande, 2019). Data Science has created many success stories (Atwal, 2020) and comes with both societal, economic and scientific influences (Braschler, Stadelmann & Stockinger, 2019).

Data Science is regarded as one of the main global trends, together with buzzwords such as Big Data, Artificial Intelligence and Digitalization (Braschler, Stadelmann & Stockinger, 2019). Terms such as Big Data have promising future results, hence it is understandable that expectations are high, however, this reality is yet to occur (Poel, Meyer, & Schroeder, 2018). The hype surrounding Data Science leads companies to draw conclusions about its potential too hastily (Braschler, Stadelmann & Stockinger, 2019). The expected results of Data Science investments are commonly not generated and there is a gap between the expectations of Data Science and how it works in organizations (Atwal, 2020). For instance, many companies consider upholding data scientists' skills to a high level as too time-consuming (Waller & Fawcett, 2013) and have difficulties in finding the non-existent Data Science unicorn, that has all desired skills (Braschler, Stadelmann & Stockinger, 2019).

2.2 Challenges in Data Science

In order to make improvements in Data Science, it is important to obtain knowledge of what the challenges are (Atwal, 2020). However, Data Science is not deeply integrated in society yet which makes it hard to address all its possible challenges (Braschler, Stadelmann & Stockinger, 2019). What is proved is that many investments in Data Science do not generate what the organizations had expected, which may be due to wrong or outdated technology, outdated information architecture and lack of organizational support for Data Science projects (Atwal, 2020). Atwal (2020) also states that the Data Science process often is managed much more linearly than it should and the focus of data scientists should be managing the right things, not only at producing a perfect product, since this might not be what the customer wants anyway. Also, Braschler, Stadelmann and Stockinger (2019) claim that many companies do not have the knowledge of how to make a sufficient risk analysis for Data Science and are therefore not aware of the potential risks until after the integration. If companies are not aware of risks or taking measures to avoid them, losses will occur in many regards (Treder, 2019).

2.2.1 Data Quality

Another problem companies might be facing when working with Data Science is data that is low in quality, incorrect or outdated (Braschler, Stadelmann & Stockinger, 2019). Also, data might be misinterpreted, incomplete, unreported or wrongly measured which will lower the quality of the data used (Smith & Cordes, 2019). For instance, sometimes analysis fails completely due to an inaccurately placed decimal point or an unintendedly positioned minus sign (Smith & Cordes, 2019) which must be dealt with to reach success in Data Science (Atwal, 2020). The amount of data and the diversity of different data sources makes it hard to integrate the data correctly (Cai & Zhu, 2015) and according to Atwal (2020) 55 percent of data scientists state that lacking quality of data is their biggest challenge when working with Data Science. Cai and Zhu (2015) states that one reason for the low-quality data is that there are no unified data quality standards. According to Cai and Zhu (2015) there are some established standards such as ISO 8000, however, there are ongoing debates that the standards need to be improved and that only a few countries are following the standards. Also, Cai and Zhu (2015) state that today many companies use data collected from other organizations than their own, which makes it even harder for companies to ensure high quality of data. Another aspect is the very short timeliness of data; if the data cannot be collected and managed in real time, the company will deal with data that is outdated and incorrect, which might lead to misleading results or useless conclusions (Cai & Zhu, 2015). Thus, data scientists need to continuously review the data carefully and possibly reevaluate surprising or unexpected results (Smith & Cordes, 2019). According to Cave (2016), Data Science tells a lot more about the skills of the data scientists, than it tells about the world the data was collected from.

2.2.2 Ethics

There are ethical issues (Passi & Jackson, 2020) as well as moral aspects to be found in Data Science (Floridi & Taddeo, 2016). Atwal (2020) claims that aspects such as privacy, autonomy and solidarity have to be considered in Data Science projects and Cao (2018) describes issues such as protection of privacy, security in systems, data security and the trust put in data. Similarly, Braschler, Stadelmann and Stockinger (2019) emphasizes security and privacy as two primary issues regarding ethics in Data Science and Passi and Jackson (2020) focuses on trust in data. Moreover, Cave (2016) focuses more on the two ethical issues of the fallacy of reification, entailing whether human behaviour can be reduced to a quantifiable field such as data structures, and the subjectivity of humans and whether this may impact data analysis. Considering such ethical aspects is increasingly important (Cao, 2018). This is especially due to companies' increased value of web browsing history, cookies and purchase patterns (Braschler, Stadelmann & Stockinger, 2019) and to the emerging technologies in smartphones and computers, as well as the extended use of IoT devices that continuously gathers personal data about where people are located, what they do and with whom (Atwal, 2020). However, there is a risk personal data may be misused in the way that the decisions made based on that data (Braschler, Stadelmann & Stockinger 2019).

Trust problems in Data Science are emphasized by Passi and Jackson (2020) who describe four problems related to this as being ambiguous numbers, intuitive knowledge, low-credibility of data and complex models. In addition, Data Science uses data which must be interpreted and analyzed to become information, which in turn will be used to take actions (Floridi & Taddeo, 2016). When using data to analyze human capabilities, behaviour and attributes and basing decisions on these results, it is therefore imperative to consider the subjective aspects of such

data (Cave, 2016). For instance, aspects such as an individual's prior knowledge and beliefs of the quality and understanding of such data, will impact the choices made (Cave, 2016).

Therefore, the data used in Data Science is neither neutral nor entirely objective (Cave, 2016) and the reliability of results generated through data algorithms, are determined by how questions are asked within the data, the algorithm itself and how models are calibrated (Passi & Jackson, 2020). Consequently, it is up to the data scientist to monitor and oversee the work in Data Science with the goal to ensure information that is trustworthy (Treder, 2019). However, since the data needs to be interpreted and this process is partially subjective according to Cave (2016), the author argues that the results generated in Data Science are a product of the data scientist and expresses their views. A difficult task for data scientists is to conduct trustworthy work and to corroborate this into their visualizations, presentations and insights especially if they lack knowledge (Passi & Jackson, 2020). To solve such issues, it is vital that Data Science has high-quality practices and governance wherein the core is trust (Passi & Jackson, 2020). In addition, Floridi and Taddeo (2016) recommend ethical analysis within how the data is managed in order to generate as much value as possible in Data Science.

2.2.3 Knowledge

There are multiple knowledge gaps in Data Science which in turn may make it hard to integrate Data Science in organizations (Atwal, 2020). Mentioned gaps are regarding the data scientist's skills, technology available within the organization, leadership of the Data Science and data literacy (Atwal, 2020). Braschler, Stadelmann and Stockinger (2019) also state that many do not have knowledge about how to make analyses of risks correctly and lack knowledge about what risks Data Science might lead to. Waller and Fawcett (2013) also claim the importance of data scientists both having analytical skills and domain knowledge. However, upholding this knowledge might be time-consuming, hence, no individual will single-handedly uphold both of these skills to a high level (Waller & Fawcett, 2013) . Also, Atwal (2020) states there are inconsistencies in what skills data scientists assume they need, compared to what they need, for instance technical skills are valued amongst data scientists but are not the driving factors of data-driven decisions and Data Science success.

Adjustments are not commonly made to the organization when integrating Data Science and companies fail at providing enough data, software and tools since leadership underestimates this (Atwal, 2020). This leads to the knowledge gap of leadership and data literacy since organizations lack knowledge of how to manage Data Science (Atwal, 2020). Also, according to Dichev and Dicheva (2017), it is important to pose meaningful questions when working with Data Science, which requires knowledge and competence of data-driven work. The knowledge and capabilities related to understanding datasets, drawing conclusions from them, communicating these conclusions, trusting decisions made from data and defining the skills needed, is defined as the term data literacy (Atwal, 2020; Dichev & Dicheva, 2017). However, Schuff (2017) claims that definitions of Data Science undermine the importance of data literacy, although data literacy is vital within organizations (Dichev & Dicheva, 2017) to draw benefits and create insights from data (Atwal, 2020).

2.2.4 Organizational Support

Atwal (2020) states that the challenge of delivering Data Science successfully is the cultural attitude of people. In fact, culture is the main aspect in every change process and therefore is the first thing that needs to be in place when going through the transformation of Data Science (Treder, 2019). Only companies that are prepared to change rapidly and that quickly can adjust to the new requirements will succeed with Data Science (Treder, 2019). It is also important to have a digital mindset throughout the whole organization and make sure the adoption of Data Science does not only affect one team (Subrahmanyan & Jalona, 2020). Additionally, problems occur when companies are employing data scientists and the managers and executives are expecting them to add value for the business without offering the support needed (Atwal, 2020). Executives and managers tend to be afraid of being slowed down when having to regard others' needs which becomes problematic since it is the key to managing data properly (Treder, 2019).

Unfortunately, it is common that organizations see data scientists as completely responsible for making the organization data-driven and are expected to overcome challenges themselves (Atwal, 2020). Companies encounter problems when the data management department and the IT department are not supporting each other and the rest of the organization in data problems (Treder, 2019). Lacking advancements in data literacy might lead to incorrect decisions and if people within the organization disregard the recommendations for Data Science (Atwal, 2020). To succeed with Data Science integration, it is critical to have an organizational culture that is built upon supporting each other with clear directives (Atwal, 2020). However, having the organizational culture in place is only one step in the Data Science process (Subrahmanyan & Jalona, 2020), otherwise the exploitation and management of data might be harmed (Atwal, 2020). If the culture is healthy there is a higher chance of creating a workforce that is engaged and are working proactively to make improvements (Subrahmanyan & Jalona, 2020). Data needs to be everybody's business (Treder, 2019).

2.3 Data Science Success

There is evidence that by investing in Data Science, organizations have created invaluable success stories and have succeeded to create advantages against competitors (Atwal, 2020). Also, using data in this way may lead to both economic and social winnings (Hoffman, 2017). To reach success in Data Science Treder (2019) mentions several influencing aspects, for instance the capabilities of an organization to quickly change, adapt and adjust their business and how data scientists lead the Data Science projects. Using Data Science, the Netflix recommendation algorithm was created leading to savings of $1 billion annually, and a hospital in London, in collaboration with Alphabet, created an artificial intelligence system making it possible to allot treatment of more than 50 diseases as good as an expert doctor would (Atwal, 2020). Data Science and predictive technology was used to predict hurricane Frances and to forecast people movement patterns (Provost & Fawcett, 2013a). Also, the investments made in Data Science by the pharmaceutical company Monsanto, has decreased their costs of global logistics and transportation by $14 million every year, which subsequently results in a reduction of 350 metric tons of CO2 emissions (Atwal, 2020).

2.3.1 Actions for Data Science Success

In order to gain data insights, companies need to ensure that the data is meaningful for the organization and that they are not only following the popular trend of managing data through Data Science (Treder, 2019). Braschler, Stadelmann and Stockinger (2019) explain that although organizations are data-driven, steps need to be taken to put their data in motion to benefit from it. In order to reach success when solving a Data Science problem, it is important to question the goals, to be willing to readjust and improve the methods and analysis (Braschler, Stadelmann & Stockinger, 2019).

Furthermore, in order to successfully integrate Data Science, it is vital to have a clear plan or a strategy that is intuitive and easy to understand (Atwal, 2020). In order to realize more benefits in Data Science and better adapt to changing requirements and assumptions, it is advantageous to work agilely (Demigha, 2019). However, depending on the objectives of the Data Science project, there are several different strategies that can be used (Treder, 2019). A method for understanding an organization's environment is mentioned by Atwal (2020) by emphasizing the creation of strategic objectives which includes five main parts; people, technology, organization, data and processes. Another strategy that can be used entails analyzing the organization's culture with the goal of minimizing the gap between how Data Science currently works in an organization, and the expectations of it, and determining which factors inhibit further data-driven efforts (Atwal, 2020). Companies can also set Data Science strategies to determine how well a process and organization is performing, for example by using time, cost and quality as metrics with separate key performance indicators (van der Aalst, 2016). To easily set a Data Science strategy and make these types of decisions, it is vital that the company knows the organization, employees, stakeholders and customers and are aware of trends and competencies in the organization and that both the internal aspects, as well as external factors, are included (Atwal, 2020). Without being aware of this, no strategy will work (Atwal, 2020).

Hence, it is also important to view the extent to which an organization's culture is prepared to have a more data-driven focus (Treder, 2019). Atwal (2020) emphasizes the importance of management enforcing data literacy within an organization which entails basing decisions correctly on data, the knowledge and capabilities of doing so and the understanding of when it is misinforming (Atwal, 2020). Data literacy is vital to be able to benefit from data and for Data Science to be successful (Atwal, 2020). Organizations need to be prepared for changes following Data Science in this sense, since it impacts that organizational culture (Treder, 2019). In addition, past technologies and methods in the organization need to be questioned, however, employees tend to refer to previous successes as a way to justify old solutions, which is a typical example of resisting change, thus makes organizational change increasingly important (Treder, 2019).

According to Braschler, Stadelmann and Stockinger (2019) even successful solutions might lead to problems and thus, it is important for companies to think one-step ahead. Braschler, Stadelmann and Stockinger (2019) states problems most commonly occur if the company only considers details and uses a far too narrow perspective when solving problems, which in turn might lead to even more problems. Hence, since it is impossible to predict the outcome of solutions and what will work, it is advantageous to test scenarios of different solutions before choosing the final solution (Atwal, 2020). As Pagán (2018) explains, it is also imperative to measure the Data Science efforts by using metrics to ease information gathering and subsequently lead to gained knowledge. For example, it might be beneficial to use automated

techniques for data wrangling, i.e. preparing the data in order to decrease the amount of manual work (Braschler, Stadelmann & Stockinger, 2019). Also, to minimize the risk of second order problems occurring, Atwal (2020) claims the importance of getting feedback from users and measure the results in order to see whether the action was successful.

2.3.2 The Organizational Responsibility

In order to integrate Data Science and a data-driven approach correctly, changes in the organizational and management support (Atwal, 2020) and the organizational culture are required (Treder, 2019). In order for employees to embrace this organizational change, communication is a must, and management must ensure the new mindset is not forgotten (Subrahmanyan & Jalona, 2020). Also, to increase the revenues, new management and new skills are required (Raguseo, 2018). Atwal (2020) explains that management tends to employ data-driven decision making in organizations but rarely use it themselves, which transmits wrong signals to employees. Making management understand how to use data to make decisions and how this can assist them is a time-consuming process (Subrahmanyan & Jalona, 2020). Therefore, it is vital for all employees in an organization to have a data-driven mindset and for management to ensure that everyone understands how they are benefited by making data-driven decisions (Subrahmanyan & Jalona, 2020). To facilitate this, Subrahmanyan and Jalona (2020) propose appointing digital officers in every department to promote this mindset. Also, it is important to continuously review the extent to which an organization's culture is ready to have a more data-driven focus (Treder, 2019). In order to reach success, the organization also has a responsibility in emphasizing the necessity of exchanging information and knowledge between experts and data scientists with different backgrounds (Braschler, Stadelmann & Stockinger, 2019) and to always offer data scientists the right tools to detect data that is incomplete, incredible or biased (Smith & Cordes, 2019).

2.3.3 The Responsibility of the Data Scientist

In order to succeed with Data Science, it is important that data scientists have the ability to get a data perspective of business problems (Provost & Fawcett, 2013a) and to completely understand the impact of Data Science (Atwal, 2020). For example, many methods for data collection come with privacy concerns and thus, data scientists may consider if there are any possibilities to limit the amount and access of personal data or if there are any methods that can be used to prevent leakage of personal information (Falgoust, 2016). Data scientists need to closely investigate the data they are analyzing in order to avoid faults in measurements or clerical errors (Smith & Cordes, 2019). The data scientists also must consider that the quality of the tools and techniques used directly impact the quality of the decisions and visualizations (Berntsson Svensson, Feldt & Torkar, 2019). However, an informal study mentioned by Braschler, Stadelmann and Stockinger (2019) showed that only about half of all data scientists take measures to ensure data quality due to the fact that management and customers do not require them to do so. In order to generate trustworthy information and get successful results of Data Science, Treder (2019) emphasizes the importance of established data scientists to monitor and lead the Data Science projects. Also, to succeed in Data Science it is required that the Data Science team has all required skills and have different profiles (Braschler, Stadelmann & Stockinger, 2019).

2.4 Literature Summary

The Literature Summary aims to sum up the literature used in the Literature Review (see Chapter 2) and show the central concepts. In the Thematic Overview (see Table 2:1) the concepts that are connected to each theme and its connection to literature is illustrated.

The concept of *Data Science* is a buzzword that is broad and varying and can range from Data Engineering, Business Intelligence, Data Analysis to Computer Science and Statistics. The data used in Data Science has several different meanings depending on the author and includes everything from numbers to observations to text collected online (see Chapter 2.1). *Value creation in Data Science* is important to make it meaningful and entails various approaches to data management. However, many companies have difficulties in this which leads to unfulfilled value of data (see Chapter 2.1.1). *Data Science can be used* in companies in various industries to solve business challenges and technical problems, and emerged to assist people in scanning large data amounts and to generate insights using data (see Chapter 2.1.2). Thus, the person who manages the data and leads these types of projects is referred to as a *data scientist* and needs to have expertise in both the technical domain and the structure of their organization (see Chapter 2.1.3). Also, *Data Science Expectations* vary and the results expected are not often generated (see Chapter 2.1.4).

When managing *Data Science, challenges* commonly occur which makes many organizations not reach their expectations of Data Science. Challenges might relate to the work methods used, insufficient technology and unawareness of risks (see Chapter 2.2). The main challenges identified in the literature review are regarding *data quality* (see Chapter 2.2.1), *ethics* (see Chapter 2.2.2), *knowledge* (see Chapter 2.2.3) and *organizational support* (see Chapter 2.2.4). However, to solve issues in Data Science, knowledge and understanding is needed about what caused them. To manage these challenges and reach *success in Data Science* (see Chapter 2.3), several *actions* that impact the expected outcome are provided. The actions identified from the literature concern goals, work methods, organizational culture, management and Data Science techniques (see Chapter 2.3.1). The chapter also states the *responsibility of the organization* and the *responsibility of the data scientist.* The organization needs to support the data scientists' work to create a data-driven mindset throughout the whole organization and offer the data scientists the right tools, for example, the ability to exchange knowledge with others (see Chapter 2.3.2). The data scientist needs to be fully aware of the impact of Data Science and to carefully review and control the data they are using in order to provide truthful information (see Chapter 2.3.3).

2.4.1 Thematic Overview

The Thematic Overview demonstrates the main themes presented in the literature review and illustrates concepts that were most occurring in literature (see Table 2:1). The table includes the following main themes: Data Science, Data Science Challenges and Data Science Success and additionally shows each concept related to each theme. The Thematic Overview (see Table 2:1) is intended to guide and structure the upcoming part of the thesis as well as providing a foundation to build the interview guide upon.

Table 2.1: Thematic Overview

Themes	Concepts	Literature
Data Science	• Value Creation in Data Science • Data Science Use Areas • The Data Scientist • Data Science Expectations	Atwal, 2020; Berntsson Svensson, Feldt & Torkar, 2019; Braschler, Stadelmann & Stockinger, 2019; Kotu & Deshpande, 2019; Shreshtha, Singh, Sahdev, Singha & Rajput, 2019; Treder, 2019; Brennan, Chiang & Ohno-Machado, 2018; Cao, 2018; Poel, Meyer & Schroeder, 2018; Saltz & Stanton, 2018; Hoffman, 2017; Müller & Guido, 2017; Sammut & Webb, 2017; Philip Chen & Zhang, 2014; Provost & Fawcett, 2013a; Provost & Fawcett, 2013b; Waller & Fawcett, 2013; Silverman, 2011; OECD, 2008.
Data Science Challenges	• Data Quality • Ethics • Knowledge • Organizational Support	Atwal, 2020; Passi & Jackson, 2020; Subrahmanyan & Jalona, 2020; Braschler, Stadelmann & Stockinger, 2019; Smith & Cordes, 2019; Treder, 2019; Cao, 2018; Dichev & Dicheva, 2017; Schuff, 2017; Cave, 2016; Floridi & Taddeo, 2016; Cai & Zhu, 2015; Waller & Fawcett, 2013.
Data Science Success	• Actions for Data Science Success • The Organizational Responsibility • The Responsibility of the Data Scientist	Atwal, 2020; Subrahmanyan & Jalona, 2020; Berntsson Svensson, Feldt & Torkar, 2019; Braschler, Stadelmann & Stockinger, 2019; Demigha, 2019; Smith & Cordes, 2019; Treder, 2019; Pagán, 2018; Raguseo, 2018; Hoffman, 2017; Falgoust, 2016; van der Aalst, 2016; Provost & Fawcett, 2013a.

3 Research Methodology

The research methodology begins with a presentation of the selected research method for this thesis and a description of how the literature review was conducted. Thereafter, the data collection strategy is stated, including how the informants were selected, how the interview guide was conducted, and how the interviews were transcribed. This is followed up by a description of the data analysis. The research methodology chapter ends with a research quality and ethics part, which describes how quality and ethics are considered throughout this thesis.

3.1 Research Strategy

In order to answer the research question, the aspects and themes derived from the literature review was the central focus of this thesis. This focus was the foundation that the empirical study was based on to determine what influenced companies' expected outcome of Data Science. A qualitative research method was selected to answer the research question posed. Using a qualitative research method is preferable to get deeper insights and knowledge from real cases (Recker, 2013) and makes it possible to gain details and explanations (Burke Johnson and Onwuegbuzie, 2004). This type of method investigates text, beliefs, knowledge, experiences or other types of social aspects (Schulte & Avital, 2011; Recker, 2013). Since our research question and thesis relied on gaining deep and detailed knowledge and insights of people's perceptions, a qualitative research method was thus considered as suitable. To gain this deep knowledge it was also considered important to interact with the informant, something that a qualitative research method allowed.

Qualitative research methods tend to be more time- and resource consuming both regarding data collection and data analysis (Burke Johnson & Onwuegbuzie, 2004). Hence, since this thesis uses a qualitative research method, it is useful to study a limited amount of cases. In order to collect data for this research, interviews were selected as a method. However, due to time and resource constraints and in order to keep high quality of the interviews, only 6 interviews were conducted. However, qualitative research risks being affected by the researchers own biases (Recker, 2013; Burke Johnson & Onwuegbuzie, 2004). In order to avoid these kinds of problems to the greatest extent possible, these types of considerations were valued regarding ethics and research quality. How these considerations were given is presented in Chapter 3.5 (*Research Quality and Ethics*).

3.2 Conducting the Literature Review

The Literature Review (see Chapter 2) of the thesis entailed conducting a review of the literature used. According to Randolph (2009) the main parts of conducting a literature review entail a justification for carrying out the review, research questions guiding it, a data collection plan, a data analysis plan and a data presentation plan. Before the literature review commenced, the research question and purpose were decided and are clearly outlined in the Introduction (see Chapter 1).

Before the data collection of the literature began, certain keywords were specified to ensure high quality of the literature review. In order to identify keywords before the literature search commences (Timmins & McCabe, 2005) it is important that they are relevant (Efron & Ravid) and correct, since incorrect keywords may lower the quality of the review (Timmins & McCabe, 2005). The keywords identified in this thesis are stated below and (Randolph, 2009) also states keywords must be documented. To broaden the search and to consider all relevant keywords of the thesis, the terms AND and OR are used. The term OR is used to search for terms that have a similar meaning and AND is used to search for new and separate terms that differ from one another. The following keywords were used in the literature search:

- "Data Science" AND "Data Science Value Creation" AND "Data Science Use Areas" AND "Data Scientist" AND "Data Science Expectations"
- "Data Science Challenges" OR "Data Science Problems" AND "Data Science Data Quality" AND "Data Science Ethics" AND "Data Science Knowledge Gaps" OR "Data Science Knowledge Issues" AND "Data Science Organization Support" OR "Data Science Management Support" OR "Data Science Organization Culture "
- "Data Science Success" OR "Data Science Success Factors" OR "Data Science Actions Success" AND "Data Science Organization Responsibility" AND "Responsibility Data Scientist"

The above-mentioned keywords were defined before the search for literature on the internet began. According to Randolph (2009), the process of gathering data usually begins with an online literature search where Timmins and McCabe (2005) mention that the internet is a good source holding many types of information and an emphasis is put on finding the appropriate sources and literature. When conducting the literature review, it was of priority to find sources that were up-to-date. Efron and Ravid (2019) state that searching for the most recent research is a good foundation to start the literature search upon. The search for sources for this thesis focused mainly on literature from 2015 and onwards. The search engines Google Scholar, LUBSearch and LUBCat were used, where the last two are search engines provided by Lund University. The use of search engines when searching for literature is discussed by Efron and Ravid (2019), who state that these assist researchers in finding literature more quickly and ensure the reliability, validity and trustworthiness of the sources found.

The literature gathered in the literature review consists mainly of journal articles, books, e-books and conference papers. To ensure the research quality, peer-reviewed papers are valued highly in this thesis. However, papers that are not peer-reviewed will also be used partially, to ascertain that no information is missed. Papers that are not peer-reviewed are mentioned by Timmins and McCabe (2005) as informative and should not be neglected.

The information found in the literature used has been read through in multiple sets to ensure its relevance. Efron and Ravid (2019) highlight that the process of searching for literature is iterative and analysis should be conducted several times. The last chapter of the Literature Review (see Chapter 2) consists of a Thematic Overview (see Table 2:1) which is divided into the most commonly occurring themes of the literature review and relating identified concepts, and what sources that encompass them. Summarizing the literature assists in organizing and structuring the information (Efron & Ravid, 2019). Furthermore, the themes and concepts used in the Thematic Overview (see Table 2:1) are linked to the purpose and research question of this thesis. Efron and Ravid (2019) explain that themes and subthemes should be built on the

foundation of the purpose, questions posed and the sources used. This is shown in the Thematic Overview (see Table 2:1) and additionally, the themes and concepts identified here, were themes and headings used consistently throughout the thesis and were the basis upon which the interviewee questions were formed. The questions used in this empirical study were derived from themes in the literature review, and by using them, this research hopes to identify similarities, differences and conflicting statements from the interviewees responses. Moreover, the literature review should act as a base for designing the questions used in the empirical part of the thesis (Efron & Ravid, 2019).

3.3 Data Collection

As earlier mentioned, interviews were the selected method for collecting data for this thesis. Interviews are the most commonly used method for data collection (Recker, 2013; Bhattacherjee, 2012) and proven to result in detailed and high-quality data since many different aspects are considered (Schultze & Avital, 2011). Since this thesis aims for a result that can be applicable for companies regardless of industry, this was considered as important. Interviews can be conducted either through focus groups, face-to-face or via telephone (Recker, 2013). Interviews were considered the best suitable method for data collection for this thesis since it made it possible to ask follow-up questions and investigate thoroughly what was regarded as important or relevant to this research and assist in answering the research question. Also, interviews were selected as a method for data collection since it made it possible to conduct them without being at the same place as the interviewees. To not have to be in the same location as the interviewees was of substantial interest due to the ongoing Covid-19 outbreak which limited both traveling and physical contact. This is also why the interviews were decided to be conducted via telephone and video-call.

Furthermore, interviews of a descriptive nature were found the most suitable for this research. The descriptive nature was chosen due to the importance of focusing on the perspective of the data scientist, which is in line with the delimitations of this research, and since knowledge was lacking about each organization and how they were managing Data Science. Recker (2013) claims descriptive interviews are used to generate detailed descriptions of how something is comprehended by individuals. It was of importance to ensure that the interviews had a clear outline and that the questions asked were relevant throughout the whole interview but also that the interviewee could elaborate more freely to shed light on previously unidentified aspects. Thus, semi-structured interviews were chosen. Semi-structured interviews were also considered as advantageous to not disregard any important information. In semi-structured interviews, the script is incomplete where the researcher only has prepared a few questions and improvisation is required (Myers & Newman, 2007). In this thesis, it was regarded as important to have detailed discussions and descriptions from the interviews to decrease vague results and findings. According to Recker (2013) semi-structured interviews makes it easier to discuss topics and obtain reasons for the answers from the interviewees.

3.3.1 Selection of Interviewees

In this thesis, finding interviewees in several different industries was aimed for. Otherwise, there would be a risk that some findings were only related to a specific industry and that these would not apply to a broader context. The interviewees were primarily selected based on their

role within the organizations and their experience. When selecting informants or interviewees the researcher must consider if it is more advantageous to interview experts that are specialized on the phenomena or subject or generalists i.e. people in the community or industry in general (Albuquerque, de Lucena & de Freitas Lins Neto, 2013). As previously stated, this thesis desired to make findings applicable in a broader context and thus, generalists has been prioritized. Also, according to Myers and Newman (2007) researchers that only interview experts or specific people with high status within the organization might fail to understand the broader context.

As demonstrated in the Interviewees table (see Table 3:1), the official title and role each interviewee has is shown, in what industry they work within, how many years they have been conducting work within the area of Data Science and the assigned letter each interviewee has been given depending on the time order that the interviews were conducted. The interviewees were required to represent an organization that works within Data Science either within their own company or by consulting other organizations about Data Science. Also, the interviewee needed to have good insights in how Data Science is managed by the organization. Since Chapter 2.1.3 (*The Data Scientist)* is a vital part of the literature review in this thesis, first and foremost interviewees with the profession data scientist have been asked to participate. If positive responses were not given as required when asking the data scientists, our strategy was to include experts and others working within the field of Data Science. However, this was not needed since enough responses from data scientists were generated.

Table 3.1: Interviewees

Interviewee	Role	Industry	Years of Experience
Interviewee A	Data Scientist	Finance	6 years
Interviewee B	Data Scientist	IT-Consulting	11 years
Interviewee C	Data Science Manager	Packaging	11 years
Interviewee D	Data Scientist	Furniture	13 years
Interviewee E	Data Scientist, CTO	IT-Consulting	20 years
Interviewee F	Data Scientist	IT-Security	15 years

3.3.2 Design of Interview Guide

The conducted Interview Guide (see Table 3:2 or Appendix A) provides several themes and questions for the interviewee. These themes follow the same structure as the Thematic Overview (see Table 2:1) and are recurrent throughout the thesis. Since a semi-structured interview was conducted these questions were not followed strictly but acted as a guide for what questions to ask in order to not miss any important parts or aspects. Therefore, the questions could be reformulated or reorganized during the interview process and questions could be added or removed. When conducting semi-structured interviews the first questions are often more general (Recker, 2013) and should include an introduction of the researchers and an explanation of the purpose and objectives of the interview (Myers & Newman, 2007).

