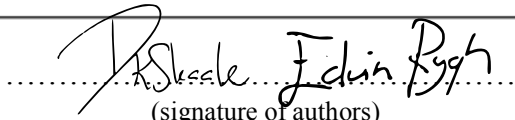




Universitetet
i Stavanger

FACULTY OF SCIENCE AND TECHNOLOGY

MASTER'S THESIS

| | |
|--|---|
| Study programme/specialisation: Industrial Economics Entrepreneurship and Technology Management | Spring 2018 Open |
| Authors: Didrik Kahrs Skaale, Edvin Rygh |  (signature of authors) |
| Internal Supervisor: Jan Frick External Supervisors: Pål Øystein Stormorken & Stefan Fürnsinn | |
| Title of master's thesis: Big Data Technology Adoption Through Digitalization in Yara International ASA | |
| Credits: 30 | |
| Keywords: Big Data, Value, Digitalization, Data Accessibility, Decision-making, Process Optimization, Micro- segmentation, Enhanced Innovation, YARA International ASA | Number of pages:98..... + supplemental material/other: ...4..... Stavanger, June 14 th 2018 date/year |

Abstract

As digital technologies keep evolving at great speed, our ability to store, extract and analyze data is rapidly improving. Big data is a popular term that is used to describe large and complex datasets that is generated through a variety of internal and external sources. As the volume, variety and velocity of data keeps increasing, traditional analytical methods are stretched to new limits. This thesis argues that increased capability to handle data can lead to great opportunities for well-established organizations. However, the thesis acknowledges that there are several barriers that organizations need to overcome to reap the benefits of their data assets. The thesis builds on the assumption that Yara has great opportunities to strengthen their big data capabilities and thus unlock large values, and aim to answer the following research questions:

RQ1: What is the current state of big data technology adoption through digitalization in Yara?

RQ2: What should be emphasized in the coming years to unlock potential value from big data?

This qualitative study answers the first research question by conducting 12 semi-structured interviews, measuring the state of five big data value drivers identified in a literature study. The findings are relatively unambiguous and indicate that Yara are at an early stage of adopting big data technologies. The thesis identifies increased data accessibility as a catalyst for other value drivers, and observes several issues related to how data is made available and accessed.

The second research question is answered by analyzing the transcripts in relation to theory and proposing suggestions. It is found that a higher level of big data technology adoption should allow Yara to utilize their data for increased value creation in five broadly applicable value drivers. Although the investigation of the current state of big data adoption in Yara indicates that Yara is at an early stage of big data technology adoption, the thesis argues that Yara could increase their utilization of data quite rapidly during the coming years. By investing in data infrastructure, data governance and interconnectivity, Yara would ensure that new systems are able to communicate, and that data is more accessible. Investments in technology is, however, not enough. Yara also need to digitally mature through digital transformation. The thesis argues that it is important that Yara develop and continually update a digital strategy encompassing the entire organization.

Table of Contents

- Abstract.....I**
- Table of Contents.....II**
- Preface V**
- List of Figures..... VI**
- List of Tables..... VII**
- Concepts..... VIII**
- 1. Introduction.....1**
 - 1.1 Objectives and Limitations2**
 - 1.2 Background.....2**
 - 1.2.1 About Yara..... 3**
 - 1.2.2 Digital Farming in Yara..... 3**
 - 1.2.3 Big Data 4**
- 2. Theory5**
 - 2.1 Digitization5**
 - 2.1.1 Fundamentals of Digital Transformation 6**
 - 2.1.2 Digital Transformation in Practice 9**
 - 2.2 Big Data12**
 - 2.2.1 Capturing Data 12**
 - 2.2.2 The Three Vs 13**
 - 2.3 The Value of Big Data 15**
 - 2.3.1 Data Governance..... 15**
 - 2.4 Big Data Analytics 16**
 - 2.4.1 Visualization 17**
 - 2.4.2 Artificial Intelligence..... 17**
 - 2.4.3 Machine Learning..... 17**
 - 2.4.4 Data Mining 20**

| | |
|---|----|
| 2.5 Cognizant’s Data Maturity Scale | 21 |
| 2.6 McKinsey Global Institute’s Five Ways of Value Creation | 23 |
| 2.6.1 Increasing Data Transparency and Accessibility | 23 |
| 2.6.2 Data-Driven Decision Making..... | 24 |
| 2.6.3 Process Optimization..... | 26 |
| 2.6.4 Precisely Tailored Products and Services | 28 |
| 2.6.5 Enhanced Innovation..... | 29 |
| 3. Methodology | 30 |
| 3.1 Research Strategy | 30 |
| 3.1.1 Research Process..... | 30 |
| 3.1.2 Selection of Literature | 32 |
| 3.1.3 Choosing a Qualitative Research Strategy | 32 |
| 3.1.4 The Relationship Between Theory and Research | 32 |
| 3.2 Research Design..... | 33 |
| 3.3 Research Method..... | 33 |
| 3.3.1 Selection of Interview Objects | 34 |
| 3.3.2 Data Collection..... | 34 |
| 3.4 Data Analysis | 37 |
| 4. Analysis..... | 38 |
| H1. Data Transparency and Accessibility | 40 |
| H1.1 Data Infrastructure..... | 40 |
| H1.2 Methods of Access | 43 |
| H1.3 Traceability | 45 |
| Insight H1 | 48 |
| H2. Data-Driven Decision Making..... | 49 |
| H2.1 The decision-making process | 49 |
| H2.2 Input and Output of Analyses | 52 |
| H2.3 The level of trust | 54 |
| Insight H2 | 57 |
| H3. Internal Process Monitoring and Optimization | 58 |

| | |
|---|-----------|
| H3.1 Internal Connectivity | 59 |
| H3.2 Internal Modeling Capability | 61 |
| H3.3 Monitoring Capability/Rate of Monitoring..... | 62 |
| H3.4 Interconnectivity | 63 |
| Insight H3 | 65 |
| H4. Precisely Tailored Products and Services | 66 |
| H4.1 Data Collection Capability from External Sources..... | 66 |
| H4.2 External Modeling Capability | 68 |
| H4.3 Generative Capability..... | 69 |
| H4.4 End-user Communication | 70 |
| Insight H4 | 71 |
| H5. Enhanced Innovation Through Product Data | 71 |
| 5. Quality of study..... | 73 |
| 5.1 External Reliability | 73 |
| 5.2 Internal Reliability..... | 74 |
| 5.3 Internal Validity | 74 |
| 5.4 External Validity..... | 74 |
| 6. Conclusion | 77 |
| 6.1 The Current State of Big Data Technology Adoption in Yara | 77 |
| 6.2 Suggestions for What Should be Emphasized in Coming Years..... | 79 |
| Bibliography..... | 82 |
| Appendix 1: Interview Guide..... | 89 |

Preface

This master thesis is the concluding work of our Master of Science at the University of Stavanger. The thesis was written during the spring of 2018 for the Department of Industrial Economics, Risk Management and Planning. The thesis was written in collaboration with, and as a study of, Yara International ASA.

We would like to thank our supervisor, Professor Jan Frick, for his valuable guidance throughout the semester. He has been a great support during our frequent meetings. We are sincerely grateful for your contributions!

Thank you to our sponsor, Stefan Fürnsinn, and our external supervisor, Pål Øystein Stormorken – we appreciate you taking the time to make this thesis a reality.

Thank you to all the interviewees at Yara. Without your help, this thesis would not have been realized.

We would also like to thank the employees in Yara that has spent time on facilitating for this thesis. Thank you, Maria Stæger-Holst, Ersin Bircan, Eline Netland, Eline Sambu and Linn Häkkinen. Your positivity and encouragement during these months are truly appreciated.

Finally, we would like to show our gratitude to our family and friends for their support and patience throughout the semester. Thank you, Jenny Kristine Mazarino, and INDØKS, for making these years such a joy.

Stavanger, June 14th, 2018



Didrik Kahrs Skaale



Edvin Rygh

List of Figures

| | |
|---|----|
| Figure 1: Illustration of Kaikaku and Kaizen and how they could be combined | 8 |
| Figure 2: Illustration of a broadly applicable framework for digitalization..... | 10 |
| Figure 3: Illustration of high bias and variance introduced by underfitting and overfitting of machine learning algorithms | 19 |
| Figure 4: Illustration showing how the value of analysis increases with the scale of data, and the depth of analysis | 21 |
| Figure 5: Illustration showing how information can flow between a digital twin and a physical process..... | 28 |
| Figure 6: Illustration of the research process of this thesis. | 31 |
| Figure 7: Illustration of how the observations are linked to measurables and value drivers ... | 39 |

List of Tables

| | |
|--|----|
| Table 1: Characteristics of companies defined as digitally early, developing or maturing | 9 |
| Table 2: Overview over sampled interviewees, their respective index and their relevance | 35 |
| Table 3: Illustrates the refinement process of the observations | 37 |
| Table 4: Semi-structured interview guide for employees in Yara..... | 90 |

Concepts

| | |
|--------------------------------|--|
| <i>Analytics:</i> | The process of refining data into useful insights and knowledge |
| <i>Data cleaning:</i> | An activity that aims to increase data integrity and usability |
| <i>Data governance:</i> | An organization-wide activity based on a set of principles that aim to preserve data integrity and usability |
| <i>Data infrastructure:</i> | The backbone allowing data to be protected, stored, transported, processed etc. A storage space for data coupled with supporting activities that promote accessibility |
| <i>Digitization:</i> | The conversion of data from analogue into a digital form |
| <i>Digitalization:</i> | The act of digitizing processes |
| <i>Digital maturity:</i> | A measure of how well companies adapt in a digital environment |
| <i>Digital transformation:</i> | An improvement journey a company undertakes in order to increase their digital maturity |
| <i>Disparate data set:</i> | A dataset from a data system that was designed to operate without exchanging data or interacting with other data systems |
| <i>Information system:</i> | A system for the collection, organization, storage and communication of data |

1. Introduction

As digital technologies rapidly evolve, and businesses are increasingly focusing on adapting the very leading edge of technology, market situations are altering faster than ever before. Maintaining competitive advantages require progressively more agile business models that improve internal efficiency and strengthen product differentiation. (Kerravala & Miller, 2017)

Through digitization, businesses are generating large volumes of high variety data at high velocities, providing great opportunities for value generation (IBM: The Big Data & Analytics Hub, u.d.). This value is unlocked by using data to gain increased knowledge (Wu, Zhu, Wu, & Ding, 2014). Big data technologies are enabling organizations to utilize data for an increasing number of applications, increasing companies' knowledge and insight into both internal and external factors. This enhances their ability to improve internal efficiency, exploit external opportunities and in extreme cases disrupt entire business domains. (Hurwitz J. S., Nugent, Halper, & Kaufman, 2013), (Parviainen, Tihinen, Kääriäinen, & Teppola, 2017)

Technology alone, however, is not enough. Organizations also need to have a digital maturity level that facilitates for organization-wide adoption of digital technologies (C. Kane, Palmer, Nguyen Phillips, Kiron, & Buckley, 2015). Many organizations may believe that they can gain more value from their big data than they are able to realize in practice as most of the discussion around the value of big data are characterized by optimism (Hurwitz J. S., Nugent, Halper, & Kaufman, 2013). Hence, the authors believe that it is important to analyze how organizations translate or fail to translate their big data into value. By being increasingly data driven, business could improve decisions and efficiency of processes, thus lowering cost and increasing profits. (Kerravala & Miller, 2017)

The thesis conducts a case study of Yara International ASA, a leading chemical manufacturer, with the purpose to identify and illustrate which factors contribute to successfully translate data into value. It builds on the assumption that Yara has great opportunities to strengthen their big data capabilities and thus unlock large values.

1.1 Objectives and Limitations

The objective of this thesis is to answer the following research questions:

RQ1: *What is the current state of big data technology adoption through digitalization in Yara?*

RQ2: *What should be emphasized in the coming years to unlock potential value from big data?*

The research questions will be answered by running diagnostics across Yara's departments. The diagnostics aim to take a snapshot of the current state in Yara and analyze it. The analysis aims to present insight about what Yara should emphasize in the coming years to unlock more value from big data. The authors' perspective from outside the company aim to benefit Yara by identifying challenges and opportunities, without the biases an employee might have. The authors have an interdisciplinary background and therefore aim to present a study that bridges the gap between technical and business personnel. The thesis functions as a call to action for several levels of the organization.

The study is conducting semi-structured qualitative interviews that require a large amount of processing work, which limits the number of interviews that are possible to conduct. The large scope of the thesis limits the literature review to focus on the most important aspects of technologies and concepts relating to big data and digitalization. The scope excludes the aspect of data security although the authors acknowledge that data security is an important factor for organizations to consider while dealing with big data.

1.2 Background

The purpose of the following chapters is presenting context to the reader about the following:

- The development of Yara and what led them to where they are today
- Yara's market situation and operations
- The evolution of big data and a brief understanding of the opportunities that arises

1.2.1 About Yara

In 1903, Sam Eyde and Kristian Birkeland had successfully developed a process for direct nitrogen fixation (Yara International ASA, u.d.). Yara's roots date back to 1905 when Sam Eyde, Kristian Birkeland and Marcus Wallenberg founded the Norwegian industrial company Norsk Hydro that produced nitrogen fertilizers from the process Eyde and Birkeland developed (Yara International ASA, u.d.). The underlying reason was the widespread famine in Europe, and especially in Norway, as one of the poorest countries at the time. (Yara International ASA, u.d.).

In the following 90 years, Hydro expanded its operations from fertilizer to metals, oil and industrial products. The agricultural division made fast success abroad, leading to a number of acquisitions and new sales offices on the continents. In 2004, the division de-merged from Hydro and was stock listed on the Oslo Stock Exchange as Yara International ASA (Yara International ASA, u.d.).

Today Yara's activities are divided into crop nutrition solutions, nitrogen application solutions and environmental solutions. They currently have 15 000 employees and are distributing to 160 countries. Today Yara has integrated their entire value chain, following acquisitions. (Yara International ASA, u.d.)

1.2.2 Digital Farming in Yara

Yara has acknowledged the need for a digital revolution in the agriculture industry to be able to keep up with the growing population and the increasing scarcity of resources (Yara International ASA, u.d.) as they strive towards the mission that states: "*Responsibly feed the world and protect the planet*" (Yara International ASA, u.d.). More food must be produced from less, and they recognize that digital tools will play an important part in tackling this the coming years. As a result, the department called "Digital Farming" has been established (Yara International ASA, u.d.). Digital Farming are centered around four "Digital Hubs" that works as competence centers, located in Germany, Brazil, USA and Singapore. The locations are chosen for being close to both core markets and digital talent (Yara International ASA, u.d.).

Yara recognize that the use of digital tools will have positive effects for both the company internally and for farmers externally. Farmers could gain insight and information for enhanced decision making, be able to micro-segment fertilization, use data- and computer-driven decision support and have easier access to information (Yara International ASA, u.d.). This would leave Yara with data so that internal processes and products could be optimized.

Yara's slogan is "knowledge grows", and one of the challenges for Digital Farming will be to translate the unique knowledge that Yara has into tools that will contribute to shape the future (Yara International ASA, u.d.).

1.2.3 Big Data

Big Data was a problem for many companies in the early 2000s. The lack of processing power and storage capacities made it hard to handle the increasing amount of data. As demand for hardware rose and technology improved, prices for components fell. This eventually made it possible for more companies to afford equipment and to actually utilize the data (Russom, 2011). As more companies are digitized and collectors of big data, powerful tools are emerging to handle and utilize it. Big data analytics is considered a game changer in the business due to its ability to improve business efficiency and effectiveness. Big data analytics is becoming increasingly important in decision making (Wamba, et al., 2017), and is a gateway to less confirmation biased decisions throughout industries (Günther, Mehrizi, Huysman, & Feldberg, 2017). There are however digitized companies that either do not see the underlying value or do not know what to do with their data.

2. Theory

The purpose of this chapter is to provide the theoretical foundation of the thesis. It introduces important terms and concepts used to answer the research questions.

2.1 Digitization

Businesses are becoming increasingly digital by both applying technology to build new operating models, processes, software and systems, and by exploiting the convergence of people, business and things. From this continuous progress, new product and service opportunities emerge and business operations are transformed, thus gaining higher revenue, efficiency and competitive advantage. Maintaining the competitive advantage is achievable by rapidly adapting to changes and exploiting opportunities as they arise. Agility and digitization are therefore closely related and leading factors for competitiveness that should be prioritized by businesses and IT leaders. (Kerravala & Miller, 2017)

As the global economy is becoming increasingly dynamic, customers continue to expect more from products and services. Product differentiation is harder than it previously was and can no longer be done through altering single factors such as price, quality, features or support. Companies therefore have to find ways to retain current customers and attract new ones, the latter often being far more expensive than the first. The sum of perceived value, functional match to requirements, top notch quality and total customer experience, now determines how the customer bases move within markets and industries. Later years, social and environmental aspects of businesses have also made a bigger impact on customer behavior. Accessibility of information has made it possible for customers to easily explore options at competitors and move their businesses. It has also led to transparency within industries and for companies to be held responsible for actions in public media. (Kerravala & Miller, 2017)

2.1.1 Fundamentals of Digital Transformation

(C. Kane, Palmer, Nguyen Phillips, Kiron, & Buckley, 2015) argues that simply digitizing processes or work products through applying new technology is not enough. Businesses also need to undergo a more fundamental change called digital transformation to reach a digital maturity level that fosters data-driven innovation and decision-making. (Parviainen, Tihinen, Kääriäinen, & Teppola, 2017) defines digital transformation as the changes in ways of working, roles and business offerings that is caused by the adoption of digital technologies in an organization. These changes come at several levels of an organization, including:

- Process level: Organizations are adopting technologies that allow them to streamline their processes, automating tasks and reducing manual labor.
- Organization level: Organizations are offering new services and discarding obsolete practices. Existing processes are offered in new, more digital ways.
- Business domain level: There is a change of roles, and value chains in ecosystems.
- Society level: There are changes in the society structures of an organization. These changes affect how people work, interact and influence decision-making. (Parviainen, Tihinen, Kääriäinen, & Teppola, 2017)

The impact of digital transformation can be identified from three viewpoints. These are:

- Internal efficiency: Digital transformation allows organizations to streamline their own internal processes through changes in roles and tasks.
- External opportunities: Digital transformation can lead to new business opportunities in existing business domains through new services, customers and insights.
- Disruptive change: Digitalization can completely change entire business domains. (Parviainen, Tihinen, Kääriäinen, & Teppola, 2017)

Digital maturity is a measure of how well a business adapts in a digital environment. Businesses with a high digital maturity generally have more success applying digital technologies. Digital transformation is an improvement process that elevate a business' digital maturity level (C. Kane, Palmer, Nguyen Phillips, Kiron, & Buckley, 2015).

(C. Kane, Palmer, Nguyen Phillips, Kiron, & Buckley, 2015) identifies digital strategy, not technology, as the most important factor for digital transformation. They claim that the strength of digital technologies stems from how organizations integrate them to transform how they do business. Less digital mature companies tend to focus on individual technologies, and the focus of the strategy is highly operational. In more digital mature companies, the strategy focuses on how technology can be used to transform the business. A lack of an overall digitalization strategy, and competing priorities, were identified as the largest obstacles to digital transformation by (C. Kane, Palmer, Nguyen Phillips, Kiron, & Buckley, 2015).

Digitally mature organizations are generally less concerned about taking risk. They see failure as a prerequisite for success and encourage employees to be less risk averse (C. Kane, Palmer, Nguyen Phillips, Kiron, & Buckley, 2015). By being comfortable taking risk, digitally mature organizations are more inclined to undergo radical transformations, requiring larger commitments in time and resources compared to smaller continuous improvements. In Lean, such large-scale more radical changes are called Kaikaku, the Japanese word for “radical improvement or change”. Smaller, continuous improvements are called Kaizen, the Japanese word for “continuous incremental improvement” (Seeliger, Awalegaonkar, Lampiris, & Bellomo, 2004). (C. Kane, Palmer, Nguyen Phillips, Kiron, & Buckley, 2015) argues that if a company sees innovation as something incremental, it will be marginalized in the coming years, and as such call for more radical changes in companies.

Kaizen and Kaikaku can be combined, utilizing Kaikaku for inducing radical changes, and using Kaizen to continuously maintain and improve the operational impact. See Figure 1. Taichii Ohno, who pioneered the Lean model at Toyota, experienced success with a combination of both Kaizen and Kaikaku. When Toyota is introducing a new car line, or creating a new factory, they still use a Kaikaku-like production preparation process. (Seeliger, Awalegaonkar, Lampiris, & Bellomo, 2004)

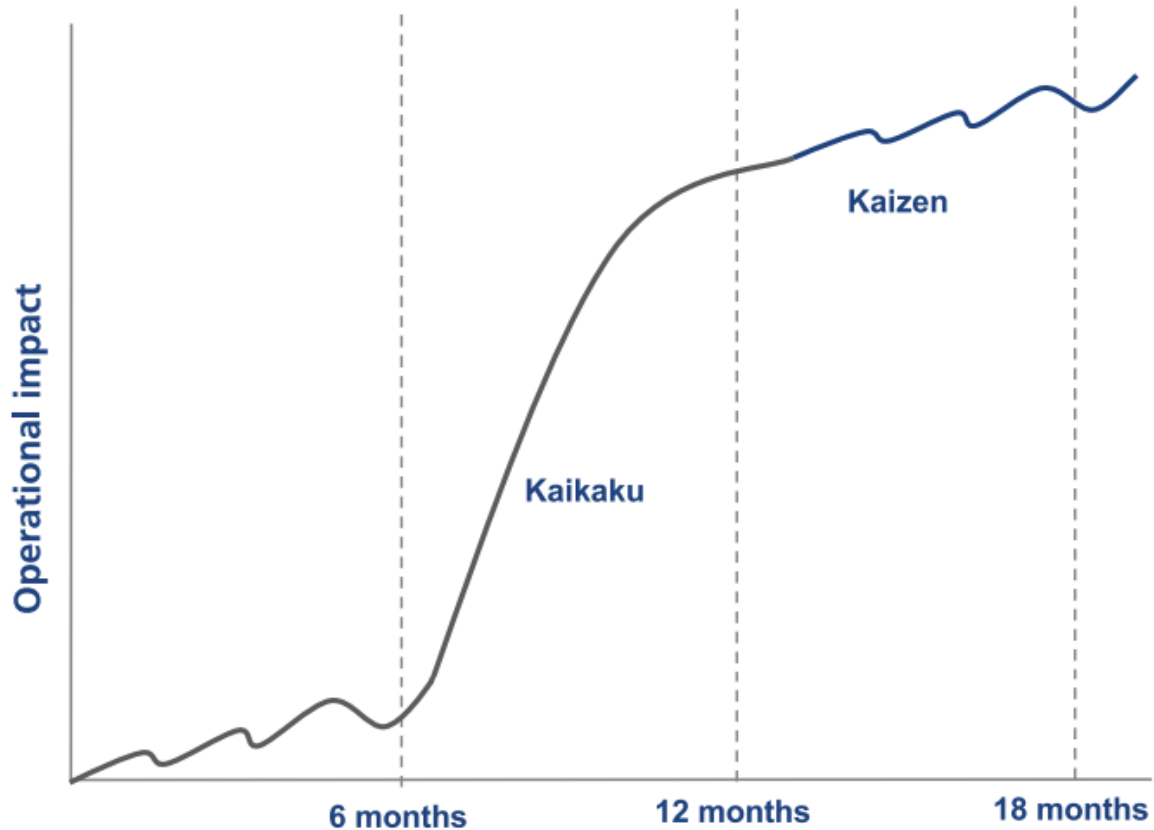


Figure 1 shows the operational impact over time of Kaikaku and Kaizen. Note how they can be combined so that Kaikaku is used to rapidly increase the organizational impact, while Kaizen maintains and incrementally increases the organizational impact over time (Seeliger, Awalegaonkar, Lampiris, & Bellomo, 2004).

(C. Kane, Palmer, Nguyen Phillips, Kiron, & Buckley, 2015) also identifies that having a culture contributive to digital transformation is a key feature of digitally maturing companies. These companies have, through digital transformation, created a culture that encourages risk taking, fosters innovation and develops collaborative work environments. Thus, organizations undergoing a digital transformation should strive to facilitate for these cultural changes. There are several ways to impact culture, for instance, by having the work environment facilitate for more multidisciplinary communication, gamification through contests and leaderboards or through internal storytelling (C. Kane, Palmer, Nguyen Phillips, Kiron, & Buckley, 2015). There are arguments stating that technology adoption is shaping the culture of an organization, and there are arguments saying that the culture shapes the adoption of technology (C. Kane, Palmer, Nguyen Phillips, Kiron, & Buckley, 2015). In any case, it is evident that culture and technology adoption is strongly codependent.

To develop and drive a strategy that fosters digital transformation through cultural and technological advancements, changing mindsets and processes, leaders that lead by example, with proficiency in digital trends and technologies, is highly beneficial (C. Kane, Palmer, Nguyen Phillips, Kiron, & Buckley, 2015). In addition to highly skilled leaders, maturing organizations are characterized by a strong focus on talent development in digital skills.

Table 1 shows characteristics of companies that are classified as having an early, developing and maturing digital maturity. Note how a lack of strategy and a siloed culture characterizes companies that are early in their digital transformation, while digitally maturing companies are characterized by a high focus on transformation, innovation, talent management and collaboration (C. Kane, Palmer, Nguyen Phillips, Kiron, & Buckley, 2015).

| | EARLY | DEVELOPING | MATURING |
|---------------------------|---|---|--|
| Barriers | <i>Lack of strategy</i> More than half cite "lack of strategy" as a top-three barrier | <i>Managing distractions</i> Nearly half indicate "too many competing priorities" is a top-three barrier, "lack of strategy" still a challenge for one-third | <i>Security focus</i> Nearly 30% cite security as a top-three barrier; managing too many competing priorities remains a top concern for 38% |
| Strategy | <i>Customer and productivity driven</i> Approximately 80% cite focus on customer experience (CX) and efficiency growth | <i>Growing vision</i> CX and efficiency growth; over 70% cite focus on transformation, innovation and decision making | <i>Transformative vision</i> Over 87% cite focus on transformation, innovation and decision making |
| Culture | <i>Siloed</i> 34% collaborative; 26% innovative compared to competitors | <i>Integrating</i> 57% collaborative; 54% innovative compared to competitors | <i>Integrated and innovative</i> 81% collaborative; 83% innovative compared to competitors |
| Talent Development | <i>Tepid interest</i> 19% say their company provides resources to obtain digital skills | <i>Investing</i> 43% say their company provides resources to obtain digital skills | <i>Committed</i> 76% say their company provides resources to obtain digital skills |
| Leadership | <i>Lacking skills</i> 15% say leadership has sufficient digital skills | <i>Learning</i> 39% say leadership has sufficient digital skills | <i>Sophisticated</i> 76% say leadership has sufficient digital skills |

2.1.2 Digital Transformation in Practice

(Parviainen, Tihinen, Kääriäinen, & Teppola, 2017) proposes a conceptual framework for digital transformation that provides a more tangible approach that companies can utilize in their digital transformation endeavors. The framework is synthesized from feedback and experiences the authors have collected during case studies in four types of industrial companies. The framework is designed to be general, and they argue that it can be applied for most companies. It follows plan-do-check-act principles for improvements at a high level. This section summarizes the framework proposed by (Parviainen, Tihinen, Kääriäinen, & Teppola, 2017).

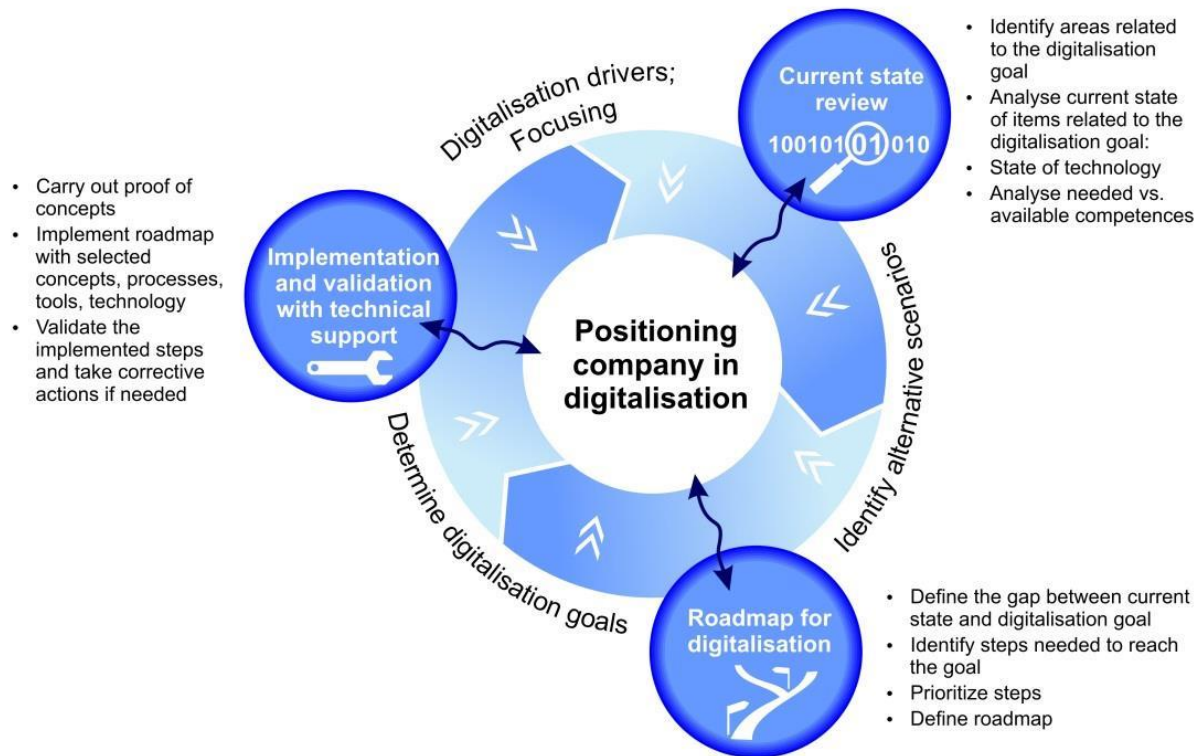


Figure 2 illustrates a broadly applicable framework for digital transformation in practice, developed through several case studies of companies from different industries. Note how the framework is iterative and presents guidelines for how a company should progress after defining their digitalization goals. (Parviainen, Tihinen, Kääriäinen, & Teppola, 2017)

Step 1, Positioning a company in digitalization: Analyze the potential impact of digital transformation. This is a rather big step that analyzes what the company stands to gain or lose through digitalization. Step 1 is divided into four sub-steps:

- Digitalization impacts: Identify and analyze current and upcoming trends and their relevance to the company's business domain. How far the business domain in general is in the adoption of these trends should also be analyzed.
- Digitalization drivers: Look at relevant trends identified above and analyze the impact each trend will have on the company. The importance of each trend to the company should be defined.
- Digitalization scenarios: Identify potential scenarios for the company's future based on the digitalization drivers. This stage evaluates the costs, risks and benefits of implementations.
- Digitalization goals: Define the company's digitalization process by analyzing selected scenarios from the previous step and their feasibility for the company. Define business-

related measurable that allow evaluation of the different parts of the defined digitalization process.

Step 2, Review of the current state: Review the current state of the company compared to the desired state after the digital transformation. The gap between the current situation and the desired future should be identified. Step 2 is divided into two sub-steps:

- Analyze impacted areas: Identify which areas of the company will be impacted by which goals defined in step 1.
- Analyze the situation in impacted areas: The current situation of the affected areas is evaluated in relation to the desired future state of the area.

Step 3, Roadmap for digitalization: Define how to close the gap identified in Step 2. Define the concrete actions that are needed to reach the desired state. Step 3 concludes with a detailed plan for reaching each goal defined in step 1. Step 3 is divided into four subsets:

- Identifying the gap: The gap between the current state, defined in Step 2, and the desired future state after completing the digitalization goal is identified.
- Identifying actions to close the gap: Actions can be taking on new technologies, optimizing existing processes or re-defining processes with the use of digital tools. An analysis should be conducted to identify which processes have the highest potential to benefit from digitalization. Key Performance Indicators should be evaluated and updated to meet new business targets.
- Analyzing the feasibility of actions: The feasibility of actions defined in subset 2 should be analyzed and the actions should be prioritized following for instance a cost-benefit analysis or an impact analysis. Trials and prototypes are helpful for gaining a deeper understanding of which actions are needed.
- Defining a digital roadmap: Once feasible actions have been defined and prioritized, they can be arranged into an actual roadmap. The roadmap should define the order, importance and responsibilities for each action.

Step 4, Implementation with technical support: Implement and validate the actions in Step 3. Return to previous steps as needed. The model is iterative, meaning that after one cycle is done,

it should be continuously repeated, building solutions and fine tuning the digital effort of the company. When technical advancements are attempted, it is often useful to first implement proof-of-concepts. The validation of the implemented actions should analyze whether the actions lead to desired impacts. In case desired impacts are not met, corrective actions should be considered.

2.2 Big Data

(Diebold, 2012) argues that the origin of the term big data is rather fuzzy, and that it “probably originated in lunch-table conversations at Silicon Graphics Inc. (SGI) in the mid 1990s, in which John Mashed figured prominently.” (Diebold, 2012, p. 5). Despite the origin in the early nineties, (Gandomi & Haider, 2015) show that the use of the term was not widespread in scientific documents until 2011.

2.2.1 Capturing Data

Companies struggled for years on how to capture information about customers, products or services. It was fairly uncomplicated while having a small number of customers and even fewer products. Over time, technology have developed rapidly, and markets have grown to become more complicated – often consisting of diversified companies with wide product lines. While the technology developed, the prices of equipment also dropped – which made it possible for considerably more companies to utilize new technologies such as big data analytics (Hurwitz J. S., Nugent, Halper, & Kaufman, 2013). The result is a larger amount of data coming from a higher number of sources. As much as 18.9 billion network connections was predicted by 2016 (IBM: The Big Data & Analytics Hub, u.d.).

Such development has led to great complexity in the digital world. Parts of the new data are structured and stored in conventional databases, while most new data are highly unstructured and harder to handle. Unstructured data could typically consist of documents, pictures, video, or human generated data as click-stream data from websites or social media uploads. The accessibility and embracement of powerful mobile devices connected to the internet is, and will

continue to be, an essential factor for the fast growth of the digital universe (Hurwitz J. S., Nugent, Halper, & Kaufman, 2013).

2.2.2 The Three Vs

What characterizes this so called 'big data', and what is it good for? (McAfee & Brynjolfsson, *Big Data: The Management Revolution*, 2012) describes the purpose of big data to gather intelligence and develop it into business advantage. Big data has for several years been defined by the three Vs: *Volume*, *Velocity* and *Variety* (Laney, 2001).

Volume is undoubtedly one of the main characteristics of big data. Big data volume could be quantified in different ways, most commonly as bytes, but also as files, tables or in terms of time (Russom, 2011). However, defining the amount where data are considered big data would be impractical as the storage capacities are continuously growing, thus what is considered big data today will not be considered that in the future. The growing storage capacity will continuously allow for bigger data sets to be captured (Gandomi & Haider, 2015). To put the enormous volume growth of data in the previous years into context; the amount of data crossing the internet every second in 2012 exceeded what were stored on the entire internet 20 years before that (McAfee & Brynjolfsson, *Big Data: The Management Revolution*, 2012). (EMC, IDC & Cyclone Interactive, 2014) estimates that the data in the digital universe increases from 4.4 zettabytes in 2010 to 44,4 in 2020.