According to Myers and Newman (2007) it is important that the informants are aware of what the interview is about and that the expectations of the informants are set correctly. Besides these things, the interviews were introduced by explaining some of the ethics that are considered in the thesis such as the importance of informed consent, voluntary participation and anonymity. Thereafter, the interview began by following the question format that was defined in the Interview Guide (see Table 3:2 or Appendix A). Firstly, some introductory questions were asked where the interviewee was expected to answer a few general questions about their role and their experiences of Data Science. Secondly, the main questions were asked following the main themes that were derived from the literature review, these questions are also presented in the Interview Guide (see Table 3:2 or Appendix A). The last questions asked were about if there were any other things that the interviewee wanted to add or if they had questions regarding the research. These last types of questions might be the most informative and might cover parts that otherwise would have been missing in the interview (Patton, 2015).

Table 3.2: Interview Guide

Introduction	**Questions**
Introductory Questions	• What does your job entail? • Explain your job tasks, position within the organization etc. • What is your experience of Data Science? Previous jobs, skill set etc.
Themes	**Questions**
Data Science • Value Creation in Data Science • Data Science Use Areas • The Data Scientist • Data Science Expectations	• In what use areas and how do you use Data Science within the organization? • What type of data do you manage in your daily work? • What are your main tasks when working with Data Science? • What skills do you believe to be most important to have when working with Data Science? • How is the Data Science department structured at your own company? • Do you usually collaborate with people with other roles or from other departments? • How do you work to make the data valuable for the organizations? • What are usually *your* expectations of using Data Science? • What are usually the *organization's* expectations of integrating Data Science? • Do you consider there to be a match between *your* expectations versus the expectations of *your organization* and the *organizations that you assist?*
Challenges in Data Science • Data Quality	• Have you, your organization or organizations you have assisted faced any challenges in Data Science? *If yes:*

• Ethics • Knowledge • Organizational Support	• How were these solved? • In the future, to avoid such challenges, what actions do you think should be taken? *If no:* • How could challenges be avoided? What sort of preventive actions can be taken according to you? • Do you consider data quality when working with Data Science? How? • Have you encountered any issues regarding low-quality data during your work? • Do you consider Ethics when working with Data Science? How? • When working with Data Science, have you ever encountered problems regarding ethics, inaccurate management of data, subjective data or lack of trust to data? If so, in what way? • Do you consider possible knowledge gaps when working with Data Science? How is this incorporated in your work? • Have you encountered any issues regarding knowledge gaps during your work? • Do you consider organizational management support when working with Data Science? How? • Have you encountered any issues regarding the organizational support in your organization or organizations you assist during your work in Data Science?
Data Science Success • Actions for Data Science Success • The Organizational Responsibility • The Responsibility of the Data Scientist	• What do you think is important to consider in order to reach success in Data Science? Why? • Do you think that *your* expectations, *your organization's* expectations and th*e expectations of organizations that you assist* are reached in Data Science? • What actions do you think your organization should take to better reach their expectations of Data Science? • What actions would you recommend other organizations to take to better reach their expectations of Data Science? • Who has the main responsibility of seeing through these actions? • Are there any actions you could take in *your* role in Data Science to help with its success?
Ending	**Questions**

Closing Questions	• Is there anything additional you would like to mention that would be relevant for our thesis? • Do you have any questions regarding this interview or our research in general?

3.3.3 Conducting the Interviews

As previously mentioned, the interviews were conducted through telephone and video-call. The program chosen for making the video-calls was Google Meet. This was since the researchers of this thesis perceived Google Meet as easy to use and since it did not require downloads of the program to be able to use it. The interviews were recorded to make sure that every word was captured, which made it easier both to transcribe and analyze the data. Also, recorders do not conceal parts of conversations or change what has been said (Patton, 2015), which means that by recording the interviews, the risk that something will be misinterpreted or misunderstood is decreased. In order to ensure even higher quality of the interviews, notes were taken during the interviews. Note-taking was first and foremost used as a tool during the interview to easier remember and iterate back to things that had been said but also to easier formulate new and relevant questions during the interview process. Patton (2015) argues that notes facilitate the analysis of the data and to find usable quotes from the interviewee, but also to use as an assurance if the recorder does not work. In this thesis, one researcher interviewed the interviewee while the other researcher took notes. However, the interviewer also took key notes in order to more easily pose questions and the researcher that took notes asked some questions as well if anything was missed or was of special interest.

Besides recording the interview and taking notes, the organizations the interviewees worked for were researched. Myers and Newman (2007) states that the researcher needs to be well prepared about the organization that the interviewee represents in order to have a more professional impression and show interest. However, showing interest was experienced as more effortless when conducting interviews via video-call, than when using telephone, since video-calls also include body language, such as nodding or smiles. This was considered as valuable information for us as researchers and thus, video-calls were the preferable choice if the interviewees were satisfied with both options.

To demonstrate the above-mentioned aspects regarding how the interviews were conducted, the Interviews table (see Table 3:3) was created. The table includes what type of interview was conducted, the date for each interview and its duration, which researcher transcribed and respectively verified each transcription, and in which appendix the responses are presented for each given interviewee. Also, to visualize the researchers work division between the transcripts and verifications of transcripts, the first letter of each researcher's surname, namely B and R, is used.

Table 3.3: Interviews

Interviewee	Interview Type	Interview Date	Interview Duration	Transcribed by	Verified by	Appendix
Interviewee A	Telephone	April 14th 2020	45 min	B	R	B
Interviewee B	Video-call	April 14th 2020	75 min	R	B	C
Interviewee C	Video-call	April 15th 2020	45 min	B	R	D
Interviewee D	Video-call	April 17th 2020	40 min	B	R	E
Interviewee E	Video-call	April 20th 2020	60 min	R	B	F
Interviewee F	Video-call	April 23th 2020	45 min	R	B	G

3.4 Transcriptions and Analysis of Interviews

After the interviews were conducted and completed, the data from the interviewees was gathered and the interviews were transcribed (see Appendix B-G). This entails that the interviews are transformed from oral form, i.e. the recordings, into written form, in order to facilitate the analysis (Kvale, 2007). Also, the transcriptions were considered as a way to ensure transparency and reliability and ease replicability of the research. When transcribing the interviews, a tool in the form of a web app was used which assisted to pause, fast-forward and rewind the recording in an effective way. The tool also allowed us to decrease the speed which slowed down the recordings and was helpful in order to not miss any information. All contact with the interviewees that was made prior to the interviews, such as first contact, interview format and meeting scheduling, were made in Swedish. Due to that all responses from the interviewees were in Swedish also, an assumption was made concerning that the interviewees could participate in an interview that had Swedish as the main language. Hence, all interviews were decided to be held in Swedish and later translated into English. To minimize the risk of misinterpretations being made during translation, while transcribing the interviews, a three-step process was used. Firstly, the interviews were transcribed word-by-word in Swedish, secondly, the transcriptions were translated into English and thirdly, the researcher not conducting the first two steps of the translation process of a certain appendix, verified the translation to ensure no misinterpretations were made. For example, the interviews that were transcribed and translated by researcher B were verified by researcher R and vice versa. When the transcriptions were finished, the data was analyzed. Analyzing and interpreting data means that the researchers make sense of the interviews by compiling and comparing the answers to find patterns (Patton, 2015). Out of the many varying techniques that exist to analyze data, the

chosen method was coding. Coding was considered as an appropriate method, since the six interviews, each exceeding 40 minutes, generated large amounts of data. This correlates with Linneberg and Korsgaard (2019) that by using coding for analyzing large quantities of data gathered, a higher quality of the data analysis can be reached. Since themes and concepts were previously identified in the Thematic Overview (see Table 2:1), the codes were created from these. Additionally, Vaughn and Turner (2016) describe themes and topics used in coding as a good method to help define and highlight different sections of the data and assist in the analysis.

Two researchers were involved in the data analysis of this thesis and used coding which assisted in the iteration of the data analysis and to ensure that no data was lost. When collaborating, coding enables researchers to go back and reevaluate the data which eases analysis when researchers cooperate and allows identification of data that may have been missed (Linneberg & Korsgaard, 2019). To ensure high credibility in the coding process, Intercoder reliability (ICR) was adapted. The two researchers of this thesis conducted the coding process through independent coding of the transcriptions and this was followed by a validation of the other researchers' coded transcriptions, to make a comparison of the results. This was adapted to ensure that the codes were interpreted correctly by both researchers. Using ICR is a way to ensure that different coders have the same perception of how the data should be coded (O'Connor & Joffe, 2020). Being conscious of that the coding scheme will be used separately by another researcher and the fact that the individual codes will be compared with one another is assumed to increase the likelihood of performing a high-quality coding process. O'Connor and Joffe (2020) mentions that ICR gives coders the incentive to perform well and be consistent in the coding. Additionally, since the coding choices were compared, discussions between the different opinions of the coders occurred which made the researchers reflect on their choices. ICR is stated by O'Connor and Joffe (2020) as a technique that allows dialogue between the researchers and permits clarification regarding conflicting interpretations.

The coding used in the data analysis in this thesis was mainly focused on themes derived from the literature examined. defined in the Thematic Overview (see Table 2:1). Using the literature as a foundation to create themes in coding is deductive coding according to Linneberg and Korsgaard (2019) and states that taking such an approach gives structure to the analysis. Furthermore, before the data was gathered, themes and concepts were identified in the Thematic Overview (see Table 2:1) and questions within these themes were subsequently decided and defined in the Interview Guide (see Table 3:2 or Appendix A). Hence, the codes used were Data Science, challenges in Data Science and Data Science success were defined in the Coding Scheme (see Table 3:4). Subcodes were also defined following the concepts identified within each main theme.

Table 3.4: Coding Scheme

Code	Code Description	Subcodes	Subcode Description
DS	Data Science	DS-VCDS	Value Creation in Data Science
		DS-DSUA	Data Science Use areas
		DS-DS	Data Scientist
		DS-DSE	Data Science Expectations
DSC	Data Science Challenges	DSC-DQ	Data Quality
		DSC-E	Ethics
		DSC-K	Knowledge
		DSC-OS	Organizational Support
DSS	Data Science Success	DSS-ADSS	Actions for Data Science Success
		DSS-OR	Organizational Responsibility
		DSS-DSR	Data Scientist Responsibility

3.5 Research Quality and Ethics

In qualitative research, many issues regarding ethical aspects are encountered (Wiles, 2012; Brinkmann & Kvale, 2005). These issues most commonly arise due to the complexities of investigating people's lives and publishing it to the public (Brinkmann & Kvale, 2005). Since this research is qualitative and uses interviews with people to generate data, efforts will be made to minimize such complexities. Moreover, Patton (2015) emphasizes the importance of considering ethics in research in order to ensure that the empirical data is of value and that the research maintains a high quality. As Wiles (2012) states, ethical aspects may be considered and integrated before research commences. Since some ethical issues may emerge during the research process, ethical aspects and research quality were two main considerations that were valued in this thesis both before, during and after the research process.

3.5.1 Research Quality and Ethics in the Literature Review

Before beginning the search for literature to review, identifying relevant keywords in relation to the research topic were valued. The risk was otherwise, as mentioned previously, that incorrect keywords could lower the quality of the literature review (Timmins & McCabe, 2005).

Three search engines were used to search for literature, as described in Chapter 3.2 (Conducting the Literature Review). Using search engines is a way to ensure the reliability, validity and trustworthiness of the literature found according to Efron and Ravid (2019).

This thesis aims to uphold high levels of credibility and reliability. When the literature was gathered and summarized, the critical analysis of it can commence which ensures the credibility and reliability of the literature (Efron & Ravid, 2019) and according to Bhattacherjee (2012) for research to be credible, it must be perceived as believable by the readers. To ensure the credibility and reliability of this research and for it to be increased, primary, peer-reviewed sources from established publishers are mainly used and determining the relevance of each source is valued. The quality of the literature review relies mainly on what type of references that are used and hence, each individual source should be scanned to see if it is peer-reviewed or written by a famous researcher within the field (Efron & Ravid, 2019). However, since only focusing on peer-reviewed sources may lead to neglecting important information from non-peer-reviewed sources (Timmins & McCabe, 2005), literature that is not peer-reviewed was also used in some cases if determined as vital for this review.

3.5.2 Research Quality and Ethics in the Data Collection

Before collecting data for this research and before interviews commence and are recorded, Bhattacherjee (2012) argues that all informants should be informed about their rights. The information should include what the study is about and inform the participant about that their participation is voluntary and that they have the right to withdraw their participation (Wiles, 2012). Also, the participants should be aware of how the researchers will manage anonymity and confidentiality (Wiles, 2012). Before starting every interview, the interviewees rights were stated clearly and were informed of what the research was about, its purpose, that notes will be taken and that the interview will be recorded if the interviewee gives us their consent. The interviewees received information that their participation was voluntary and that their name and organization are not mentioned in the thesis, but that their role and the industry they are working within is mentioned. The researchers of this thesis informed the interviewees that the result of the interview will be compiled into a master thesis at Lund University and that this master thesis can be accessed online.

Due to this and that the interviewees were asked prior to the commencement of the interviews if a copy of the thesis was desired, the interviewees can confirm the findings discovered in this thesis. Confirmability is important when regarding research quality and entails the extent to which a participant for instance confirms the research results (Bhattacherjee, 2012). Additionally, the quality of the research may be increased if transferability is considered which is being able to generalize findings to other settings (Bhattacherjee, 2012). Transferability in this thesis was considered by providing thorough and detailed descriptions of how the data was collected in the given context and structured tables to visualize this, which makes it easier for other researchers to repeat this study. These are measures that were considered which will improve the quality and ethics in this research. If the participants are not well informed about such things before the interview starts, there is a risk bias might be a problem for the interview (Bhattacherjee, 2012).

3.5.3 Research Quality and Ethics in the Data Analysis and Empirical Results

The process after the interviews are conducted is critical in order to prove that the data is valid and high in quality (Patton, 2015). For this research, the first step after the interviews was to go through the recordings to make sure everything was recorded. This step was regarded as an important step since if the recordings would not work the notes made during the interview could be checked and more notes could be added of what was remembered. If there would be any uncertainties, it was of importance to get in contact with the interviewee as fast as possible after the interview was finished to complete the notes and receive clarification. After the recordings were checked, the interview was evaluated to determine if all predefined questions were asked and if anything could be improved until the next interview. According to Patton (2015), evaluating the interviews are critical to make sure that the data collected is useful in order to answer the research question.

The interviewees were informed about how anonymity and confidentiality was considered before the interview began. When the data collected from the interviews was used in the empirical results, the interviewees roles and the industry they worked within were mentioned. This was regarded as important in order for the reader to understand the context. However, the interviewees name and organizational name were not mentioned. These measures were taken to protect the interviewee and its organization and not risk publishing any sensitive or secret information about the interviewee or its organization. Anonymization is usually used in order to protect the informant, which means that the researcher uses pseudonyms instead of facts such as name or organization that can be connected to an individual (Bhattacherjee, 2012; Wiles, 2012). However, complete anonymization was not considered as an option for this research since the researchers of this thesis know the identity of the interviewees. Instead, confidentiality was used. Confidentiality is used when it is not possible to ensure anonymity, for example in interviews when the interviewer and interviewee meets or in other ways sees each other, and means that the researcher can identify the data from an informant, but the reader cannot (Bhattacherjee, 2012).

During the transcriptions, ICR was used entailing that both researchers coded the transcriptions individually and then later compared them with each other to ensure that the codes were perceived the same way which enabled discussion of conflicting interpretations and increased the dialogue within the research team. O'Connor and Joffe (2020) explains that ICR can be used in the transcription process as a method to ensure high credibility and Bhattacherjee (2012) describes that to increase the dependability of research, it is of value if researchers investigating the same thing independently reach the same conclusion. Hence, using ICR increases the dependability of this research. Using ICR and transcribing the interviews directly from the recordings also increases the credibility of this research since it increases the likelihood of documenting the data in the most accurate way possible. Credibility is mentioned by Bhattacherjee (2012) and can be improved by having word by word transcriptions and having correct records of the interviews.

4 Empirical Results

The empirical results present the result of the empirical investigation. The interviewees views and reflections on Data Science are described, including views on value creation in Data Science, use areas, the data scientist and Data Science expectations. Furthermore, a description of the interviewees' perception of challenges in Data Science and what challenges the interviewees have experienced regarding data quality, ethics, knowledge and organizational support are described. Thereafter, the empirical results are finalized with a presentation of what actions the interviewees regard as important for succeeding in Data Science and who is responsible for taking these actions. The empirical results are concluded by an empirical summary, summing up the interviewees answers regarding each theme and concept.

4.1 Data Science

What was clarified during the empirical investigation is that Data Science is a concept that includes a lot of things. According to interviewee D, many organizations see companies such as Google, invest in Data Science and how they are doing interesting things with it, but in reality, it is not always that easy (D.6). Interviewee E describes how Data Science is presented as a magic box that accomplishes revolutionary things but that there are big gaps between this perception and the reality of Data Science (E.12). There are also many different definitions of Data Science but whether it is called Data Science, AI, logistics optimization, risk analysis or something else does not really matter as long as the data is being managed (E.14). This statement also reflects other interviewee responses and what they include in the term Data Science. For example, interviewee B and C view Data Science as tightly connected to machine learning, prescriptive analysis and text- and image Analysis (B.2; C.17) whilst the interviewee F states Data Science is more towards Business Intelligence (F.2). Interviewee C states that even though they are not working with Business Intelligence, it might still happen that they create dashboards to visualize their solution but that is not the biggest focus (C.19).

However, all interviewees state that they use data in their daily work and according to interviewee C, 90 percent of the data is structured data that is found in relational databases (C.4). What type of data the interviewees are managing varies however depending on the industry they are working in. Interviewee A who works within the finance industry, is exclusively handling risk-related data such as data about credit risks (A.4). Interviewee E works at an IT consulting company and works with all kinds of data that is related to a customer-specific problem and thus varies. However, interviewee E states that much of the data they manage is sensor data from IoT devices along with data from ERP systems (E.8) and interviewee D primarily uses large amounts of data gathered from the supply chain (D.2). Interviewee F believes that regardless of the industry, the most common type of data to manage is sales data:

> My experiences relate to this kind of data, starting with sales data and then financial data and stock data. After that, it can be a little anything (F.4).

4.1.1 Value Creation in Data Science

It is important for interviewee A to value data in order to increase the quality and make the systems faster (A.8). Companies may invest in advanced platforms and systems that are used only to create nice graphs but without delivering any value, according to interviewee F (F.6). The organization needs to have an expressed need interviewee F continues, in order to make the data valuable and to know what answers they are looking for (F.6). Looking at what data is available and what can be done to solve the problem is the first thing that needs to be done since otherwise it otherwise gets difficult to link the data to value, according to interviewee C (C.10). Adding new data does not mean that you will get more information and it is important to know what new information the added data may offer, as interviewee B describes (B.6). However, interviewee A states that it is important for them to make the data valuable and manage it correctly since if they do not, there is a high risk that the company will lose a lot of money (A.8). Hard work and time is spent trying to understand, format and clear the data into something that can be used, as described by interviewee D (D.2).

4.1.2 Data Science Use Areas

The empirical results show that Data Science has different use areas. Interviewee B claims Data Science to be used in everything from sales and advertising to helping a customer find the perfect product (B.10). According to interviewee D Data Science is advantageous when there is a need for automation or when there is an automated solution that could replace manual work (D.2). For example, the company that interviewee C works for uses Data Science for financial services, to better deal with financial risks or to ensure that the pay system is designed fairly for employees (C.14). Interviewee D also claims that Data Science is used to solve pure optimization problems, such as what price should be put on a product in order to sell as much as possible or to minimize the costs (D.2).

One of the main use areas of Data Science that has been identified during the empirical investigation is different types of text-and image analysis and recognition and according to interviewee B, Data Science is also connected to many other fields such as face recognition and voice recognition (B.6). Both interviewee B and C claim that their work with Data Science includes text analysis (B.2, C.4). Interviewee B refers to this by stating their use of methods to translate text to speech (B.2) and interviewee C explains about the use of text analysis to search and classify internal documents (C.4). This is used by interviewee C to for example detect faults in machines and to detect errors automatically in the production area (C.4). Also, interviewee E describes the use of image analysis and sensor data from IoT (Internet of Things) devices (E.8). These kinds of techniques can also be used to set up recommendation engines where user data is used to connect a customer or a user to a specific item, something that interviewee B has experienced and witnessed (B.2; B.10).

Another use area for Data Science that has been identified during the empirical investigation is to forecast and find patterns. Interviewee C uses Data Science to better forecast the behaviour of customers and its consumption, such as how will the customer act or react to this packaging, what customers buy from them regularly and what customers they risk losing

(C.12). To forecast and anticipate sales is a use area for interviewee B and D too (B.14; D.12). Interviewee B states that Data Science is used to calculate the probability that a customer buys a product it just has clicked on and, if they are consulting for a real estate company, Data Science is used to calculate what price should be put and how the house should be marketed (B.14). Interviewee C also states that Data Science is used in the logistics and manufacturing areas to better control the stocks and ensure that they can deliver if someone places an order (C.12). The ability to predict maintenance is something interviewee E also mentions as a use area of Data Science (E.20). Based on data from past projects, the company that interviewee C works for, also uses Data Science to better anticipate the probability of winning new projects and what employees that risk quitting (C.14).

4.1.3 The Data Scientist

It might be complex to specify what the exact tasks are for a data scientist according to interviewee B (B.22). Interviewee C states that they work as internal consultants, with the exception that they do not charge for these internal services, and they usually work on four to five different projects simultaneously (C.2). Both interviewee A, B and F describe that a big part of their work is to process and structure the data they use. This includes answering questions such as if the data is dated correctly, if the data can be relied on, if the right granularity exists and how key A relates to key B, as explained by interviewee A (A.2). To further demonstrate this, interviewee B states:

> [...] first you have to go in and see what data there is, and what data you would like to have. And then you have to decide for a project, which project is somewhere at the intersection, between being technically feasible and that it adds value to the company (B.6).

Data scientists also have to collect the data and ensure that the data is available for the project as interviewee B describes (B.6; B.10). Interviewee F also states that data scientists help in analyzing the data and visualizing the data for the customers (F.2). According to interviewee A and C, the data scientists might also be the ones that produce code (A.2), look at code requirements or review code (C.2).

Another substantial part in the role of a data scientist, is educating and advising customers or users. Interviewee A explains that a large part of their work is to explain the data and help colleagues and customers to understand what the area entails (A.2). This is something interviewee D agrees upon and describes that one mission is to educate about what Data Science is and what it is for, such as what is possible to measure, what does it mean to build models or how to predict something with Data Science (D.6). Interviewee E claims that about half of their time is spent on coaching new employees, making customer presentations, talking at conferences and educating their customers (E.4). This kind of advisory or educational position can for example concern what kind of data will be needed for the project, what benefits Data Science integration will lead to or if the project should be built from scratch or if it should be built on something that already exists, according to interviewee B (B.6). Interviewee C continues by stating that loads of their time is spent on discussing with internal customers about what their decision-making looks like and in what way they expect to use the results of the products (C.2). According to interviewee D, there are many people within their organization that send requests and ask if their problems can be solved by using Data Science (D.2). It is important that the data scientists work to give their customers an understanding of what they

can do and what they cannot do with Data Science (E.10). Data scientists also have the responsibility to monitor the results of the decisions made, measure if the expected value is reached, and deal with problems that might have arisen as interviewee (C.2) Commonly this is about the economic impact of decisions made, such as how a decision impact the balance sheet (C.2).

During the interviews, it has been described that a data scientist usually does not work alone but in a team. Interviewee F states that a single data scientist will never be able to do the whole job and that there must be a mix between different parts of the business to be able to understand the needs (F.10). Interviewee A works in a team with about seven to nine people where about 80 percent are data scientists and the others are managers, programmers or some form of specialists (A.16). The team that interviewee B is included in, represents the whole business and are divided into different areas, such as frontend, backend, AI, Machine Learning, sales, marketing and management (B.3). To have people from the management group represented in the Data Science team is good according to interviewee F, and it is important that the team both has domain knowledge and data knowledge (F.30). According to interviewee B and D, data scientists often collaborate with other resources, such as data engineers and solution architects (D.2) but that they often are responsible for specific parts of the project (B.3). Making this division is sometimes difficult and for example knowing what a task for a data scientist is and what a task for a data engineer is, as interviewee B explains (B.22). Moreover, interviewee A has good experiences of collaborating with other roles and employees (A.18):

> [...] when I do these types of teamworks I would say that I think everything works very well. Everyone is always curious and positive about each other's work, everyone always respects each other and so on. So actually, I only have good experiences with this (A.18).

4.1.4 Data Science Expectations

According to the empirical results, there seems to be no real definition of what Data Science entails or what is included in the term which leads to varying expectations of what you can do with Data Science. The expectations are often high according to interviewee D, and it is often difficult for business leaders to know what to expect from these investments (D.6). Also, interviewee B claims customers have unbelievably high goals and expectations on how Data Science will work (B.34). All interviewees agree that the expectations that their organization has on Data Science usually are not achieved. Interviewee C states that organizations always think that the data is good and in perfect condition, which is often not the case (C.35). Interviewee C explains that organizations tend to think that a solution to the problems they have are found by inserting numbers and many do not understand the complexity of Data Science (C.35). Interviewee E mentions that many organizations view Data Science as popular to have and that many big companies are investing in it, which might lead to too high expectations since their own organization might not have the same resources and capabilities (E.12). Also, interviewee F states that organizations usually have too high expectations on the data scientists and what experiences a good data scientist needs to have, expectations that no one lives up to in reality (F.28).

However, interviewee D states that the expectations on Data Science has become more realistic during the years since more experience exists of what you can do and what you cannot do

(D.17). When an organization has had Data Science integrated for some time, both organizations and the data scientist realizes that the projects take both more time and are more complex than they first had thought (D.17). Interviewee F states that the expectations they themself have on Data Science are usually achieved but that the term and field is too wide to be able to say if the expectations are achieved or not (F.28). Interviewee F continues by stating that this is because two data scientists might have completely different skills and experiences but both can still be called a data scientist (F.28).

4.2 Challenges in Data Science

There are several challenges to be found in Data Science according to all interviewees, even though there are varying opinions to what kind of challenges occur. For instance, interviewee A emphasizes main challenges in Data Science as there being poor documentation and that much information is solely dependent on an individual, and explains the fragility of having many employees rely on a bunch of people for gathering valid information (A.2; A.6). Also, Interviewee D mentions several problems and highlights that the major challenge is a lack of people who have the knowledge and experience to deal with Data Science (D.8). Interviewee C focuses on the challenge of not knowing what data is needed and how to use it and adds that one of the major problems is accessing data and receiving the data in the desired format (C.2; C.23). This is also addressed by interviewee E who states that there are many challenges related to not having any data and that the data lacks a connection with the rest of the business (E.20). The challenge of last mile delivery is something described by interviewee C, namely that even though there is a fully developed and finished product, a challenge is making the user understand the value of what they have done and understand the weaknesses of it (C.23). Interviewee B claims that there are many challenges related to Data Science (B.36) whereas interviewee F does not mention any overall challenge related to Data Science specifically but states that there are several problems that emerge constantly (F.14).

4.2.1 Data Quality

All interviewees mention challenges in Data Science related to data quality. Interviewee C explains that the quality of the data is always an issue and that there simply are no good data sources, the data always needs to be cleaned (C.29). Additionally, interviewee D emphasizes that data quality is a considerable problem but that it mostly concerns the structured data already existing in the system (D.10). Sometimes when working with data in projects, interviewee F explains that the data is too poor and will make the Data Science project fail (F.20). Interviewee A exemplifies this problem by explaining that mistakes are made since different countries input data to the data fields, data manipulation occurrence, empty fields and data points and inconsistent terminology used in different databases which makes it hard to find connections (A.26). Similar problems are stated by interviewee B, such as when there are inconsistent ways of filling in the data in excel sheets or in databases (B.38).

Interviewee D continues by stating that a more substantial challenge related to data quality is when there is not enough data to work with which results in gathering new data where new types of questions emerge, regarding for instance if the source can be trusted (D.10). This topic

is also approached by interviewee E that states the example of a doctor of mathematics employed by a company to work wonders with their data, but the problem was that the company barely had started to generate any data, which made it hard for this employee to do anything (E.20). The maturity of companies in such matters is something that interviewee E explains as varying (E.20). Interviewee D also mentions that data quality is a major part in the processing of data, and although data is the new oil, many organizations collect data not knowing what to do with it (D.10). Although the data is of low quality and not clean enough, interviewee B emphasizes there must be an awareness of the value you can find in that data in order to know what data to prioritize (B.40). As interviewee B states:

> [...] it is a problem that the data is not clean enough, [...] but it is also somewhere that you have to be aware where you can bring some kind of value [...], what to keep, what to throw (B.40).

Both interviewee A, B and C state the activities in managing challenges that are related to data quality. According to interviewee B, since the data can be more or less structured, one typical task that falls upon the data scientist is to make the data structured and ensure that the data is in the correct form in order to be useful (B.38; B.40). As interviewee C describes, about 70 percent of the data scientists job tasks entail washing and transforming the format of the data, and that, together with understanding the data, consists of 90 percent of the workload (C.29). All of interviewee A's time is spent on working with the data and that data quality issues lead to major delays and rerouting of job tasks (A.22; A.26). However, although interviewee F also witnesses challenges related to data quality, the production of data is far better today than it was 10 to 15 years ago due to that systems are being used to input data instead of using manual inputs via excel (F.20).