Variety is a big data characteristic that is defined by the increasing variety of sources data are gathered from. Big data can derive from social network uploads, sensor data, GPS signals and more (Russom, 2011). As (McAfee & Brynjolfsson, *Big Data: The Management Revolution*, 2012, p. 5) put it: "*Each of us is now a walking data generator.*". This is illustrated by (IBM: *The Big Data & Analytics Hub*, u.d.), claiming there were 420 million wearable, wireless health monitors in use by 2014 and that 6 billion out of the total 7 billion people in the world have cellphones. For several years, companies have simply been hoarding consumer data without realizing its value. Technology has made it possible to analyze it in a more complex way, even though several of the new data sources creates more noise than previous structured business data. There are undoubtedly high technical demands to big data analytic tools as they have to find signal in highly unstructured data (Russom, 2011).

Velocity of big data has been increasingly relevant as the real-time data streaming and information transferring has grown (Russom, 2011). In previous years, the technology did not keep up with the velocity of the data and thus could not be analyzed within a sufficient time frame (Hurwitz J. S., Nugent, Halper, & Kaufman, 2013). As (Hurwitz J. S., Nugent, Halper, & Kaufman, 2013, p. 21) put it: “In the end, those who really wanted to go to the enormous effort of analyzing this data were forced to work with snapshots of data. This has the undesirable effect of missing important events because they were not in a particular snapshot.”. The real-time information transfer makes it possible for companies to be more agile and thus make decisions and respond to changes faster, given that they are able to use analytic tools in real-time as well. This could clearly lead to a competitive advantage in the market (McAfee & Brynjolfsson, *Big Data: The Management Revolution*, 2012).

In later years, other Vs have been suggested as additional big data defining characteristics. IBM presented the fourth V, Veracity, as measurement of unreliability in data sources. The data might not be trustworthy, in example social media data which entail human judgement, and thus might be imprecise and uncertain (Gandomi & Haider, 2015). As much as one third of business leaders do not trust their data according to IBM’s infographics (IBM: *The Big Data & Analytics Hub*, u.d.). A fifth V, Variability, was introduced by SAS. Variability is the change in data flow rates, as the velocity of big data might not be consistent (Gandomi & Haider, 2015).

However, it is important to consider that the characteristics are dependent of each other, and a change in one of them would most likely influence others. Despite the defining Vs, it is claimed that the true internal limits of big data are continuously evolving with technology development, and are dependent upon factors such as size, sector and location of the firm (Gandomi & Haider, 2015). (Gandomi & Haider, 2015) also claim that every firm has its “three-V tipping point”, the point where traditional data management and analysis technologies are insufficient for gathering value adding business intelligence. Passing this point, means that the firm are entering the world of big data, and should trade-off implementation cost against expected future value extracted from big data technologies. (Gandomi & Haider, 2015)

2.3 The Value of Big Data

Organizations are collecting, mining and exploiting data from both internal and external sources at increasing rates (Loebbecke & Picot, 2015). This has led to many organizations finding themselves at the tree-V tipping point where they start dealing with big data (Hurwitz J. , Nugent, Halper, & Kaufman, 2013). Thus, big data has become a focus point for organizations and academics. This is mostly due to its perceived potential for generating business value in the form of operational and strategic enhancements (Wamba, Akter, Edwards, Chopin, & Gnanzou, 2015).

This trend of organizations capturing more and more data in combination with improvements to the analytics capability of most organizations has allowed for analyzing bigger and bigger datasets, often consisting of multiple databases (Hurwitz J. S., Nugent, Halper, & Kaufman, 2013). By purposely combining disparate datasets (i.e. multiple separate databases designed not to communicate with other information systems), data scientists are able to uncover previously unknown correlations. This makes it possible to get a more nuanced picture of the factors that influence for instance an event, improving predictive accuracy (Aaltonen & Tempini, 2014). This combined with the internet of things, giving machines, sensors and devices connectivity thus allowing organizations to capture real-time data from sensors and systems, lets organizations predict the present as well as the future (Loebbecke & Picot, 2015).

Literature on big data is characterized by optimism, and with good reason. However, most of the value of data is not realized until an organization is able to do something with it (Hurwitz J. , Nugent, Halper, & Kaufman, 2013). It is therefore crucial to understand how organizations can leverage their big data resources to generate value.

2.3.1 Data Governance

Many companies have for some time perceived data just as a part of doing business, and have not leveraged it properly. As markets are becoming increasingly data driven, there is a risk of being overtaken by competitors if data is not treated seriously (Tupper, 2011). Unfortunately, efforts in improving integrity of corporate data tends to be initiated at the point where the data is such a low quality that decisions are noticeably worse. This effort in improving data integrity

is called "data governance". Data governance is said to be a corporate wide activity (Linstedt & Inmon, 2014) committed to by all levels of management (Tupper, 2011). (Tupper, 2011) defines the foundation of data governance by eight policy principles. These principles emphasize reusability, quality, structure, ownership, ethics and internal sharing of data. Committing to such policy principles at all company levels encourage the use of data by making it more accessible, reliable, traceable and authentic (Infosys Limited, 2017), thus boosting competitive advantage (Tupper, 2011).

2.4 Big Data Analytics

Like crude oil was the catalyst for much of the technological advancements we enjoy today, data is predicted to be the catalyst for many of the technological advancements we will enjoy tomorrow. While crude oil is not very useful to the most of us, we acknowledge it as a valuable resource because we know we can refine it into something that is useful. The process of realizing value from crude oil can be seen in three steps. First crude oil is extracted. Then it is refined, increasing its potential value. This potential value is then realized either as fuel or as some sort of product that holds value for a customer. (Schmarzo, Economic Value of Data (EvD) Challenges , 2017)

Similarly, data holds limited value until it can be utilized. Realizing value from data can also be described by a similar three-stage process of collection, refinement and realization. While crude oil and data makes for very entertaining analogies, there is a couple of major differences. Crude oil is physical and will be consumed to realize value, while data can be reused infinitely. Crude oil is also a commodity. It is traded with the fundamental understanding that each barrel of a certain grade of oil is *exactly* the same. This is not true for data. As such, the value of data is not very tangible. (Schmarzo, Economic Value of Data (EvD) Challenges , 2017)

Analytics can be considered the refinement process of data (Schmarzo & Sidaoui, n.d). By performing different analyses on datasets, one can increase the potential value and usability of data. There are a wide variety of analytic tools, ranging in sophistication and extent from simple data visualization like plotting a line graph or a pie chart to deep learning algorithms. As data can be reused infinitely, the same pieces of data can be refined through multiple analyses. Increased analytics capability has allowed organizations to perform analyses on truly massive

data sets. Big data analytics is a common name for methods used to perform analyses on such vast amounts of data (Hurwitz J. , Nugent, Halper, & Kaufman, 2013).

2.4.1 Visualization

The simplest form of big data analytics is visualization. There are a wide range of different visualization tools with varying sophistication. Some tools, can build dynamic dashboards and reports, while others are great at producing graphs and tables that can be manipulated to extract actionable information from huge datasets. Dynamic dashboards and reports replace common static reports that tend to be made ad hoc and expire rather quickly, and are especially viable for visualizing rapidly changing data (dataPARC, 2017). Visualization software is typically also used in combination with more advanced analytics. Most mature organizations already use some sort of visualization software, especially to extract knowledge from business intelligence data (Ohara, 2012).

2.4.2 Artificial Intelligence

Artificial intelligence is a term encompassing several technologies used to simulate intelligence. Today different kinds of artificial intelligence are used to solve problems that was earlier believed to be impossible to solve without human interference. Some types of artificial intelligence are highly useful for big data analytics. The most commonly types of artificial intelligence used for big data analytics include machine learning and data mining (Russell & Norvig, 2009).

2.4.3 Machine Learning

Machine learning algorithms are typically classified as supervised or unsupervised learning algorithms (Megahed & Jones-Farmer, 2015). Supervised learning algorithms are used for two types of tasks; regression and classification. In supervised learning, a training set of data, \mathbf{x} , with correct labels, \mathbf{Y} , are iterated upon.

The goal is to find the function:

$$Y = f(x) + \epsilon$$

where: $Y = \text{output}$ (label or a continuous numerical value from regression)
 $x = \text{input}$ (dataset – simple or complex)
 $f = \text{function describing the relationship between input and output}$
 $\epsilon = \text{random error term with mean zero}$

that gives the most precise output Y for all the values in the dataset. Once a sufficient precision has been reached, this function can then be used to predict Y values for new x values. (Maini, Machine Learning for Humans, Part 2.1: Supervised Learning, 2017)

When supervised learning algorithms are used for regression, the output is typically a continuous numerical value. Examples could be predicting how much a house will sell for, or what yield one can expect from a given field. Training is done through labeled inputs, and once training is done, the algorithm will output a continuous approximation based on the input. The accuracy of the algorithm is typically tested by running the algorithm on a test set of data. In all supervised learning algorithms, there is a bias-variance tradeoff. Bias is the amount of error that is introduced by attempting to approximate real-world phenomena with a simplified model. Variance is how much the model's test error changes based on variance in the training data. High bias occurs if a model is *underfitted*, and high variance occurs if it is *overfitted*. Underfitting occurs if your model is not complex enough to capture the underlying trend in the data, while overfitting happens when the algorithm “overlearns” the test data. It starts picking up trends that are not representable to what's happening in the real world. In order to have a good model, you need to have a model with low bias and low variance, one that is fitted just right. (Maini, Machine Learning for Humans, Part 2.1: Supervised Learning, 2017).

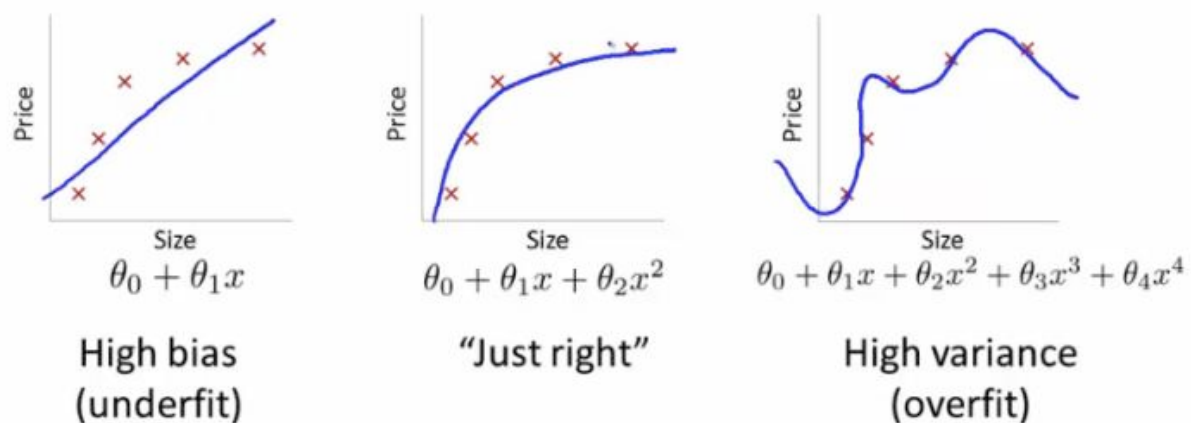


Figure 3 illustrates how training data influences algorithm performance. Note how underfitted algorithms lead to high bias in output data, and how overfitted algorithms lead to high variance in the output data. (Maini, Machine Learning for Humans, Part 2.1: Supervised Learning, 2017)

When supervised learning algorithms are used for classification, the algorithm outputs a label that is assigned to the input data. Typical use-cases are labeling images or sorting data. One example of applied supervised learning algorithms are labeling pornographic content to avoid pornographic content being shown to people who are not supposed to see it. (Maini, Machine Learning for Humans, Part 2.1: Supervised Learning, 2017)

After pornography started being distributed over internet, the world faced a problem – pornographic content would pop up in search engines, in ads etc. To conquer this problem, various initiatives were taken to try and label content portraying naked people. To begin with, the most successful method was man-made image classifying algorithms. These were made by people concocting smart methods for telling whether an image was of a naked person. (Fleck, Forsyth, & Bregler, 1996) used pixel colors that are likely to be of nude skin as their first screening. If more than 30% of the image contained pigments of yellow, red and brown, the image passed on to step two. This step consisted of trying to recognize limbs as connected lines. If certain patterns of limbs were discovered, and if these did not form unnatural angles, the image was classified as pornography. This method managed to correctly label and sensor 43% of images containing naked people, while only falsely labeling 4% of the non-pornographic control images.

Supervised learning algorithms replaces these man-made algorithms. Instead of the algorithm relying on what clever rules we can imagine using for classifying the image with, the algorithm

learns what it should look for itself. Today, state of the art Convolutional Neural Networks, a type of supervised learning, are classifying pornographic content with an accuracy of 97,2%. (Zhou, Zhuo, Geng, Zhang, & Li, 2016)

Unsupervised learning algorithms uses unlabeled data, meaning there are no Y component to the data, x . Unsupervised learning algorithms are typically used for clustering data into groups based on similarity, and for reducing dimensionality, essentially compressing the data while maintaining its structure and usefulness. Clustering is highly useful for segmenting, and for “cleaning up” databases, while dimensionality reduction can help simplify models, and reduce file sizes (Maini, Machine Learning for Humans, 2017). There are several open-source alternatives for machine learning that can be used for creating free proof-of-concepts and for training both supervised and unsupervised machine learning algorithms. For instance, Python with libraries like NumPy, SciKit-Learn, Keras, Tensorflow or Theano (Bobriakov, 2017).

2.4.4 Data Mining

Exploring large volumes of data, and extracting useful insight and knowledge, is the most fundamental task for big data applications. Data mining is a process that identifies and extract useful data in huge datasets. Much like mining for valuable resources, it *mines* out *ores* of useful data from huge volumes of “*dirt*” data (Wu, Zhu, Wu, & Ding, 2014). The process must be automatic or (more usually) semiautomatic (Witten & Frank, 2005). The patterns discovered must be meaningful, and lead to an advantage, usually in the form of economic value. Data mining usually uses techniques from machine learning, but these techniques are put to different means. Data mining is carried out by a person in a specific situation, on a specific dataset and with a specific purpose or goal in mind. Typically, this person uses visualization methods or the different pattern recognizing techniques from data mining such as clustering labeling and regression to discover new insight from a huge dataset or to predict future observations (Amatriain, 2016), (Gung, 2016).

2.5 Cognizant’s Data Maturity Scale

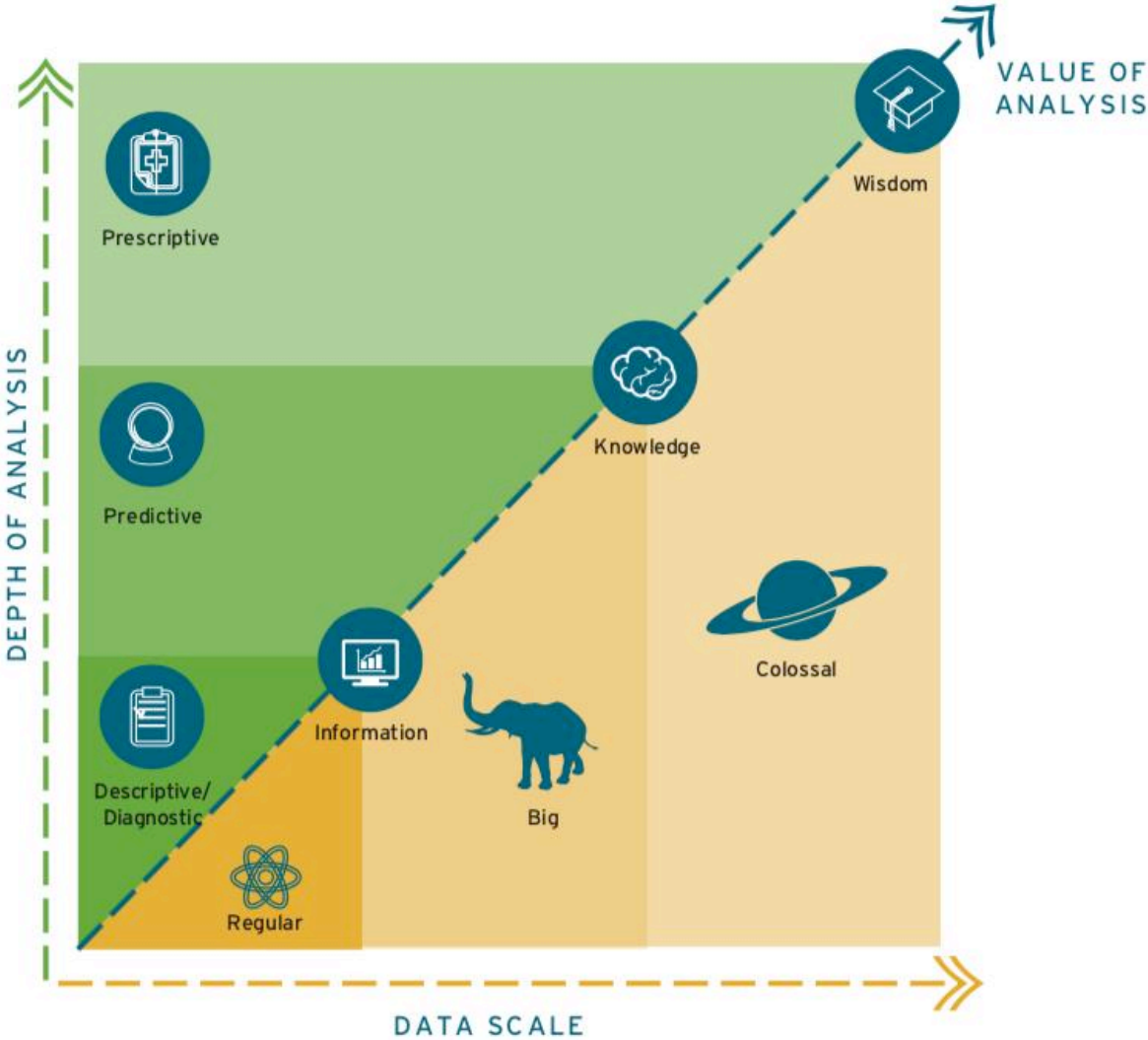


Figure 4 illustrates how the scale of data, and the depth of analysis impacts the degree of insight an organization can extract from their data. Note that the organization is dependent on both a large scale of data, and a high depth of analysis to increase the value of analysis. (Cognizant, 2016)

Data, analytics and value are closely related to each other. With data growing from regular data to big data and to “colossal” data, increased analytic capabilities unlocks value as seen in figure 4. Strengthening these analytic capabilities and then realizing value from the data often requires investments in new technology, digital tools, complex modeling and skilled personnel. Large companies with vast amounts of data would in addition have to see this potential value in a bigger picture, and not only in case specific projects within departments (Cognizant, 2016). (Cognizant, 2016, p. 2) says that: “Many organizations are overwhelmed by data because they fail to develop the right strategy to derive benefits from it. In fact, most organizations fail to look at all the aspects of colossal data collectively, and instead seek to implement individual

point solutions to address specific issues.”. (Cognizant, 2016) categorize this value into three elements:

- Information – details about something that has happened
- Knowledge – insights that explain why something has happened or future outcomes
- Wisdom – informs about future actions based on historic data (Cognizant, 2016).