4.2.2 Ethics

Interviewee B, C, E and F acknowledge challenges regarding ethics in Data Science. Challenges have been seen by interviewee A regarding wrong data but not about ethics in that sense (A.26) and interviewee D did not answer this question. Interviewee B explains that in situations where there is too much confidence in what is to be solved, ethics can be a problem and states that the inaccuracy comes from having visions that are too big without any realistic goals (B.46). Ethics is also a problem according to interviewee C and explains that it depends on what data is managed and what decisions are based on that data (C.29). Ethics is something that is considered by interviewee E but depending on the situation it is not always easy or obvious what choice to make (E.22). Ethical issues are also explained by interviewee F since the task of a data scientist is to look at the numbers that the data shows, but realizations emerge of that the data is sensitive emerge at a later stage when discussing the needs and results (F.22), for instance:

> [...] You realize that every line in this data set is a natural person who visited a health center, it is quite sensitive (F.22).

The nature of the ethical issues that may occur in Data Science is described by interviewee E. An example is interviewee E's work with image recognition through camera streams with construction companies where the goal was to identify people with helmets in dangerous places at construction sites (E.22). However, interviewee E realized that the model based on the data

used was from the construction workers and all these people were white, middle-aged men (E.22). Hence, if someone not matching that description would be in a dangerous place, the risk is they would not be identified and solving these types of issues is difficult (E.22). Interviewee F also gives the example of when working with data the realization is reached that this is not a machine that has randomized some numbers, but the data symbolizes physical people with illnesses for instance, and it is important to use it with caution and dehumanize it (F.22).

There are no easy answers to these types of ethical questions and reasoning is needed according to interviewee E (E.22). Interviewee B states the importance of solving these types of issues and mentions the use of mathematical methods to deal with bias in order to generate a fair result (B.46). It is important having bias in mind when selecting what data to use in the first place as described by interviewee B (B.46). Interviewee A explains that their team and the organization they belong to treat data with respect and measures are taken to not view data belonging to colleagues, family members and public figures and not share the data with anyone (A.26). Additionally, security and ethics of data are tasks that fall upon the data scientist, as interviewee B explains (B.22), however, interviewee F usually ignores what the data describes and focuses on the numbers that the data shows (F.22).

4.2.3 Knowledge

Interviewees A, C, D, E and F all explain that gaps in knowledge have been encountered in Data Science. However, interviewee B did not answer this question.

A common knowledge gap seems to be that there is lacking competence amongst data scientists. Interviewee C explains that it is a fundamental problem not getting enough competent data scientists since it takes time to build up the skills needed (C.23). Interviewee D continues by describing that although data scientists are good with data, understanding is lacking in what they are working with, how the data is collected and in what context (D.10). What usually is lacking according to interviewee A is programming skills (A.27), computer knowledge and not having the right tools and skills to use modeling tools (A.12). . Interviewee D mentions this also:

> So I would say that traditionally there are knowledge gaps, there are more people who are good analysts than those who can code. (D.14).

Interviewee C explains that data scientists need skills related to business (C.23). Interviewee D mentions that software developers often lack skills relating to math and statistics whereas people from the business side are good analysts but lack a software mindset (D.14). This is corroborated by interviewee E who explains that there is a knowledge gap that exists from both sides which is unfortunate (E.12; E.32). For instance, talented mathematicians working with probability techniques do the same work that is conducted in Data Science, but realization is lacking of this as well as data scientists believing their techniques to be best and not realizing that their work tasks are the same as the mathematicians, is problematic according to interviewee E (E.12; E.32). Interviewee F mentions that a good data scientist is often expected to have ten years of statistical experience, knowledge of Machine Learning models and have worked with business analysis for 30 years, but there are none that live up to that specification (F.28).

Interviewees A and F also focus on aspects regarding skills in project management and communication. Interviewee A describes how they themselves need to step in and ask for completion of documentation of code and ask senior programmers to comment more in the codes to make them understandable (A.27). Interviewee A claims difficulties in lifting their eyes since not everyone wants to share and explains bad communication as a problem (A.27). Project management in Data Science is lacking according to interviewee F (F.24). Interviewee F claims that they seldom are involved in projects with project management professionals and is something that businesses are quick at downsizing in budgets (F.24). Interviewee F also states:

> You probably underestimate quality or how it could have been with professional management (F.24).

4.2.4 Organizational Support

Support from the organization and related aspects is mentioned as a challenge according to interviewee A, C, D, E and F whereas interviewee B has not answered this question.

A challenge met with organizational support is how mature the company is according to interviewee E (E.20). The work with data may take longer than expected which causes problems with management since they want to see results as described by interviewee A (A.34). Interviewee D describes that the Data Science team does not always create useful data products in a continuous way and explains that there is not a linear development (D.12). This in turn makes it difficult sometimes to manage the expectations of the work data scientists are expected to do and leads to a mismatch between the calculations showing the return of investments before and after employing data scientists (D.12).

Moreover, interviewee C, E and F mentions the culture of the company where interviewee F states that Data Science should be a supporting role in an organization (F.30). Interviewee C explains that the Data Science division of their company is seen as a peripheral part of the organization rather than the central part of it which interviewee C states as a strategic problem, in contrast to interviewee F, and more attention and focus should be put on what they are doing (C.25), For instance, their Data Science team has chosen to follow the CFO, since the CFO is influential, employees listen to what that person has to say and this person also works with numbers and facts (C.27). Interviewee F witness situations where the culture is not healthy where for instance there have been too strict hierarchical divisions of responsibilities and closed doors (F.16). The consequences of having such a culture and climate is elaborated on by interviewee F:

> If you have such a climate, it is really hard to understand what the need is if there even is one so to say. And it is difficult to get ahead, it is difficult to find key people within the company who can answer specific questions if one is not allowed to ask the question (F.16).

Interviewee C and E also describe resistance in companies. When working at companies that have advanced NASA techniques and state-of-the-art engineering methods for instance, it is hard to show that Data Science will do magic with these things without undercutting how advanced their techniques are, interviewee E mentions (E.20). Also, much resistance is met

when trying to transform an organization into more data and fact oriented which is a difficult task according to interviewee C (C.27).

4.3 Data Science Success

All interviewee mention that there are many different aspects impacting the success of Data Science which according to interviewee B can be described as the following:

> The concrete success factor is to think of two circles, one is technically feasible, the other to create value and interest. [...] these two circles usually have a small intersection, where there is an area where it is technically feasible and that there is a general interest and financial incentive. Try to find that interface (B.48).

However, if reaching success in Data Science was an easy task, it would have been accomplished already according to interviewee D, explaining that few projects succeed (D.25). Success in Data Science is not easy and a solution is not reached only by inputting numbers, which many expect according to interviewee C (C.35). Interviewee B also describes that generating results in Data Science is more complicated than spending a budget and monitoring results (B.26). A difficulty with determining its success is because it is a creative project with many roles (B.26) and according to interviewee E, Data Science is not an exact science (E.24).

Hence, companies have different levels of maturity in Data Science according to interviewee B (B.26). Some companies believe that Data Science only is about computer science, but interviewee F mentions that these companies are only in the beginning of their data journey (F.35). According to interviewee E, Data Science is an unclear concept since it can be changed and include different fields and as long as the focus is on the actual definition of Data Science, advances are not made (E.14). Data Science was already being conducted 20-30 years ago according to interviewee F, but it is rebranded in a modernization effort to fit machine learning into the definition since that is something new (F.35). Whether Data Science is called Data Science, logistics optimization or risk analysis or not does not matter according to interviewee E, as long as companies implementing it become better at using and working with their data and using probability-based mathematical techniques (E.24).

4.3.1 Actions for Data Science Success

Several actions for Data Science success have been mentioned by all interviewees where interviewee B and C relate Data Science success to how value can be generated. Interviewee B explains actions for success in Data Science as:

> I think that success in Data Science is, thinking it through, what are the problems we can solve, what is the value we can get from it and then based on that, building some kind of model, which can then be built into a app, which can then be selected on a phone as well as generate a value (B.50).

Interviewee B continues by explaining that the value can be generated differently depending on how it is perceived, for example it can be making money (B.50), but this requires that the technology is mature enough (B.26). One key to success according to interviewee C is linking the result to value and measuring the effect or the effect of the decisions made from this value (C.2). Decision making is focused on by interviewee E by emphasizing the importance of data quality during data washing and processing, and having the ability to make good decisions and assumptions from the data (E.24). Moreover, this depends on whether the data is collected correctly, which interviewee C mentions as a success factor by stating the importance of being purposeful when starting data collection (C.31). For instance, insoluble data examples may emerge where there is no possibility to predict the data or the data might not even exist (C.31). However, according to interviewee F, the data is secondary and can always be found somewhere (F.30).

Interviewee F focuses on success being attributed to clear needs specification, clear mix in the project teams between domain- and data knowledge and having good project management (F.26). Interviewee C states that having a problem and formulating it in a clear way makes it more likely to succeed with a project (C.31). The importance of having realistic milestones and goals is also mentioned by interviewee B, who claims that showing a small result is better than nothing (B.36). Interviewee D states a similar aspect as not trying to solve all problems simultaneously and explains that it takes time setting up a team with enough experience and thirst to learn (D.16). As a help on the way, interviewee D explains that finding someone that has experienced the same problems before is a good place to start since a problem is that many people do not have the right expertise nor are able to lead (D.16). Interviewee F mentions consultants as a good example for solving problems since they do not have any history with the company and explains it as the following:

> It is easier as an outsider to be able to ask stupid and hard questions and be able to question decisions than someone internally does (F.18).

Consultants can help companies with drawing out frameworks and formulate the needs and goals of a project which may be helpful according to interviewee F (F.18). Interviewee D describes a way to solve problems and share knowledge of these is by meeting people with similar roles that have had experience with similar problems but in a different context and maturity (D.25). However, interviewee D continues by stating that there is no easy way to fix a problem since it must be applied in the organization and its context (D.25). Hence, interviewee C and E emphasize the importance of communication to reach success in Data Science. There is no universal way of reaching success according to interviewee E, but a good first step is to have conversations that are disconnected from the hype and sales discussion surrounding Data Science (E.14). Also, interviewee C explains that having clear communication with the customer is a success factor since it is important to ensure that the customer understands the results of the delivery and problems that may follow, otherwise they may not use it (C.31). The communicative ability of the data scientist is also described by interviewee E as important to reach success, to be able to explain to a CEO how things work in an understandable way (E.18). Another important aspect in reaching success mentioned by interviewee A is being active with what needs to be learned, seeing things from a bigger perspective, and constantly showing that you are good since this is otherwise taken for granted or forgotten (A.36). Interviewee A claims that for the data scientists to be seen and receive higher pay and more responsibility, it is important to succeed with Data science (A.36). In addition, having a representative from the

Data Science team in management and steering groups is vital according to interviewee F as a way to help prepare and plan for upcoming changes in the organization (F.30).

4.3.2 The Organizational Responsibility

There are various aspects that all interviewees mention that fall under the responsibility of the organization. To succeed with Data Science, interviewee B explains that projects need to be technically feasible and that there is a general interest and financial incentive (B.48). An organizational challenge is not having an ultimate goal with Data Science according to interviewee E (E.26). It is also important to have a clear picture of what the data scientist's role is in the company and the company's maturity level of Data Science as mentioned by interviewee F (F.30).

Interviewee B, D and E focus on the Data Science team and data scientists. A good mix of employees in the data scientist teams that complement each other is valued by interviewee D, and that they are allowed to implement and test things, which requires support from the top (D.29). Before investments are made in Data Science projects, interviewee B stresses that awareness is needed about analyzing the problem itself and not applying quick fixes because those will in the end work poorly (B.50). For example, hiring more data scientists to solve a problem is not the solution, according to interviewee B (B.50). Even though you have the best team of data scientists, the best tools, knowledge and easy problems to solve, if the predictive power of the data does not exist, it does not matter according to interviewee E (E.26).

Other aspects mentioned regarding organizational responsibility are mentioned by interviewee A, C, D, E and F and relate to management. A measure for reaching success in Data Science according to interviewee C is regarding resource allocation and how the company sets their budget (C.37). For instance, if the company sets the budgets in the autumn, the whole next year relies on that budget which reduces the possibility of receiving new resources and in this sense, interviewee C stresses the difficulty of an old traditional organization in becoming agile (C.37). Interviewee E focuses on the agility and iterative work that Data Science entails and results generated from Data Science need to be explained to top level managers in an understandable way (E.25). It is required for management to also come with demands regarding delivery times and interviewee D values short projects with light profits since results can be shown quickly (D.29). The importance of management support is explained by interviewee D in the following way:

> So if you have a manager who thinks that this with Machine Learning and Data Science is just rubbish, it will never work [...] (D.29).

Interviewee D also describes issues between the organization and the Data Science team with other employees expecting the data scientists to automate all their jobs, which is not the case (D.29). Hence, for Data Science to work, there must be ownership and managerial responsibility constantly overseeing the work over time (E.28). Interviewee F also values good project management of Data Science (F.24). Interviewee A believes that the organization has the main responsibility of leading Data Science to success but that the responsibility also lies upon the data scientists themselves (A.40).

4.3.3 The Responsibility of the Data Scientist

All interviewees mention the responsibility of the data scientist when it comes to reaching success in Data Science. As mentioned previously by interviewee A, the responsibility of reaching success in Data Science depends partly on the organization, but also on the data scientists (A.40). However, since data scientists have many other work tasks to complete, it is difficult for them to give suggestions and solutions to problems (A.40).

The skills data scientists and its team are responsible for having and upholding are described by interviewee B, C, D, E and F. Interviewee F describes the challenge of finding good people with the skills needed to be a good data scientist (F.10). For instance, technical understanding and skills are needed but it does not have to be software or platform specific and there needs to be understanding of how and in what ways data can be stored (F.10). The difficulty of finding data scientists with the right mix is stated by interviewee F and explains that their company did not grow or move forward due to that difficulty (F.10). Interviewee E also values data scientists with a breadth in skills and explains that it is important to have people that are very technical, systems scientists that understand how things actually work in a business and economists for understanding the society and context in which things will be applied (E.18). There are some fundamental areas that interviewee E expects that a data scientist should know, for instance, maximum likelihood optimization and know what a vector is (E.18).

Hence, many skills are focused on by the interviewees B, C, D, E and F which the data scientists must ensure that they have. Interviewee C and D focus mainly on technical skills. Interviewee C explains that skills needed are basic programming knowledge, being able to program production code, handling version management, understanding Machine Learning, good understanding of statistics and understanding the variations of distributions that exist (C.17). Also, interviewee D describes skills needed in math, statistics, programming and understanding of the business operations (D.2). Skills that interviewee B explains are important for data scientists to have are to be flexible, varied and be willing to learn (B.26). Interviewees E and F focus on the skill of communication. Interviewee E mentions communicative ability, to be able to describe to someone in a management position what has been done, how it works and describe this in an easily understandable way (E.18). This is also mentioned by interviewee F who describes communication, being able to listen to other opinions and formulating a need for the data as skills imperative to reach success (F.10).

Furthermore, interviewee B, C and E focus on other aspects related to the data scientist and its team's responsibility in reaching success in Data Science. Interviewee C states the responsibility of being better integrated with data engineers for Data Science to work and the value of a closely integrated team of data engineers, data scientists and business analysts is emphasized (C.39). The importance of collaborating with other roles and using traditional techniques is mentioned by interviewee E and explains this collaboration is not currently being realized but it should be (E.32). However, interviewee B states that it is difficult to distinguish between what a task for a data scientist is and what a task for a data engineer is, therefore it is important to realize what problems a data scientist will be able to solve (B.20). Hence, according to interviewee B, being a good data scientist and reaching success in Data Science is based on having visions and being able to implement them (B.30).

4.4 Empirical Summary

The Empirical Summary aims to sum up the empirical results connected to the central concepts shown in *2.4 Literature Summary* and that have been coded using the Coding Scheme (Table 3:3) in the Research Methodology (see Chapter 3).

Data Science is an unclear concept that may include several aspects, the definition varies depending on interviewee, D and E have similar definitions, as well as B and C, whilst F for instance has another take on it (see Chapter 4.1). *Value creation in Data Science* is important according to interviewees A, B, C, D and F but not focused on by interviewee E. Data is used in Data Science according to all interviewees in their daily work but the kind of data varies from risk-related data, customer-specific data, IoT sensor data, ERP system data, sales data, financial data, supply chain data, data for text analysis and image-, voice-, and face recognition (see Chapter 4.1.1). Data Science has many *use areas* according to the interviewees. Data Science can be used in sales, advertising, manufacturing, logistics, automation, optimization, financial services, predictive maintenance, recommendation engines, text- and image analysis and face- and voice recognition (see Chapter 4.1.2). The tasks of the data scientist vary as described in *The Data Scientist*. Tasks mentioned are internal consulting, accessing, collecting and structuring data, advising and educating users and customers, and monitoring results of decisions made based on the data. Interviewees A, B, D and F also describe the team that the data scientist works within, skills that are required and the tasks to be completed (see Chapter 4.1.3). *The Expectations on Data Science* are by all interviewees stated as high and are usually not achieved (see Chapter 4.1.4).

According to all interviewees, there are several *Challenges in Data Science*. General challenges mentioned relate to poor documentation, lacking knowledge, data accessibility, setting realistic goals and milestones and users not understanding how to use the finished product (see Chapter 4.2). All interviewees explain challenges that exist related to *Data Quality* such as empty fields and data points, data inconsistency, poor data sources, unclean data and not having enough data (see Chapter 4.2.1). Interviewee B, C, E and F acknowledge challenges regarding *Ethics*, interviewee D did not elaborate on this and interviewee A understands there are challenges regarding wrong data but not ethics in that sense. Ethical issues may concern for instance sensitive data and biased data collection but there are no easy answers to solve this, although, reasoning and using mathematical methods eliminating bias may help (see Chapter 4.2.2). Interviewees A, C, D, E and F encountered challenges related to *Knowledge*. Challenges are the knowledge gap between the roles in Data Science and the data scientists' lacking knowledge, competence and skills related to business, mathematics, statistics, software, project management and communication (see Chapter 4.2.3). *Organizational Support* is mentioned as a challenge by interviewees A, C, D, E and F where aspects such as companies' resistance, management, culture, support, maturity and expectations on Data Science are stated (see Chapter 4.2.4).

All interviewees mention aspects impacting the *success of Data Science*. The interviewees thoughts on reaching success in Data Science relate for instance to the unclear term of data science, difficulties determining its success and the complications of getting a result (see Chapter 4.3). *Actions for Data Science Success* are elaborated on by all interviewees and examples of actions are clear goals, needs specification and problem formulation, generating value from the data, data predictability, maturity of technology, clear communication, having a data scientist in managerial groups and having a mix of skills in the Data Science team. External

parties are explained by interviewee D and F as valuable assistance (see Chapter 4.3.1). All interviewees explain *the organizational responsibility* and its part in Data Science Success. Aspects mentioned are that there needs to be an interest, goals, financial incentives, clear picture of the data scientists' role, maturity of Data Science, mix of skills within the Data Science team, the predictability of the data, agile and iterative work, ownership and responsibility of management (see Chapter 4.3.2) *The Responsibility of the Data Scientist* is mentioned by all interviewees also. Data scientists are responsible for having a vision and have technical, business and communicative skills and there needs to be a mix of skills in the Data Science team as well as integration between the different roles (see Chapter 4.3.3).

5 Discussion

In the following discussion, the literature presented in the literature review and the empirical result are concluded. As in the previous chapters, this chapter commences with a presentation of Data Science, value creation in Data Science, the use areas for Data Science, the role of the data scientist and what expectations there are on Data Science. The chapter will continue discussing the challenges with Data Science, including data quality, ethics, knowledge and organizational support. The chapter ends with a discussion of Data Science success, what actions can be made to reach success and what responsibility the organization and the data scientist has on taking these actions.

5.1 Data Science

Both the empirical results and the literature review showed that Data Science is a broad and vague term that includes many areas, fields that may both be conflicting and confusing. The empirical results show that Data Science consists of varying terms from Machine Learning and text- and image analysis to Business Intelligence, which is also mentioned as fields interfused with Data Science by Kotu and Deshpande (2019). Moreover, Sammut and Webb (2017) states different techniques are often used in Data Science, which can explain the above-mentioned differing interpretations of Data Science. The empirical results indicate that many companies invest in Data Science due to success stories of famous companies and aim to achieve the same result, in reality however, it is not as easy. This might be why Kotu and Deshpande (2019) claim that Data Science has increased in popularity and many organizations now consider Data Science as a vital resource.

Furthermore, OECD (2008) states that the data most commonly are of numerical characteristics and are collected from observations, however, when the interviewees were asked what data they use in their daily work, examples such as risk-related data, data related to customer-specific problems, sales data and sensor data from IoT-devices were mentioned. Since it is hard to determine what is included in the terms numerical characteristics and observations that OECD (2008) use, it is possible that the data provided by the interviewees may fall under those definitions. For instance, the interviewees might use data of numerical characteristics, even though they did not focus on this in the empirical results. Also, Silverman (2011) states that data may be collected from interviews and conferences which is something that does not correlate with any of the statements in the empirical results. This may be due to the fact that data scientists primarily work with types of data that are suggested by the interviewees, namely data related to company-specific activities for example sales data and customer data, which usually are not gathered through interviews and conferences.

To conclude, the empirical results suggest extensive use of data in Data Science. However, the types of data vary depending on industry, and the literature has several varying definitions of data which do not completely correlate with the definitions provided by the interviewees. What is important to remember is that Data Science is a field that is still relatively novel (Cao, 2018) which indicates the possibility of Data Science as a concept might become clearer or more well-defined in the future. However, since Data Science is not yet a clear concept, which is shown

both in the literature review and empirical results, an advantage may be to identify its use areas in order to ensure that it is integrated correctly and for the right purpose, and that resources exist to facilitate it.

5.1.1 Value Creation in Data Science

In order to make data valuable, both the empirical results and the literature review address the importance of having an expressed need, to know what the data might offer and how it may help the organization. The empirical results indicate that many believe that the more data you have the better results you get, and that companies tend to invest in platforms and systems with high-end techniques that can create advanced graphs but do not offer any value for the organization. Hence, if companies believe the above-mentioned, the competitive advantage that correct data management may result in (Treder, 2019), and the social and economic gain from creating value from data (Hoffman, 207), may be lost. The empirical results also show this explaining that by not managing data correctly nor making it valuable leads to the risk of companies witnessing economic losses. Since data is the foundation of work in Data Science, it is vital that value can be created from it. Companies have difficulties in using their data correctly and the potential of data if often not fulfilled (Berntsson Svensson, Feldt & Torkar, 2019) which is something the empirical results also emphasize. Hence, the empirical results and literature review show the importance of companies ensuring their capabilities of generating value from the data before investments in Data Science are made.

Conclusively, both the empirical result and the literature review explain the importance of generating value from the data used in Data Science. If companies do not have the capability of generating meaningful data and using it correctly, they may witness losses.

5.1.2 Data Science Use Areas

What has been witnessed both in the literature review and the empirical results is that the use areas of Data Science are broad. Data Science can be used in everything from manufacturing to advertising (Provost & Fawcett, 2013b) and the empirical results explain that its use ranges from sales and advertising to helping a customer find the right product. As seen in the empirical results Data Science is also used for solving problems in financial services and optimization problems and where automation is needed. Being able to solve all of these issues might be the reason why, as Shreshtha et al. (2019) state, Data Science is being integrated into all industries, which is corroborated in the empirical results. However, all industries are not represented by the interviewees, hence, the empirical results are not able to entirely prove the statement of Shreshtha et al. (2019). Also, many investments are made in Data Science to improve marketing and advertising, forecast market trends, customer relationship management and maximize the customer value (Provost & Fawcett, 2013b) and similar aspects such as to view trends, track behaviours and make forecasts are also indicated in the empirical results.

In the literature review, Shrestha et al., (2019) states that the use areas of Data Science can be addressed as expectation, security, computer vision and natural language processing. Expectation, security and computer vision are fairly recurrent in the empirical results, where Data Science is used to anticipate sales, to forecast behaviours of customers and to manage risk-

handling of data, such as credit risk data (demonstrated in Chapter 4.1 Data Science). However, the last factor, natural language processing, as mentioned by Shrestha et al. (2019) is not stated in the empirical results. Hence, a parallel can be drawn that natural language processing is intended to be used in Data Science as explained in the literature review, but it is not used in practice in accordance with the results from the empirical study.

Conclusively, there are different use areas of Data Science in many industries, which both the literature review and empirical results demonstrate. However, some use areas are not equally focused on or at all in the empirical results compared to the literature review, for instance, the use area of natural language processing is not mentioned by any interviewees. This chapter also clarified that Data Science is useful in many contexts which possibly has contributed to its popularity.

5.1.3 The Data Scientist

The empirical results show that a substantial part of a data scientists' work concerns structuring and processing data, ensuring validity of data and presenting data for users and customers, which correlates with Atwal's (2020) and Saltz and Stanton's (2018) perceptions in the literature review. However, the literature review and the empirical results disagree on several parts. For example, Saltz and Stanton (2018) claim that data archiving is a part of the data scientists tasks but no interviewees have explained this as one of their tasks. The empirical results also claim education and advisory as an important work task for a data scientist, such as explaining data for customers and colleagues and educating others about the possibilities of Data Science. In contrast, the educational aspects are not stated in the same way in the literature review, apart from Saltz and Stanton (2018) that mentions that the data scientist should enable communication with users and have knowledge of data visualization. The empirical results also addressed team work as an important aspect of the data scientists work explaining that a data scientist will never have the ability to do all the work alone. However, the only literature that in some way reflects on this is Atwal (2020), Saltz and Stanton (2018) and Waller and Fawcett (2013) who state that data scientists need to understand the business and application domain. However, that statement does not include working in teams, which is mentioned more frequently in the empirical results.

To summarize, the tasks of a data scientist that commonly occurred in the empirical results, correlates with the tasks identified in the literature review. However, several tasks that the empirical results mention as part of a data scientists' work are not stated in the literature. This may be due to the difficulties of determining what is included in the broad concept of Data Science, as previously stated, which may result in varying and conflicting statements regarding the data scientist's tasks. Nonetheless, it is clear that there is a knowledge gap in the literature in this sense and knowledge lacking between what is relevant in practice versus in theory.

5.1.4 Data Science Expectations

In the literature review, Provost and Fawcett (2013b) states Data Science as sought after in all industries. This is reflected in the empirical results showing that Data Science is popular to invest in by many companies. However, as seen in both the literature review and in the empirical

results, these investments tend to come with high expectations on Data Science and expected results are not always realized.

The empirical results also claim that customers often have unbelievably high expectations and goals on both Data Science as a field but also on the capabilities of the data scientists. Having high expectations of the data scientists' skills is also mentioned in the literature review (Braschler, Stadelmann & Stockinger, 2019; Waller & Fawcett, 2013) and finding the unicorn in Data Science who does possess all desired skills is not possible (Braschler, Stadelmann & Stockinger, 2019). Moreover, the empirical results indicate that the term and field of Data Science is too novel to be able to determine if the expectations are reached or not and the varying expectations may be due to its lack of definition. This is also established in the literature that boundaries of Data Science are difficult to determine (Braschler, Stadelmann & Stockinger, 2019). Hence, the vague definitions and meanings of Data Science might be problematic when setting its expectations and goals and thus might make it difficult for companies to reach the expected outcome of Data Science.

5.2 Challenges in Data Science

There are several challenges in Data Science both according to the literature review and the empirical results. For example, the empirical results suggest there is a lack of people with the right knowledge and experience to manage Data Science, which Braschler, Stadelmann and Stockinger (2019) state is required to be able to conduct risk analyses of Data Science. Challenges stated in the empirical results also relate to not knowing what data is required, how to use it, lack of data, data accessibility and connecting the data to the business. These challenges are not stated in the literature review, apart from Atwal (2020) mentioning there must be knowledge about what data is required, its use and that companies might find it difficult to manage Data Science if the technology is outdated or flawed. However, the empirical results describe difficulties in making the user understand what they want, what value has been created by using Data Science and if there are any weaknesses related to that. Hence, this might make it hard for data scientists to determine if they in fact are managing the right things if clear specifications have not been provided. Unclear task division and lack of organizational support is also mentioned in the literature review (Atwal, 2020).

This chapter shows a variety of different challenges found in Data Science, explained both in the literature review and empirical results. Even though the literature and the empirical results tend to focus on different challenges, it is clear that problems regarding Data Science occur in all areas of the field. However, the field of Data Science is emerging and thus it is possible that more challenges may be shown gradually. As Braschler, Stadelmann and Stockinger (2019) also state it is difficult to address these challenges since Data Science is not deeply integrated in society yet. Although, its maturity and integration may vary in an organizational context, and hence, it may be easier to address challenges in companies that have integrated Data Science on a deeper level.

5.2.1 Data Quality

All interviewees in the empirical results claim challenges related to data quality whilst in the literature review, Atwal (2020) argues that low quality of data is described as a challenge by 55 percent of data scientists. Hence, it is clear that the quality of data used in Data Science might be problematic. Challenges occur due to inconsistency, mistakes or manipulation in data as stated in the empirical results, or due to incorrect or outdated data as stated in literature (Braschler, Stadelmann & Stockinger, 2019; Smith & Cordes, 2019). Also, Smith and Cordes (2019) claim unreported data and incomplete datasets to be challenging which may be the reason there is not enough data to work with, which is a challenge described in the empirical results. Also, the lack of good data sources and the data must always be cleaned have been specified as challenges in the empirical results. This may be due to the large data amounts and the many different data sources available as described by Cai and Zhu (2015), however, questions regarding trust may arise since this increases the need for new data from external sources, leading to difficulties in companies ensuring high data quality.

Problems in data quality lead to the failure of data analysis (Smith & Cordes, 2019) or misleading results and useless conclusions (Cai & Zhu, 2015) and the empirical results explain that poor data quality may result in delays and rerouting of work tasks or even lead to the project failure. Hence, addressing and solving challenges in data quality is important in order to generate correct results and conclusions and avoid Data Science project failure. However, the reasons for low quality data are partially conflicting in the empirical results and the literature review. For example, Cai and Zhu (2015) describe that low data quality is due to the lack of unified standards for data quality whereas the empirical results focus on that data collection is conducted without a goal or lacking awareness of the value in data. According to the empirical results, the maturity of a company plays a vital role regarding this and hence, the parallel can be drawn to that the more mature a company is, the less problems arise concerning data quality. Furthermore, an aspect that the literature and empirical results agree on, are the data scientist's tasks of data quality which includes ensuring that the data is correct and valuable, structuring the data, washing, transforming and formatting it. As stated in the literature review, data quality relies both on the data and on the skills and capabilities of the data scientist (Cave, 2016). Since the data scientist needs to understand the data to review it to see if adjustments need to be made, the risk is that this may reduce the data quality if any errors or misinterpretations are made (Smith & Cordes, 2019). Hence, the parallel can be drawn to that the data scientist has an important part to play regarding upholding the quality of data in Data Science.

5.2.2 Ethics

Ethical challenges may occur in Data Science concerning privacy, security and sensitive data, which is stated both in the literature review (Passi & Jackson, 2020; Braschler, Stadelmann & Stockinger, 2019; Cao, 2018; Cave, 2016; Floridi & Taddeo, 2016) and in the empirical results. These challenges are related to the misuse of personal data and basing decisions on data without producing value to whom it concerns (Braschler, Stadelmann & Stockinger, 2019). However, both the literature and the empirical results focus highly on the role of the data scientist in ethics. For instance, interviewees claim that the data scientist needs to realize the sensitivity of the data that is being dealt with by valuing ethics and managing data securely. Also, a problem addressed in the literature is that Data Science uses data that requires interpretation and analysis which in turn mirror the views of the data scientist (Passi & Jackson, 2020; Floridi & Taddeo, 2016;

Cave, 2016). Potential bias in Data Science is also mentioned as a challenge in the empirical results which for example might occur when using image recognition where techniques only identify people of certain appearances.

Both the literature review and the empirical results claim that these types of ethical challenges and moral issues are complex to handle and solve which requires good governance where data is valued highly. For example, the empirical results show that ethical considerations are valued when data belonging to colleagues, family members and public figures is not monitored and ensure the data is not shared. However, empirical results claim that problems related to bias can be solved by using mathematical methods to increase fairness and using ethical analysis to manage the data and generate as much value as possible is recommended by Floridi and Taddeo (2016). Moreover, the empirical results show that ethical issues emerge when too much confidence is put in what Data Science can solve and from having too vast visions without realistic goals. Hence, valuing realistic goal setting and planning a manageable vision of Data Science would minimize ethical issues.

5.2.3 Knowledge

Both according to the empirical results and the literature review knowledge gaps occur in Data Science which can make it difficult to integrate Data Science in companies. Atwal (2020) explains that there are inconsistencies in the skills that data scientists think they need compared to what they need. However, the empirical results show that companies have expectations of data scientists having fundamental skills. Since, there is a risk these skills may be taken for granted and subsequently may become neglected. Instances where knowledge is lacking might also partly be explained in the empirical results saying that data scientists with good software skills usually lack in analyst skills and vice versa. Also, both the empirical results and the literature review (Waller & Fawcett, 2013) claims that building both analytical skills and domain knowledge requires a lot of time which makes no individual uphold the skill sets required to a high level.

Atwal (2020) mentions that organizations do not know how to manage IT investments and do not provide the right tools. Also, the empirical results argue the difficulties of not having the right skills to use Data Science tools and it is therefore of importance that organizations value and ensure that data scientists have the needed skills. To do this, it could be beneficial for companies to provide training in certain tools if the data scientist is lacking knowledge in that. However, the empirical results address that people managing Data Science tend to for example to lack computer knowledge and skills in communication and project management, which is not in agreement with Atwal (2020) who claims that technical skills usually are valued by data scientists. Having people with both computational and analytical skills in Data Science is indicated in both the literature review and the empirical results as something companies should embrace. Hence, the value of data literacy in an organization is emphasized. However, for data literacy to be efficient, it may be beneficial for organizations being certain of the work tasks and problems that the data scientist will aim to solve, and to recruit data scientists depending on this.

5.2.4 Organizational Support

Organizational support and management has been focused on as a challenge by literature as well as the empirical results. A problem stated in the literature review is that the management and IT departments do not support each other in data problems (Treder, 2019) and that organizations see data scientists as the only ones to make an organization data-driven (Atwal, 2020). This has also been witnessed by the interviewees and the empirical show that resistance is met when data scientists try to transform an organization into more data oriented.

The culture and attitude of employees is a challenge lying in the way of delivering success in Data Science (Atwal, 2020), which agrees with the empirical results. Therefore, before undergoing the transformation that Data Science requires, an organization's culture needs to be in place to be able to adapt to the change and adjust to the new requirements (Treder, 2019). However, in the empirical results, cultural problems have been witnessed in situations where there have been hierarchies with too strict divisions of responsibility and closed doors, which is in direct contrast to what Subrahmanyan and Jalona (2020) explain as a good culture, with an engaged workforce that works to make improvements proactively. Also, the empirical results show that challenges in organizational support and a negative climate might lead to difficulties of finding people with the knowledge to answer questions, especially if questions are not allowed to be asked. This proves the importance of data being everybody’s business and implementing a digital mindset throughout the whole organization, just as (Treder, 2019) and Subrahmanyan and Jalona (2020) state. However, the interviewees describe that a company's maturity is of importance to overcome the challenge of organizational support. Therefore, having an organizational culture in place that embraces Data Science may minimize the challenge of organizational support, and in turn increase the maturity level of a company.

5.3 Data Science Success

There are several success stories of organizations that have used Data Science to create competitive edge towards competitors, as described in the literature review. Hence, as Atwal (2020) demonstrates, Data Science has brought success stories in several different and varying industries. However, there are several different aspects that impact Data Science success according to the empirical results and the literature review. The empirical results show that an organization's success in Data Science depends on the organization's maturity level and since Treder (2019) mentions an organization's capabilities to change as an aspect influencing success, this may increase their maturity level. An example regarding maturity is stated in the empirical results, companies that view Data Science as computer science are on the beginning of their data journey, and hence will have more difficulties in reaching success. However, the empirical results show that Data Science was already being conducted 20-30 years ago, due to the techniques it uses, but has been rebranded to fit Machine Learning into the definition. Since the concept is unclear, as can be seen throughout the literature review, being hung up on the definition of Data Science only inhibits advances.

However, reaching success in Data Science investments is no easy task according to the empirical results, which becomes clarified since many projects fail. The empirical results show that since Data Science is not an exact science, it may be difficult to determine its success, and there must be a balance between being technically feasible, a general interest and having

financial incentives for it to succeed. Reaching success in Data Science may be that a company becomes more aware and more efficient with their work with data and whether that is called logistics optimization, risk analysis or Data Science is not important, according to the empirical results. Due to the vague, unclear and broad definitions of Data Science and its possibility of being applied in many different industries for varying purposes, there is no one way or a good way to reach success in Data Science, which may be attributed to the statement previously mentioned in the empirical results regarding that Data Science is not an exact science.

5.3.1 Actions for Data Science Success

Actions for reaching success in Data Science are mentioned in the literature review as well as in the empirical results. In order to reach success in Data Science, Treder (2019) demonstrates the importance of not only following the data management trend but ensuring that data is meaningful in order to create insights from it, something that the empirical results agree on. For instance, interviewees state the importance of purposeful data collection and focusing on the value that can be generated from the data. Braschler, Stadelmann and Stockinger (2019) claim that although organizations are data-driven, steps need to be taken to ensure that the data generates benefits. However, according to the empirical results, value can be generated differently, to make money for instance, but this varies depending on context and requires mature technology. Hence, this emphasizes awareness in organizations of knowing what value can be created from Data Science and having mature technology.

Also, to create success in Data Science both the literature and the empirical results claim to always strive to question methods and analyses to be able to adapt to changing requirements (Braschler, Stadelmann & Stockinger, 2019; Demigha, 2019) and to have clear goals and problem formulation, as stated in the empirical results. In order to do this, clear strategies can be used which may assist in how to take these actions. However, problems can still emerge and therefore it is important for companies to think one step ahead (Braschler, Stadelmann & Stockinger, 2019) and to assist, the empirical results values finding expertise that has experienced similar problems previously, since a problem is the lack of people with leadership and expertise. In this instance, consultants can help by solving problems more easily than someone internally may do and help formulate frameworks, needs and goals, and asking questions no internal staff would ask, as stated in the empirical results. This might be valuable to consider since Braschler, Stadelmann and Stockinger (2019) mentions a problem commonly occurring is an organization's narrow perspective when solving problems, which correlates with the empirical results stating that clear communication and to have realistic conversations about Data Science is required to succeed. For example, to solve these problems, it might be beneficial to communicate with people that have encountered similar problems in different contexts and to have data scientist representatives in management and steering groups.

Conclusively, as stated in the empirical results, there is no universal way of reaching success. Due to the varying applications of Data Science and industries in which it can be used, the success varies depending on the context. The literature review and empirical results indicate that success in Data Science is reached by linking the data result to a value, measuring the effect of decisions made from it and having a clear plan or strategy. Also, seeking help and expertise externally to investigate problems from another perspective may be of value, however, to do this, knowledge about how to create value from data, and have a goal with Data Science is required.

5.3.2 The Organizational Responsibility

To reach success in Data Science, both literature and empirical results indicate that someone or some department needs to take responsibility for taking actions. What is common in the empirical results are that all interviewees state a few aspects relating to the success of Data Science that either direct or indirect fall under the responsibility of the organization. For example, the interviewees state it is the organization's responsibility that there is a general interest in the Data Science projects and that they are technically feasible. Also, both the empirical results and Braschler, Stadelmann and Stockinger (2019) explain that in order to reach success, the organization has the responsibility of seeing to it that there is a mix in the Data Science team between experts and data scientists that complement each other with knowledge. Also, the organization is responsible for providing data scientists with the right tools in order to identify incomplete, incredible or biased data (Smith & Cordes, 2019). This in turn, may help with solving challenges related to data quality. However, according to the empirical results organizations have to be aware of the problem that needs to be solved by Data Science before making investments, since quick fixes, such as just hiring more data scientists, is not a solution. This further underlines the importance of organizations understanding what problems they have and in what way they expect Data Science to solve these.

Furthermore, Subrahmanyan and Jalona (2020) state that communication through leadership is required to ensure that data-driven mindsets are adapted into organizations for the employees to embrace organizational change. This correlates with the empirical results stating that the responsibility of leading Data Science to success belongs to the organization since it will not work without management support. However, according to the empirical results, lacking a Data Science mindset in companies is a critical challenge where several interviewees witness employees within companies tending to think Data Science will automate and replace their jobs and thus work against the Data Science efforts. Bringing Data Science to success requires management, ownership and seeing over these aspects over time according to the interviewees, which correlates to Treder (2019) mentioning the importance of an organization to constantly review if the culture is data-driven.

However, integrating this new mindset might be a time-consuming process for management (Subrahmanyan & Jalona, 2020) which could be one reason that this mindset is not always a part of the company and in turn could lead to Data Science failure. Hence, the organization has a noteworthy responsibility when it comes to reaching Data Science success, as found in the empirical results and literature review. As mentioned, having a general interest in Data Science throughout the whole company is important according to the empirical results. However, for something to be technically feasible does require more effort and covers many different aspects of an organization, for instance ensuring a mix of skills in the Data Science team. In addition, this chapter illustrates the vital support of management and the organization, which without, will make it increasingly difficult to take actions toward reaching Data Science success.

5.3.3 The Responsibility of the Data Scientist

Although the organization has a considerable responsibility in leading Data Science to success, the data scientist also has a part of that responsibility, which is something that both the literature and the empirical results agrees on. In order to succeed with Data Science, data scientists should understand how it impacts an organization (Atwal, 2020), however, what may come in the way of understanding this impact, are problems associated with what is included in the data scientists' responsibility according to the empirical results. One responsibility Treder (2019) mentions however, is that established data scientists are important to lead Data Science projects and oversee the work which is also indicated in the empirical results. For instance, the empirical results show difficulties in distinguishing between tasks of different roles in the Data Science team which could be partly solved through integrated collaboration where the team works together to create more value and assist in solving such a problem. A clear work task division may be a good solution in this case where leadership and the responsibility of the data scientist is key. Also, other responsibilities of the data scientist and the Data Science team are required to succeed in Data Science, as mentioned in the empirical results. For instance, these responsibilities entail keeping a high level of skills, having an eagerness to learn, and ensuring that there are a mix of different skills within the team. However, although it is mostly up to the data scientists to ensure they have the required skills, aspects such as there being a mix in skills for instance depend on an organization's recruitment of data scientists and having an eagerness to learn is more likely to be increased in a healthy organizational culture.

The data scientists also heavily rely on the data they are working with and thus Smith and Cordes (2019) state that the data scientist needs to have the right competence to be able to investigate the data to detect faulty measurements or errors. Therefore, as stated in the literature review, another skill data scientists must possess is to evaluate the quality of tools and techniques and the decisions and visualizations created based on them (Berntsson Svensson, Feldt & Torkar, 2019). Since Data Science success both requires responsibility of the organization and of the data scientist, a joint effort is required to generate a result of high quality. For instance, to solve problems related to unclear work tasks and roles within the Data Science team, the organization could support this from the top-down and raise awareness amongst the data scientists regarding the intended work tasks.

6 Conclusion

The aim of this thesis is to identify what may influence companies in order to reach their expected outcomes of Data Science, by answering the following research question: What influences companies' expected outcome of Data Science?

To conclude, all themes and concepts that have been stated throughout this thesis influence the expected outcome of Data Science. Since Data Science is a broad concept with many areas of application, the expected outcome of Data Science is subjective, and depends on the organizational context and maturity. Hence, there is no universal approach that can be applied for companies to reach expected outcomes in Data Science. However, this thesis has identified that there are both negative and positive influences on the expected outcomes of Data Science which are presented below.

Negative influences on the expected outcome of Data Science that have been identified in this thesis include various things. Many negative influences relate to the work method, technology and unawareness of risks. Also, data quality issues refer to incorrect, inconsistent, manipulated, unreported and incomplete data and not knowing how to use the data, are a negative influence since it leads to difficulties of analyzing the data correctly, misleading results and useless conclusions. Another aspect that influences the expected outcome negatively when not managed correctly, is ethics. This concerns the privacy, security and sensitivity of the data and becomes a negative influence leading to biased data and misused data. Lacking knowledge has been identified as a negative influence also, concerning insufficient skills and competence of data scientists, which results in difficulties in Data Science integration and incorrect decision-making. A negative influence may also be issues in organizational support including culture, management and employees which may lead to unclear directives and various types of resistance within the organization. Another negative influence affecting the whole outcome of Data Science is that the concept is perceived as relatively unclear and includes a variety of different techniques which may impact how it is integrated and adapted in an organization.

Positive influences on the expected outcome of Data Science that have been identified in this thesis include several aspects. One positive influence is the ability to create and generate data that is meaningful and valuable for the company. Also, the capabilities of being able to adapt to change, continuously evaluate and question past methods is a positive influence on the expected outcome of Data Science which in turn requires realistic goals and that a clearly stated problem exists that can be solved by Data Science. Hence, thinking one step ahead and having strategies are a positive influence on the expected outcome also. What also influences the outcomes of Data Science positively is being open-minded and receiving input from different perspectives, hiring consultants or receiving help from external parties. To reach the expected outcome of Data Science someone needs to be responsible for taking actions and thus divided responsibility is proved to be a positive influence. The organization is responsible to enforce leadership that values communication, being able to adapt to organizational change, provide management- and organizational support, having a data-based mindset and culture, and a general interest in Data Science throughout the organization. Also, the organization is responsible for ensuring that the projects are technically feasible, the right tools are provided and enforce knowledge sharing and a mix between roles. Moreover, the data scientist has the responsibility of overseeing the Data Science work and managing their work correctly which requires a clear task division and upholding a high skill-level.

6.1 Future Research

The concept of Data Science is relatively new and not deeply integrated in society yet, therefore, more research regarding the concept is required to contribute to this field. A suggestion for future research is to create a framework or method to describe the concept of Data Science more in detail, since an aspect that has emerged during the research process is that a clear concept of Data Science is lacking.

To validate or falsify the conclusion of this thesis, suggested future research may be to investigate and compare what may influence companies' expected outcome of Data Science by including the perspective of other stakeholders. For instance, executive managers, Data engineers and data analysts. Also, the negative and positive influences identified and described in the conclusion, may be a foundation for future research when investigating what actions companies can take in order to succeed in Data Science projects.

Appendix A

Interview Guide

Before the interview commences:

- Presentation of us and the goal of this master thesis/what we are investigating
- Explain objectives, contents and time limit of the interview
- Explain that notes will be taken during the interview and recorded if consent is given
- Inform that the participation is voluntary
- Explain that names will be left anonymous but that job descriptions will be mentioned
- The result of this interview will be presented in a master thesis at Lund University and will be available online. If the interviewee desires, a copy of the thesis can be sent to them and they are most welcome to contact us regarding this.

Introduction	**Questions**
Introductory Questions	• What does your job entail? • Explain your job tasks, position within the organization etc. • What is your experience of Data Science? Previous jobs, skill set etc.
Themes	**Questions**
Data Science • Value Creation in Data Science • Data Science Use Areas • The Data Scientist • Data Science Expectations	• In what use areas and how do you use Data Science within the organization? • What type of data do you manage in your daily work? • What are your main tasks when working with Data Science? • What skills do you believe to be most important to have when working with Data Science? • How is the Data Science department structured at your own company? • Do you usually collaborate with people with other roles or from other departments? • How do you work to make the data valuable for the organizations? • What are usually *your* expectations of using Data Science? • What are usually the *organization's* expectations of integrating Data Science? • Do you consider there to be a match between *your* expectations versus the expectations of *your organization* and the *organizations that you assist?*

Challenges in Data Science
• Data Quality
• Ethics
• Knowledge
• Organizational Support
	• Have you, your organization or organizations you have assisted faced any challenges in Data Science?
	If yes:
	  • How were these solved?
	  • In the future, to avoid such challenges, what actions do you think should be taken?
	If no:
	  • How could challenges be avoided? What sort of preventive actions can be taken according to you?
	• Do you consider data quality when working with Data Science? How?
	• Have you encountered any issues regarding low-quality data during your work?
	• Do you consider Ethics when working with Data Science? How?
	• When working with Data Science, have you ever encountered problems regarding ethics, inaccurate management of data, subjective data or lack of trust to data? If so, in what way?
	• Do you consider possible knowledge gaps when working with Data Science? How is this incorporated in your work?
	• Have you encountered any issues regarding knowledge gaps during your work?
	• Do you consider organizational management support when working with Data Science? How?
	• Have you encountered any issues regarding the organizational support in your organization or organizations you assist during your work in Data Science?

Data Science Success • Actions for Data Science Success • The Organizational Responsibility • The Responsibility of the Data Scientist	• What do you think is important to consider in order to reach success in Data Science? Why? • Do you think that *your* expectations, *your organization's* expectations and th*e expectations of organizations that you assist* are reached in Data Science? • What actions do you think your organization should take to better reach their expectations of Data Science? • What actions would you recommend other organizations to take to better reach their expectations of Data Science? • Who has the main responsibility of seeing through these actions? • Are there any actions you could take in *your* role in Data Science to help with its success?

Ending	Questions
Closing Questions	• Is there anything additional you would like to mention that would be relevant for our thesis? • Do you have any questions regarding this interview or our research in general?

Appendix B

Interviewee A

Interviewee: Interviewee A
Title/Role: Data Scientist
Date and Time: 2020.04.14

I: Interviewer
A: Interviewee A

Reference number	Person	Questions and Answers	Code
A.1	I	Okay, then we'll run. Can please start by explaining your experiences of Data Science and what the job you have today entails? What are your usual tasks?	
A.2	A	I started with taking technical mathematics so from there I have a good sense of numbers and logic and that kind of thinking and also that things can be incredibly difficult to understand at first but that it is then not impossible. I have also worked as an IT consultant for almost two years before I switched to Data Scientist. However, I found that the IT consultant was a little too fluffy and not as much of that concrete job I have always wanted. I have always had that "data profile" in jobs as I realized that no one was really good at checking this out. I understood that one must understand the data in order to build models which is why I wanted to look deeper into Data Science. So, ehm, my job that I have today is mainly about processing data and structuring data. I look at questions such as have we dated, can we rely on the content, have dated the right granularity, how does date A relate to date B in terms of keys, ID and granularity and so on. When structuring data, I make sure that it is user-friendly for my colleagues within the team. In addition to this, I encode some and document the solutions we have. Documentation is often not prioritized otherwise and even if it is not on my agenda it is very important that it is done. But I also do a lot of explaining about data and help my colleagues understand what data is about. So a part of my work is like, my team colleagues make the models and I help and explain the data. But I mean, it's both.	DS, DS-VCDS, DS-DS, DSC

A.3	I	Okay, but what types of data do you manage in your daily work?	
A.4	A	Since I work with risk the data I manage is mainly, it is risk-related data. Mainly I am working in SAS Enterprise guide which means that I have most of the data in the SAS dataset. Of course, a little Excel sometimes occurs. But I mean, we have tons of data, but I only work with risk and credit risk so I only have access to this data. This is like that for security reasons. So I do only have access to the data I manage, which is risk-related data.	DS
A.5	I	Okay, I see. Do you see any challenges regarding your organizations' data management?	
A.6	A	Unfortunately, a lot of people are dependent on other people and it is pretty fragile to have it that way. There is also very poor documentation where all the info is in the people's head. I mean, since only some people have access to certain data, it is hard when people do not documentate it correctly and when only these key persons know about changes and stuff. So if something in the data should be changed, this is communicated very poorly.	DSC
A.7	I	Do you have any process or strategy for how you should work in order to make this data valuable or do you work in any way in order to make the data valuable? How would you say you can make data valuable? Or do you and your organization even think data is valuable for you?	
A.8	A	Yes, definitely. All discussions are about the data and that it is the future. You want better quality, faster systems and so. And that is part of my job. I mean, my main tasks are to help with the documentation and communication. It's also not that effective if we all sit in each corner and investigate and fight with the same thing. Then I would say that I am fast in data, in coding and that I am good at communicating and structuring. So in that way I would say that I create value from the data we use. But I mean as I earlier mentioned I mainly only control my own division and department which is the risk department. But here we all work as a data scientist or quantum with data. And here everyone is struggling with the same things , such as granularity, the scope, content and so on. So, we have to work with data, otherwise the company will lose a lot of money.	DS-VCDS

		Like, if we don't have any models then we have to pay more as a company, which is also affecting the value of the data.	
A.9	I	You have mentioned some of your tasks when working with Data Science, I wonder, are there any tasks that take more time or are more important than other tasks? According to you.	
A. 10	A	Hmm, regarding time I would say that processing data, coding and ensuring a good structure, like these are tasks that often take a lot more time than my other tasks. But I mean, I don't think this is the most important part of my work, what I think at least. Like, I am a good communicator who makes sure everyone knows what to do and what types of tasks or parts of a task that can be complicated or hard to execute. So, for me, it is important that anyone in the team can do another person's job, which is often achieved through good structure, documentation and communication. I think it is very important that the work I and the people in my team do is not person dependent. Unfortunately, not everyone in the team thinks so .. but I would say that according to me communication and documentation are the most critical parts of my work but it depends on who you are asking of course.	DS-DS, DSC, DSS
A.11	I	Okay, are there any other skills or knowledge that people in your team or other people working in Data Science often lack?	
A.12	A	I would say that computer knowledge is generally lacking. In addition, I think that the right modeling tools and skills on how to use the modeling tools are often lacking. How do you model like that?	DSC-K
A.13	I	Okay, do you think your views on this are shared with the rest of the organization or the other people working within your team? Or do you think it differs how you look at it?	
A.14	A	No, I would say that we are on the same stage here and that we share this opinion, actually.	DSC-K
A.15	I	I understand. Are there many different roles that exist within your Data Science team or within your organization and how are your department structured?	
A.16	A	We have many people who work with data and in my department we are divided into teams of about seven to	DS-DS

		nine people each time. I would say that maybe 80 percent of these people are Data Scientists. The others are managers or specialists in adjacent tasks that are not as data-heavy in terms of programming and so. But, these roles can be data-heavy in other ways, such as having the right data, but how it is then implanted is more for us Data Scientists. So most commonly we are working within our team, like the main team. But sometimes we do also collaborate with other departments and stakeholders. Unfortunately I do not do much of this, but it happens.	
A.17	I	Even though you say that you don't make that many collaborations or team works with other departments or teams, do you know if these collaborations usually work well or are there any challenges here?	
A.18	A	Hmm, I mean, I can only talk for myself since I do not know that much about their experiences regarding this specifically. But when I do these types of teamworks I would say that I think everything works very well. Everyone is always curious and positive about each other's work, everyone always respects each other and so on. So actually, I only have good experiences with this.	DS-DS
A.19	I	Okay, but if we move on a bit to your expectations of Data Science and so. What expectations do you have for Data Science in your job and would you say that these expectations are the same for the rest of the organization?	
A.20	A	I would say that my expectations are mainly about finding tracks in the data that can be used in models and decisions, but on the other hand the organization is probably more looking to make money, I guess. That is their main goal. So in that sense I would not say that the expectations are the same. But otherwise, I would say that expectations match well if you compare what we as Data Scientists think and what the organization thinks. Absolutely.	DS-DSE
A.21	I	We talked a little about it before, but do you see challenges in your work with Data Science? Or are there any challenges you encountered when working with Data Science?	
A.22	A	We always try to work proactively and set pessimistic goals. But it always gets difficult when the data you need to use lies in different databases, when it comes	DSC, DSC-DQ

		from different systems, has different granularity and so on. Sometimes the data is not compiled and the IT department often has a lot to do, which leads to major delays and workarounds and in the end there are large delays. Which does not only affect my work and that I get stressed, but it also affects the organization when a lot of work is unnecessarily. However, I would say that we learn a lot from it anyway. Unfortunately, it can lead to a lot of extra jobs during certain periods, which can be quite long. But I know that we are trying our best and everybody is trying their best to work solution-oriented and proactively.	
A.23	I	Okay, do you think you have any role in solving problems that arise in your work? And what role do you think the organization has?	
A.24	A	The managers make decisions and they are the ones that schedule our time. But of course they have to listen to us employees, usually it is we who know best how long it takes to program a certain thing and such.	DSS
A.25	I	I see, do you consider ethics, subjective data, misinterpreted data, wrong data or even bad quality of the data over all, when you are working with Data Science? Have you encountered any problems or challenges regarding this?	
A.26	A	Yes, I would say that, both we in the team and the organization as a whole where we all treat the data with respect and nothing else is accepted. So high quality of both the data and the management of data is really important for us, I mean crap in, crap out. For example we are very careful not to share data with anyone and we are not allowed to check the data of colleagues and friends in our databases, nor public figures. But, I mean sometimes our team has realized that we have used incorrect data or data that has changed over time or that we have made mistakes in our implementations etc. It is also poor documentation, data fields that have proven to be filled in by different countries, data that is sometimes manipulated and so. Sometimes there may also be empty fields, empty data points, unreasonable data points, the same contract is called different in different databases which means that they can not connect properly. Among other things, there can be many more things than that. So, at the periods when such problems arise I get nothing done by my usual work, but all the time is spent just digging into the data. Even though we,	DSC-E, DSC-DQ

		in the end, hopefully have something that is ish okay to work with. But this could be a challenge, of course, so I would say that I have encountered challenges regarding the data quality and wrong data but not about ethics in that sense. Or at least not as I can remember. But as I said, we always work proactively and we all think it is important that if you realize that you have done wrong we are blamed not on anyone else but we help to find out and look ahead together.	
A.27	I	Have you experienced any knowledge gaps when working with Data Science? Or like areas where knowledge has been lacking in some way and that it has become a problem?	
A.28	A	Yes, unfortunately, not many people are equally teased about documenting their code well, so I often have to ask for completion. However, this is something my colleagues usually appreciate that I ask for. We also have all different programming levels and those who are really seniors usually do actually too hard code to understand. Then I have to ask them to comment even more and go through step by step what they have done so that we can understand. We are many talented but sometimes we have a hard time lifting our eyes which is clearly a disadvantage. It is also true that everyone has different interests in sharing and some are not so good at communicating. They do not understand the whole.	DSC-K
A.29	I	Would you say that your company tries to fill these knowledge gaps? Is it something they value?	
A.30	A	Yes, I think this is something that all in our organization values, regardless of position. But, I mean as long as we don't have the knife on the throat, they encourage knowledge sharing. But, for example, if we have a tight deadline, it is more and more about just getting there and then regardless of how. Often I am a bit of a handbrake that says “now we can all sit for an hour like that person X can explain to us what they have done ”. I also try to get the managers to explain in a good way what our purpose is with the work. A good manager is good at explaining such. Because sometimes it annoys me that people sit in their own corners and freestyles and not talk to anyone else at all. If that person is going to, for example, parental leave and be gone six or eight months and no hand-over has been done, then I get annoyed. But this has only happened when I've been a parent myself, I'm like a pain in the	DSC-K

		ass when it is important that we share and not just sit in our own room and drive. But, to answer your question I would say that the company values knowledge exchange, it is important.	
A.31	I	Would you say that support from your organization or the corporate culture is something you take into consideration in your work? Is it affecting the way you can manage your tasks?	
A.32	A	Well, it is a hard question. I don't really know. Like, we help each other and that is part of our culture at risk anyway. If that is what you mean. And I have support from the managers as well. Our CRO will gladly come by and ask how it is going. He keeps an eye on us all and is happy to provide encouragement and such. And this is important for the culture at the company but also because you feel that you are a little seen anyway. I think this kind of support is important in order to do a good work even though it is not directly computer related.	DSC-OS
A.33	I	Okay, have you ever any problems regarding this?	
A.34	A	Well, yes. Sometimes the data only takes so much longer than expected. But then we are usually helped to fix a work around. It is important. We help each other through creative solutions. Although sometimes there are quick fixes. But we all help each other anyway. And it can also occur issues with the management, especially since there are managers who want to see results but who can't, which is not really fun, of course, but it usually settles later. Also, we can get into trouble if the management has not planned or planned the resources well enough. But then there has often been a change of project to get back on track.	DSC-OS
A.35	I	Okay, I see. Do you have any tips and tricks for how to make Data Science and its projects successful? Are there any success factors?	
A.36	A	Ehm. Yes, I think you have to constantly show that you are good, otherwise it is taken for granted and forgotten. It is good if you both can code, see the whole picture and build models. Then you are flexible and able to do whatever is needed and put on whatever is needed right now. If I see it from my own perspective, I think it is also important to be active with what I want to learn and that I can lift my eyes and see the bigger perspective. You may not always think about the best	DSS-ADSS

		for the team, but also of your own best. So to be smart, to be seen and to receive higher pay and more responsibility. I think that is important in order to succeed as an organisation but also as a Data Scientist.	
A.37	I	Do you think that both your and your organization's expectations of Data Science are usually achieved?	
A.38	A	No, I wouldn't say that. I think we should get better education and that we should be certified to show what we can and that we know the things we can or the things that are expected of us. I think the company needs to come up with better educational proposals and ensure that we have the time required to attend training so we can be even better. It is also important to exchange knowledge and to exchange knowledge within the organization.	DS-DS, DSC, DSC-K
A.39	I	Okay, is there anyone you think has the main responsibility for reviewing the actions within the organization? So you can achieve better success then?	
A.40	A	Yes, it is the organization as a whole, but we ourselves too. But unfortunately, it is very difficult for us to come up with good suggestions on such problems as long as we have plenty to do that just needs to be clear.	DSS-OR, DSS-DSR
A.41	I	That was all the questions. Thank you so much for putting up this interview. It helps us do something huge. Is there anything else you think of that could be relevant for us to know? Or do you have any other questions about our essay or the questions?	
A.42	A	Thank you, no, I don't think so actually, I think we have covered many important parts here.	