If the foundation of data is sufficient, analytics should be run to obtain insights. (Cognizant, 2016) classifies the range of analytic capabilities as following:

Descriptive

Commonly consisting of static reports that present information about what has happened. This form of analytics requires small amounts of data.

Diagnostic

Diagnostic analytic mines data to figure out why something happened, and this requires some more data as foundation.

Predictive

This type of analytics uses statistical models and algorithms to improve the understanding of previous actions so that a prediction of what is going to happen is possible to make.

Prescriptive

Prescriptive analytics is the most advanced form of analytics – it defines the best course of future action based on a synergy of data, business and mathematics.

2.6 McKinsey Global Institute's Five Ways of Value Creation

McKinsey Global Institute (Manyika, et al., 2011) have identified five broadly applicable ways to leverage big data for value creation. These are:

- Increasing data transparency and accessibility
- Adopting data-driven decision making to improve decision making through sophisticated analytics
- Using more accurate and detailed data for process optimization
- Offering precisely tailored products or services through highly specific segmentation
- Using data from current products and processes to improve the development of the next generation of goods and services

2.6.1 Increasing Data Transparency and Accessibility

Simply making data more easily accessible to relevant stakeholders can create value. Only accessible information is useful (Tikhonov, Little, & Gregor, 2015), and more accessible data across separated departments and platforms can reduce search and processing time (Manyika, et al., 2011). This enhances cooperation and alignment between departments. Improved internal transparency also allows for better measurability within the organization. This leads to increased managerial knowledge, boosting decision making, resource planning and process optimization.

Physical data infrastructure is the prerequisite for several data-driven approaches as it makes data accessible, while security infrastructure is important to protect the data (Hurwitz J. S., Nugent, Halper, & Kaufman, 2013). A data lake acts as both physical infrastructure and security infrastructure and is an important tool for increasing the data transparency and accessibility. As opposed to a data warehouse where data is stored in a predetermined structure, data in a data lake is stored in its native format. This means that data governance, metadata and computer security are important supporting factors that needs to be in place for an organization to adopt a functioning data lake (Schmarzo & Sidaoui, n.d).

For organizations to utilize data with low quality, low degree of metadata or data governance, data cleaning is often required. Data cleaning is a process that detects, corrects, replaces, modifies, or removes messy data from a dataset or a database. Data cleaning is especially needed when integrating disparate datasets with others. Prevention is also typically more effective than curing, and it is therefore important for organizations to plan how their datasets could be combined. (Acaps, 2017)

One important aspect of data transparency is data traceability. Data traceability ensures that details are kept throughout the process where data is sourced and validated, triggering a respective workflow. This benefits companies in two ways, by informing the management and informing the employees. (Sentance, 2016)

The management are able to use these details to optimize processes and potentially fix inaccurate data. It could also improve decisions and drive better results by increasing efficiency and data quality. (Sentance, 2016)

Employees could gain better insight by knowing where data was sourced from, who executed the process and by who it is validated, derived, interpolated or normalized. Data may have been praised as good by colleagues, but someone still may want to figure out what process the data passed to achieve the given status, particularly if the data does not correspond with another source. (Sentance, 2016)

As traceability improves, visualization tools are becoming increasingly important. Visualization helps to spot patterns and anomalies in datasets, and thus turn data into insightful information that can be used in decision making. The combination of data, metadata and visualization tools is important for creating feedback loops, and ultimately, for improving quality of market data and efficiency of internal processes (Sentance, 2016).

2.6.2 Data-Driven Decision Making

Decisions depend on insight, and the degree of insight one can get depend on the detail level and quality of analyzed data. If the quality of input data is poor, the output of the analysis will be poor, providing less insight. If the level of detail in the input data is low, the output of the analysis will have a low level of detail (Gupta, 2014). Big data technologies enable managers

to gain new valuable business insights that may be directly translated into improved decision making and business performance, while allowing more accurate predictions and precise interventions. In most cases, it requires that data scientists translate patterns in data into business information and for decision makers then to embrace this information as evidence while making decisions (McAfee, Brynjolfsson, Davenport, Patil, & Barton, 2012). Evidence-based decision-making is reliant on being able to process high volumes of data with high velocity (Gandomi & Haider, 2015).

The technology has evolved to the point where retailers not only know what customers buy, but they can sophisticatedly analyze patterns and behavior to predict what the customer might need or want to buy at what time. The algorithms are continuously improving for each customer interaction, making it extremely hard for retailers that do not embrace this technology to stay in business with competitors that do. Companies that are born digital have made great accomplishments the last couple of years, but the more traditional businesses might have the greatest opportunities for creating new competitive advantages in their markets by transforming their businesses and improving their big data capabilities. (McAfee, Brynjolfsson, Davenport, Patil, & Barton, 2012)

(McAfee, Brynjolfsson, Davenport, Patil, & Barton, 2012, p. 5) says that "Data-driven decisions are better decisions—it's as simple as that. Using big data enables managers to decide on the basis of evidence rather than intuition. For that reason it has the potential to revolutionize management."

For this big data managerial revolution to be possible, a fundamental change in how decisions are made and who makes them might be necessary. In traditional companies digitized data is often scarce, and then it makes perfect sense to have executives or employees using their intuition and experience to make decisions. The more important the decisions are, the higher ranked the decision maker tend to be. This phenomenon is in the big data community recognized as "HiPPO" - the highest-paid person's opinion. (McAfee, Brynjolfsson, Davenport, Patil, & Barton, 2012) claim that executives often spice up reports with data to support and justify the decision that has already been made using HiPPOs. They point out that there are many executives that lets data override their opinion, but that intuition-driven decision-making is still too widespread in businesses. This assertion is based on their work with testing the hypothesis that data-driven companies are better performers. They conducted 330 structured interviews in

that regard, and the results were clear: The more companies identified themselves as data-driven, the better they performed on objective measures, such as financial and operational results. This resulted in measurable increase in stock market valuations for the data-driven companies in the study. (McAfee, Brynjolfsson, Davenport, Patil, & Barton, 2012)

Proper leadership seem to be a necessity while going through a big data transition, as habits and ways of working tend to influence from leaders to employees. There are simple techniques that leaders could follow, such as letting data overrule intuition in decision making and getting in the habit of asking "What do we know" instead of "What do we think?". It could be just as important to ask the right questions as to have more or better data. Human insight is undoubtedly needed in combination with big data capabilities and it will continue to be essential to have executives that spot opportunities, understand markets and have soft skills to facilitate for cross-functional cooperation and to handle stakeholders. (McAfee, Brynjolfsson, Davenport, Patil, & Barton, 2012)

2.6.3 Process Optimization

Utilizing big data for process optimization can be a great way to translate data into value. Most organizations strive for consistency in their repeated processes. This is especially true for manufacturing and production processes (Rice, 2017). Big data allows for better monitoring of processes, providing improved process control thus increasing process consistency. Improved data capabilities combined with internet of things-connected machines and sensors allow continuous monitoring and near real-time data analysis (Gillon, Aral, Ching-Yung, Mithas, & Zozulia, 2014) of data collected from massive systems. This data often consists of a multitude of data types, structured or unstructured (Qin, 2014). These massive systems in a complex value chain should be interconnected for the full value of the value chain to be unlocked (Heckler & Gates, 2017). This would make data available from across the value chain and empower new partnerships, collaborations (Günther, Mehrizi, Huysman, & Feldberg, 2017) and synergies across departments. Synergies particularly emerge from integrating acquired companies in the existing interconnected value chain (Corporate Finance Institute, u.d.). For the interconnectivity to be achievable, there should be a focus on both technological interconnectivity of systems, platforms and data, and on interconnectivity of controls, data governance and cyber security (Heckler & Gates, 2017).

Before big data, most process control measures were snapshots or samples. This means that only a subset of important information was collected to represent larger periods or batches (Hurwitz J. , Nugent, Halper, & Kaufman, 2013). Increased data capabilities now allow continuous monitoring of entire processes and can therefore detect anomalies that might have previously not been detected as they occurred outside the snapshot or sample. Big data technology also allows for analyzing historic data to identify in-control baseline samples that further monitoring can be compared to (Megahed & Jones-Farmer, 2013). This combination allows for better understanding of possible issues with the process at a faster rate compared to traditional process control methods, allowing earlier and more precise corrections.

One data-driven technology that is increasingly used for process monitoring and optimization, are digital twins. A digital twin is a digital representation of a physical object. Mirroring a physical process offers a powerful way to monitor and optimize the process if paired with connected sensors (Petty, 2017). Digital twins offer value through five benefits (Oracle, 2017):

- Increased visibility in the operations of machines and in large interconnected systems
- Being used to predict the future state of machines, which can be used for optimization of the process or maintenance
- One can use digital twins for what if-analyses where one can simulate certain scenarios
- Digital twins can be used as a communication and documentation mechanism
- If designed correctly, digital twins can be used to connect disparate systems in the value chain.

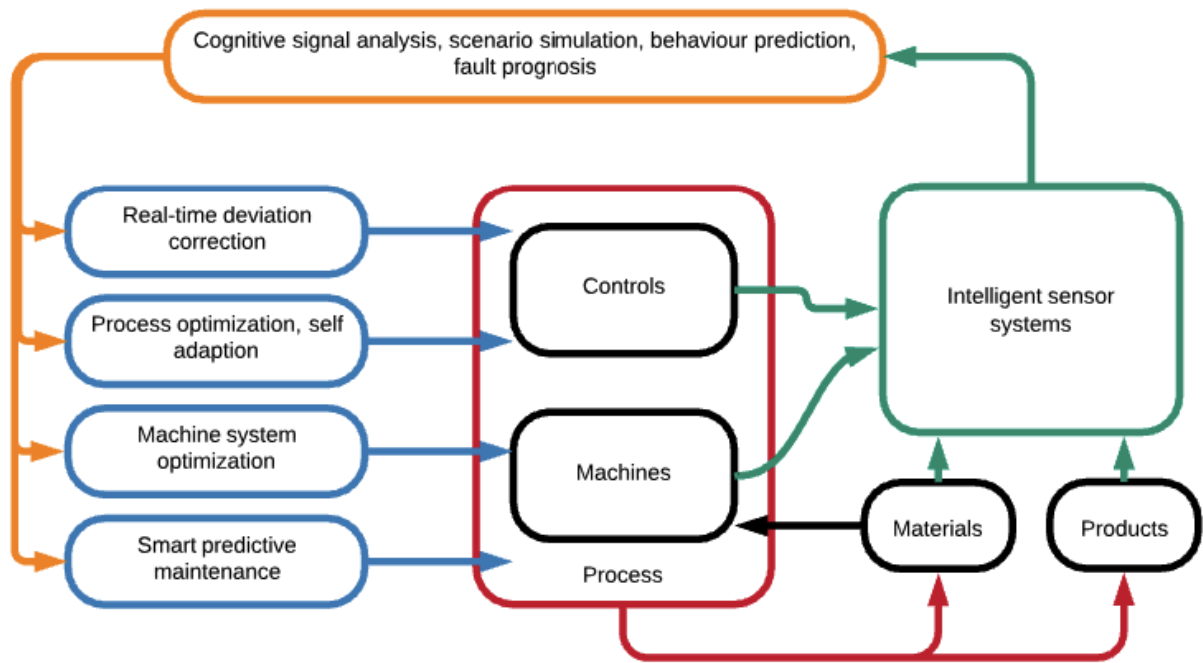


Figure 5 illustrates how information can flow between a digital twin and a physical process following the IFaCOM system for process monitoring and correction. The orange box is the digital twin. Note how intelligent sensor systems provide input data to the digital twin, and how this data is used to perform analyses and simulation, which in turn lead to corrective actions at the process level (Eleftheriadis & Myklebust, 2016).

2.6.4 Precisely Tailored Products and Services

Big customer data combined with sophisticated analytics can provide organizations with highly specific segmentations, called micro-segments, and detailed customer insight. This deep knowledge of past, current and potential customers, allows organizations to offer precisely tailored products and services that suit the individual customer's needs, adding customer value (Manyika, et al., 2011). Micro segmentation through big data has been used to great extent in the marketing and risk management scene, where actors have been able to produce highly detailed ads and recommendations for years. There is a wide range of external data that can be relevant for specialized segmenting, for instance previous shopping habits, customer visits or product usage (Datafloq, 2016).

Handling all this data and turning it into more customized offers and recommendations is a complex task, requiring more effort than less customized content. (Drew, 2017) argues that the increase in volume, cost and complexity is justified by increased performance. It also identifies creating massive quantities of personalized products, offers and recommendations to be a significant burden for scalability. In other words, there is a trade-off between how precisely one

is able to tailor ones' products and services, and scalability. If an organization has a high capability of producing personalized products, offers and recommendations, this trade-off becomes less significant. As such an organization's generative capability is an important factor for how well they can scale precisely tailored products and services. (Drew, 2017), (Ariker, Heller, Diaz, & Perrey, 2015)

2.6.5 Enhanced Innovation

Utilizing data from current products to understand how they are performing and how customers actually use the products can lead to great insight on which features and services customers want in future products, enhancing innovation (Manyika, et al., 2011). Manufacturers are using data from current products to enhance their existing products, develop new products and to innovate business models. The deep insight that big data analysis offers on increasingly dynamic markets, can help organizations set pricing strategies, manage their product portfolio and provide after-sale service offerings to customers.

3. Methodology

The purpose of this chapter is to explain the method of the thesis. First the research strategy and process (3.1) are presented. Further, research design (3.2) and research method (3.3) are described before data analysis (3.4) concludes the chapter.

3.1 Research Strategy

This section describes the research process, the selection of literature, the research strategy and the relationship between theory and research.

3.1.1 Research Process

The research process is illustrated in Figure 6. The process started with conducting a literature review on big data, which will be further elaborated in (3.1.2). After the literature review and discussions with Yara, it was possible to develop the research questions (RQ1), (RQ2) and objectives (1.1). After considering the literature review, a preliminary consideration of Yara's big data maturity and the research questions, the five value drivers (2.6) the thesis is built around were identified. The value drivers were subdivided into measurables to allow for increased measurability. Further the research design and methods were developed, allowing to initiate the process of booking interviewees in Yara. With interviewees in place, somewhat tailored interview guides could be developed. The interviews were conducted according to interview guides and then transcribed from sound recordings. Transcripts were analyzed, resulting in 28 observations within the five value drivers. The observations were discussed to identify and present causes, effects, and suggestions for the respective observations. These were then used to conclude with the current state of big data technology adoption in Yara, and suggestions for their future digital endeavors.

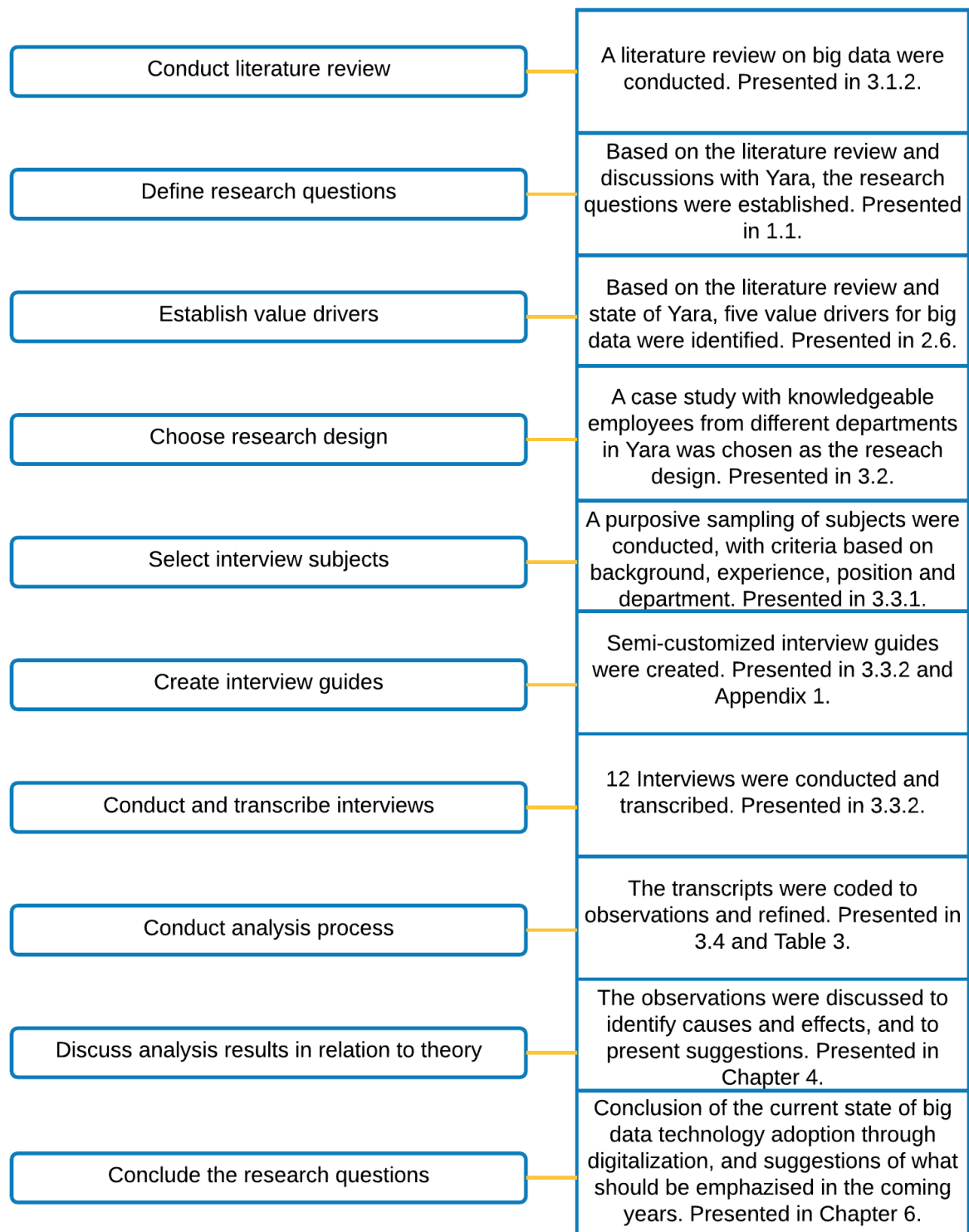


Figure 6 illustrates the research process of this thesis. The illustration is chronological from top to bottom and aims to show how the research progressed step-by-step from literature study and research questions through analysis to conclusion.