Appendix C

Interviewee B

Interviewee: Interviewee B
Title/Role: Data Scientist
Date and Time: 2020.04.14

I: Interviewee
B: Interviewee B

Reference number	Person	Questions and Answers	Code
B.1	I	Would you like to explain what your job entails, what are your main tasks?	
B.2	B	Yes, ehm, I've had a few different jobs over the years that include Data Science, and you could say that my career, if you should call it that, in Data Science, began in 2009, when I joined a startup, and where I became employed and worked with text analysis. At that time it was so innovative and you had to do a lot from scratch, a lot from the beginning and at that time I worked mostly with a tool called MatLab. Since I have this mathematical background in Theoretical Physics, there were many applications of linear algebra, much to create what you might call a kind of map of a language. You found some methods to translate text to speech, so if you look at wire mesh with a map, each point can be described by two numbers, x and y, but a map for two may need much more dimensions than two, so that eh you can't imagine those dimensions but you can do them in a computer. So I worked with that, text analysis for 5-6 years, and then I started working at a media company that started using my knowledge to recommend articles. So that eh, they linked users and articles, so, a bit like when you go on facebook, they have your personal information, and those ads are relevant to you, etc. You want to connect a user to an item.	DS-DSUA
B.3	I	I understand, interesting, and your experience in Data Science?	
B.4	B	Yes ehm, then I worked as a consultant in Data Science. And that means helping with the sales support, finding the companies that can be helped by a Data Science	DS-DS, DS-DSUA

		project to some extent. A lot has to do with finding the companies you can help, and being able to do some kind of proof of concept, and then trying to have some ideas in the pipe that we can create. Then try to show these companies some value. How can you profit from an optimization, an automation, an investigation and many of the customers have been within eg. Healthcare, Region Skåne eg. Ehm, companies that deal with automation, and then of course there are different degrees of maturity of what it looks like at these companies. E.g. There are those who already have established data science initiatives, such as Ikea, TetraPak, Axis, the big ones. And they are usually so big and mature that they would rather expand their own Data Science team and AI team than hire a consultant. And if you end up in such a type of company, you are often part of such a venture, and then it is more so that they have a fairly mature process, in how far they have come.	
B.5	I	What kind of tasks did you usually get to do?	
B.6	B	Typically, they want you to engage in a very specific task. And then there is quite a lot of data, you can say that there is a kind of cycle in Data Science, or AI applications, first you have to go in and see what data there is, and what data you would like to have. And then you have to decide for a project, which project is somewhere at the intersection, between being technically feasible and that it adds value to the company. Usually these are the parameters that are played on. Because many people who deal with AI have heard many nice buzzwords, so you can do face recognition, voice recognition, you can do a lot with image recognition at all so maybe you get some very big visions, and then you think you can solve everything with this. Many people believe that if you have more and more data, you will get better results. But this is not always given. Somewhere you come to a limit where new data does not mean adding new information or more information, but you need to know more about how good information this new data has. Somewhere in the project you are also a kind of advisor about what kind of data will be needed and what benefit it will do.	DS-DS, DS-VCDS, DS-DSUA, DSC
B.7	I	Hm I understand, do you have any examples of that?	

B.8	B	An example is then in home healthcare. There we have a lot of problems with that a lot of data that would be available is regulated by the GDPR law, it is sensitive data with the person's name, identity, age, illness and so on. If you want to do medical research, for example, it is a big problem that the data may be available but that you do not have access to it. So that it is a role I have, to try to explain how to run projects. To also talk to sellers who are out there, how to find data. Later, when you are about to start a project, it is very much about gathering information and data that you can use. And, sure, it's great to be able to access data, you want to be able to access as much data as you can. There you can be very creative after all, you can have data such as companies have unique access to like sales data, production data, user data and so on..	DSC-E, DS-DS, DS-VCDS
B.9	I	Interesting, do you have any more examples of that?	
B.10	B	Yes, for example, Spotify is a very AI driven company. They have a lot of information about the customer, what the customer listens to, what songs have they played, what songs they have started playing, what searches they have done, what songs they play often, what songs they have recorded on but then skipped forward, there is, like, a lot of information. Then, with that recommendation, you can make a playlist that you think this user might like. Usually they are very good. Like songs that were played for a long time ago that are forgotten. There are algorithms that look at each other similar users have used. Ahh okay, those who have listened to this and that song, they've probably listened to these as well, and usually it sounds really good. So this about collecting data, it's important. And sometimes it may be that you can think about which data is important. If you look at spotify then, for example, it is a subscription, you pay a fixed fee every month for us to provide a service that you like. And many of these IT initiatives by Data Science are based on this principle. A single user pays a relatively small amount, but many people do so, so they accumulate their value there. Ehm, since then we have network sales, advertising, how should a user find the right product. You can look at Data in all these examples. Then we have some typical tasks that I could have done, collecting data, then the data must also be made available for a project, which perhaps is not that I have a lot of data, if it is not possible to plug in an algorithm eventually. So what	DS-DSUA, DS-VCDS, DS-DS

		many Data Scientists do, too, is reminiscent of what a restaurant would be like.	
B.11	I	Okay in what ways are these similar?	
B.12	B	Yes, if we say that you are a passionate chef, and have a dream of opening your own restaurant, because you have a kind of vision of what it would look like with your food then, you think cooking is the most fun, but maybe the cooking really only seems to be a small part of the business itself. You have to find raw materials, you have to find a venue, you have to do marketing, you have to do job interviews, you have to do the decorating and you have to follow up. So the thing about running a restaurant and running Data Science projects actually has many parallels. If you find that finding an algorithm is fun, then I can compare it to the cooking itself. Let's say then that we have these goods that are in the form of data and now we should do the cooking, which I think is really fun. So it's about applying some interest in mathematics or seeing patterns, or finding stuff, then it's good not to have to reinvent the wheel. So some may think that yes but I'm so good at finding dishes and so I'll do it on a continuous basis and that my restaurant will build on. And perhaps it might be so! but! A restaurant for example, what would you bet on, I have eg a concept here that is very exciting or should I eg. open a pizzeria. Everyone knows how it tastes, how to do it, you can only vary to a certain extent, have a certain type of cheese, own tomato sauce, etc. Or - should I completely apply a new food concept from the beginning. And it may or may not succeed. And so this is the case with Data Science as well.	DS-DS, DS-DSUA
B.13	I	In what way??	
B.14	B	Like should you build things from scratch or should you build on something that already exists and make them a little bit better. There you have, as well how it should work, of course, if you come up with any concept from nowhere, maybe not just a dish but a way to cook that was completely unique. A smart concept where one focuses on the raw materials and the environment rather than the dishes themselves. But what do I know, it can be very varying. Then we typically have the Data Scientist who wants to find an algorithm and maybe improve it and then eventually you get something that takes in some kind of input and delivers some kind of	DS-DS, DS-DSUA

		output. And okay, the input / output can be that a customer has clicked on a product, how much is the probability that they will buy it, or we are a real estate company and we have got some houses, should we put any price idea, what should we sell the house for and what should we market to customers, it is a feature that we want to use. So data in - data out. There we usually have the role more for those who are data engineers. Where they have interest or are passionate about how we should get this down to a mobile app, I as a person am not so interested in that kind of thing. But there are steps in it that I could not solve, therefore you want to join a team, who are good at doing it. And ehm, then you have to put together so that there is a data engineer, a data scientist typically deals with the parts of this chain that have more to do with retrieving the data, insert it into a formula that makes it useful, look at this algorithm, or by all means, make it your own algorithm, but if you have to make your own algorithm, and it really should be useful in some context. Then you really have to believe in your idea, and compare this to cooking and restaurant as I said, there is like a community that is such that there is very open software and open source, and I would say that it is a quite a small probability that from scratch you would create any algorithm that beats what is standard in the market and community. Do not do it.	
B.15	I	Yeah interesting, why?	
B.16	B	It's really the same thing when doing research. Actually, what is important is knowing what somewhere is the future now, what is the best thing to use now. It will be like this that we will come to some kind of level that everyone thinks is normal. But I think, how long back we think since it was all cell phones had buttons. It's ten years ago maybe. I remember this myself, but now nobody would use a button phone. So you have to have an understanding of what phase you are in. So it is, as in Data Science, we have access to information, we have access to data, we have so much data that it is difficult for us humans, our brains, have a hard time handling this amount. It is very difficult to sort out what is important, what is noise in all this. And a successful Data Science project, has some kind of, ehm, metric in what value it generates. And it is something that is a key figure, which somehow shows how good an algorithm is. E.g. Here's how, in an advertising campaign, how can I find users or customers who will	DSC, DSS-OR, DSS-DSR, DSS-ADSS, DS-DSUA

		click on this ad. Or in an image analysis project, How do I find or diagnose various diseases using the X-ray image. How do I know if there is a tumor here, or what kind of disease it is. And, I mean, today it is more so that it feels like, this is a little bit the problem with self-driving cars today that it feels uncertain that technology is not far enough ahead to make decisions compared to a person. But soon, in a certain number of years, we will say how can we let a person make such decisions, as well as we can not analyze all information in a good way.	
B.17	I	Would you like to elaborate it a little more?	
B.18	B	Yes a little if you look at web screens, historically, ehm so then I like the comparison with the best chess player. Ehm 40 years ago there were chess computers that could beat hobby players, but they could not beat the best people. And I think it was some 20 or 15 years ago, I don't know, that a chess program beat the best human. And now we have a history of the time when the best human beat the best computer, it was quite a long time ago now, and now it is long since it has developed the method of learning in a way that is no longer accessible or understandable to our brain.	
B.19	I	Interesting.	
B.20	B	Yes there are many aspects there like how, how ehm, what is a task for a data scientist, what is a task for a data engineer, I think then you have to apply this out here in the companies then you still have to be good at zooming in, as well as what projects or problems I will be able to solve for you so that it is interesting to start even. Because it comes down to so much! If it is technically possible, there is some new algorithm or technology or for all hardware, a graphics card that can count very quickly, now we have eg. these various cloud services so that a company can test ideas without having to invest very big money in hardware, you can test a thing up on the cloud then, an idea that you may have on a smaller scale then. Then you realize that, but we needed the equivalent to a process and a graphics card and then we pay amazon or google or microsoft azure then who are the big players, and then the things concerning sensitive information come in again then, okay, for example now we are a medical company that is looking for research or a finance company where it does not work to have all bank accounts out without our	DS-DS, DSS-DSR, DSC-E

		control, etc. etc. Basically, take such a company as Spotify, they are not so data sensitive, it may not be the world's worst issue for anyone to find out, this and that person's favorite music is this and this, but I do not know, it may be sensitive also. But you understand the principle, there are different degrees of sensitivity, you absolutely cannot have national security issues and you have to have some kind of backup, and these cloud giants do have this. They put a lot of money on who has access to what and then there are laws that regulate this.	
B.21	I	Are there any other tasks that fall on the Data Scientist?	
B.22	B	Yes, there is another role that exists then, the security of it all, how it should be used, what kind of ethics are there. So from my point of view,, I've always liked to see a pattern where most people see chaos. These things that I have done before may not have to do with each other you might think, being trained in physics and now you sit here and have worked with a company that deals with call center and alarms for old people. But the line of thought is that I try to find some kind of pattern where most people see disorder. And that's probably what has somehow driven my career in that way. Okay so, if you ask what I'm doing, what roles, it's very different. Because, ehm, okay, a quick answer is to in general try to find some kind of value in knowing a certain data and analyzing it in a certain way.	DS-DS, DS-VCDS
B.23	I	Do you collaborate with a lot of other people or roles in your work?	
B.24	B	Hmm yes and when you say my job now, I have just now switched employers, when you just get into a project it is a lot of cooperation with others. And it is very important to know what this AI project chain looks like, from having like data to analyze to having a product that does something and then the question is whether that product will be something that a single user to this company will use such as an app e.g. a weather forecast or depending on the answer, some recipe, Ica food bag, which recipe should I suggest to sell the goods I have an estimated surplus on, it can be a link between this eg. Ehm, sometimes you don't even know what data is important because we may not have this ability to make these connections. That's what you want to solve with this technology. How should I best get flow in my sales, how should I best and most efficiently organize ambulances, what personnel or staff	DS-DS, DS-DSUA, DS-VCDS

		should we have in an emergency, when can we calculate many approaching healthcare, is there anything, a pattern we can recognize, it is important that e.g. with the emergency to have a nurse who knows e.g. CPR, is this more important than someone who can take care of and manage dementia, that you do not know. Take for example these times when you know that coronavirus has a more negative impact on older than younger, you have to take that into account when dealing with this first contact. But now we are in an extreme situation as there is no model that could have looked back how an emergency was manned today so you have to adapt. So it's really a good example of how new data can take over to change a model. I like these examples in healthcare, restaurant and spotify, they are like three different aspects of data driven companies. A restaurant does not have to be data driven at all, it can simply be as it was 20-30 years ago, where a restaurant had to use word of mouth to show that it has opened something new exciting here, 20 years ago sushi became popular, but now sushi exists in every corner. Ehm, and so it is, like, it is always a process that is feedback to itself, you get new data where you improve this data, and ehm, then you want to be able to use it as well. But, I would like to say, for good and for bad, that some think so that if I have an education then I can provide for myself, for example, that I want to become a craftsman. I want to learn how to install something in a house, how to set up an alarm in a house, and then I want the education that wants to take me through an entire professional life. So I have a feeling that the world doesn't work that way anymore.	
B.25	I	Yes that is very true. Regarding competencies and skills of the Data Scientist, is this something you see is particularly important there?	
B.26	B	Yeah ehm, what I would come up with is that I think those who have the best conditions to succeed in their career and in their life are those who are most willing to be flexible and learn new things. I would like to emphasize that this is the case within Data Science as well, where it is so important to address this variability. To realize that it can add some value, one must be able to be at the forefront. If you want to make money from it, the technology must be mature enough to work, it is like a very nice balance there. I do not know if I have reached the conclusion in advance, but in that case you may say so, but if you do not have those questions then	DS-DS, DSS-DSR, DSS-ADSS, DSS-OR, DSC

		maybe you should ask them. What problems or what are the problems with, yes I have heard that you should invest in becoming a data-driven company, this with AI can solve a lot, and it is important that we jump on this train now, I have heard these beautiful words about machine learning and artificial intelligence, big data, if you only have enough data then you can solve a lot of problems. But many companies, and especially now, when there are such uncertain economic times and it will be in the future now for quite a long time. Because we were already entering a kind of recession and now it is a total reboot of society that is happening, and it is a little difficult now for you and for everyone. Because we really come as a whole society, from individuals to how the whole world works, because we have relied on this globalization and that it is good and that there is an endless access and everything may have to be reviewed with this mindset. It's a bit difficult to define, I can say like this, that there is a bit of a different maturity for this with data science, it's still like this is a creative project. And there are certain roles, but at the same time it is difficult then that just as I mentioned that it is a craft that you have to do in a certain way. Getting the result is more complicated than just spending a budget and getting results from that.	
B.27	I	I understand, you have touched upon so many different parts of what we were going to ask here, I was thinking a little bit about, when you are out at companies in different assignments, do you work in teams with others or what does it look like ?	
B.28	B	Yes, but it all depends on what the assignment is and how mature their data science process is. I may not be able to talk so much from my own part because on assignments I have actually never worked in a team personally.	
B.29	I	I understand, how have you worked otherwise?	
B.30	B	Well I have been employed by a media company and then I worked in a team. At the time when I was working in teams, I was most responsible for the mathematical bit of finding algorithms and then it was very focused on text analysis. How could I find information from text and then do something relevant with it. And so then I worked with a team that was divided into FrontEnd, BackEnd, AI, Machine Learning and a sales team, and a marketing team and	DS-DSUA, DS-DS, DSC

		management then, everything that exists. But then I was also hired as a consultant for a company that wanted to be more data-driven, but they did not know for themselves how they wanted it, or how it would create a success in it, and it can usually be that way to be a good Data Scientist or success in Data Science is based on being able to have visions and being able to implement them. Because if you have no visions, then you don't get anywhere. E.g. If we would have a camera that records what's going on in town, everyone can get suggestions for what to do to have fun. Like then it gets so flimsy, I think it's good if you think you can find a task that you can solve. That is so limited that there is a solution but so interesting that it is innovative and that it will generate some kind of financial profit.	
B.31	I	Mm, is it the leadership of an organization that should have this vision or who should have it?	
B.32	B	Hmm yes it varies, for example working as a data science consultant, you then try to prove a value that it could have for a company and then maybe you usually work on some kind of, ehm, a demo, a proof of concept, a kind of kick start with maybe a limited amount of data in the beginning, what values you could have in a project, it's not that you have millions with rows since then you might get some code that makes it like okay, but I have seen that there are certain trends and some patterns and so, but this data could be useful, this data might even decide, what color should this button have on the website, should it be red or should it be green, we might get the best result from our marketing campaign depending on the color. There are different roles but if you work independently or maybe with some team, then you probably think it is important that you show some benefits, not just that you use AI for the sake of it, because then it usually is not good.	DSC, DSC-OS
B.33	I	Do you have an example?	
B.34	B	Yes, let me draw a parallel again. Ehm, the dating world, now I'm going to get a partner for the sake of it. Everyone else has one and it is great fun and then you get happy. Then you test it, and I do not get very happy just because I met a person, maybe should be a certain type of person, if you want it so very much but you have no idea why you want it, then maybe it will not be good once you get there. But if you have some sort of thought with it eg. what does happiness mean to me,	DSC, DSS-OR, DSS-ADSS, DS-DS

		what should i do to achieve it, I should take a little step so maybe you can start then you eventually get much better at this selection, what is the goal then and is there some kind of dealbreaker, may have very high expectations but you may not think about what you add to others for it to work. You do not think that you are part of a whole, of something bigger. This is the case with this company. Okay we want some service that you have to pay for because we have used some kind of AI, both as a customer and as the company that provides this service. The customer may have set unbelievably high goals or their expectations for something to work, I should have a mobile that only works in the middle of the forest and it should only work, and should be able to deliver this answer directly, it should know before I answer which I looking for, it's as realistic as finding this dream prince or princess. Anyway, so should the company then, I think a good kind of data product delivers some kind of Aha experience, in fact, ehm, and a little bit so that it shows that it serves a purpose. When I looked at these pictures I saw some kind of context, they somehow belonged together. Sorted in a smart way and shows information in a way that made me get this Aha experience. I think both as a consumer, that you download a software that is easy to get started, easy enough so it should be ehm feel simple but advanced enough to do something useful for you. That balance. I think both companies have to think that way and that customers have to think that way. And this is something of a kind where a good job as a Data Scientist who has some visions but at the same time must stand with two feet on the ground.	
B.35	I	Mm, Have you seen any challenges regarding that?	
B.36	B	Yes, there are a lot of challenges to it. Just as often you can be like this, my advice to those who initiate data science projects is that you should have some milestones and goals that are realistic, and that can show on a result, show on a small result rather than nothing. If you are going to introduce some kind of digitalized sales, or marketing, or automation somewhere in any production line or within an organization, within an image analysis, but stick to coming up on some kind of step and rather something that is good enough than something that is amazing. I think that is the way to go, because if you just come into any organization and say that now we should do something really smart, that there can be some kind of	DSC, DSS-DSR, DSS-ADSS, DS-DSUA,

		superstition on this with artificial intelligence, then it will not be so good.	
B.37	I	But eh, you mentioned before that there is a lot of data in organizations today and that a lot of data is handled. Are problems regarding the quality of the data something you have encountered? Regarding misinterpretations eg.	
B.38	B	Yes, that's a good question you ask there. It is very specifically directed to those in Data Science, it is a typical task to make sure that the data is in the form that is useful at all. A common task is, for example, usually you get data in the form of an excel sheet or it may be in a database, it may be filled in a little inconsistent way. Sometimes someone has said that it is a yes / no question and in another place in the company there is a zero, or sometimes there is no answer at all and some fields are left blank. Sometimes a database is constructed with fields that are designed for the user to have future use of them, and all that there is so that there are slightly different forms of methods that are used today. It may be that you have some form of method that given that you put something in, you get a response. It is usually the case here that eg. recommendations work. You put in users, you put in a range of products, and you maybe put in some kind of history. Then you get out; we think this customer will be interested in this item based on what it has bought in the past, and people who have bought similar products have bought products in the past. So it is a way to structure the information in an intelligent way or arranged according to the recommendations of certain algorithms.	DSC-DQ, DS-DS, DS-DSUA,
B.39	I	Do you have an example of that?	
B.40	B	And ehm, yes, if you say this if you look at this interview and record it to get concrete data, then the data can come in more or less structured form, and the task that the Data Scientist often has, is to get it structured in some way. If you then, let's say then a common example is then this online evaluation of a service. how would you rate this movie? 1-5, it's pretty simple. Okay, what did you think of the plot, what did you think of the actor's efforts? Do you have any own words to describe any comments? And then you come across that there can be slightly different types of commitment when you have come so far that you have	DS-VCDS, DS-DS, DS-DSUA

		to fill in your own words. Some just skip that bit, and think it's enough that they have graded, and some do a whole essay as well. And ehm, that's the art of doing some sort of evaluation of how valuable this information is. For some, just fill in the highest grade or lowest grade there. And then you should get a nuanced picture. And that is again back to your question it is a problem that the data is not clean enough, and then you can say that yes, but it is also somewhere that you have to be aware where you can bring some kind of value it is just when you can find, what to clean out at home, what to keep - what to throw. After all, it is like a small process to handle data. If we should say when to clean the fridge, it will eventually result in you thinking through and, okay, this I absolutely can not eat, this I have to cook tomorrow, so I have to buy some more ingredients so that I can do something sensible with this, okay this maybe I do not need to have in the refrigerator, it only takes up space we can put it in the pantry, a bit so is the case with data.	
B.41	I	What an interesting way to see it.	
B.42	B	Yes e.g. The data, if I do it in any way, if it is an image, this image may contain something that may be of interest, but one must still make it available in some form. E.g. a typical project there is then that we have images and satellite images or close-ups, with these flying cameras you have nowadays, you can take pictures of a field. And so you want to be able to identify only in this crop, if we have come as far as we expected, when can the crop be expected, if it will be ready by the time we thought, or later, we can plan for it, and ehm, how are we doing with weeds here, is this a plant we have planted or is it something we might need to use pesticides on. And if we can do more targeted control then, we can know what weeds are here and what pesticides we can use right here, on this weed, using the drone's information to locally combat and approach this or should we just spray it with something that is bad for the environment and all the crop and food, we have to get rid of the weeds at all costs. That's what I think about cleaning the attic, cleaning the fridge, and that's the case with the data.	DS-VCDS, DS-DSUA, DS-DS
B.43	I	Mm, and then is it the data scientist, or you in this case, making such decisions?	

B.44	B	Yes indeed it is! It's a very good description of the role of a data scientist, just because you are this creative chef who thinks it's fun to cook but somewhere you have to think that okay I would think it would be very interesting to cook such a dish, and it is accessible right now. You have to ask yourself, I may think it is very fun but it may not give any rushing sales figures, if we do not put it in any French name and market it under the heading rustic medieval food so that people feel it is exciting. For example, I mean, when I think about it, a museum project with an exhibition, of something that is thought provoking or provocative or so there is an art that everyone thinks is beautiful, then there is too abstract eg. photo art, or exciting, and some may think that the image can be taken by anyone. Nothing special about it, so that is the case with data too. There is some data that conveys information that is quite obvious, but then there is some data that is a bit hidden. I think a good data scientist can do things that can get people to respond and convey this Aha experience, help and be a tool, an instrument to convey this Aha experience. Actually you can summarize a good Data Scientist like that. Tie it together. There are really many different steps, you have to show value, you have the data you have, you predict what more data would be interesting, suggest somewhere in the management to allocate money for this.	DS-DS, DS-VCDS, DSS-DSR, DSS-ADSS, DS-DSUA
B.45	I	Have you ever experienced wrongful use of data or subjective data in any way, is it something you take into account in your work?	
B.46	B	Okay, yes. Yes, again, the inaccuracy can come from the fact that there is too much confidence in what is to be solved, if you do not define what problems to solve if it is for loose bridles, for big visions without having realistic intermediate goals, then it can be a problem. I have seen that. Now we solve this by just maybe ehm, yes eg. in text analysis. We want to find some connection between how much is written about a particular topic, a certain keyword. If you search for a person or party or something and want to link the success of a political party to how much is written about it in the newspapers in the last weeks, months, and what is written then maybe you take in more data, do polls where people are allowed to write their own opinions then. If you then take as I mentioned if you should go out with a survey that is based on voluntary participation, then it is not the random, it is not	DSC-E, DSC-DQ

		voluntary participation, if some have received the question and some have refused, take if you do not take that into consideration, it will be wrong for those who answer the questionnaire to take it. It's not the entire population, it's a sample of the original data set. So you have to take into account the type of data you have worked with. A bit like this is now:: Eg. with corona: If you look at it, it is terrible for all those who are affected, and how many of those who are ill and actually die and who are most at risk, that is one side of the coin. But then we also have the whole social price, how many are so infected by the whole virus, we do not know this because we do not test the entire population, how big is the consequence that the entire society shuts down and that some workers can not perform their jobs, some lose their jobs. Some might have needed a treatment or surgery that they can't get because of such great pressure on health care. Which is the mental cost to society, some get depressed and go into some kind of wall, because of what is happening. Maybe many more than those affected by the virus. And maybe those who suffer from it as much as suffer from an influence or cold, they carry it but they do not know it themselves and they infect others. That's exactly what consequences you have to think about, in a broader perspective, so yes, this with the data is somehow the result of wishful thinking, it's very important to get around it. And there are also mathematical methods for looking at bias, it is very important to keep that in mind to get a fair result. With the statistics you can present a little what you want. It is usually the case that there is an inquiry that has been ordered, and it is often ordered with a certain result in mind already, but then it is important to have bias in mind when selecting their data. To have some type of data, it must give some kind of Aha experience.	
B.47	I	Yes, we thought a little more about how to actually achieve success with Data Science, if there are any concrete factors that you think are important to consider in order to do so?	
B.48	B	Yes it does. The concrete success factor is to think of two circles, one is technically feasible, the other to create value and interest. And then like, sometimes these circles overlap completely, sometimes there is no common interface, but usually, it is that these two circles usually have a small intersection, where there is an area where it is technically feasible and that there is	DSS-ADSS, DSS-DSR, DSS-OR

		a general interest and financial incentive. Try to find that interface.	
B.49	I	Mm, do you think expectations will then be reached at Data Science, if one were to do so?	
B.50	B	I think the important thing is that before you go in and put resources and money into Data Science projects you have to be aware that there are like quick fixes as usual, with quick fixes, they work quite poorly. If you have not analyzed the problem in itself, what is the problem I will solve, with a quick fix eg. now we hire some more data scientists because we have a good budget, is not the solution. I think that success in Data Science is, thinking it through, what are the problems we can solve, what is the value we can get from it and then based on that, building some kind of model, which can then be built into a app, which can then be selected on a phone as well as generate a value. And the value can be generated in many different ways, there are many different ways to show a value. A value can be making money, how much I earn, and a value can be when you think about why Facebook became so successful. Nobody paid for a Facebook account, but still Mark Zuckerberg could become one of the richest people in the world, how it went. It's just that many people use it, and when many use it, it's like already a way to expose things and control what is to be displayed, and there is a value. And then maybe value is not money or so, it can be convenience, it can be a benefit to the environment, to humanity, what is valuable is a question that one must ask in the context.	DS-DSE, DSC, DSS-OR, DS-VCDS, DSS-DSR
B.51	I	I: Well then, I actually think that was our last question. You've covered so much in everything you've said. Thank You so much	
B.52	B	B: Great, just fun to help.	
B.53	I	I: Would you like to have the paper later when the thesis is published and completed?	
B.54	B	B: Yes, it would have been fun to see.	
B.55	I	I: Perfect, then we'll arrange that. Thanks so much for the interview and for wanting to assist us.	

Appendix D

Interviewee C

Interviewee: Interviewee C
Title/Role: Data Science Manager
Date and Time: 2020.04.15

I: Interviewer
C: Interviewee C

Reference number	Person	Questions and Answers	Code
C.1	I	Okay, can you please start by telling us a little about what your job entails? What kind of work do you have? What does an ordinary day look like?	
C.2	C	Yes, it's a bit of a wrong time and ask what a normal day looks like right now. But we can say like this, I've had quite a few roles through the years and I've had many different roles that have been more or less data intensive. I've been working as a Data Scientist for five years or so. And now most recently, since the beginning of December, I work as a manager in Data Science which means that today my relationship looks a little different than it has done before. Now I focus more on project management, coordination and work a lot towards our internal customers. And my group primarily focuses on decision making and improving decision making within *, so we work with all possible organizations within *. You can say that we work as internal consultants but we do not charge for our internal services, some do, but then you easily end up in a trap where you deliver a product that no one wants to use because they have not understood what they have ordered or so. So we are also responsible for monitoring the results of our products and reporting how much impact all our models have and how much value we contribute. What I would say is one of the keys to success, is to link the result to value, ehm or measure the effect anyway ehm, if it is difficult to measure the value then at least measure the effect in terms of decision making. Have the decisions been changed in any way. And to be a little clearer, in our case, is it that we have increased sales in some way or have we made savings in some way, or have we lowered the risk in some way? Where the risk can be	DS, DS-DSUA, DS-DS, DSS-ADSS

		interpreted a little differently in different situations, and partly it can be the breadth of a distribution, if we can reduce the spread, it is a way to reduce the risk and thus improve process quality. But it can also be that we reduce the risk of being sued by having control over things before things go wrong. And then it can be difficult to measure the value, but then we can see that we reduce the risk of unforeseen events in some way. But yes, back to the question of what an ordinary day looks like then. An ordinary week looks like I am working on four, five projects in parallel for which I am responsible, so much of my time is spent on keeping track of the different product groups and seeing what problems have arisen, looking at results, interpreting preliminary results. , have discussions if we go ahead and things like that. So I don't work as operational anymore, I don't code that much anymore. However, we look a lot at code requirements, review code, look at data and the results quite a bit. But a lot of the time goes to discuss with our internal customers what their decision-making looks like, in what way they expect to use the results of our products and models and ehm, how the decision they think they will make, what it has for impact on the balance sheet. So what will be the economic impact on the decisions? And if you do not know exactly what it is to decide what to make and how to use the data, it is often a warning flag that you will fail with the project, you do not know what to do with it but you dig only a little to see what you find and see if you find something interesting. And it is one, yes, but it is a little risk-taking.	
C.3	I	Yes absolutely. But what kind of data are you most often handling in your role?	
C.4	C	I would say that 90 percent of it is about structured data, in relational databases, that we retrieve data in relational databases. And since 10 percent is probably about image analysis, we have looked at a bit .. no we have a project that is quite interesting also where we have text analysis too. So text analysis today is very much about internal documents that we search and classify and also when you look at something that can go wrong in a machine out in the field, there is something that enters an error description and then you should try to find a subprocess as a patient can do, to diagnose the patient then you can not diagnose the machine with the help of these diffuse error symptoms	DS, DS-DSUA, DS-DS

		that occur. Trying to classify what this can be, how big problem it is and then look at comparable cases historically. So that kind of text data we looked at a little. And then image data where we shoot in the production and try to detect errors automatically. But this is how 85, 90 percent is relational data from SQL databases. And of that, I would say that ehm, at least 95 percent of it is about supervised learning. And there is very little clustering or unsupervised learning, if you know the differences?	
C.5	I	Yes, so I have heard the differences a bit before but It would be nice of you to explain it one more time.	
C.6	C	Yes, but okay, supervised learning, then we know the answer to what we are looking for, so for example it may be a classification question, is this package damaged or not. For example, this milk package is damaged or not, so it gets a label that is damaged or not damaged. So we collect lots of data where we have this label, damaged or not damaged. And then we can train the computer to distinguish between what is damaged or not damaged. It is supervised learning. Unsupervised learning would be that you only show lots of pictures and say find different packages. But then it could very well be those who are super-perfect that jumps up instead of those who are damaged.	DS-DSUA
C.7	I	Okay but..	
C.8	C	There is also clustering where you divide all these pictures into five different classes and you get the ones that are most similar to each other in the same groups. It is also something that is not used very much because it is very difficult to interpret that data, what it means, why things end up in the same group.	DS-DSUA
C.9	I	Mm, but how do you work to create value and to make the data you use in your role as valuable as possible? What do you do to make the data you use as valuable as possible?	
C.10	C	Yes, we go on a decision. What kind of decisions are being made and mostly, or really I can say that we start from the balance sheet a bit and see, where do we have large costs, where is the money in the organization, what decisions are made that play a role in this cash flow. And then in the next step, what data do we have that describes this, or is there any data around that we can use to facilitate this decision. So, that's it, we	DS-VCDS

		always assume the monetary value in decision-making and then we look at whether we have data and what we can do about it. Instead of assuming what data we have and what we can do with this. Because then it is often very difficult to find the link to value, it is not impossible, but it is not at all clear.	
C.11	I	Ehm, the areas of application of Data Science within your organization? You mentioned earlier that it is used within the organization to see inaccuracies and such. But in what more areas do you use it?	
C.12	C	Yes, we use it in every possible area. And if we see how we at *, we have customers and so in the dairy industry and packaging and food producers all over the world. So we look at customer analysis of these customers but we also look at consumers. How do consumers act in the end, how do consumers react to packaging? And since we are, we start from consumption and consumers, consumption patterns, consumption trends and then we look at how our customers act, what problems do they have in their production, which customers buy regularly from us and which buy irregularly. Is there any customer that we see risk losing or not, ehm, we look at the logistics stage, so where are the costs at the logistics stage. For example, if we have very large stocks that cost a lot, we have one, we have to deliver within a certain security of delivery, so if someone places an order, it must not come too late because then they can not produce. So it's a commitment as to how often we get deviations in delivery times and so, so we have to have a stock of raw material in order to produce packaging material. But this stock of raw material costs quite a lot of money, so we would like to have it as optimized as possible so that we do not lose too much. The same is true when we manufacture machines, so machine parts take a very long time to produce, so there are long lead times on it. And then we sometimes want to be able to place an order before we have won our order out to the customer, but then we have to be relatively sure that we have something going on so that we can minimize the risks in the entire supply chain.	DS-DSUA
C.13	I	Okay,	
C.14	C	Another thing we look at is the probability of winning projects, based on as soon as we hear about a project that a customer has, for example that they are going to	DS-DSUA

		build a new factory, we know a number of things or can find out a number things whether we have good opportunities to win this project or not, based on past projects, historical projects. Yes, but we also look at the financial side, what do we have for financial risks, what do we have for currency risks, ehm, we look within HR, does the pay system work or is it fairly set. We can also look at which employees are at risk of quitting. It is not a project I have completed yet but it is a project we have talked about that one can do.	
C.15	I	It's really impressive, it seems to be used everywhere as well. In all...	
C.16	C	Yes, wherever you have data. And right, it is important to have the money too. Because it is quite long projects, I would say that an average project for us takes, ehm, yes it is really another problem that there is not, it is very difficult to judge in the beginning how long a project will take. But the minimum for the simplest I would say is three months. And we have had projects that have been going on for a year and a half and are still not launched. And it may be then that in those countries we may be working half-time. While on these more difficult ones, we may work three four people full time for one and a half years. And it will be very expensive. Data Scientist is not the cheapest resource, there are not that many. So you have to manage your resources pretty much. So yes, that's why it's important to look at the money.	DSC
C.16	I	Yes exactly, absolutely. I am thinking about your role as a Data Scientist, what skills do you think are important to have in order to succeed as a Data Scientist?	
C.17	C	There are some basic things to keep in mind. Then there is a large spread of things that are good to have and that you need within the team but that maybe not every person needs to have. A basic thing is that you have to be able to have basic programming skills, maybe not as a developer or software developer but you need to feel safe with programming production code and be able to handle version management with Git and so on. You have to understand machine learning fairly thoroughly or need to understand anyway, you don't need to be able to build machine learning models but you need to be able to understand how they are used and what they have for weaknesses.	DS-DS, DSC

		They are usually pre-packaged quite good in the models themselves, so it's not that we build viral networks from the ground up, but you use ready-made packages. Ehm, as you develop further. But you have to understand the weaknesses and what are the pitfalls and what are the strengths and so on. You need to have a good understanding of statistics, above all you need to understand the distribution and how the distributions look, or what different types of distributions exist and how they function and why they occur. Such as normal distributions and what assumptions one can make around distributions. Often, the problem with what we are doing is the odd cases that do not look normal. It can cost a lot of money if you miss them. Then you have to be aware of it. So much is about understanding the statistics around the data, but also being able to write code that handles. So it is like the core, you have to have it. And it's like, I mean the kind of Data Scientists we have. Data Scientist is a bit of a loose concept but with us we mean someone who works with machine learning and prescriptive analytics, so we, if you speak purely descriptive, is not what our group does. Without, there's something called Business Intelligence, are you, do you know it?	
C.18	I	Yes,	
C.19	C	Yes, so Business Intelligence is not something that our group works with. We can definitely create dashboards to visualize our solutions, but that's not where the focus is. Without there are other groups within * who work full-time in producing reports and dashboards and stuff. And then we have Data Engineers who are not in our group either, but it is another group that works to produce the actual data flow, when we are going to put a model into production they work as well as moving data from a database into For our model, the result is stored and only those who have access to the data have the right to see it and so on. All the technology around managing data is someone else who takes care of it. But in some groups, Data Scientists do it, but not with us.	DS-DS, DS-DSUA
C.20	I	Yes, no it seems to be a little different how all teams are made up. But, I think ...	
C. 21	C	Yes, another role that Data Scientist may have is Business Analyst which is more about business understanding and someone who may have some	DS-DS

		programming skills. It is also something that the organization is thinking about contributing to our project, not what our group is focusing on, although of course we have to have a fairly good understanding of the business problem itself in order to solve it for them.	
C.22	I	Yes. Regarding the challenges in Data Science, what are the biggest challenges? We were into it a little before but what are the biggest challenges you see or that you meet in your job as well?	
C.23	C	Yes, we'll see. We have a fairly large Data Science group, we may be upwards of 20 Data Scientists, which is not normal yet, not in Sweden anyway. But, one problem for many is to get enough competent Data Scientists, which is a fundamental problem. Because even though you can get hold of people who can perform tasks, this combination is, it takes quite a long time to build up. Most preferably when it comes to business understanding, but also being able to navigate within a company, thus navigating with all stakeholders and all decision makers and being able to communicate the result and so on. It takes some time to build up those skills. So providing skills is a fundamental problem. But if we ignore that, the biggest problem we have is accessing data and the administration around accessing data, it's often secret. It takes quite a while before we get the data in the format we want. So data sources and access to data is a problem. I would say that maybe a bigger problem for our group is, but also for others, because we are organized as we are, so we do not have Data Engineers in our group but we are incredibly separated from them. And, that is how organizing a group has a very big impact on what, what kind of problems you get. But the biggest problem we have is what we call last mile delivery, so it is when we have a finished product, fully developed, everything is in place and get someone to use it, use the result. And to make the user understand the value of what we have done but also understand the weaknesses of what we have done. Many people expect us to come up with some miracle solution that answers all problems, ehm, anyway and that we never make any mistakes. But what we do is perhaps to be marginally better, we may perform two three percent better than human decisions do. And then it is clear that we make many mistakes. But, percent over time does a lot.	DSC, DSC-DQ, DSC-K

C.24	I	Support from the management and the organization itself, is it good in your organization and that you still feel that you get support from the organization?	
C.25	C	I would say that we are somewhere in between where we are an old Swedish industry with a lot of legacy in the organization and you have, you shit in developing. At the same time, we have come a long way in investing and building this group as we are. The fact that we are 20 people is not common, at least not in Sweden. In this way, there is an understanding of the value of what we do, but it is still the case that we are seen as a peripheral part of the organization rather than a central part of the company and how we should work in the future. So it is a strategic problem that we struggle with quite a lot, to get more focus from the organization and more attention from the organization. So that is why we have changed places now so now we are not, many are under IT, but we are under the CFO.	DSC-OS
C.26	I	Hmm, okay	
C.27	C	So we have chosen to follow the CFO because he is a person of great influence and people listen to what he says. And he is also number-oriented, fact-based, fact-oriented, so he wants everyone to listen to what we say. So we have him allied there. But getting the organization to become data and fact-based is difficult. There is a great deal of resistance, it is a very strong confidence in the gut feeling and expert experience but without putting the CFO on it it will be very qualitative decisions instead of quantitative decisions. So like, this should be good, then we do it or this I feel wrong and so you do not. Which is found in all organizations more or less, I would say that if you have a startup today then it is likely that you are much more data oriented than we are in * and other large companies.	DSC-OS
C.28	I	I also think about ethics, because there are some others who have mentioned data quality as a pretty big problem and to have good quality of the data so that it does not become subjective, that you still keep good ethical guidelines as well as with regard to everything like that too. It goes hand in hand. Is it something you have come across?	
C.29	C	Yes, no, but data quality is always a problem. I would say it's so deeply rooted so I don't see it as much of a problem, but that's how the world is. There are no good	DSC-DQ, DSC-E

		data sources really, everything has to be cleaned up quite properly and, but many who come in new are surprised by how poor the quality is and how many errors there are and, ehm, so that is definitely a problem. And I would say that 70 percent of our working time is spent on washing data and making it fit in a format that fits our modeling. So, I would say that understanding the data and washing the data is 90 percent of the job. Then the modeling itself is something that you do a week maybe out of these three months. I exaggerate a bit, but in principle it is something that is fairly quickly clarified and that is not the difficulty in Data Science, provided you have these basic prerequisites to understand modeling and so you have done it a few times so go it pretty quickly. You also mention ethics, yes it is a problem. What I would say is that it depends a little on what data you have and what decisions are made based on this data. If we have decisions that affect people then you have to be very careful about what you do, but I would not say that this is such a big problem for us. If we see what is the probability that we will win a project then there are not so many ethical positions to make, ehm, it is our customers and our competitors that we may be interested in, not individual people. And if there is any kind of bias in that model, it's not the whole world. However, if we work in HR and look at salaries or see the likelihood of someone quitting or so you become very much more sensitive and important that you focus on ethical aspects also and so and integrity issues and so on. So, it is a big and important problem but not so big for us.	
C.30	I	Ehm, if we are going to leave the boring stuff with all the problems and things like that, I think about how do you achieve success in Data Science? Are there some factors that are also important for you to benefit from these investments that you have made?	
C.31	C	Yes, I think I may repeat myself a bit, but again it is based on what are the problems we are trying to solve and what is the decision we are trying to optimize. Often, there is often some form of optimization problem. We have a resource, like a scarce resource, what is it called in Swedish now again .. an unusual resource, it can be an employee who will work with something, which projects should he choose to work with. But it can also be a machine that produces parts of the production, when to maintain it to optimize the	DSS-ADSS

		uptime, so to speak. So usually there are some such optimization KPIs in the basics that the decision is about. And it is important to find that problem and formulate it really clearly. If you have done so, you are much more likely to succeed with the project. But then, even if you have done all that, it can be that the data does not exist, it can be that you can not predict, but you are faced with a problem that is insoluble because there is nothing in the data that can predict what you are looking for. Ehm, so having the right data collected is also a success factor and being purposeful when you start collecting data. But you have to do it on a fairly long time horizon, because you have to come up with a critical mass before you can solve problems with machine learning and it has to be representative over time, at least one year you have to collect data in order to start solving problems in general. way. Obviously, that's not true of everything, but I would still say that .. you have to have enough data to represent the entire problem space at least that covers all possible cases that are common. Another success factor I would say is .... yes but to be clear in their communication with the customer, to explain results but also problems along the way and to communicate in a way that the customer understands is A and O because if they do not understand what we deliver they will not use it.	
C.32	I	I believe it is, or does it at least feel like there are, often pretty high expectations of what to achieve when investing in Data Science?	
C.33	C	Mmm..	
C.34	I	Do you think that you usually live up to those expectations or if you in the Data Science team have completely different expectations than the organization has on these investments? Or, how do you look at it?	
C.35	C	Yes, yes, but it is always the case that the organization believes that, or first, the organization always thinks that the data is good. They always think the data is great and in perfect condition when they deliver. If they say that they have bad data then it is probably really very bad. It may be that they have good control over their data as well, but immature organizations always believe that they have great data. That's the first thing. The second is that they think the problem is solved by inserting numbers at one end and then an answer comes out at the other end and that the problem	DS-DSE, DS-DQ

		is solved. So often it is about having a process together with the customer where they get to take part in the whole development process and understand the complexity of the result and explain what they can do and what they can not do.	
C.36	I	Is it something or some measures that you feel that * should put in place to succeed even better in your investments in Data Science or in your work?	
C.37	C	Yes, I would say that one would be much faster when it comes to resource allocation. That * as a company sets a budget for next year, yes you may start budget work in the summer and then you set the budget in the autumn and then the whole next year is run. We cannot get any new resources added, so we are a very heavy organization. And not only does it affect us, but we try to work very quickly and easily. But being agile in a big old organization is difficult.	DSS-ADSS, DSS-OR
C.38	I	If we consider your team and in your work with Data Science, is it something you can influence or something you can do to succeed even better?	
C.39	C	Yes, I would say that we needed to be better integrated with our Data Engineers so that we had, because it is also a scarce resource, they are in the IT organization, the few we have and they are always overloaded and that is a problem. We need Data Engineers, we need Data Scientists and we need Business Analysts who work very closely together. And I would say it's something that doesn't work well on *. And that is largely an organizational issue.	DSS-ADSS, DSS-DSR
C.40	I	Yes, we are beginning to reach the end here. Is there anything more we need to add Anna Charlotta?	
C.41	C	No, I think we actually got answers to most things actually. It is very interesting to hear. We have read a lot of literature on how it works in theory but it is interesting to hear how it works in practice as well, it is cool to hear.	

Appendix E

Interviewee D

Interviewee: Interviewee D
Title/Role: Data Scientist
Date and Time: 2020.04.17

I: Interviewer
D: Interviewee D

Reference number	Person	Questions and Answers	Code
D.1	I	But you can start by telling us what your job entails, like what's your work and what does an ordinary day look like?	
D.2	D	Yes, the part of * that I sit on is called * and our company consists of several different business groups and several different companies. So the part that I sit on is the part that designs all products, there are also those that consist of Supply Chain where you source all products in the entire value flow chain. And Range and Supply is a wholesaler who then sells the products to the part of * who runs the department stores. And the part of the company that I sit at is called * and we are based in * and there are all these product designers and those who design the value flow chain and we sit in a support organization so there is a team called Advanced Analytics. We are a group consisting of a mix of Data Scientists and Data Engineers and then we also have project managers. We work project-based to the extent that the part of Data Science that we work with is most linked to business processes, for example we may need to answer "do we have the right products in our range" and then it is in relation to what. And then you can of course ask those who have worked for a long time * whether it is the right product or not. So it can also be something that is close to our heart, for example "do we have products in the right price segment" and then maybe it is more from a strategic perspective where you want to understand how price sensitive our products are. We have strategies where you say you want at least fifty percent of the products to be or be perceived as cheap. And it is of course subjective depending on who you ask, like cheap in Norway may not be the same as cheap in another	DS, DS-DSUA, DS-DS, DSC

		country. But what we do then is that before we do a Data Science or Advanced Analytics project there must be a problem, a business problem. So we don't sit down and do any Data Science solution, because we think it is exciting or that there is some interesting technology that we want to use. Instead, it must be a need. These needs often consist of wanting to answer a question, for example wanting to answer something complex, such as how things work together. Then you may need to build a model so that you can make a model that is explanatory. We sometimes do. However, it is often about wanting to explain something, but also anticipating something. A classic example is anticipating sales. So we do a lot of forecasting and you might want to build different scenarios, and then you can use Data Science. Another thing could be that you want to automate something, then you may also need to build a Data Science solution, something you do manually you might want to automate or do it more efficiently. It could be, for example, that you get something that you handle manually but if you can create a logic that is data-driven or rule-based then you could automate it. Another thing that we also work with is pure optimization problems, for example what price do we want to put on a product so that we can sell as much as possible or if we want to minimize costs. And the way we work is that we have been allocated in these projects so you have a certain number of Data Scientists that you divide and then in for example in my case, I have worked as a Data Scientist for just over two years in the role I have. And then I have a project that I am signed to and then we also have other resources, such as Data Engineers, Solution Architects and the way we work is that we try to quickly build something that can show value to the business owner, so all the the products we build, we do not build a search engine or a recommendation engine that, for example, you as customers use *. Without all of our Data Science projects, it is basically about supporting any person or system in this business process. It may be product designers, it may be someone who has the responsibility to put a price on an article, someone who has some planning function and so on. So that's the kind of business problem. And there must be a clear business problem, if there is, then you contact our department. There are many requests, so there are many who ask "hey, can you use Data Science to solve	

		our business problem?" and "what can you do?" and so on. Sometimes the question is relevant and may be worth a lot to answer but there is no advanced component in it. Sometimes it may be that there are five lines of SQL code where you get a mean or something and that we can help them with but we want to work with something where there is some kind of advanced component if you say that there is any kind of model. So much is about understanding, sometimes it is for example sales forecasting or you want to predict the sale where the question is obvious or you know that we want to be able to understand something a little further in the future, for example how can the salesman look with some uncertainty but it can also be a bit more unspecific business problems. For example, it may be about which product mix should be where. And then it is about formulating the business problem or sometimes the business problem is more of an opinion that is a bit fluffy and diffuse so then we spend quite a lot of time understanding the problem and how to reformulate it into a Data Science problem, that is to say that it there must be some form of data and some form of math. So maybe we should maximize some function or something. And then we can wonder if we have done this before, sometimes we have done it and sometimes we have not done it. And then we also work like so that some problems we manage to do ourselves but sometimes there are complex problems that we may not have any experience around and we have a limited team. We currently have three Data Science seats and about three, four Data Engineers, Solution Architects and then we have some project managers and then we also have Business Analysts. We have more projects than we can do and some we could do ourselves but we can't. Sometimes it may be that there is a very specific need and then you need to bring in someone from outside who helps. So we work project-based, we may have a team of between five and ten people. There may be three people from this Advanced Analytics team that we sit in and then we have representatives from Business who are representatives of those who own the problem or process. So let's say that there would be some forecasting problem so they understand the business context, they are included as objects in the project, that you follow up often and try to make sure you are not sitting for three months and try to develop an app and then it turns out that it not at all this was what you have. So at a high level it looks	

		like that and that means there are several parts to it, partly to turn the problem into a Data Science problem and the whole point is that we want to use data to answer this question and then it's about maybe create or use an existing algorithm that can then be used to map this problem to the data and possibly create a forecast or model where you might generate coefficients which you can then use as explanation. So we work in teams, quite small teams. We also have a methodology that is inspired by software development, agile methods and a method called Scrum so Data Science at * is about having some kind of knowledge in math and statistics, you have to be able to write code and you have to understand the business operations. It is not traditional software development, but we are very inspired by how you work in software development such as agile methods and Scum and so on who come from there. Then it is the progress we make step by step and we work hard to try to get the data into a format where it is possible to do something about it. So we spend quite a lot of time, we have large amounts of data that we collect mainly from the Supply Chain. However, it is not always a design idea for you to put it into an analysis tool, but you have to spend quite a bit of time understanding the data, combining it, clearing the data and gradually improving it so that you can use it to answer the question.	
D.3	I	Would you say that it is the task that takes the most time in your daily work?	
D.4	D	It depends a little, sometimes it is. Now it is that we have been doing this for a couple of years, some of the problems we work with are well known where we know what to do, but some of the problems are a little more diffuse. We're pretty new to this at * so we've been doing this for a couple of years, we've come a bit. But we are not like, for example, Spotify, which has the majority of developers who have a single mission that is aimed at optimizing this recommendation engine, it is more like a well-oiled machine where basically everyone is a developer and music enthusiast. So we spend a lot of time discussing with stakeholders, explaining things and what can be done. We spend a lot of time kneading, searching and clearing data and then we spend a little less time on the fun, that is to do these models, get a value and then we also spend a lot of time trying to explain the outcome and follow up.	DS-DS, DSC

		Some of the stuff we do, we notice that this seemed to be great but it probably wasn't that good anyway. So the difference between Data Science projects and traditional engineering projects is that you can see the benefits of it, how to build something. And you can see that no, but this will be useful. But if, for example, you have to foresee something or you want to optimize something, then maybe we spend a lot of time and then it turns out in the end that no but this was not so good, maybe we learned something from it. So it is sometimes difficult sometimes to know the benefit and then very high expectations from the business and management and thus Data Science is the new holy grail and this should be the answer to everything.	
D.5	I	Would you say that the expectations that exist within the organization and from the management that it fits well with the expectations that you as Data Scientists often have? Or does it usually differ?	
D.6	D	Yes, I would say that it differs quite a bit. It's hard to know what you can get. There are some problems, for example, to predict sales, that every company wants to be able to do. So you can plan what the sales should look like, it is much easier to plan your flow of goods and so on. Then comes a Covid-19 or something that allows all forecasts to be thrown out the window. Some problems, I think, are what the clients and these managers understand, others they do not really know what they can get out of it and above all they do not know how long it takes to do something. For example, to optimize a product portfolio, it is not tribal and especially not if you have done it before. So I would say that expectations are often high and you may not know from a business leadership perspective what you can get out of it. You've heard a lot, you've heard a lot of buzzwords, and you've seen what Google and these cool companies are doing so you expect this can't be that hard. But I would say that part of our mission is also to try to explain what Data Science means because it is a fairly broad concept and it is not always very easy. So our mission is also that we try to educate and we try to explain, we do not try to do Data Scientist by everyone but we try to explain what you can do, what does it mean when you build models, how do you handle this with predicting something, what means uncertainty, one can measure it and so on. So I would say that there may be a mismatch between expectations from the top of what can be delivered.	DS-DSE, DSC

D.7	I	The biggest or the main challenges with Data Science, what do you think that is?	
D.8	D	I would probably say that there are many problems to solve. And it depends on what you can do, which problem we most benefit from solving. And then it is about utilizing the resources you have, like resources in Advanced Analytics and Data Science in the best way. There are always problems that could be solved, but above all I would say that a major problem is the lack of people who may know this and who have experience with it. So, for example, we have had some problems when we have tried to recruit that it is difficult to find people who are good. Then there are many in the academy who are well educated, brilliant and extremely smart and they often have good knowledge of the latest Data Science techniques and methods. For example, I was studying for a long time, I studied macros and statistics and then there was nothing called Data Science. Since like five six years ago, all Statisticians changed title to Data Scientist while now there are courses that are focused. So, I am absolutely convinced that many well-educated smart people will become very capable Data Scientists. And what we and many others are doing now is that, of course, you want to find that rock star from Google who has been doing this for ten years and who knows everything, but they do not grow on trees and above all they are very difficult to recruit. The shortage of resources means that you want to recruit from the outside but then you compete with all other companies so everyone like Spotify and King and everyone searches for Data Scientists. Then you try to hire people directly from the academy and then you need to teach them a little about what it is like to work in an organization. If you come directly from the academy it is a little different to land in a large company and then you need some time to get warm in your clothes. And then there is a third thing that we try to do which is that we try to educate internally, analysts and so on, who are interested in learning more. So I would say that the lack of resources and how to deal with these problems is the biggest challenge.	DSC, DSC-K
D.9	I	Yes, we have seen some before and we have read literature before that sometimes there are quite a lot of problems regarding ethics, subjectivity, lack of data quality and so, is that something you have experienced?	