3.1.2 Selection of Literature

The initial literature review on big data and the identification of the five value drivers revealed which topics were necessary to further investigate, and how they were related. Most of the literature were collected from ScienceDirect and Google Scholar, and some from white papers. The white papers worked as a supplement in topics that had a lack of literature at academic databases such as ScienceDirect and Google Scholar. These two databases were chosen due to their high quality interdisciplinary content. They both cover most of this thesis' content, both business-wise and technical-wise. Both databases also support AND-/OR functionality and are considered high integrity, which has been important aspect for the authors. To find the most relevant literature, both key word search in these databases and reference search within high integrity sources found in key word searches were used.

3.1.3 Choosing a Qualitative Research Strategy

A qualitative research strategy has been undertaken, meaning that words have been emphasized rather than numbers during the collection and analysis of data. The authors chose an interpretivist epistemological and constructivist ontological stance in this qualitative study. Ontology refers to reality and what is there, and epistemology refers to the relationship between the authors and the reality. Constructivism means that each individual's construct of reality is influenced by social interactions and settings and is subject of continuous revision. As the interviewees share the same social settings and interact, their meanings and constructs are influenced by each other and increasingly aligned. The interpretivist stance led the authors to be involved throughout the interview process, interviewing a small number of employees in each department and to generate observations with causes that are rejected as absolute facts, but rather viewed as the perceived truth. (Bryman, 2012), (Carson, Gilmore, Perry, & Gronhaug, 2001)

3.1.4 The Relationship Between Theory and Research

The thesis has mainly had a deductive approach, as qualitative studies tends to have. A deductive approach means that the observations are based on theory, contrary to an inductive approach where the theory is based on observations. The deductive approach in the thesis has,

however, not been as linear as deductive models may illustrate (Bryman, 2012). (Bryman, 2012) explains how a fully linear approach may not be applicable as the researchers view on the literature could be altered from the result on the analysis, and thus utilizing some elements of inductive approach to elaborate certain relevant topics.

3.2 Research Design

The unit of analysis in this thesis is employees in Yara with knowledge about Yara's current or future big data adoption and a specific knowledge about topics that connects to one or more of the value drivers.

This thesis can be considered a case study. (Bryman, 2012) describes a case study as a study associated with an intensive examination of the setting in a location, such as a community or organization. It also states that the case study design tends to favor qualitative methods, such as unstructured interviewing, as these methods are important in the generation of an intensive and detailed examination of a case (Bryman, 2012).

Interview subjects and their role in Yara have been treated anonymously and given a randomized number in Table 2. The numbers were utilized to make sure the observations could be linked to the transcripts, without it being traceable to specific employees. Anonymity was chosen so that the interviewees were comfortable to share their subjective thoughts and experiences regarding big data adoption in Yara. A vague description of their roles and expertise has been presented in Table 2 to provide context to the analysis.

3.3 Research Method

This section describes how the interview objects was selected, and how, and from what sources, data was collected.

3.3.1 Selection of Interview Objects

To achieve the most accurate results, a purposive sampling of interviewees was conducted. Contrary to probability sampling that randomly pick subjects, the purposive sampling strategically chooses subjects relevant to the research questions that differ from each other in terms of key characteristics (Bryman, 2012). Even though the authors sampled with the research goals in mind, indicating a purposive sampling, the process had hints of being a convenience sampling at times. A convenience sample pick the subjects that are currently available to the researcher (Bryman, 2012). Certain employees in Yara had little or no available time to participate in such a study, leading the authors to sample the subjects within the criteria that was available for an interview.

The criteria for sampling was that an interviewee should have some relevant insight and experience in topics related to the value drivers and the research questions. To evaluate whether different employees had such insight, the authors considered their education, their previous work experience, what department they work in and what type of position they have. If some of these aspects indicated that they could provide valuable insight to the study, they were invited to an interview. For the analysis to present valid insight from across the company, interviewees from all departments were purposely sampled.

3.3.2 Data Collection

Semi-structured interviews have been the main source of information in the thesis. Besides the interviews, the authors have also supplemented with information from Yara's internal documents and information from Yara's public web pages.

Interviews

Eleven interviews were conducted with one interviewee, and one with two interviewees. Table 2 does not show the number of interviewees to preserve anonymity. The authors sought interviews with one interviewee to encourage the interviewee to speak freely and to reveal honest opinions and thoughts. A semi-structured interview form further enhances the chances of revealing the interviewees concerns or their view on important aspects, as the sequence of questions and the frame of reference can be somewhat tailored to the situation and specific

interviewee (Bryman, 2012). The authors chose interviewees from every department and with varying roles and responsibility to get a broad insight in the company as a whole, and to identify possible internal variations. The semi-structured format made it possible to steer the interview in the direction of the value drivers, while allowing the interviewees to further elaborate what important aspects that should be investigated. The interviews were recorded and transcribed with consent from the interviewees, as recommended by (Bryman, 2012). It ensures validity and allow for repeated examinations of the interview (Bryman, 2012). Table 2 illustrates what type of position and what field of expertise the interviewees have. It also illustrates the duration of each interview and the number of words each transcript contains, which respectively sums up to approximately 7,5 hours and 21 500 words.

Table 2 shows an overview of the sampled interviewees. Interviewees are indexed for anonymity. The indexes are applied to Figure 7 to illustrate which interviewees lead to which observation. This table allows the reader to link observations to the relevance and field of expertise of each interviewee.

| Interviewee number | Number of interviewees | Place and date | Relevance (Field of expertise) | Duration [min] | Number of words |
|---------------------------|-------------------------------|-----------------------|---|-----------------------|------------------------|
| 1 | X | Stavanger 23.04.18 | Scientist (Crop nutrition, External modeling and External data collection) | 55 | 1110 |
| 2 | X | Stavanger 26.03.18 | Manager (Product development, Product portfolio and External data collection) | 20 | 1293 |
| 3 | X | Oslo 13.04.18 | Manager (Project management, External data collection and User experience) | 35 | 1355 |
| 4 | X | Oslo 12.04.18 | Data architect (ERP Systems, Data governance, Infrastructure and Big Data) | 45 | 2735 |
| 5 | X | Stavanger 25.04.18 | Manager (Innovation, Digital transformation and Strategy) | 23 | 971 |
| 6 | X | Oslo 13.04.18 | Manager (Artificial intelligence, Data infrastructure and Process monitoring) | 59 | 3257 |

| | | | | | |
|------------|-----------|-----------------------|--|------------|--------------|
| 7 | X | Oslo 12.04.18 | Manager (Data infrastructure, Innovation and digital transformation) | 54 | 2400 |
| 8 | X | Oslo 12.04.18 | Manager (Business intelligence, Analytics and Finance) | 32 | 2522 |
| 9 | X | Stavanger 28.02.18 | Scientist (Analytics, Big data and Data infrastructure) | 18 | 533 |
| 10 | X | Oslo 13.04.18 | Manager (Process monitoring, Lean and Internal data collection) | 36 | 1761 |
| 11 | X | Oslo 06.04.18 | Manager (Customer experience, Artificial intelligence and Information systems) | 36 | 1681 |
| 12 | X | Oslo 12.04.18 | Manager (Project management, Digital Transformation and Supply chain) | 33 | 1877 |
| Sum | 13 | N/A | | 446 | 21495 |

Interview Guide

(Bryman, 2012) points out that contrary to a quantitative interview that has a strict structure and sequence of the questions, the qualitative interview guide can function as a brief list of memory prompts to ensure the relevant topics are covered. The interview guide in this thesis ([Appendix 1](#)) worked as a bank of mid-30 questions that was used to steer the interviews in the right direction and as a reminder for the interviewers of what topics to cover, not as an exact guide. The interview guide also contained of some specific questions for specific employees to get a deeper insight in the topics where the interviewee was the most knowledgeable. Several of the questions were rather open and facilitated for the interviewee to talk about a broad spectrum of concerns within a given value driver.

Other sources of information

Besides semi-structured interviews, two sources of documentary information were used to support certain topics of the analysis. The authors have received internal documents, such as

strategies and presentation slides, from supervisors in Yara to support the authors’ deeper understanding of internal projects and investments. These are however subject of a confidentially agreement and will not be published. Publicly available information from Yara’s web pages has also been used to provide information and context to the reader about Yara’s operations and market situation.

3.4 Data Analysis

This thesis uses a variation of coding as the analytic technique to process data. Coding is a part of the commonly used qualitative analytic framework called grounded theory. During the data collection, in this case the interviewing process, the authors utilized memos to ensure that important nuances of each interview were included in the later process. (Bryman, 2012) underlines that coding as you go along is an important aspect to prevent being swamped with data at the end of the data collection. The process of transcribing interviews from the recordings was initiated early, to ensure the best results as the interviews was fresh in mind. The transcripts were coded into observations, in this case short sentences that describe a situation, and not into single words, as coding more commonly does. Then observations were cross-checked with all transcripts. As recommended by (Bryman, 2012), the authors initially coded what seemed to be too many observations and refined these twice. In the refinement process, the authors analyzed the observations by examining what numbers of interviewees backed the observation and how relevant the observation was to the research questions. Table 3 illustrates that the initial coding led to 59 observations, which later was refined to 40 and then to a final number of 28 observations. (Bryman, 2012)

Table 3 illustrates the refinement process of the observations. Note how the number of observations was reduced from 59 to 28 after two refinement steps.

| Refinement process of observations | |
|--|----|
| Initial number of observations | 59 |
| Observations after first refinement | 40 |
| Final number of observations after second refinement | 28 |

4. Analysis

The analysis will present insights from the semi-structured interviews and discuss them in relation to the literature. The chapter reviews Yara's performance in each of the five value drivers, labeled H1 through H5, and summarizes them respectively. In each value driver, the relevant observations are stated and explained, and then discussed by analyzing the causes, effects and suggestions. Some observations are missing either causes, effects or suggestions as the authors deemed them not applicable.

Figure 7 illustrates the state of big data adoption through digitalization in Yara. It is sectioned from left to right with:

- A: The assumption the thesis is based on (1, para. 3)
- B: The five big data value drivers (2.6)
- C: The measurables of each value driver
- D: The observations from the interviews
- E: The main causes for selected objectives
- F: Relevant theory topics for respective causes

Figure 7 illustrates how H1 is a catalyst for H2 through H5. By investigating the measurables in the interviews, the observations emerged. The observations are underlined with randomized numbers of who stated it. This is done to preserve anonymity and facilitate for verifiability. The results have been analyzed to identify the most important underlying causes. The authors have the perception that some of the causes may apply for several observations and are therefore cross-referenced accordingly. The figure illustrates a snapshot, and an overview of, Yara's current big data maturity, and thus works as a visualization of the answer to RQ1.

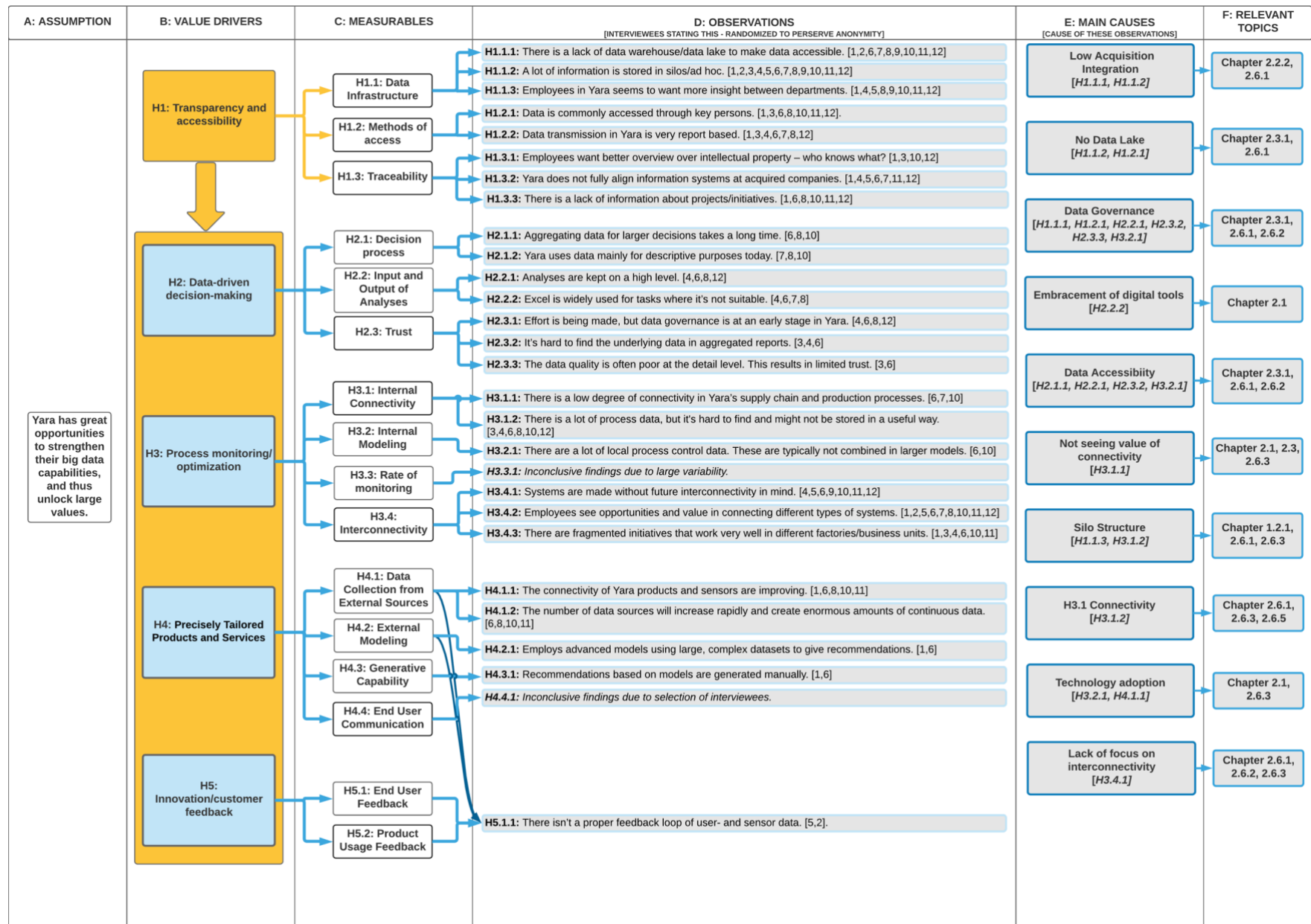


Figure 7 illustrates how the observations are linked to measurables and value drivers. It also identifies potential causes for selected observations and the relevant theory for each cause. Note that H1 is a catalyst for H2 through H5 and that observations are numerated with indexes from Table 2.

H1. Data Transparency and Accessibility

Simply making data more accessible throughout the organization can create value (2.6.1, para. 1). It also enables further value generation in the parameters H2 through H5. Improvements in H1 generally requires investments in data infrastructure and data governance (2.6.1, para. 2). To measure the current state of data transparency and accessibility in Yara, the following measurables were evaluated:

H1.1 Data Infrastructure

H1.2 Methods of access

H1.3 Traceability

H1.1 Data Infrastructure

Identifying the current state of data infrastructure is essential to give further recommendations on how to increase the accessibility of data and organizational transparency. Having the necessary infrastructure in place is the prerequisite for several data-driven approaches for value generation (2.6.1, para. 2). Through the interviews, the following observations were made:

H1.1.1 There is a lack of data warehouse/data lake to make data accessible.

Explanation:

Most interviewees expressed that they are missing systems to let them access data from both their own department, but also data from across departments. They pointed out that there are many fragmented systems and databases, and not an easy-to-access common platform or a data lake in place.

Causes:

This observation is a result of a lack in data infrastructure following a low focus on data governance. Disparate databases are not pooled in a data lake, both because of a lack of integration following acquisitions and because there might not previously have been that much of an incentive to pool data from different regions and departments. Yara grows through

acquisitions, and when they are acquiring a company, the system integration seems to be kept at a minimal effort basis (H1.3.2). The fact that Yara has no platform for pooling global data could be seen as both a cause and a result of the minimal integration efforts that are performed when Yara grows. If there is no way to pool the data, there is no immediate reason for integrating the systems better. As data governance and metadata are important supporting functions for a data lake (2.6.1, para. 2), pooling the data might not be useful before the systems are more integrated.