D.10	D	Yes, I guess I just have to admit that this is part of data processing. That said, we work mostly with internal data. Those in other parts of * work more with customers and gather customer information, such as unstructured data, Twitter feeds, manage images. In our case, it is more about structured data that already exists in systems, but the data is not, when you have collected the data you have not had Data Science in mind. So yes, data quality is a problem and it is big. Most Data Scientists are good at data, but they may not understand what they are looking at. How it is collected, in what context.. We have a lot of tables with data from different parts of the value flow chain and sometimes it is very difficult to know what it is you are looking at. Even though you have been working on it for a while, you learn every day that "yeah, I had no idea about this". So yes, it is true that there are challenges with the data. Then I think there are even bigger challenges if you don't have that much data and you need to collect it, and then you have to think about this with a measurement system, how it is collected, is there something that makes you need to review the source, can you trust the source and so on. So there are many Data Science projects where you might use public APIs, you collect data from different places and then you have this question if you can trust the data. So data they say is the new oil, and that is true. But when just collecting data and not knowing what to do with it, I think also, most organizations are there at some point.	DSC-DQ, DSC-E, DSC-K
D.11	I	Yes exactly, but yes we were in it a little before with the expectations of the organization and so on. Are there any challenges with just support from management and company culture and what you have encountered when working with Data Science?	
D.12	D	Yes, I think that's a great question. I think in our case it is so that you believe in Data Science because it has been a hot topic and you are interested in it. But it can be a bit, often we notice that the development when we, for example, start a Data Science team or similar, so it is not so that you can start spitting out useful data products on a continuous band and the development may not be linear. So in the beginning it might be difficult, we work for example and many other companies work on having a problem and having many different phases. For example, you have something you call POV, a Proof of Value, where you	DS-DSE, DSC, DSC-OS

		say that okay we will solve this problem. And then you take a limited part of the scope, that is to say you have a hundred products, but we can say that you have two products so you try to prove that this approach is good. Then you evaluate and to do things like so-called POVs there are usually no problems. Most organizations usually manage to do this and say that this is good, we should do this. And then to scale it up and often it is for our part that when you have done these POVs, you might be working locally on your computer and you might be sitting and getting something in quite a short time. But then if you take the next step in trying to implement it, maybe it is then that you want to build some solution that dozens or hundreds of employees should use. So some people say that it is often more complex to build it, it takes more time to maintain and it is something that you often come into after a while, that you want these models to spit out some kind of result tables. That they roll on automatically. And then maybe it will be more a software engineering perspective. So I think the expectations of how difficult or easy it is to build things, that there is a little mismatch. And sometimes, even as I said, you do not know what the benefit will be. So I think many executives have said in many companies that we should invest in this and this is good, but if you have to do such ROI calculations, return on investment, then what do we get out of it if we employ ten Data Scientists, what do we get. I think that is uncertain. And then some people choose to employ a hundred people and others choose to no but we take a few. So it's always about balance.	
D.13	I	Yes exactly. We did also discuss knowledge a little before and that you felt that there are many who are not right or that you do not have enough resources. Is it something special that you see that people who work as Data Scientists or apply for, that there is some knowledge that they often lack?	
D.14	D	Yes, I think so. I think there are many good analysts but I think if you think of this kind of venn diagram where you have areas that overlap. So I think you need to have several things. It is very common that you may be a software developer and that you may want to go a bit more to the Data Science-group and then maybe you have some problems with math and statistics. But it is also common for people to perhaps come from businesses that are quite good at analyzing things, they	DSC-K

		can do the business and they may have worked at the company for a long time and can do a lot, but then maybe they lack this piece of software instead. So that is where we are concerned that those who want to become Data Scientists with us need to learn how to program and that type of programming is often a little different from traditional programming. So I would say that traditionally there are knowledge gaps, there are more people who are good analysts than those who can code. So I think it's the mix that you now offer at Data Science courses where you learn a little bit of everything and then of course there are different parts of the university and depending a little on where you come from. To know a little more than most people in one area is, so you do not have to be the best, then of course everyone wants that unicorn that is the best on everything and they are the ones who stick to Google and you will not get it.	
D.15	I	But success factors, what factors would you say are the main success factors for companies to succeed with Data Science?	
D.16	D	Yes, that's a great question. If it would be an easy answer it could be sold to everyone. But I think there are different ways to do it. One approach is to try to find someone who has done this before. If you have done it a hundred times and you can do it in your sleep, there is no problem. The problem is that not many people have the expertise. And I do not think it is enough to hire only one Data Scientist because then that person will quit because they realize that this is no further, no one wants to work for themselves. So you have to look a bit at what it looks like, if you have resources, but personally I do not believe that you have to hire a whole army directly but you need to be a few at least. You should be happy to find someone who has done this before. If you find someone who might be able to lead it. And then, don't try to solve all the problems at once. You need to set up a team where some have experience and some who have a willingness to learn and then you need some time. Then there is always the question of whether to centralize or to decentralize. After all, there are organizations that may have many departments and so there is a Data Scientist in each department. Or it is as with us that you are centrally located and it probably depends a little on the challenges and what kind of business you have, but I think you need a fairly long-	DSS-ADSS

		term investment. So I think you can't do much in six months so you can probably think a couple of three years ahead and then maybe tone down, you will probably overestimate how much you find to do in the short term but then maybe you underestimate how much you will learn and how much you will get done in three years' time.	
D.16	I	Absolutely, but the expectations you often have on your Data Science projects, are they usually achieved or what does it usually look like?	
D.17	D	Yes, I can say that at the beginning of the first few years we thought that things would go faster and that we could say more. So what often happens is that when we do such a POV you say that this is great, we can add this, expand the scope so that there will be more data and then we will be ready in six weeks. And so it turns out that no it was not possible, but then you have to scale it up another problem that you have not thought of. So now I think expectations are more realistic, we probably have higher expectations to achieve more in the beginning. And now we have learned how much you can do and that is what you do when you have done this for a while, then you will learn that it takes a little more time and that it is a little more complex.	DS-DSE
D.18	I	Are there any steps that you think you should * take to succeed even better with Data Science than you do now and reach your expectations even better?	
D.19	D	What did you think of action then?	
D.20	I	Well, measures to reach your expectations even better and to reach your expectations even better, is there something that you as Data Scientists * even better in your job?	
D. 21	D	Yes, it sure is. I would think we would need to be some more and to find a good mix of people who complement each other. So sometimes it may be that it tends to be men with glasses and plaid shirts and I think they might be good at doing something but maybe a more mixed team and I think of course, everyone wants more resources but I think to bring in someone who may have done it before and who can share so that you can avoid these biggest mines that you are probably going for at first. So that would be the case.	DSS-OR, DSS-DSR

D.22	I	Yes, no but this is great information and we have really covered a lot. Is there anything else you are thinking about that might be relevant for us to know or something that you have been thinking about?	
D.23	D	It probably depends a little on what the ambition is. Have you thought about how you should, so to speak, package this and a little about the method. Do you have any hypothesis or is it more unpredictable research?	
D.24	I	No, but we have seen that there are certain challenges and from what we have read it is mainly within ethics, knowledge, data quality and also corporate culture and support from the organization. But then we have also seen that there are few Data Science projects that really succeed. And this is where we want to try to find what you can improve and what you can develop to succeed more in your projects.	
D.25	D	Yes, but I suppose that's right. I have also heard that very few projects succeed. But the problem is also that most things don't work. So most ideas that seem good when you sit in the car on the way to work and you come up with something on the highway and think that this is what we should do. Then going from there to action can be tricky and it turns out it was not so easy. So obvious things, if it was easy to do, someone had already done it. But then I think you need to go on some mines and what you can do to avoid and go on unnecessary mines is of course that you meet people. So before the Corona outbreak we went to conferences and talked to other similar people who have had similar roles. And then you notice that maybe not the most, but that many people have similar issues but in a different context and maybe in a different maturity. That's when you might come across someone who has already built their company from the beginning on Data Science and then maybe you meet someone from Ericson or some engineering company and then maybe you meet someone from some cool company like Zalando or something like that. So they have similar challenges but different context and maturity where some have come very far and then you can get some tips, then maybe they say that "Aha, but we discussed the questions you are talking about two, three years ago and then came we got to this ". But you also notice that there is no easy way to do the same as anyone else, but then you have to bring it to your own organization	DSC, DSS

		and think about how this fits our context. But no, it's not easy and I believe in testing things and failing that you try and then you learn something from it. And so you try to avoid making the same mistakes several times.	
D.26	I	But do you work much like that, to * try to share experiences with others in the same roles, but at other companies?	
D.27	D	What we do is we meet within * and there are many different companies within * so we have internal Data Science conferences like this. Then we also go to conferences and meet other companies so we do and it is extremely interesting. Above all, you want to see what is the latest out there and then it is inspiring, then it can be difficult to make concrete collaborations with other companies because of legal stuff. You can't share data with another company anyway, so exchanges and collaborations are there from a conference perspective. So there it is. And it is mainly to inspire and to build networks. So we do and you want to do this because you learn a lot from others.	DS
D.28	I	Well, I also just thought a little bit about what you think about, like who should introduce measures to be successful with Data Science, is it you or your team sitting with Data Science or is it more someone in a management position who should make sure everything works out?	
D.29	D	I think it depends a little on where you sit in an organization but often it is so in larger organizations that you have split the responsibility. And if you have chosen to put these resources into a support organization, then it is their role to support core business. So it is possible to operate and design products without us, it is possible to sell them without us and so on, while without those who design the products it is not possible to do this. But then it is that you have to have a clear responsibility and a clear vision and a clear goal, what is the purpose of this, after all, that they should support each other. But it is not anchored from the top, because it also costs a lot of money to hire Data Scientists and then you have to have it anchored in some way and from the top you have to have some form of directive in some way. Then I think this team working on this needs some freedom to implement things and test so it is both and.	DSS-OR, DSS-DSR

		So I think you need to have support from the top. So if you have a manager who thinks that this with Machine Learning and Data Science is just rubbish, it will never work, but nobody really thinks about it like that now, it is more about how much you want to invest. But also what are the expectations, what to do. And then it is important that you in the organization, we are very much for us to go out and meet them in the core business and show that we do this and then they wonder if it is we who will automate their job but no it is not what it is about but we want to help with tools. So you need to get responsibility but then you also need to be able to deliver. And then I think it is important that you get some projects, some light profits so you can show pretty quickly that they in this team are good and they do what they do. So you don't sit locked in your bubble and sit and code for five years and then you come up with a 3D product like no one, where people just know what you are doing, this is a waste of time. I am not saying that it is happening but there is a risk that, so I think you have to have it sanctioned from the top and then they also have to make demands and say that now is the time to start delivering. So you have to start and try to meet somewhere on the road.	

Appendix F

Interviewee E

Interviewee: Interviewee E
Title/Role: Data Scientist, CTO
Date and Time: 2020.04.20

I: Interviewee
E: Interviewee E

Reference number	Person	Questions and Answers	Code
E.1	I	To begin with, Could you tell us a little briefly what your everyday duties are?	
E.2	E	Yeah, right, if I start with...um.	
E.3	I	Because you're CTO too, right?	
E.4	E	Yes, for my job at*, at their Data Science division one can say where we work with, where we have a number of Data Scientists, Data engineers, who work. So my daily work, it's pretty broad. I work with about half of my time practically, hand-on in the projects I work with. And about half of my time I spend on coaching the new and more junior employees, making customer presentations, talking at conferences, uh, well, just going around and talking to different customers and companies etc etc. Um, well, a lot of discussion and a lot of what to say, a little more advisory position as well. Um, kind of. So it's pretty broad, what I do kind of.	DS-DS, DS-DSUA
E.5	I	Do you work with the technical aspects of Data Science?	
E.6	E	Yes, exactly, so you can say ehm, who said that about half of my time I put in the practical work of working as a senior data scientist in the projects we do then. But uh, then I have almost 20 years of experience now, so it's like then it's easy to put these puzzle pieces and understand yes but okay you do that over there then it won't work with this. So then you have to do something to make things fit in and so. So that, a little more overall perspective too. But as I said, much	DS-DSUA, DS-DS

		of my work is about actually doing general Data Science tasks.	
E.7	I	What kind of Data are you using?	
E.8	E	It can be just everything between heaven and Earth. We're doing a lot of imaging right now, with image analysis. So automated image analysis. We work a lot with sensor data, so sensor data from connected IoT devices and, we also work a lot with ERP system data, as well as classic like this, invoices, purchase orders, that kind of stuff, it's pretty sprawling. After all, we are a consulting business, so to a very large extent, the data we use is related to a customer-specific problem so it can be very different as well.	DS-VCDS, DS-DSUA
E.9	I	When you are out and advising on other companies and consultants as well, what kind of problem companies need help with then?	
E.10	E	Um, it's very different. It depends very much on the degree of maturity one might well say. But one discussion we are in right now is that you should try to get everything to end up right and help the customer understand what it is they can and what they can not do. Yes, firstly, only what I came from here before the meeting, one of the major airlines in the world was discussed. Ehm, who needs to work more with health screening for passengers who step on board the planes as well. You want to be able to identify the right people so to speak that you should not let on board the plan, etc. And it is a huge problem, and I am not the expert on everything around, but there are a lot of regulations that passengers have bought a ticket so you have certain rights as passengers, as well as being on board the plane given that certain conditions are met etc. so that then it comes in there with the Data Science problem today to work Yes, the questions then how it relates to when they own the right to get on board the plane if I bought a ticket, etc. and also the technical details, what reasoning, what conclusions, can be drawn at all through such kinds of health screening and how the probability affects the specific business models in the whole. So those are the kind of exciting reasoning you get to hear.	DS-DSUA, DS-DS
E.11	I	I was thinking, these companies or customers that you're out at, Would you say that your expectations of Data Science compared to their expectations of how to	

		help those with Data Science, would you say that they match or is there any gap there?	
E.12	E	Ehm, there are often very large gaps. I think I have to say that. To a very large extent, Data Science and AI, the whole AI, is very popular to use at the present time, and it is like, a little bit that magic box that we do not really understand but that we hear everywhere from that it is so amazing and it accomplishes revolutionary stuff. Then we want that, want to buy that AI box or the Data Science box. So please can't just sell Data Science to us, sell AI to us. There are very large gaps. And it's a bit where my advice is connected, to take the customers from that initial, We don't really know what we want but we understand that we need AI, to actually crystallize this into something that can become much more concrete. So that's the kind of typical thing I work with. But definitely, so there are very very large gaps, almost a bit in an unfortunate way. I have come across extremely, terribly talented mathematicians in many industries, ranging from logistics to banking / finance / insurance as well as yes but we use these and these techniques, super advanced and sophisticated as well as mathematical methods. Who uses very much probability, just the same as in Data Science as well, but who says, this with AI and Data Science, we haven't looked at it now, but that we believe in, I don't really know what it is, I haven't put myself into it completely, but there, it's going to revolutionize our world. You're working on this daily, this is what you can do, and you don't understand that that's exactly what you're working on over there. And it's somewhere, then there's been some feedback in the whole, throughout the AI and Data Science World, and it's a bit problematic then I can feel.	DS-DSE, DSC-K, DS-DSUA, DSC
E.13	I	Yes, indeed. How to solve such kinds of problems as well or what do you think is the best way to go about dealing with it?	
E.14	E	Yeah, that's a great question. I am not sure there is any universal way of solving it, but I would say that the first step is actually to have, a sober debate and conversations that are a bit disconnected from both the hype and the sales discussion, as well. And a little bit drop this trying to put a stamp on it. Yes one of the most interesting meetings I have been at was then at a financial institution where IT, so sat and built giant	DSS-ADSS, DSS-DSR, DS-DSUA, DS-DS

		interesting machine learning models, yesyes exciting but how does it relate to the other mathematical activities of the bank as well. Njae, they don't do anything like this, they just work with statistics. And me coming from a mathematical, statistics background, machine learning and statistics are just the same thing as well. Sure, there are some concepts that wouldn't be called machine learning and there are machine learning stuff that couldn't be called statistics. But it is like the extremes, the overlap I would say is 95% exactly the same as. So that it is this way that you try to put a lot of stamp on it, that is this, it is not that, as long as we remain in that, I do not really think that you are going forward either. It's the same thing with Data Science and AI, it's so fuzzy concepts, you can change them, you can make them include everything. And then the definitions, then as well as the definitions do not mean anything either. So that the faster you realize it, the faster you can realize as well as then that if Data Science is revolutionary, but the very term Data Science in fact does not mean anything, what is in that case that is so revolutionary? We take two steps back. It may be that we may always be better at working with our data, we may always be better at using probability-based and mathematical techniques in how we use our data to get better. Then if we call it AI or Data Science or as well as logistics optimization or risk analysis, it doesn't matter.	
E.15	I	Mm. I was thinking about this you mentioned with, you work so much with data, is it something, Have you come across that there has not been high data quality in customers for example?	
E.16	E	That's always the way it is. It's also an exciting area. Data quality in general is a huge problem, so it is. If you look at the people who are newly educated and who come to us who may have studied statistics and mathematics, etc. and have received a lot of training in these techniques you work with these datasets, etc. then you get out into the real world and then you think this okay these datasets, I just can't do anything with this based on what I have learnt in school. It's the same thing there, it's about taking two steps back and actually starting to wash and clean and structure this data and start doing lots of assumptions. For example, seeing here we do not have data so then you simply have to assume that it looks like this for example. So that yes but absolutely, one usually says that	DSC-DQ, DS-DS

		somewhere between 70% -90% of the time in a data science project goes to and goes out on washing and organizing data. Yeah, that's it. That's it. that's it. Then, it's the same thing there. You might ask yourself, what really means poor data quality. It's not like this that someone in the purchasing department sits and tabled junk characters in fields in SAP just because it's funny. Most often there is a reason why it becomes this way. One assumes, for example. that a customer who resides in Sweden has a Social Security number, but if we then have a customer who resides in Sweden but does not have a Swedish Social Security number. Well, what are you doing? The system sets a Social Security number, yes, okay, 0000000000, alright. Yes, but there will be junk data, poor quality of the data. From some point of view. On the other hand, there is plenty of information in it that can be used. So it's the same thing where that poor data quality as well is easy to somehow, here is poor quality "point". It's the same thing there, it's a giant world with a thousand different things that could happen as well.	
E.17	I	Snce you have the title of CTO and a little more management perspective on it, are there any competencies that you consider or value most at Data Scientists? Or that you consider most important in all the work that you do?	
E.18	E	Well, that's a difficult question. Is there something I value more hm. It's really hard, so there are, when it comes to Data Science and Data Scientists, there are some fundamental areas that I, like, assume if I discuss with a Data Scientist, I expect that person to know what a vector is, that person knows what maximum likelihood optimization is. Like, there are a number of things that I would like to say are base factors. But then, then, I would say that since what I'm trying to accomplish when I do a lot of recruitment and stuff, I'm not a manager or anything like that, like no staff responsibility, but what matters to me when I look over my colleagues and the team that I have then, it's that we ensure a breadth in the skills. So that I, I want us to have people who come from very very technical as well as very mathematical areas, ehm who can really really go in depth with some stuff, but also have economists who can somehow understand the great brushstrokes for how things fit into society and I want System scientists who understand how things	DS-DS, DS-DSUA

		actually work in a business or I just want to make sure that we can actually build all of these perspectives, and so that we get a really wide range of what we do because I think Data Science and AI are like, because the area can be made so extremely wide and so extremely large then there is also not like a person who can know everything. Or this, if only everyone has this skill, then we have the best team in the world. I kind of Don't believe that. It is a cooperation and you have to twist and turn this around from very many different perspectives. I believe that. So I probably wouldn't say there is, something more important than anything. Possibly, communicative ability. To be able to describe what it is you do ehm, and certainly the risk is that it gets a little snowed in when you get into the technical details, and somehow then be able to lift your eyes and to explain to a CEO or a purchasing manager or financial manager, or whatever it is for something, this analysis we have done or this model we have done it works like this and describe things in an understandable and easily understandable way. Possibly. But at the same time it is nothing I expect that there is something everyone must be able to, if at least there is someone so is good at it then you always have it to fall back on.	
E.19	I	Ehm, at companies and customers you are with, is the organizational culture or that you have some kind of digital mindset, or management support something that you have encountered any challenges with?	
E.20	E	Yeah, um. Yes, I absolutely do. And, it's a little bit again this, Well, I might say that there are challenges in two different directions maybe. On the one hand, it is the challenges where you see, take and look at, say, the Swedish industrial companies based in some kind of Swedish engineering, etc., sophisticated engineering techniques if you say so then. And it is clear that engineering in some sense is mathematical based. It is somehow where it has its core. Okay, so this is how we say that now we have some digital team here to work with AI and Data Science. They can go in here and build advanced models for the entire business. They can go in and start building, I don't know, models that describe when machines need to be maintained for example when they should be stopped sleeping. And that's how you say all right, we use techniques here from NASA and a number of other major, state-of-the-art techniques in engineering. And	DSC-OS, DS-DSUA, DSC, DSC-DQ

		then you come here and tell me that you will do magic here with your AI and Data Science. Um, and often it's like you're undercutting how advanced you still have it over here. Just predictive maintenance is just one of those things, for example. Maintenance engineers and machine engineers have very sophisticated ways, they do it correctly, to have very advanced and sophisticated ways to count on planned maintenance and to do maintenance optimization. When you somehow come in from the side and say yes but we do not know about maintenance optimization, but, we are good at Data Science and AI so just input a lot of data and so we get out a magic answer, then the magic answer will be, it will kind of be completely useless. And there's, like, the dynamism that a little bit crushes. The second thing that I can see is a bit the other way around. Where you're very, very immature. Among other things, I encountered one, I was going to coach a girl who came in as a Data Scientist on a company like and ehm, she sort of had a hard time finding a really good Use Case, to be able to take on this digitalization or innovation part as well. I kind of said, what's your data to work with. So she answers Nah I have no data yet, no data had started to generate yet, and did she get any data, there were only seven lines in an excelark with someone who had written some notes as well. Approximately at that level as well. And it's like this, but the CEO would've sort of said You're like a doctor of mathematics so come in here and work wonders. We have understood that Doctor of mathematics with a focus on machine learning as well as it is, you will do magic. But if you have no data to work with and if you have no connection with the business then there is like nothing to do. So to say that there are many challenges around these pieces, definitely. Ehm, then like this an observation is well that it may be a bit, often a bit too much Playhouse on these innovation and digitalization devices as well. Um you focus more on doing fun and advanced stuff. It's more focused on doing as advanced stuff as possible, rather than getting things done, even though they're simple and actually delivering a lot of business value as well. But it's more of a reflection, perhaps.	
E.21	I	Well, what I also thought about was, we've read some literature about this, too. And have read some about ethics and subjective data, partly at the data scientist once it should take out or select the data or the data	

		itself that is inside the system. Have you encountered any problem in it or is it something you consider when working and so?	
E.22	E	Yeah, but that's definitely the way it is. Um, just ethical issues, we're working on it. And like this, in some cases it's not quite easy or obvious, ehm we work a lot with construction companies among others, where we work with image recognition in camera streams, to find people who make out a helmet or people who are in dangerous areas and things like that. This is a little insight then that I myself had and then it was like that we started working with the case like this and ethically, this is just ethically good, we are trying to protect people here. There really are no ethical problems here then. Then you start thinking like another step. And all the training materials that we have used to train these models have been with the construction workers who actually work at the construction site right now. And where one realizes, all the models are based on those construction workers where everyone is white middle-aged men. Which means our model will be the best at recognizing white, middle-aged men. So if suddenly a young, colored girl who works at the construction site starts, we will be worse at detecting and protecting her which is, of course, a little ethical dilemma. So then the question is, well how do we solve it? It's not as simple as telling that construction company that yes now you have to go out and hire young girls who are going in and work at the construction site because we need training materials. It's never gonna work. Without one might think about whether you hire extras that start wandering around as well. It is not quite easy questions as well, then you have to start to ponder and apply technical methods. The airline that I mentioned just now is also such an example. Where ethical issues are encountered. Yes but if you develop a screening to be allowed to board the plane because you do not have a fever, what to do with false positives etc. what happens to it, we will potentially stop a random number of other people who were not allowed to board the plane then, is it right or wrong etc. For it is clear that it is a test that catches most of those who are sick to protect yet those who are allowed to board the plane, we protect those more than those who are sick then. So that ehm, there are issues definitely. But most often there are no easy answers to the questions but it is reasoning one has to twist and turn on.	DSC-E

E.23	I	Um, yeah. Regarding how to achieve success in Data Science, are there any factors you think are important there? Or that affects the success of Data Science more or less?	
E.24	E	Um, yeah. I would say that one of the absolutely most important factors I come across is the understanding that Data Science is not, or AI, is not an exact science. Um, being like being able to, um, I have some good examples. If you have a person doing a work, it's also not perfect, but if something goes wrong, then you can always go to that person and say this, well, but when you did that, you did it wrong, do not do it again and at worst even dismiss that person. A little bit roughly put, but a little bit like that. There is always someone who somehow performs something. If you instead run this from models, AI models and Data Science models, then it is clear that there is no one the same way you can say this is your fault as well. Things just happen. And as I said understanding that, let me say this. If you take the data science glasses on, but still that it works. Self-driving cars are a great example when you see okay if a person drives on another person then it has the driving car, it was your fault, you made a mistake or whatever it is. And then as a society you can say it was a mistake don't do it or okay it wasn't a mistake you get a punishment. As soon as we have AI models doing this, it's not quite possible to say the same way as well. Then it is extremely important to be able to say, Well, what is it we compare to. We can not compare with 0 power. We can not say that in the best of worlds with self-driving cars there are 0 accidents there never happens anything. It's kind of what we compare to. That is not the way it is, but you have to compare with something. You have to assume that even in a world where people drive, we have accidents, okay but self-driving cars. Are they better or worse than humans? And it's somewhere there. Then again is the problem with ethics then, you get back to it. Say if we were to roll out self-driving cars tomorrow, and that we would, the number of fatalities would go down. Then we have a moral obligation to actually roll out self-driving cars tomorrow. Because we're gonna save a lot of lives. But those who die in traffic, without us being able to say that there is someone we can say it is your fault. We have like no one, society has like no one to say that it is you that we can somehow reason around, did you make a mistake or did you not make a mistake.	DSS-OR, DSS-DSR, DSC, DSC-E, DS-DSUA

		It's just like this, was it a mistake or not. There's, like, no one to answer. And very much if you then connect to Data Science generally, you then look very much at recommendations, customers and telephone operators who are going to get an offer for us to think that they are going to terminate their subscription. Yes it is the same thing, either it is up to the individual customer service employee. With experience has come up with this with reasonably good fingertip Sony, this person should probably have an offer as well. But if you have not come to the right person or not as well, that's its thing. But to then talk about Data Science, a model that, no we can not guarantee that all those that we have sent a message to absolutely would have left. You'll never be able to do that. But just as said that the knowledge and the realization that it is probability based as well, it is not an exact science. That is the most important factor then. And then there's a lot of other stuff like that. The importance of data, and data quality, and the importance of as well as the ability to make good assumptions where you have lots of data washing and data processing that must be done but it is a bit at a detailed and lower level.	
E.25	I	You've touched on quite a lot of different stuff here, but is there some kind of specific strategy or plan that you work along with the team or when you're out with the customer or organizations you assist that can help you respond to this in some way or all of these different types of issues or aspects of Data Science.	
E.26	E	Well, that's a great question. It can be said that what is the big challenge is, of course, a little about the fact that the ultimate goal is unclear to say the least. If you are thinking of a traditional system development project as well, you should go from Point A to point B to Point C, you have as well as requirements and you know what the system should do and given that you have people who know their things and given that you have the right tools in place etc, then there is nothing stopping you from reaching from point A to point B. But in Data Science, it can be as if you have the world's best people who have the world's best knowledge and skills and tools, the world's funniest and lightest problems. But if you do not have the predictive power of the data, it does not matter, then you will still not be able to describe it there, however much you would like. So that it is like a giant challenge to get something out, you are not in advance	DSS-OR, DSS-DSR, DSC, DS-DS, DS-DSUA

		if you are actually going to come up with a result. A zero result that shows, nah but then it shows nothing, we can not show on this connection, or models say that it resembles a coin toss. You don't know that in advance then, what we do and what I think or as you hope others do also is that you work very iteratively, very agile so you work in pretty small cycles. You start to look at the data and start to produce something kind of rough. And so thinking Okay seems there are some indications that there is as well as predictive power in this data. Okay if yes, then we can go ahead and start crystallizing this a bit and work quite iteratively that way and start implementing and presenting stuff as well as for their counterparts and for users and okay, if we reason this way, this is understandable, you know what I mean like. Also because you should not like not to snow themselves on for technical details as well as just can I get all the way up and explain this in an easily understandable way. That's it, that's a lot of stuff like that, so yes, very agile and very iterative.	
E.27	I	Who do you think like, has this main responsibility as well as making sure that everything flows on, that one can take action, is it you in a Data science team or is it the organization you are with?	
E.28	E	That's a great question, great question. I would say that this is because everyone has, their respective tasks throughout the interaction as well. I as a Data Scientist and my entire team as well, well, there are demands, or like this. It should and should be required of much of what we are doing to be able to take height for changing scenarios over time so as to ensure a well-functioning model in the long term. But it's also not the answer to the whole question because no matter how much you do at first, the world around you is changeable. Like it is never possible to increase everything, there must, as well, be ownership and a management in the organization and a managerial responsibility in ensuring that these things work, including over time. That's how it definitely is.	DS-DS, DSS-DSR, DSS-OR
E.29	I	Well, actually, I think we've got answers to most of the things we've been thinking about. Is there anything more that you think would be relevant to our essay that you want to add?	

E.30	E	Yes it is a super exciting area and it is super valuable to start studying it in more detail as well, and I think as well that there is not one single reason for some Data Science projects to fail, but one of the absolutely major reasons so I would definitely say that it is this problem simplification that you most often encounter. I and my colleagues as well as we love to like go way down in the most advanced techniques and the most odd algorithms etc. that's what we think is so damn funny, when it is then put forward a problem for us so usually we think yes but absolutely would you be able to run this or this etc and so you Predictive maintenance as well as that system engineers, automation engineers are concerned with, it is the same thing there as well. If we say that a model is produced that says that there is a high probability that the machine will break, we are, as in a probability-based world, so we can not guarantee that the machine will break. What will the company do, should one stop their entire production, or should one take income loss at potentially millions of kronor an hour, just because one model says that the machine seems to start collapse soon, no I will not do that, because so much I do not trust this model. It can be how advanced algorithms like, it is more profitable if you count on it so, and then it turns out that it is more profitable to run these machines until they collapse, and run them until they crash. But it is clear that all sellers and all as well, Gartner and Accenture and all, predictive maintenance gives you this much money, put forward use cases and examples etc, but all that is a bubble. If you're going to do it in some sensible way, then you should hook arms properly with the automation engineers who can weibull distributions and survival techniques on their five fingers and have you seen now that we can use neural networks in the survival analysis, oh that’s exciting as well. But it is extremely rare, I never think I have seen that such real-life and productive dialogues have actually occurred.	DSS-DSR, DSS-OR, DSS-ADSS, DS-DSUA
E.31	I	Is there any kind of cooperation with these systems engineers or automation engineers?	
E.32	E	Yeah! But then you want to say, for example. they know logistics regression on their five fingers, and so we come from Data Science and AI and say yes but logistic regression, it has been around since the 19th century. Now we will work with cool techniques, such as neural networks. Well, what does the insurance	DS-DSUA, DSC, DSS-DSR

		mathematician say about that, I will never be able to interpret it, we will never be able to have a pricing analysis based on what the neural network says. Okay, but hook your arm and say check this out! Here are some new techniques in how we can interpret and compress a neural network down to logistic regression. So from there you can start pricing. So the thing about a real hook arm but with these traditional techniques that have always been there as well as it's just the same thing as Data Science stuck with other glasses. That is a bit what has to be done, and it is learned far too little about this at the present time.	
E.33	I	Would you say there's some kind of knowledge gap in it?	
E.34	E	Yeah, but exactly. I think the knowledge gap exists and there is it from both sides as well. For example: we don't really believe that, but from the point of view of Data scientists, it is that my techniques are much better, I can have much better accuracy in my models. Well, you don't think about the big picture. So that yes, it's a bit what I think is missing.	DSC-K
E.35	I	Really interesting to hear.	
E.36	E	Of course there are a lot of failed stuff as well. Then you can say that those prominent examples are not what you read in daily time. The drive stone or drive motor in something you like. Netflix recommendation engine and Spotify's recommendations, as well as Amazon's recommendations or as well as logistics optimization, and logistics management, as well as the internal stuff there as well, it's those boring stuff. It doesn't make itself very good to write about it in lots of articles but it's the thing that succeeds. It's the boring stuff that succeeds where you really get down on the details.	DS-DSUA
E.37	I	Perfect, thank you very much for your answers.	