Effects:

As making data accessible is both a value driver on its own and a catalyst for the other value drivers (2.6.1, para. 1), H1.1.1 has a lot of impact on how much value Yara is able to generate from their data. If data is not accessible, it loses a lot of its usefulness (2.6.1, para. 1). Data not being accessible also affects the organization's knowledge of what kind of data it has. There might be very useful databases that is not being used as other parts of Yara has no knowledge they exist.

Suggestions:

A data lake should be a sound investment for Yara as it is an effective way of increasing data accessibility and transparency (2.6.1, para. 2). For a data lake to be successful, Yara must first have a clear vision on what the use-cases of a data lake would be. Yara also need to have a strong focus on improving data governance and computer security as these are important supporting factors of a data lake (2.6.1, para. 2). By implementing a data lake, Yara could also see potential improvements to collaboration and innovation across departments, as it is made easier by having a designated space for interdepartmental data storage. Digitally maturing organizations are characterized by having a more collaborative and innovative culture (2.1.1, para. 7). This increase in collaboration and innovation in combination with facilitating for easier access to measurables for leaders, increasing their ability to evaluate different parts of the digital transformation process (2.1.2, para. 2) and tweaking digital strategies, could drive digital transformation in Yara. This would increase Yara's digital maturity level, improving their ability to further adopt digital solutions (2.1.1, para. 1-3).

H1.1.2 A lot of information is stored in silos/ad hoc.

Explanation:

All interviewees pointed out that data is stored in silos or at an ad hoc basis from project to project. The sharing of data between departments is particularly difficult and prevents data from being transparent within the organization.

Causes:

H1.1.1 is an important cause for why data is stored in silos. According to interviewees, different departments and their day to day business has traditionally been very distant. This has led to data being stored in the means most efficient to the respective department, without too much thought for how these silos could be connected. Low system integration following acquisitions (H1.3.2) is part of the cause why information is stored in silos/ad hoc.

Effects:

Data being stored in silos, poses several challenges for Yara in relation to value generation from data, as making data accessible is a prerequisite for several of the value drivers (2.6.1, para. 1). It makes it harder for employees to know what kind of databases Yara has in different regions and departments. Data might be stored in different structures across silos, posing problems when one sees the value in combining datasets from different silos. For integrated systems to be functioning across silos, data cleaning (2.6.1, para. 3) might be required. Although allowing some data to be stored in silos might be profitable in the short term following acquisitions and expansions, it acts as an obstacle for Yara to be able to utilize their data for organization-wide purposes (2.6.1, para. 1). Having data stored in silos could also negatively affect Yara's culture in relation to digital transformation, reducing collaboration and innovation across departments. Companies that are early in their digital transformation process is generally characterized by being siloed. See Table 1.

Suggestions:

Yara should consider following the framework proposed in 2.1.2, making bridging silos a digitalization goal. They should further investigate the current state that causes this observation and identify and prioritize feasible actions that can be taken to bridge the gap between the current state and a desired future state of interdepartmental accessibility of data. One such action could be investing in a data lake while focusing on data governance and computer

security. They should focus on bridging new data as it is generally easier to prevent data cleaning than to cure data with low integrity (2.6.1, para. 3).

H1.1.3 Employees in Yara seems to want more insight between departments.

Explanation:

Employees sought better insight in projects and results from other departments. They were clear that better connection between departments relying on another - such as supply chain and production - would unlock potential value by the improved collaboration and increased efficiency.

Effects:

Employees want to perform at their best. All additional efforts that employees must make to access data they need from other departments can be considered wasteful. This can affect the motivation of employees and might discourage them from being data-driven. Failing to make relevant data available can inhibit the knowledge employees use for decision making (2.6.2, para. 4). It also makes it harder for the different departments to be aligned towards common goals (2.6.1, para. 1). If employees are unsatisfied with the interdepartmental insight they have, this could have a negative effect on the desired collaborative and innovative culture that drives digital transformation (2.1.1, para. 7).

Suggestions:

To facilitate for employees being increasingly data-driven, collaborative and innovative, Yara should investigate what data employees need easy access to. Granting access could be achieved through a data lake or similar (2.6.1, para. 2). There should be an increased focus on transparency between departments to align employees towards common goals.

H1.2 Methods of Access

How employees access data is an important measurable when evaluating the data transparency and accessibility. If there is a lack of knowledge or smart systems that lets employees find data they need, they might have to turn to less effective means of gathering data useful for them. When measuring how employees in Yara access data, the following observations were made:

H1.2.1 Data is commonly accessed through key persons.

Explanation:

Interviewees expressed that they need to establish a network of key persons to be able to access data. This is said to be difficult for new employees that have not had the chance to establish this type of network.

Causes:

Yara has no data lake (H1.1.1), which results in employees having to access data by other means. Even though employees might have the right authority to access data, the easiest way is often through key persons. This could be caused by not knowing where to find data or how it is structured. It could also be due to not all employees being able to refine raw data into something useful, or simply that the user-friendliness of the methods of access is too low.

Effects:

When employees need to go through key persons to get the data they need, Yara becomes reliant on key employees. If one of these resigns, the knowledge that they possess might be lost. Being reliant on key persons is also quite inefficient as one ends up with more people on a work task that would be avoided if the employee could access the data on their own. The person who needs the data probably also must wait before receiving the data, increasing idle time.

Suggestions:

Improved analytical capability, paired with a data lake allowing employees to access data, could facilitate for employees accessing already refined data (2.4, para. 3). If refined data can be accessed on an as-need basis instead of a know-how basis, it would aid decision making (2.6.2, para. 1) and the alignment of the organization while increasing organizational transparency (2.6.1, para. 1). If Yara believe data is sufficiently accessible, they should increase their focus on talent development in digital tools, as this could help employees find and use the data they need on their own. A strong focus on talent development is also a key feature of digitally maturing companies. See Table 1. Hence there is reason to believe that further training of employees and leaders in digital skills could drive digital transformation in Yara.

H1.2.2 Data transmission in Yara is very report based.

Explanation:

Data is commonly transmitted through static reports and not with real-time dashboards. The static reports may not include all necessary information, and the data tends to be outdated quickly.

Effects:

Interviewees expressed that reports often require a substantial effort to make, and typically only serves one purpose (2.4.1). The more in-depth the report, the bigger the workload required. Reports also tend to mainly show aggregated data, possibly washing out the underlying causes for an aggregated result, and thereby important data that could be used in decision making.

Suggestions:

Yara should investigate whether static reports are the most beneficial way of transmitting data. For frequently expiring data, more dynamic alternatives might be more suitable (2.4.1). Investments in infrastructure (2.6.1, para. 2) and analytical capability, could potentially allow Yara to transmit data through interactive dashboards (2.4.1), or at least through more dynamic reports that could update themselves based on new data. If Yara's digitalization goals require more dynamic ways of conveying information than their current state offers, Yara should consider increasing their dynamic capability in reports to avoid static reports being a bottleneck for further digital transformation efforts. This should be done through a set of feasible actions rooted in their roadmap for digitalization (2.1.2, para. 2-4).

H1.3 Traceability

It is important for the data infrastructure and methods of access to facilitate for traceability of data (2.6.1, para. 4). If one is unable to trace the origins of manipulated data, details that might be useful are washed away. While measuring the traceability of data in Yara, the following observations were made:

H1.3.1 Employees want better overview over intellectual property – who knows what?

Explanation:

Employees expressed that it is difficult to know where to get what data. As data is regularly accessed through key persons ([H1.2.1](#)), they said it would be beneficial to have an overview of intellectual property, as both collaboration and traceability of data would improve.

Effects:

Not being aware of what data exists, and what other parts of Yara are working on, inhibits the organizational alignment ([2.6.1, para. 1](#)). One benefit of owning a large value chain consisting of many acquired companies, is the opportunity of generating positive internal synergies between several departments ([2.6.3, para. 1](#)). If employees are unaware of useful intellectual property in Yara, they might need to redo experiments or in other ways expend resources to acquire intellectual property that Yara already owns.

Suggestions:

As long as data is being commonly accessed through key persons ([H1.2.1](#)), Yara should increase the focus on transparency of intellectual property as it could help avoid parallel efforts while boosting positive synergy effects. It would also allow employees to more efficiently access information, possibly boosting collaboration and innovation. This may drive Yara's ability to adapt to opportunities and threats produced by digitalization, through having a culture more contributive to digital transformation (Table 1).

H1.3.2 Yara does not fully align information systems at acquired companies.

Explanation:

It appears that Yara tends to align information systems at acquired companies based on present value principles. At the current state of big data adoption, this often results in minimal integration efforts.

Causes:

At the point of acquisition, it often makes more sense financially to opt for a minimum effort solution when it comes to information systems integration. This is because it requires the smallest investment, both economically and time-wise.

Effects:

Through minimal system alignment, Yara can reach the break-even point on acquisitions sooner than if they would invest more time and money into fully aligning the information systems following acquisitions. By not fully aligning information systems after acquisitions, Yara might find themselves in a situation where they need to manipulate a lot of data before they are able to use it. This cleaning of data could potentially outweigh the cost of complete alignment, as curing data is less effective than prevention (2.6.1, para. 3), while hindering the roll-out of new digital tools.

Suggestions:

If Yara is approaching the “Three-V tipping point” (2.2.2, para. 6) of big data, they should add the future value of data to the equation when they are considering the degree of information system alignment. As the value of data is not very tangible (2.4, para. 2), this is challenging, and it is not always given that complete alignment of information systems is more right for new acquisitions than minimal effort alignment. Yara should decide on how much effort they should put into aligning information systems following acquisitions after conducting cost-benefit analysis or impact analysis of complete systems alignment (2.1.2, para. 4). This should be done after they have identified clear digitalization goals and the gap between current state and their desired future state following the framework presented in 2.1.2. Increased alignment of information systems following acquisitions would most likely encourage interdepartmental collaboration and innovation, and hence improve Yara’s digital maturity level (2.1.1, para. 7).

H1.3.3 There is a lack of information about projects/initiatives.

Explanation:

Several employees sought more information about the different projects and initiatives in the organization. They indicated that a better overview would enable positive synergies and prevent parallel work tasks.

Effects:

Employees not being aware of what efforts are being made could potentially lead to projects in different departments unknowingly performing redundant work, or a decrease in positive synergy effects. One of the benefits of owning the entire value chain is unlocking ways to positively synergize (2.6.3, para. 1), but to do this one must know what is going on in other parts of the organization.

Suggestions:

By developing and iterating on a digital roadmap by using the framework proposed in 2.1.2, Yara would have a clear overview of current and future digital initiatives. By making this roadmap increasingly accessible for employees, or at least managers, Yara could avoid parallel efforts, and enhance alignment on digital endeavors across the organization. Making news-stories about exciting digital endeavors increasingly available over the intranet, could also induce cultural change in Yara, increasing collaboration and innovativeness amongst employees (2.1.1, para. 7). This would in turn be conducive to digital transformation in Yara, as innovativeness and collaboration are traits that typically characterize digitally maturing companies (Table 1).

Insight H1

As making data accessible is both a value driver on its own, and a catalyst for several value drivers (2.6.1, para. 1), Yara has a lot to gain by having an increased focus on making data accessible. There is currently no efficient system for pooling disparate datasets (H1.1.1), and this is both a cause and an effect of data being stored in silos/ad hoc (H1.1.2). Data might be stored in different structures across silos, and this can pose challenges when Yara see the value in combining disparate datasets. Employees commonly access data through key persons (H1.2.1). This leads to Yara being reliant on these employees and is usually less effective than if employees can access information on an as-need basis. As Yara tend to grow through acquisition (1.2.1, para. 2), they need to consider increasing the effort they put into aligning systems following acquisitions (H1.3.2), especially if they are approaching the “Three-V tipping point” (2.2.2, para. 6). The Digital Farming division in Yara is focusing on giving their customers access to data to empower them with more well-informed decisions and increased insight in their operations (1.2.2, para. 2), in essence increasing the customer’s perceived value

of Yara’s products and services. Hopefully, the observations in H1 can function as a reminder that giving employees easier access to data empowers employees in the same way. Through empowering employees, Yara can gain economic value through increased efficiency and improved decision making.

Most of the issues that are uncovered in H1 will most likely be identified as gaps that needs to be overcome for Yara to reach their digitalization goals, if they decide to follow the framework proposed in 2.1.2. This is due to data transparency and accessibility being a prerequisite for several data-driven approaches (2.6.1, para. 1). The sooner Yara starts improving their data transparency and accessibility, the less clean-up they need to do on their data to reach bolder digitalization goals in the future. Advancements in H1 coupled with a sound strategy and a culture that drives digital transformation, would act as a catalyst for Yara to gain value through increased internal efficiency, increased ability to exploit external opportunities and disruptive change (2.1.1).

H2. Data-Driven Decision Making

By leveraging data, decision makers can base more of their decisions on evidence instead of intuition or experience (2.6.2, para. 1). Having good data transparency and accessibility in combination with high analytics capability can help an employee make more informed decisions, quicker (2.2.2, para. 4), (2.6.2, para. 1). To measure the current state of data-driven decision making in Yara, the following parameters were investigated:

H2.1 The decision-making process

H2.2 Input and Output of Analyses

H2.3 The level of trust

H2.1 The Decision-making Process

In order to perform an analysis that can be used in decision making, employees need first to gather the data they are going to perform the analysis on. The faster one can gather the data and analyze it, the more time one can spend on making a well-informed decision. At the same time

– the more data (ie. the longer one spends gathering data and analyzing it) the better informed the decision maker will typically be (2.3, para. 2), (2.6.2, para. 1). While investigating the decision-making process, the following observations were made:

H2.1.1 Aggregating data for larger decisions takes a long time.

Explanation:

Decision makers outlined that most of the time in the decision-making process is spent on aggregating the necessary data. A short amount of time is then spent on making the decision itself. Thus, it is reason to believe that better decisions could be made if aggregating data was easier.

Causes:

This observation is rooted back to the challenges regarding accessibility of data in Yara (H1.1.1), (H1.1.2), (H1.1.3), (H1.2.1), (H1.3.2). For a company to be data-driven, there must be sufficient data as a basis for decisions (2.6.2, para. 1). There is currently a tradeoff at Yara between spending time on aggregating the data and spending time on reviewing the analyses and making the decision. From the interviews, it seems that decision-makers in Yara is data-driven to the extent that is currently possible.

Effects:

Yara is currently spending more resources than necessary on aggregating data. This is at the expense of the allocated time for reviewing the analysis and making the decision, which again may lead to worse decision outcomes.

Suggestions:

To increase the rate at which Yara aggregates data, they should focus on improving methods for accessing data and performing analyses. With that in place, there is reason to believe that Yara would become increasingly data-driven as the attitude of employees regarding data use seems to be satisfactory. There is also reason to believe that Yara could become more profitable by becoming more data-driven (2.6.2, para. 4). It seems like Yara's culture is driving data-driven decision making. This means that there might be technical or process-related issues that need to be addressed to improve the process of gathering data. Yara should consider making improvements to the data aggregation process a digitalization goal following the framework

proposed in [2.1.2](#). They should investigate whether there is a technical gap that needs to be bridged for a change at the process level to occur, and what impact bridging this gap would have at the process- and organization level ([2.1.2, para. 3](#)). They should then investigate the feasibility of actions to close this gap ([2.1.2, para. 4](#)). This thesis suggests considering the feasibility of implementing a data lake, increasing the focus on data governance as an important supporting function of a data lake ([2.6.1, para. 2](#)), and investigating whether there are other factors that result in low efficiency in the data aggregation process.

H2.1.2 Yara uses data mainly for descriptive purposes today.

Explanation:

From the interviews, it is made clear that Yara is at Cognizant's descriptive level, where data is used to describe what has happened or what is currently happening – and not to describe or make decisions on what is going to happen ([2.5, para. 3](#)).

Effects:

When data is used for descriptive purposes, it is used to describe a situation. The description is typically not detailed enough to find the underlying causes of the observed events, which will inhibit Yara's ability to diagnose conditions ([2.5, para. 4](#)). Yara misses out on the more advanced use-cases of data, which includes predictive analyses by being at the descriptive level ([2.5, para. 5](#)).

Suggestions:

In order for Yara to reach Cognizant's next level of big data maturity, diagnostic, they need to ensure that they facilitate for traceability of data ([2.6.1, para. 4](#)). This is done through investments in data infrastructure ([2.6.1, para. 1](#)) and data governance ([2.3.1](#)). Yara should focus on improving the integrity of their data, so they can make decisions on more detailed reports, allowing them to easier identify the underlying causes for observed events. Yara's digital transformation effort appears to be driven through Kaizen events. This means that they are evolving their digital maturity incrementally ([2.1.1, para. 5](#)). Yara should analyze the impact digitalization has at the process level, organization level, and business domain level. If they see that the potential gains from a digital transformation outweigh the costs, they should consider more radical improvements. Kaikaku events are typically riskier than Kaizen, but in return they usually impact the organization far more than Kaizen events ([2.1.1, para. 5](#)), (Figure 1). Toyota

had a lot of success with a combination of Kaizen and Kaikaku (2.1.1, para. 6). It is reasonable to assume that a combination between Kaikaku and Kaizen could be a sound model for digital transformation as well. Kaikaku could be used to quickly elevate Yara's digital maturity, while Kaizen ensures continuous transformation, maintaining and increasing the digital maturity level. With an elevated digital maturity level, Yara should see themselves adapting more quickly to big data (2.1.1, para. 2), increasing their ability to climb Cognizant's digital maturity ladder (2.5).