Appendix G

Interviewee F

Interviewee: Interviewee F
Title/Role: Data Scientist
Date and Time: 2020.04.23

I: Interviewee
F: Interviewee F

Reference number	Person	Questions and Answers	Code
F.1	I	Explain a little briefly what your job actually entails, what daily tasks you have.	
F.2	F	Yes, absolutely. I'm part of a small team of my company officially called Business Intelligence and what we do and what I think, too, the concept of Business Intelligence has pushed forward the role of Data Scientist or the field of Data Science maybe, ehm. BI or Business Intelligence you could say was a buzzword in the industry that began to appear about 15 years ago, and it was also about then that I started working with Business Intelligence, at a local level back then. Then you just got out of the customer relationship hype, that everyone would have a CRM system, keep track of their customer stock. All sellers would know everything that happens in all customer meetings etc. Then the next thing to jump on was business intelligence. To understand your business, look at the data you have historically generated. And it was a certain gap right from the beginning, 15 years ago. It was probably not until really the last 5 years where the concept of Data Science began to appear. Or a few years earlier. If you Google Data Science, you will still find articles and blog posts that date back to 2012 even. But in Sweden it is still a relatively new concept, I would say. Data Scientist as a professional group did not begin to appear in job ads until a couple of 2-3 years ago. That was just a little more background on what I'm doing. But what we try to do or what our mission is at my company is that we are a hub in different parts of the business and can be turned to for help in analyzing the data that we have collected and processed. Or that they need help to analyze their own data simply. It is just	DS-DSUA, DS-DS, DS-VCDS

		like our task is to be able to provide data, and to be able to manage data in an accessible way. And being able to offer an analysis platform out of the books and visualize it.	
F.3	I	Mm. What kind of data are you used to dealing with?	
F.4	F	Um, I think the most common type of data, basically across the industry, is sales data above all. It is there that you can quickly get back your investment as you do in different tools to be able to process data etc. Then you have to check on their sales or customer stock so there is also either more money to earn or costs you can draw down on, you see why do we have inventory on these articles that we have bad sales of. So sales data is absolutely the most common. Financial data will come after that I would say. Where you may be part of a group with several different companies that do not have common systems and then you may be allowed to consolidate their figures in some form of financial reporting platform. And then inventory data usually comes. Now I usually say because right now I work * and there I have been working and been for the last three years but I have a 15-year background in the Data Science profession. Mostly as a consultant so I have jumped around on assignments most in southern Sweden but also out in Europe. My experiences relate to this kind of data, starting with sales data and then financial data and stock data. After that, it can be a little anything. This is also where it will be fun because there will be completely different types of new challenges then.	DS-VCDS, DS-DS
F.5	I	How do you work for this data to generate some type of value?	
F.6	F	In order for there to be some value of the work that you are asked to do, you need to have some kind of expressed need. What usually happens when you invest in a Business Intelligence platform then or Qlik, Tableau or so, it is the biggest two on the Swedish market, that you notice that your colleagues or clients suddenly become like children in a candy store. When you see how easy it is to go from some kind of data swamp to then end up in a nice shop full of colorful graphs that can show both something and the other and also the top ten customers. But what there is, there will be no value in it, since you have to	DS-VCDS, DSC-K, DSC

		start by thinking more about what kind of need I have. What's my problem? And it is perhaps often a question to a business-minded person. Anyone who should ask themselves those questions or their department those questions, is a little bit of what you call domain knowledge. In the first place, it's not the actual computer knowledge that matters or that you're a python Ninja that can script you to anything. There won't come any value out of your data analysis if you don't have any domain knowledge. You need to know what it is you want answers to or what your problem is. Do you know that usually, using people like us then, they know what kind of data we need to look at in order to answer your question or to be able to help you out of this problem. So that I would say is the most important basic idea, that there is an expressed question which is somehow quite specific as well and almost the smaller and the simpler, the better. It is easier to start with a question that you feel that this question, it can not be answered if you look at this data, because once you get started with that bit it is quite easy to iterate yourself out to a solution to a major problem as well. And the best solutions that save the most money are just to be able to answer the simple and concrete questions in the data. There are short projects, Short efforts of 2-3-4 days, if the conditions are there. Say you come from question to answer. Well I think that quite a lot of these Data Science projects fail because there is a decision-maker far out on one side who may indeed have some kind of domain knowledge but does not have the ability to concretize their needs. And don't get help or ask for help from someone like a data scientist who should be somewhere between just domain knowledge and pure data engineering. You almost have to have a foot in both camps to understand both languages. So I think that most of the projects or many of the projects have just because you do not have a clear target picture. What is it we want to achieve with the project simply. Without it you just say that you may have a political agenda that makes it like yes but I need a starting point in my career. Now I have this project manager role and the title and now I have six months to do something. However, one does not really know what it is, as one should anchor their work in the business or in reality even.	
F.7	I	I understood. Yeah, it seems like you've been in different kinds of roles and companies as a	

		consultant, huh? What is Data Science used in organizations?	
F.8	F	Very different, I would say. Um, very different. Data Science as a concept it's almost like saying that 15 or 20 years ago one worked in IT, can be anything. Or that I work in health care. I think there are so many different kinds of applications of Data Science, that there are companies or google searches or employees working in the field. But if you're going to remove all the fluff and philosophy behind Data Science and go back to nuclear power or the driving force in why there is Data Science, then it's still the case that somewhere you've found out that if you look at past experiences of something then maybe we can learn a bit from how we should act in the future to achieve the same positive results or avoid going into the same trap once more. And since we don't have any crystal balls, we have to look at the story and 20 years ago we may not have talked about extrapolation or modeling of Data and neural networks, etc. It was enough for a fairly classic linear regression and if we behave this way that we have seen historically this set of conditions that we then think will continue this way. And quite often it has been good enough, and you know that if you historically look at the customer stock that sell very little and rarely if we do not engage in them but instead look at a much smaller part of the customer stock that trades very frequently or in large volumes, that's also when we will make our money. If we look historically, 80% of our income has come from this small group of people. And it will probably be true even tomorrow. But there are also quite a few sets of variables, and you may not think so often as to be the same as it was 3 years ago as it is today. For example, there is a very big difference between before and after the financial crisis of 2008. I believe that the need for Data Science arises precisely because we have come to the realization that we have data, if we look at it how we have acted historically, then perhaps we understand how we should behave in the future too.	DS-DSUA
F.9	I	What kind of skills would you say are most important when working with Data Science?	
F.10	F	Yeah, that's a good question. I have for some of my years worked as a BI consultant for Qlik, software platform or favorite weapons if you say so. I have	DS-DS, DSS-ADSS,

		worked in many different types of consulting firms and at Qlik himself and something that we in our little niche have always agreed on is what hard it was to get hold of good people who could as well become our colleagues simply. It was hard to grow because it was hard to find good people. I think it's hard to find this special mix that might make you a data scientist then. You need to have some technical skills of course, now there are a billion different technical tools and platforms out there and to decide to become a Data Scientist and then look at what tool I need to know, then you can almost go and do something else instead. One will not have time to learn all the technology and much of the technology will already be outdated once you understand it. Some technical understanding is good but maybe not software specific or platform specific more than understanding what and how to store data, today and how to do ten years ago, because it is still up to date. Well databases, tables, relational databases etc but perhaps more on a technical level. Actually, it's enough to have a good base. Anyone who is going to work with data needs to know SQL, to be able to ask a question to data, but you may not need to have more excellence than that. All companies have their own tools and their own processes so that there you will learn when you start working at that particular company. Each company has its own stack and you will learn. The second part I mentioned is a little closer to this with formulating a need for Data Science projects to have some kind of success, it is a factor that is at least as important, to be able to communicate. Being able to listen to what other people think, it can be difficult with domain knowledge at the beginning of their career. Things like turnover or financial instruments, or whatever it may be, have not been exposed to anything. But, if you can talk to other people and listen, understand what other people need, I would say, a little journalism is probably not entirely wrong. Be a little curious, a little inquisitive, dare to ask the same question 30000 times and be able to understand what is said I would say is probably at least as important as being able to ask this SQL statement. A data scientist himself will never be able to do the whole job of a slightly larger company, one must be this mix between the business and the data part of the business itself, to be able to understand this need exists, then	DSC-K

		maybe we need to map this need towards this data, maybe be able to talk to system owners or if we ourselves already have data for example. How are we going to get the data out and should we model it to be able to do some kind of analysis on it. So somewhere in between, I would say to be fine.	
F.11	I	I understood. Would you say that your expectations compared to the expectations that organizations you've been working with are in line with one another?	
F.12	F	Well, that's a difficult question, I think. In Sweden, Data Science is still quite new and in the university world it is certainly much already more known than out in the corporate world. I was falling off the hill for half a year ago when I saw that Ikea was looking for a data scientist for example. They don't tend to be very quick to embrace the latest trends, but they usually run on a little bit. But I was the first of my company to get the role of Data Scientist and I got that because I asked for it. I think that the entire data science label is very much so probably no from the beginning science but initially tried to cover in some machine learning and of testing for it to be in a scientific way and be able to apply different methods to test if the only change in a process gives a different result than a control group eg. However, it is extremely scientific to start talking in those terms and there are very few companies in Sweden that sit with this in a scientific way related to or close to the Data. Without I think most people who call themselves data scientists were the ones who ten years ago called themselves BI developers or data engineers, although data engineers even today have some kind of scratch next to Data scientists, I still think we do about the same things.	DS-DSE, DS-DSUA, DS-DS
F.13	I	Thank you, you mentioned a little earlier as with this need that you have encountered some problems with it, are there any more challenges you have encountered when you have worked with Data Science?	
F.14	F	Yes, you run into problems all the time but sometimes it has kind of worked, are you looking for an example where it really hasn't worked the way you imagine?	

F.15	I	Yeah.	
F.16	F	Um, yeah. Where it has not worked it is enough where culture, corporate culture has not been healthy, I would say. Where, perhaps, there has been far too strict hierarchical division of responsibilities or tasks and where there is a great deal of policy. Um, and closed doors. Are you afraid to expose your own skills, for example. then it is much more often that you ask other people in your organization who do your job for you. In order not to expose your own incompetence, you might be for what you want. You may not really know but try to package their requirements professionally but if you scrape a little on the surface then the need is really just warm air. A business climate that does not invite dialogue or that it would be okay for a coworker to become a manager, for example. it's more common than you think. If you have such a climate, it's really hard to understand what the need is if there even is oneso to say. And it is difficult to get ahead, it is difficult to find key people within the company who can answer specific questions if one is not allowed to ask the question. It is that feeling when you come in to the office, it's probably not okay to ask questions, it might be okay. But perhaps the general feeling is that you should take care of your own.	DSC-OS, DSC-K
F.17	I	Do you think there are some types of actions that could have solved these kinds of problems?	
F.18	F	Um, yeah. Sometimes it can. Consultants actually tend to be pretty good to take into these cases. Not pure engineering modules that sit and program and code but you can find a good project manager from outside for example. So it can be someone who comes in and is a breath of fresh air without any history with the company. It's easier as an outsider to be able to ask stupid and hard questions and be able to question decisions than someone internally does. I've had the advantage during much of my career been a consultant and be able to step right in and be quite so square and play a little bit stupid because you usually get much better answers to their questions if you show, yes but I can actually nothing, I know the tool that you have invested in but I don’t know your business so you have to explain your business to me as if I was five. As an outsider, I can say that they won't throw me out for that reason, but they invest in	DSS-ADSS, DS-DS

		me there because I'm good at the tool. So as to be able to venture consultants into the projects, and get help in drawing out the frameworks and be able to delineate the project, and to perhaps be able to help formulate the needs and some kind of goals. I think that can be very helpful.	
F.19	I	When working with the data, do you take into account the data quality in any way or consider it when working?	
F.20	F	Yes, I do. Maybe it's not something that I do today so consciously. But you can see that quite often I can sort of correct the data for analysis how bad it is. I think that in general we have become better at producing data today than we were 10/15 years ago. Today we have systems like this where we can get the data that we entered for reporting and analysis, etc. than we had 15 years ago. 15 years ago, there was much more excel and people who sat and manually clipped and pasted numbers because there were no systems that you could afford or that you had the knowledge of could support your needs. So that I notice the quality anyway when you work and sometimes it may be that you realize at the beginning of a project when you start processing the data we do not have the preconditions to row ashore this project, with the objectives that the project has started up because the data is too poor. Or we thought we had something but it turned out that this is something completely different. Sometimes you can re-direct the project to okay, narrow it down to that we could not do this this and this, but that it still seems to be good enough data to be able to present this part of the business for example.	DSC-DQ, DSC
F.21	I	Ethical aspects of the data and e.g. subjective data, is it something you have encountered or are thinking about while working?	
F.22	F	Yeah, there's something learned every once in a while. I'm a bit asperger when it comes to data that you just ignore what the data actually describes because right now I'm just interested in the fact that these columns contain just these numbers like that. But sometimes you take a step back above all when you start discussing the needs and the results, you realize that every line in this data set is a natural person who visited a health center, it is quite	DSC-E, DS-DS, DS-DSUA

		sensitive. This is not some machine that has randomized some numbers as well as without this is actually physical people that have generated this data, moreover, it is sick people who gave this data. Therefore, one really needs to use it with caution and try to dehumanize it as much as it goes and aggregate it so that it is not able to identify specific people, etc. But it can also be so when you look at sales data in itself that you say Okay, we have sold technical equipment to saudi Arabia or North korea, now it is not so, but sometimes you realize that here there is actually a lot of data that you yourself are not so proud of. Then you might have to step back and focus more on this just data. In this particular situation, it does not help me to know that this data has sold to a dictatorship, for example. But it happens to look like it does.	
F.23	I	Is there any type of knowledge that you have encountered that it is lacking in Data Science projects?	
F.24	F	Yes, I think it's actually just that project management is still something that's the hardest thing I think to get to a good thing. Most of the project managers I have worked with are those who have a solid list of actual project management certificates simply. It is not someone who has just decided that now I will become a project manager and run a checklist in excel, but someone who can do real project management. It is quite rare to be involved in projects with professionals in project management. That, I think, is something that, from a business perspective, is something that you are fastest at budgeting away. It will be very different in the result if you buy in a person who has 15 years of experience writing this program for example. or if we take someone who has just started scripting. So sure we need to spend money on someone who actually knows the job. But the project management is as well as I can take it myself, I may need to have a small GTD list next door and we're on the track. You probably underestimate quality or how it could have been with professional project management.	DSC-K, DSS-OR, DSS-ADSS, DSS-DSR
F.25	I	To then be able to achieve success in Data Science projects, what do you think is important then?	

F.26	F	Yes, clear specification of the needs, a clear mix in the project team between just domain knowledge and data knowledge, I would say, and good management in the project. If you can have these 3 or 4 points then you have great conditions really no matter what you take yourself for. Data is always secondary, it is always possible to find somewhere. I mean somewhere if there is some kind of data that you can start building something on even if it is only a small embryo in the project then it is iaf that as long as you have good defaults in the team then there will be something like that.	DSS-ADSS, DSS-OR, DSS-DSR
F.27	I	Do you think the expectations of Data Science are generally achieved?	
F.28	F	Um I think the expectations are achieved in what I do, but when you look and read blogs about what Data Science is I don't know if to be honest. It is some kind of accepted truth that if you are a good data scientist then you should have ten years of statistical in your back while being able to all about all sorts of machine learning models and be able to all the new different machine learning libraries and additionally have worked with Business analysis for 30 years. There are none that live up to that specification at all. I believe that the people who have, as well as contributed to this general perception, are probably people who are either pissed off that you yourself have been entitled Data scientist and are sour that everyone else has also received it and are trying to distinguish yourself in the crowd in some way. But then there are those who are much more nuanced who say it's called science but in fact it's about being able to look at data and listen to a need, and being able to visualize something out of this mix. That said I think the field is too wide to say that Data Science is successful or not, it's probably both and I think. Even two pieces of data scientists in the same company can have two completely different skill sets and assignments simply. But we still call them data scientists and ten years ago maybe they had been software engineers or data architects or something like that.	DS-DSE, DS-DS, DS-DSUA
F.29	I	As an organization, are there any types of actions that you think one should take to reach expectations of Data Science?	

F.30	F	Nah, that's a good question, but I don't think so. I think Data Science is a supporting role, I think it feels best when it is a Supporting Role. I think it might be good if some kind of representation in the Data Science team is featured in various forms of steering groups and management groups to have an ear on the rails to hear what is going on in the organization and make sure to be prepared for what changes the company might be planning to make within the year or coming years to plan for. But yes, the changes within the organization, I do not know, it already depends on how long you have come with the Data Science within the company but it is probably good that those who work with it are on the right course and know what the data scientist's role is on the company so that you have a clear picture of it.	DSS-OR, DSS-ADSS
F.31	I	Mm. Yes, Sanna do you have any questions?	
F.32	I	Nah, I don't think so, most of it's answered.	
F.33	F	Yes, it's easy to get this feeling of panic when you just read the data Sciene literature that Shit I know nothing of this. I don't have any university points in statistics, but you just might as well try not to compare yourself with all this theory. But back to this we talked a little about in the beginning. What is the reason that Data Science exists in the corporate world. Well it is to make money somehow, and as well as the driving force behind it is like being able to be as good as possible and understand their history and as well as drawing lessons from their experience. Each company will have as well different ways to look at its own history. So that is perhaps 95% of the theories still will not fit in, but you may still be able to bring it into the new company to make them change their view of their history or their approach to it. The theory and reality will differ.	DSS-OR
F.34	I	I understand.	
F.35	F	Now I do not know, but I think there are quite a few companies in Sweden that have a special department of employees called Data Science that deals only with machine learning. The companies believe that Data Science is only about computer science is the company that has not gone so far with their own data trip, but try to keep an ear on the ground trying to keep up with the concept of the world, but you may	DS-DSUA, DSS-ADSS, DSS-OR

		not have seen so much of themselves in his own company than that. It was a blog post a few years ago on LinkedIn I think it was like an Indian guy wrote I think it was that all his buddies were "DEDS" thus data engineers or data scientists and that all his friends on linkedin had changed their titles from software developers, engineers and architects and so suddenly all the data engineers, data scientists etc. were having it in mind. That, like, we do the same thing today as we did ten years ago, 20 years ago, 30 years ago, but that it was called another thing. Just like that cleaners once were just cleaners but then changed to hygiene technicians etc. to modernize themselves or distinguish themselves sometimes have to do the same thing to change skins, and that's what Data Science is right now. Machine learning is now something new but then you fit to get it into this old one that you already made but change the skins. I think it's fun to have that title, to call yourself a Data Scientist than to call yourself a Business Intelligence Developer, even though it might be the same thing you've been doing anyway.	
F.36	I	Thank you so much for your participation and answers.	

References

Alter, S. (2010). Viewing Systems as Services: A Fresh Approach in the IS Field. *Communications of AIS*, *26*(11), pp. 195-224.

Albuquerque, U.P. de Lucena, R.F.P. & de Freitas Lins Neto, E.M. (2013). Selection of Research Participants, in Albuquerque, U.P. Cruz da Cunha, L.V.F. de Lucena, R.F.P. & Alves, R.R.N. (eds), *Methods and Techniques in Ethnobiology and Ethnoecology*, New York: Humana Press, pp.1-13

Atwal, H. (2020). Practical DataOps: Delivering Agile Data Science at Scale, Berkeley: Apress

Berntsson Svensson, R. Feldt, R. & Torkar, R. (2019). The unfulfilled potential of data-driven decision making in agile software development, Lecture Notes in Business Information Processing, p. 69, vol. 355. Cham : Springer

Bhattacherjee, A. (2012). Social Science Research: Principles, Methods, and Practices, 2nd edn, Tampa: A. Bhattacherjee

Braschler, M., Stadelmann, T., & Stockinger, K. (2019). Applied Data Science, Cham: Springer

Brennan, P. F., Chiang, M. F. & Ohno-Machado, L. (2018). Biomedical informatics and data science: evolving fields with significant overlap, *Journal of the American Medical Informatics Association : JAMIA*, [e-journal] vol. 25, no. 1, Available through: LUSEM Library website http://www.lusem.lu.se/library [Accessed 25 March 2020]

Burke Johnson, R., & Onwuegbuzie, A. J. (2004). Mixed methods research: A research paradigm whose time has come, SAGE Journals, vol. 33, no. 7, pp. 14-26, Available through: LUSEM Library website http://www.lusem.lu.se/library [Accessed 8 April 2020]

Cai, L., & Zhu, Y. (2015). The challenges of data quality and data quality assessment in the big data era, *Data Science Journal*, [e-journal] vol. 14, Available through: LUSEM Library website http://www.lusem.lu.se/library [Accessed 26 March 2020]

Cao, L. (2018). Data Science Thinking : The Next Scientific, Technological and Economic Revolution, Cham : Springer International Publishing.

Cave, J. (2016). The ethics of data and of data science: an economist's perspective, *Philosophical Transactions: Mathematical, Physical and Engineering Sciences*, [e-journal] vol. 374, no. 2083, Available through: LUSEM Library website http://www.lusem.lu.se/library [Accessed 2 April 2020]

Demigha, S. (2019). Agile Projects and big Data', Proceedings of the International Conference on Intellectual Capital, *Knowledge Management & Organizational Learning*, Available through: LUSEM Library website http://www.lusem.lu.se/library [Accessed 3 April 2020]

Dichev, C. & Dicheva, D. (2017). Towards Data Science Literacy, *International Conference on Computational Science*, [e-journal] vol. 108, Available through: LUSEM Library website http://www.lusem.lu.se/library [Accessed 25 March 2020]

Donoho, D. (2017). 50 Years of Data Science, *Journal of Computational and Graphical Statistics*, [e-journal], vol. 26, no. 4, Available through: LUSEM Library website http://www.lusem.lu.se/library [Accessed 28 March 2020]

Efron, S.E. & Ravid, R. (2019). Writing the Literature Review: A practical guide. New York : The Guilford Press

Falgoust, M. (2016). Data Science and Designing for Privacy, *Techne: Research in Philosophy & Technology*, [e-journal] vol. 20, no. 1, Available through: LUSEM Library website http://www.lusem.lu.se/library [Accessed 3 April 2020]

Floridi, L. & Taddeo, M. (2016). Introduction: What is data ethics?. *Philosophical Transactions: Mathematical, Physical and Engineering Sciences*, [e-journal] vol. 374, no. 2083, Available through: LUSEM Library website http://www.lusem.lu.se/library [Accessed 2 April 2020]

Forrester. (2016). Data Science Platforms Help Companies Turn Data Into Business Value: A Forrester Consulting Thought Leadership Paper Commissioned By DataScience, Available online: https://cdn2.hubspot.net/hubfs/532045/Forrester-white-paper-data-science-platforms-deliver-value.pdf [Accessed 20 March 2020]

Hoffmann, A.L. (2017). Making Data Valuable: Political, Economic, and Conceptual Bases of Big Data, *Philosophy & Technology*, [e-journal] vol. 31, no. 2, Available through: LUSEM Library website http://www.lusem.lu.se/library [Accessed 19 March 2020]

Kotu, V. & Deshpande, B. (2019). Data Science: Concepts and Practice, 2nd edn, Cambridge: Elsevier Science

Kvale, S. (2007). Doing interviews. Los Angeles: SAGE

LinkedIn Economic Graph (2017). LinkedIn's 2017 U.S. Emerging Jobs Report, Available online: https://economicgraph.linkedin.com/research/LinkedIns-2017-US-Emerging-Jobs-Report [Accessed 24 March 2020]

Linneberg, M.S., & Korsgaard, S. (2019). Coding qualitative data: a synthesis guiding the novice, *Qualitative Research Journal*, [e-journal] vol. 19, no. 3, Available through: LUSEM Library website http://www.lusem.lu.se/library [Accessed 10 April 2020]

Müller, A.C. & Guido, S. (2017). Introduction to Machine Learning with Python: A Guide for Data Scientists. Sebastopol : O'Reilly Media Inc.

Myers, M. D., & Newman, M. (2007). The qualitative interview in IS research: Examining the craft. Information and Organization, *Information and Organization*, [e-journal] vol. 17, no. 1, Available through: LUSEM Library website http://www.lusem.lu.se/library [Accessed 11 April 2020]

O'Connor, C. & Joffe, H. (2020). Intercoder Reliability in Qualitative Research: Debates and Practical Guidelines, *International Journal of Qualitative Methods*, [e-journal] vol. 19, Available through: LUSEM Library website http://www.lusem.lu.se/library [Accessed 29 April 2020]

OECD (2008). Glossary of Statistical Terms, OECD Publications Centre, 2008.

Pagán, J. (2018). The Path Toward Data Insights, *US Black Engineer and Information Technology*, [e-journal] vol. 42, no. 2, Available through: LUSEM Library website http://www.lusem.lu.se/library [Accessed 18 March 2020]

Passi, S. & Jackson, S. J. (2020). Trust in Data Science: Collaboration, Translation, and Accountability in Corporate Data Science Projects, *Proceedings of the ACM on Human-Computer Interaction*, [e-journal], vol. 2, Available through: LUSEM Library website http://www.lusem.lu.se/library [Accessed 2 April 2020]

Patton, M.Q. (2015). Qualitative research and evaluation methods, Thousand Oaks: SAGE Publications

Philip Chen, C. L. & Zhang, C.Y. (2014). Data-intensive applications, challenges, techniques and technologies: A survey on Big Data, Information Sciences, [e-journal] vol. 275, Available through: LUSEM Library website http://www.lusem.lu.se/library [Accessed 30 March 2020]

Poel, M. Meyer, E.T. & Schroeder, R. (2018). Big Data for Policymaking: Great Expectations, but with Limited Progress? *Policy & Internet*,[e-journal] vol. 10, no. 3, Available through: LUSEM Library website http://www.lusem.lu.se/library [Accessed 28 March 2020]

Provost, F. & Fawcett, T. (2013a). Data Science and its Relationship to Big Data and Data-Driven Decision Making, *Big Data*, [e-journal], vol. 1, no.1, Available through: LUSEM Library website http://www.lusem.lu.se/library [Accessed 30 March 2020]

Provost, F. & Fawcett, T. (2013b). Data Science for Business: What you need to know about data mining and data-analytic thinking, Sebastopol: O'Reilly Media

Raguseo, E. (2018). Big data technologies: An empirical investigation on their adoption, benefits and risks for companies, *International Journal of Information Management*, [e-journal] vol. 38, no.

1, Available through: LUSEM Library website http://www.lusem.lu.se/library [Accessed 18 March 2020]

Randolph, J.J. (2009). A Guide to Writing the Dissertation Literature Review, *Practical Assessment, Research & Evaluation*, [e-journal] vol. 14, no. 13, Available through: LUSEM Library website http://www.lusem.lu.se/library [Accessed 14 April 2020]

Raviv, N., Jain, S., & Bruck, J. (2020). What is the Value of Data? On Mathematical Methods for Data Quality Estimation, Available through: LUSEM Library website http://www.lusem.lu.se/library [Accessed 18 March 2020]

Recker, J. (2013). Scientific Research in Information Systems: A Beginner's Guide, Berlin: Springer

Saltz, J.S., & Stanton, J.M. (2018). An Introduction to Data Science, Los Angeles: SAGE Publications

Sammut, C. & Webb, G.I. (2017). Encyclopedia of Machine Learning and Data Mining, 2nd edn. New York : Springer Publishing Company, Incorporated.

Schuff, D. (2017). Data Science for All: A University-Wide Course in Data Literacy, In: Deokar A., Gupta A., Iyer L., Jones M. (eds), *Analytics and Data Science, Annals of Information Systems*, Cham : Springer, pp. 281-297

Schultze, U., & Avital, M. (2011). Designing interviews to generate rich data for information systems research, Information and Organization, vol. 21, no. 1, Available through: LUSEM Library website http://www.lusem.lu.se/library [Accessed 8 April 2020]

Silverman, D. (2011). What is naturally occurring data? SAGE Publications, [video online], Available at: LUSEM Library website http://www.lusem.lu.se/library [Accessed 26 March 2020]

Smith, G., & Cordes, J. (2019). The 9 Pitfalls of Data Science, Oxford: Oxford University Press

Subrahmanyan, S. N. & Jalona, S. (2020). Building a Data-Driven Culture from the Ground Up, Harvard Business Review Digital Articles, 28 February 2020, p2-5, Available through: LUSEM Library website http://www.lusem.lu.se/library [Accessed 2 April 2020]

Syed Fiaz, A. S. Asha, N. Sumathi, D. Syed Navaz, A.S.(2016). Data Visualization: Enhancing Big Data More Adaptable and Valuable. *International Journal of Applied Engineering Research*, [e-journal], vol. 11, no. 4, pp. 2801-2804, Available online: https://www.researchgate.net/publication/299391071_Data_Visualization_Enhancing_Big_Data_More_Adaptable_and_Valuable [Accessed March 19 2020]

The Economist. (2017). The world's most valuable resource is no longer oil, but data, 6 May, Available online: https://www.economist.com/leaders/2017/05/06/the-worlds-most-valuable-resource-is-no-longer-oil-but-data [Accessed 18 March 2020]

Timmins, F. & McCabe, C. (2005). How to Conduct an Effective Literature Search, *Nursing Standard*, [e-journal] vol. 20, no.11, Available through: LUSEM Library website http://www.lusem.lu.se/library [Accessed 16 April 2020]

Treder, M. (2019). Becoming a data-driven Organisation: Unlock the value of data. Berlin Heidelberg : Springer

van der Aalst W. (2016). Data Science in Action. In: Process Mining, Berlin: Springer

Vaughn, P., & Turner, C. (2016). Decoding via Coding: Analyzing Qualitative Text Data Through Thematic Coding and Survey Methodologies, *Journal of Library Administration*, [e-journal], vol. 56, no. 1, Available through: LUSEM Library website http://www.lusem.lu.se/library [Accessed 16 April 2020]

Veeramachaneni, K. (2016). Why You're Not Getting Value from Your Data Science, Harvard Business Review, Available online: https://hbr.org/2016/12/why-youre-not-getting-value-from-your-data-science [Accessed 28 March 2020]

Waller, M.A., &, Fawcett, S.E. (2013). Data Science, Predictive Analytics, and Big Data: A Revolution That Will Transform Supply Chain Design and Management, *Journal of Business Logistics*, vol. 34, no. 2, pp. 77-84, Available through: LUSEM Library website http://www.lusem.lu.se/library [Accessed 30 March 2020]

Wiles, R. (2012). What are Qualitative Research Ethics?, London : Bloomsbury Publishing