H2.2 Input and Output of Analyses

Performing analysis can be considered the refinement process of raw data (2.4, para. 3). The size and types of data one is able to perform analysis on is generally determined by one's analytical capability (2.3, para. 2). The output of the analysis is generally dependent on the input (2.6.2, para. 1). If the quality of input data for the analysis is poor, the quality of the output will most likely also be poor. The level of insight one can get from an analysis is also strongly dependent on the detail level of the output (2.6.2, para. 1). While investigating what inputs and outputs Yara typically has for their analysis, the following observations were made:

H2.2.1 Analyses are kept on a high level.

Explanation:

Employees indicated that analyses in Yara is kept on a high level, and that few of the underlying causes of aggregated data is presented to decision-makers.

Causes:

Data quality at the detail level seems to be poor (H2.3.3). This leads to Yara having to run analyses on highly accumulated data to improve the integrity of the analysis. The limited accessibility of detailed data is most likely what is keeping Yara from having more detailed outputs of their analyses. Yara's analytical capability also probably plays a role. With more advanced use of data mining, Yara could uncover previously unknown patterns that could be used to improve the detail level of analyses (2.4.4).

Effects:

If the output of analyses contains limited details, it is hard to identify the underlying causes of the bigger trends that are uncovered (H2.3.2). This is especially true if the traceability of data is poor. This acts as a barrier for Yara to reach Cognizant's diagnostic level (2.5, para. 4). It also affects the insight decision-makers base their decisions on (2.6.1, para. 5-6).

Suggestions:

For the output of the analyses to be more detailed, one is dependent on better quality of input data. This is because the quality of the output generally is reliant on the quality of the input (2.6.2, para. 1). Yara should focus their attention on increasing the integrity of their detailed data. Data governance (2.3.1) and data cleaning (2.6.1, para. 3) are important activities that Yara need to undertake in order to get more detailed outputs from their analyses.

H2.2.2 Excel is widely used for tasks where it's not suitable.

Explanation:

It is observed that Excel is used for tasks that could be solved better and more efficiently with other tools. Employees said that Excel works great for visualization and sorting but requires quite a lot of manual work for the result to add value. It does not work equally good for processing large data sets, manage projects in a larger scale or running large financial analyses.

Causes:

It seems that Yara has not fully embraced more specialized tools that outperforms excel in their specific use-case. Excel has a lot of use-cases, which means that employees are familiar with excel, and might choose it over more specialized software that they have less experience with. It could be that employees have not received the necessary training, or that the tools simply are not available.

Effects:

This observation is potentially causing employees to spend more time on tasks, or simply not completing tasks, as Excel might not be capable for their needs. It could also initiate a ripple effect where results from tasks are poor and thus lead to lower quality decisions and analyses.

Suggestions:

Yara could review their digital toolbox and investigate how employees use digital tools and for what purposes. If more advanced and case specific tools are available without employees using them, workshops or guides could be beneficial. If the tools are not available, the management and IT department should investigate where there are needs for tools superior to Excel and consider embracing them. To maintain their competitive advantage, Yara is reliant on adapting to the different needs in analytical capability that big data brings ([2.3, para. 2](#)). This to allow them to better exploit the opportunities of big data ([2.1, para. 1](#)). Through increased focus on training employees in digital skills, employees would most likely be more inclined to embrace more advanced tools if Yara see it beneficial to make such tools available. This would help drive digital transformation in Yara as digitally maturing companies are typically committed to making training resources in digital skills available (Table 1).

H2.3 The Level of Trust

The level of trust decision-makers have in data, is an important aspect to consider when it comes to evaluating how data-driven the decision making is, as they need to embrace information from the data as evidence while making decisions ([2.6.2, para. 1](#)). Having good analytical capabilities and high-quality data that enables data-driven decision support is not very useful if the findings are undermined by a lack of trust. There are many factors that play into how much employees trust data. For instance, the degree of context or metadata there are for the data, and the level of detail the analysis presents. ([H2.3.1](#)), ([H2.3.3](#)). While investigating the level of trust Yara employees have in data, the following observations were made:

H2.3.1 Effort is being made, but data governance is at an early stage in Yara.

Explanation:

Yara seems to be aware of the importance of data governance. They are some initiatives for improving the governance, however, these are at early stages.

Effects:

If there is a low degree of data governance, data tend to be hard to find and use ([2.3.1](#)). There is also often a lack of metadata, which results in low data integrity. Data governance is an

important activity for increasing the traceability of data, and thus a low degree of data governance makes it difficult to find the underlying causes in analyses (2.3.1), (H2.3.2). This results in decision-makers in Yara having less trust in data. Low traceability and poor data integrity also lead to decision-makers having less details to base their decisions on (2.6.1, para. 4-6). Poor data governance acts as an obstacle to combining disparate data sets as substantial data cleaning might be needed to make the datasets compatible (2.6.1, para. 3).

Suggestions:

Yara should consider accelerating their data governance efforts as improved data governance is a suggestion for several of the observations that are identified in this thesis (H1.1.1), (H1.2.1), (H2.2.1), (H2.3.2), (H2.3.3), (H3.2.1). This could be done through a mix of Kaikaku and Kaizen events (2.1.1, para. 6). Kaikaku events typically requires a larger commitment of time and resources but will usually induce more drastic impacts in a company (2.1.1, para. 5).

H2.3.2 It's hard to find the underlying data in aggregated reports.

Explanation:

Interviewees pointed out that the connection between underlying data and aggregated reports is poor. The underlying data could be old or not present the full picture, so having the ability to check the data for oneself would in many cases be beneficial.

Causes:

Underlying data is not sufficiently accessible, or structured in an efficient way (H1.1.1), (H1.1.2). There is most likely a low degree of traceability in the data. This does not facilitate for decision-makers to investigate the origins of the data in the aggregated reports (2.6.1, para. 5-7).

Effects:

This can result in less trust in reports and decision-makers utilizing an increased amount of intuition in the decision-making process. Decision-makers having to trust reports blindly without the opportunity to investigate underlying data could harm decision outcomes, and Yara's journey to become increasingly data-driven (2.6.1, para. 6-7), (2.6.2, para. 1).

Suggestions:

If underlying data is more easily accessible, decision-makers may be more incentivized to further investigate the causes of trends that are identified in analyses. This could lead to more nuanced decisions as the decision makers have more insight (2.6.1, para. 5-6), (2.6.2, para. 5). Yara should address this issue following the framework proposed in 2.1.2. They should use the effects section of this observation as a starting point, and identify the impact this observation has for Yara's ability to be data-driven in their decision making. Next, they should identify the gap between the current state in Yara, essentially further investigating the causes identified in the causes section of this observation, and the desired future state. Once the gap has been defined, they should prioritize feasible actions that can bridge the gap in a roadmap. Such actions could be implementing a data lake (2.6.1, para. 2), conducting data cleaning (2.6.1, para. 3), increasing their data governance efforts (2.3.1), or increasing their effort in information system alignment following acquisitions (H1.3.2). They should also focus on digital talent development as this is a driver of digital transformation (Table 1). IT should investigate whether there are opportunities for linking underlying data with reports with current systems or facilitate better for employees to find this data themselves.

H2.3.3 The data quality is often poor at the detail level. This results in limited trust.

Explanation:

Interviewees made it clear that if the underlying data in reports is available, the data quality tends to be poor. This makes decisions based on more detailed data low integrity.

Causes:

It is hard to identify distinct causes for why data quality at the detail level is perceived as poor. Data governance and data traceability typically plays an important part in the perceived quality of data (2.3.1), (2.6.1, para. 6).

Effects:

An employee pointed out that a missing time stamp had led them to not trust, or even refrain from using data that was available and relevant for their work. This lack of trust in the data weakens Yara's efforts in being knowledge driven, and thus the competitive advantage Yara has from this (2.6.2, para. 4). If data quality is poor, employees end up not using detailed data for their decisions. Hence the effects of H2.3.2 is also applicable.

Suggestions:

Yara should investigate why employees perceive data quality at the detail level to be unsatisfactory and focus on increasing the quality of detailed data. This could be achieved through accelerated efforts in data governance and data cleaning (2.3.1), (2.6.1, para. 3). Low data quality affects Yara's ability to adopt big data technologies and hence their big data maturity level. Employees perceiving data quality as low is an internal issue that Yara should consider addressing through utilization of the framework proposed in 2.1.2. If Yara believes that the quality of data at the detail level is sufficient, they should focus on training employees in digital skills such as more basic forms of data mining (2.4.4) and data cleaning (2.6.1, para. 3), so employees themselves can improve the insight they are able to get from data as needed. If the employees are able to better refine data themselves, they might be able to utilize more detailed data for decision making and modeling.

Insight H2

For the decision making in Yara to improve, there are some technological barriers to overcome – many of those relevant for parameter H1. There should be an increased focus on data governance, improved infrastructure, and analytic tools. Improved data governance could increase both the quality of detailed data and the trust in both reports and the data themselves (2.3.1), (H2.3.1), (H2.3.2), (H2.3.3). Improved analytic capabilities could then lead Yara to Cognizant's predictive and prescriptive levels – unlocking knowledge and wisdom that is the advantage of being an international company with the full value chain. Improved infrastructure and data governance would lead to improved data accessibility and transparency for decision makers and thus increase efficiency in decision-making processes and accuracy of decisions. It would prevent intuition in decision making and encourage the use of data. In sum – and as proved by McAfee and Brynjolfsson - Yara could gain value through increased efficiency and profitability by becoming increasingly data-driven. Thus, potentially increasing their stock market value (2.6.2, para. 4). This would translate to economic value for Yara's shareholders.

Yara should set clear digitalization goals rooted in a digital strategy that drives data-driven decision making. By further investigating their current state and comparing it to their desired future state needed to reach their digitalization goals, they will most likely identify several

technical issues that need to be addressed. If Yara are able to prioritize feasible actions to bridge the technical gap in a roadmap, they would most likely increase their ability to gain value through increased internal efficiency and exploiting external opportunities. (2.1.1), (2.1.2)

The more digitally mature Yara is able to become through digital transformation, the more data-driven they will be, as they adopt digital tools at a faster rate (2.1.1, para. 3). The more data-driven they are, the more data there will be for decision makers to gain insight from. The more insight decision makers have, the better their decisions will typically be. Innovation is also enhanced through a high degree of insight (2.6.5). Hence, this thesis argues that increased data-driven decision making and higher digital maturity will create a positive interference, increasing each other.

H3. Internal Process Monitoring and Optimization

The internet of things has sparked a trend where everything has connectivity (2.3, para. 2). This combined with a general increase in data capability (2.2.1, para. 1) enables large improvements in internal process monitoring capability (2.6.3, para. 1). Being able to monitor your own processes at an increased rate and scale combined with increasing modeling capability allows organizations to streamline their processes (2.6.3, para. 1-2), increasing productivity and efficiency. A large degree of connectivity in internal processes acts as a foundation for these optimizations (2.6.3, para. 1). If the connectivity is high for one system, an organization can optimize that system. However, there could be benefits in interconnecting several systems in the value chain (2.6.3, para. 1). To measure the current state of internal data-driven process optimization in Yara, the following parameters were investigated:

H3.1 Internal Connectivity

H3.2 Internal Modeling Capability

H3.3 Monitoring Capability/Rate of Monitoring

H3.4 Interconnectivity

H3.1 Internal Connectivity

Connectivity in internal processes is important for data-driven process optimization (2.6.3, para. 1). If you don't collect process data, you have no input to your analyses or models. Internet of things technology connects machines, sensors and devices to monitoring and controlling systems, enabling opportunities for more insight to be extracted and acted upon – faster (2.6.3, para. 1-2). While investigating the current state of internal connectivity in Yara, the following observations were made:

H3.1.1: There is a low degree of connectivity in Yara's supply chain and production processes.

Explanation:

Yara's supply chain and production departments are enormous (1.2.1, para. 3), so a very detailed observation regarding their internal connectivity is hard to present, but employees made it clear that most of their internal processes could be better connected than they currently are.

Causes:

Yara might not have previously seen the value in increasing the connectivity of their processes. This has led to few investments towards more connectivity. Investing in connectivity would probably not be profitable immediately, but over time the potential value should outweigh the cost as it allows Yara to employ advanced process optimization (2.2.2, para. 6), (2.6.3, para. 1).

Effects:

A low degree of connectivity has no tangible negative effect on Yara's everyday business. However, it inhibits Yara's opportunities for value generation through process optimization. It also results in Yara having less insight in internal processes. Less insight could lead to less nuanced decisions (2.6.2, para. 1).

Suggestions:

Due to the immense scale of Yara's operations, even marginal improvements in efficiency would result in substantial cost reductions. This means that Yara has great opportunities for value generation through process optimization if high connectivity is paired with high analytic capability (2.6.3, para. 1). Yara should consider investing in internal connectivity as it would

facilitate for more advanced process optimization. Investing in technology, however, is not enough. Yara's supply chain and production processes also need to undergo a digital transformation to gain the full value of the technological advances (2.1.1, para. 1). This in order to change the processes and roles within the value chain of Yara to make it more suited for digital applications. For increased digitalization to be a success, Yara needs to have a clear strategy driven by highly skilled leaders that lead by example, driving a culture that is less risk averse and willing to evolve their digital skills as these are all key features of digitally maturing organizations (Table 1).

H3.1.2: There is a lot of process data, but it's hard to find and might not be stored in a useful way.

Explanation:

From interviews, it was clear that process data is not stored in one place, and it is rarely stored in formats that is compatible with optimization analyses.

Causes:

Due to the silo-structure (H1.1.2), a lack of purposeful ways of pooling data (H1.1.1), and a low degree of internal connectivity (H3.1.1), process data is mainly stored at site. Interviewees expressed that sometimes data from production processes are not stored digitally, making it hard to both access and use for more advanced purposes.

Effects:

To access process data today, you would have to know where to look for the data, or most often manually ask supervisors of the respective processes or sites, and then clean the process data for it to be useful in a larger context. This acts as a barrier for both data-driven process optimization and higher process insight.

Suggestions:

The value of process data from different sites should be analyzed to confirm whether smaller or larger investments in connectivity and process data governance is profitable in the coming years. As Yara is approaching their "Three-V tipping point", they should trade off implementation cost against expected future value and pay less attention to instant profitability (2.2.2, para. 6). Increased connectivity and process data governance will most likely come as a

side-effect of bolder digitalization goals in Yara. If Yara for instance makes a digitalization goal out of increasing the degree of automation in a production facility following the framework proposed in [2.1.2](#), increased connectivity and process data governance for that specific facility would most likely be a gap that they need to bridge to reach their desired future state.

H3.2: Internal Modeling Capability

Once there is a high degree of connectivity in internal processes, one can start employing advanced models, essentially digitally mirroring the monitored process. Such models are called digital twins ([2.6.3, para. 3](#)) and can be used to digitally analyze and optimize the process. Through high internal analytic capability, organizations can also take the step from descriptive to diagnostic and predictive process monitoring ([2.5, para. 1-6](#)). This allows adjusting the process based on the current state and predicted outcomes. Production, maintenance and logistics are typical processes that can benefit from a high internal modeling capability as they are constantly repeated processes ([2.6.3, para. 1](#)). While investigating the internal modeling capability in Yara, the following observations were made:

H3.2.1 There are a lot of local process control data. These are typically not combined in larger models.

Explanation:

This observation is closely related to [H3.1.2](#). From the interviews, it is made clear that even though there are lots of local process control data at the production facilities, it is not combined in larger internal models either at site or centrally.

Causes:

There is a huge span of technology adoption at sites, which makes it challenging to standardize an optimization structure for Yara's production sites. It seems like there has not been a strong enough focus on structuring data and making data available for larger models, essentially data governance and data accessibility.

Effects:

Not employing larger models for process optimization is not very noticeable during day to day operations, however, it leads to Yara missing out on the opportunities data-driven process optimization can bring. If competitors start utilizing data-driven process optimization at a larger scale before Yara, they could potentially alter the market shares (2.1, para. 1).

Suggestions:

Yara should focus on incrementally increasing modeling capabilities through Kaizen events at individual sites, as sites become increasingly connected. This should be done with awareness of how each model would fit in larger models. As modeling capabilities at sites improve, it will make more sense to combine these in larger models, and it is therefore also important to focus on process data governance as it is less costly to prevent data cleaning efforts than to cure low quality data. (2.6.3, para. 1), (2.6.1, para. 3). Once individual sites have a sufficient modeling capability and connectivity, Yara should consider connecting individual sites together in a global monitoring and modeling system. This would most likely be more beneficial to do through Kaikaku events (2.1.1, para. 5). Yara should consider having a digitalization goal that revolves around implementing digital twins (2.6.3, para. 3) at feasible production facilities and logistics units. To ensure that the right actions are being taken to get the desired state, they should develop proof-of-concepts with small-scale testing for subsets of a given facility or unit, and upscale as they are satisfied with the impact these actions have (2.1.2, para. 5).

H3.3 Monitoring Capability/Rate of Monitoring

The rate in which these models can be updated based on new information is reliant on the rate of monitoring. The rate of monitoring measures how often new data is captured and transmitted. With today's technology, it is possible to capture and transmit data at rates that seem continuous (2.6.3, para. 1). Being able to continuously collect sensor data from processes, facilitates for far more details to be captured compared to capturing conventional snapshots of data (2.6.3, para. 2). While measuring the rate of monitoring in Yara, the following observations were made:

H3.3.1 Inconclusive findings due to large variability.

The interviews explained that there is such a large span of sensor connectivity and technology maturity in Yara's processes around the world that a general observation of the rate of monitoring is inapplicable. Some of the factories utilize more advanced sensor technology, but many have a high degree of analog and manual processes.

H3.4: Interconnectivity

To unlock the full potential of process optimization in an organization with a large and complex value chain, one needs to interconnect systems (2.6.3, para. 1). A lot of process monitoring systems are made somewhat ad hoc for a given production facility, a given logistics unit or so on. This is completely fine in most cases and ensures that the system suits the given unit. However, just like the unit that is being monitored plays an important part in the value chain, the system monitoring the unit plays an important part in the overall monitoring system. Therefore, it is important to ensure that ad hoc systems are made with a plan on how they fit into the overall system when Yara finds value in connecting data from different systems (2.6.3, para. 1). This because it is less costly to prevent incompatible data than to cure it (2.6.1, para. 3). While measuring the degree of interconnectivity in systems supporting Yara's processes, the following observations were made:

H3.4.1 Systems are made without future interconnectivity in mind.

Explanation

It was observed that new systems in different departments tend to be developed without keeping in mind that they could benefit from being interconnected with systems in other departments. They might be optimized for a given production facility or logistics unit, but without awareness on how they would fit in an interconnected system.

Causes:

It could appear that the management have not had a high focus of interconnected systems across departments and silos. It is of course difficult to interconnect systems that come from various

technological ages, and are created by different suppliers in different countries, but there could be a stronger focus on interconnectivity when creating new systems.

Effects:

When new systems are made without interconnectivity in mind, the silos (H1.1.2) between regions and departments in Yara are sustained. The full advantage of being such a large company owning the whole value chain is therefore not exploited (1.2.1, para. 3), (2.6.3, para. 1). By not facilitating for future interconnectivity, Yara might have to make substantial data cleaning efforts if they decide to interconnect systems at a later point (2.6.1, para. 3).

Suggestions:

Yara should investigate process dependencies to identify where positive synergies could be made and focus on interconnecting systems where they see potential value. When building new systems, Yara should be aware of how these could interconnect with existing and future systems. A digital roadmap planning how information could flow between systems to create positive synergy effects should be developed and iterated continuously. Yara should consider developing a more comprehensive digital strategy that encompasses the entire organization. Strategy drives digital transformation (2.1.1, para. 1), hence a detailed global digital strategy would drive both Yara's digital maturity and possibly the alignment of departments and regional offices. This thesis suggests that Yara take the proposed framework in 2.1.2 into consideration and develop and update a global digital roadmap with feasible actions that serve to bridge the gap between their current state and the desired future state needed to reach their digitalization goals.

H3.4.2 Employees see opportunities and value in connecting different types of systems.

Explanation:

This observation is closely related to previous observation H3.4.1. Most interviewees identified that interconnectivity of systems is currently limited, and they recognized that there is potential value in connecting systems both within and across departments.

H3.4.3 There are fragmented initiatives that work very well in different factories/business units.

Explanation:

Interviewees pointed out that there are lots of successful digital initiatives spread in the organization, but they seem to be initiated on an ad hoc basis and without being fully anchored in a digital roadmap.

Suggestions:

The fact that ad hoc initiated projects in Yara are successful is truly great, but somewhere along the line these systems may need to communicate. Yara could potentially face large data cleaning challenges ([2.6.1, para. 3](#)) if they store data generated from these initiatives without interconnectivity in mind. It is important that Yara focus on interconnectivity, communication and innovation to make sure that they will reach a digital maturity level that lets them tackle the challenges of the future ([2.1, para. 1](#)) ([2.1.1, para. 7](#)).

Insight H3

Yara have a potential to improve the monitoring and optimization of processes in the production and supply chain departments. For this improvement to be possible, it is recommended that there should be an increased focus on connectivity of sensors, process data governance and interconnectivity of systems. Improved sensor connectivity and process data governance could drastically improve efficiency and decrease downtime at sites – as processes and maintenance could be optimized based on the combination of real time data and historic data. It would also facilitate for great opportunities in advanced modeling such as digital twins. Improved interconnectivity of systems would be an important step towards tearing the walls of the silo structure down. As previously mentioned, Yara has enormous untapped opportunities for optimizing their value chain due to its scale. Exploiting this by tying the departments closer together and having systems that are compatible and collaborative could unlock great values.

Improvements in H3 would generally allow Yara to gain value through increased internal efficiency. Yara should adopt the framework proposed in [2.1.2](#), and develop digitalization goals to increase the efficiency of their production and logistics units. By utilizing Kaizen to continuously upgrade their digitalization efforts in process optimization, Yara can gradually

increase their efficiency with minimal risk involved. However, Yara might need to consider utilizing Kaikaku to more radically change the way production and logistics units work (2.1.1, para. 5). Through binding local changes together in a more global efficiency system, Yara could standardize process optimization. This would cut costs and reduce risks of process optimization efforts while facilitating for increased collaboration across departments and regions. Increased collaboration is a driving factor for digital transformation and would as such further accelerate Yara's ability to adapt in a digital environment. See Table 1.

H4. Precisely Tailored Products and Services

The more insight an organization has in the market they operate within, the better they can tailor offers, products and services to their segments 2.6.4, para. 1. Big data technology allows organizations to micro-segment their customers, meaning that they can gain deep real-time knowledge of specific subsets of their segments. For Yara, such a subset could be farms with a specific set of conditions, a specific farm, or even a specific square meter (1.2.2, para. 2). The more detailed data one is able to collect, the more specific the segments can become (2.6.2, para. 1). The more specific the segments, the more precisely one can tailor one's offer to suit the specific customer's needs (2.6.4, para. 1). An organization's capability to micro-segment is dependent on several factors. To measure the current state of data usage for segmentation in Yara, the following parameters were investigated:

H4.1 Data collection capability from external sources

H4.2 External modeling capability

H4.3 Generative capability

H4.4 Communication with end-user

H4.1 Data Collection Capability from External Sources

To perform real-time micro-segmentation on customers, big customer data is needed (2.6.4, para. 1). Thus, being able to collect data from external sources is a prerequisite to be able to micro-segment based on data. There is a wide range of external sources that one can collect useful data from (2.6.4, para. 1). The more data one has on the markets one operates in, the

better one can segment and the more precisely tailored products and services one can offer (2.6.4, para. 1). While investigating Yara's data collection capability from external sources, the following observations were made:

H4.1.1 The connectivity of Yara products and sensors are improving.

Explanation:

Interviewees points out that Yara is aware of the importance of getting sensors connected and that the connectivity of sensors in all segments is continuously improving.

Causes:

Data from sensors, apps and so on has previously not been collected due to a low degree of connectivity. It is suspected that this is due to the maturity of internet of things technology, and Yara not previously seeing as much use for collecting data from external sources. Yara now appear to see the opportunities collecting external data can offer.

Effects:

H4.1.2 is a direct effect of H4.1.1. A higher degree of connectivity increases Yara's opportunities for micro-segmentation 2.6.4, para. 1. Collecting sensor data could facilitate for building enhanced models for recommendations, market pull, user habits and product performance.

Suggestions:

As Yara increases their capability for collecting external data, they should develop thorough plans for its use-cases. They need to make sure that their analytical capability is at a high enough level to exploit the data they collect by refining it through analysis (2.4, para. 3). As models are developed, they need to consider what information could be relevant to share between models and facilitate for this at an early stage, as it is easier to prevent incompatible data than it is to cure it through data cleaning (2.6.1, para. 3). Yara should make sure that they have digitalization goals that require increased connectivity in products and sensors to increase the usefulness of increased connectivity.

H4.1.2 The number of data sources will increase rapidly and create enormous amounts of continuous data.

Explanation:

Some of the interviewees were also aware that increased sensor connectivity will lead to huge volumes of data with high velocity, and that this poses new challenges that Yara needs to address.

Causes:

H4.1.1 is a direct cause of H4.1.2. A growing degree of connectivity in sensors and apps enables Yara to potentially capture continuous data (2.6.3, para. 1) from a huge network of units (1.2.1, para. 3).

Suggestions:

With a growing degree of connectivity, Yara faces both challenges and benefits. For the data to be collected in a feasible way, systems need to be in place, both in terms of data storage capacity, analytical capacity, and data governance. Data governance need to be a continuous priority as it is easier to prevent low integrity data than to clean it (2.6.1, para. 3). Once Yara has the necessary infrastructure and supportive systems in place, they can start analyzing the external data to give better customer recommendations and to better foresee changes in the various segments (2.6.4, para. 1). External data from sensors and apps will produce enormous databases, requiring continuous upgrades in big data capability (H4.1.2).

H4.2 External Modeling Capability

Simply collecting the data is not very useful if you are not going to use it for something (2.3, para. 3). Just like modeling capability is an important part of internal process optimization (H3.2), the external modeling capability dictates how one is able to use data collected externally. By employing models, an organization can anticipate what a given segment needs, allowing them to tailor offers, products and services (2.6.4, para. 1). The higher modeling capability, the more diverse data one can add to the models (2.3, para. 2). For Yara, strong modeling capability could be highly beneficial for producing recommendations for farmers and industrial customers. While investigating Yara's external modeling capability, the following observations were made:

H4.2.1: Yara employs advanced models using large, complex datasets to give recommendations.

Explanation:

In two of the interviews it was made clear that Yara collect and use a large span of external data for modeling. These models seem to be made, iterated and tweaked by modelers and not by machines.

Suggestion:

Yara already employs advanced models to give recommendations. These, however, appear to be man-made models where experts define a set of rules that are used to analyze each given dataset. The way this appears to be done in Yara today is comparative to the way one censored pornographic material before convolutional neural networks (2.4.3, para. 4). By employing Machine Learning techniques, the algorithms these models are based upon could teach themselves which trends lead to which result, possibly drastically improving model accuracy (2.4.3, para. 5). Since convolutional neural networks were taken in use for censoring pornographic content, the accuracy of this censoring has increased from 43% to 97.2% (2.4.3, para. 4-5). If Yara would see an increase in accuracy like this, expert agronomists would be able to use more accurate models as basis for their farmer recommendations. Such models could also potentially help Yara better foresee field yield and market pull. Yara should consider making more advanced modeling for farmer recommendations a digitalization goal. To reach this goal, they would most likely have to bridge gaps in both skill and technology. They should start small and make proof-of-concepts utilizing open-source machine learning software such as Python with libraries like NumPy, SciKit-Learn, Keras, Tensorflow or Theano (2.4.3, para. 6). Yara should consider having a student project with the objective of developing a proof-of-concept utilizing machine learning algorithms for yield-estimation, product application recommendations, and other predictions that might be useful for Yara.

H4.3: Generative Capability

It is more resource intensive to tailor products and services to a lot of micro segments than for larger conventional segments (2.6.4, para. 2). Therefore, it is important to assess one's generative capability. The generative capability dictates how well an organization is able to

generate specialized offers, products and services, while maintaining scalability ([2.6.4, para. 2](#)). Having a high generative capability can also allow reports on markets and micro-segments to be generated more autonomously, freeing up time spent doing repetitive analysis on different markets and segments. While investigating Yara's generative capability, the following observations were made:

H4.3.1 Recommendations based on models are generated manually.

Explanation:

The recommendations for micro-segments seem to be generated manually after modeling.

Effects:

Manually created tailored services is not very scalable ([2.6.4, para. 2](#)). If Yara is able to increase the degree of automation either in model development or for building actual recommendations or offers, they can avoid their generative capability being the bottleneck for micro-segmentation.

Suggestions:

Even though Yara should look to increase their generative capability for recommendations and offers in the long run, it should not be a major focus for Yara at this point. This is because there most likely will be other bottlenecks that Yara should prioritize before their generative capability is inhibiting scalability. Depending on their digitalization goals, Yara might find their generative capability to be a bottleneck as they increase their ability to microsegment. Yara should keep their generative capability in mind and consider increasing it if they see it becoming a bottleneck for the scalability of tailoring products and services to more specific segments.

H4.4 End-user Communication

For micro-segmentation to be truly beneficial, purposeful ways of interacting with the customer is needed. Many companies solve this through own customer logins on their web-page. On these customer dashboards, customers can access relevant information and interact with the organization. Sometimes the customer is not the end-user, and it might therefore be necessary with other systems complementing the customer login. Being able to portray information to the customer on portable devices such as tablets and cell phones is increasingly useful ([2.2.2, para.](#)

3.) This can be done through apps, web-dashboards etc. While investigating how Yara communicates with the end-user, the following observations were made:

H4.4.1 Inconclusive findings due to sample of interviewees.

Insight H4

The external connectivity in Yara seems to be improving. This means that there are huge opportunities for micro-segmentation for Yara in the coming years. Great value could be unlocked by adapting a scalable machine learning system that precisely and efficiently recommends products for micro-segments (2.6.4, para. 2). For such a system to be realized it is important that there is a strong focus on how to handle the rapidly increasing volumes and velocity of data. Machine learning algorithms are reliant on high integrity data and customers are reliant on a communication platform that delivers the right information at the right time. Yara might have to be willing to commit to changes in business models and processes (2.1.1, para. 1) to gain more value through precisely tailored products and services as they get insight into more customized needs for more specific segments. Perhaps it would be wise to offer more subscription-based services incorporating sensors, providing more insight for farmers and commodity traders to act upon. Value gained through H4 will most likely be realized through increased exploitation of external opportunities, either by increasing the profitability of existing customers, or through enhancing Yara's ability to reach new customers (2.1.1, para. 2).

H5. Enhanced Innovation Through Product Data

Utilizing data from current products to understand how they are performing, and how customers actually use the products, can lead to great insight on which features and services the customers value (2.6.5, para. 1). This insight can then be used while innovating the next generation of products. By collecting data, one can also uncover potential faulty use of products. This makes it possible to avoid poor customer satisfaction due to user errors. Collecting data on which features are being valued the most also allows the organization to see which features they should focus on while marketing their products.

To measure the state of enhanced innovation through product data in Yara, the following parameters were investigated:

H5.1 End user feedback

H5.2 Product data feedback

H5.1.1 There isn't a proper feedback loop of user- and sensor data.

Explanation:

Yara appear to currently not be utilizing user- and sensor data in a feedback loop to optimize current products and features, or develop new products.

Causes:

The importance of a proper feedback loop collecting product usage data is debatable as Yara mainly produce consumable products. This might be why Yara has not previously prioritized a proper feedback loop. The degree of connectivity in products has not previously been very high (H3.1.1), which means that even if a feedback loop was in place, there would be limited and manual opportunities for feedback collection.

Effects:

Not having a proper feedback loop affects how much knowledge Yara is able to base further product development on. It might also hinder Yara from building certain types of predictive models where it could be interesting to know exactly how their products are being used.

Suggestions:

If Yara continue to improve the degree of connectivity in their products, while making sure that information is stored purposefully, and that it can be accessed efficiently, they should be able to derive product usage data from several of their sensors. This could then be used for innovation of new products and services, or for more complex analyses where product usage data is useful. Yara should strongly consider adding increased data-driven innovation through a proper feedback loop of product and user data as one of their digitalization goals.

5. Quality of Study

(Bryman, 2012) explains that reliability and validity can be assimilated from quantitative to qualitative research by adjusting the measurement issues. This chapter will assess the quality of the research design of the thesis, by addressing the internal and external reliability and validity of the study.

5.1 External Reliability

(Bryman, 2012) describes external reliability as the degree in which the study can be replicated by other researchers, something that is hard to do in qualitative work. (Bryman, 2012) also says that the best way to replicate such a study is to adopt the social role of the original researcher. This study is conducted by master students that has gained insight in Yara through semi-structured qualitative interviewing, public information and internal documents. The latter is a subject of a non-disclosure agreement and can therefore not be published. The interviews are anonymous, which means that a researcher would experience difficulties in getting in touch with the interviewees from this study, unless Yara and the interviewees themselves approve such information to be distributed. These aspects would hinder an identic replication. The authors have, however, purposive sampled representative employee, so there is reason to believe that a researcher would achieve same results with a different interviewee sample that fulfills the same criteria.

The research process in the study has been thoroughly documented. To improve reliability, a case study database was created early in the process. The database is stored electronically and consists of documents such as interview guides, transcripts, audio files, information about the interviewees and memos. The study does, however, present a snapshot of Yara's big data adoption, thus replicating the study would identify the development in which Yara has undergone during the time between the original study and the replication.

5.2 Internal Reliability

Internal reliability, or inter-observer consistency as (Bryman, 2012) states it, could be a challenge in qualitative studies that has more than one researcher. This thesis has two authors, and the authors have been aware of the challenge with consistency throughout the study. The risk of having inconsistency in observations and categorization between the authors have been mitigated by collaborating on all relevant aspects of observations and categorization. Any disagreement has been debated and not settled until all parties have agreed on the outcome. The authors have been working closely and collocated throughout the study, something that has continuously aligned the parties and led the number of disagreements to a bare minimum.

5.3 Internal Validity

(Bryman, 2012) outlines that internal validity mainly relates to causality and its relationship with conclusions. This thesis identifies a number of causes for the observations and thus raises the question of whether the conclusion holds water. The causes in this thesis are a subject of qualitative work with an interpretivist approach, meaning that the authors have analyzed the transcripts and used the literature study to connect causes to observations. The conclusions are the authors perceived truth and should not be seen as absolute truths. The scope of the study has also prevented the authors of a deeper investigation of specific observations and causes, which make the causes in this thesis perceptions rather than absolute truths.

5.4 External Validity

(Bryman, 2012) refers to external validity as of which the insight can be generalized across social settings. The generalization of social settings will in this thesis be considered both across Yara and across the industry. To discuss the two elements, a series of questions will be answered.

Has the study interviewed the right employees?

The quantitative analysis aimed to take a snapshot of big data adoption across the company. This means that the authors sought to conduct interviewees with employees from all departments. A total of 13 interviewees participated in the study. The fact that Yara has more

than 15 000 employees indicates that there is a lot of information that has not been captured in the snapshot. With that said, the questions were open, and the interviewees were able to reveal their experience and knowledge in a number of topics. The findings were repetitive throughout all interviews, and combining this with public information and internal documents, the authors are fairly confident that the insight is applicable across the company.

Has the study interviewed enough employees?

As pointed out in (1.1, para. 3), time was the constraint in the number of conducted interviews. It can be discussed whether 12 interviews were sufficient to extract relevant findings. Considering the nature of a qualitative study with an interpretivist approach (3.1.3), the authors are once again confident that the research process was able to synthesize trustworthy insights from the company. The authors do however acknowledge that the sampling had hints of a convenience sampling, resulting in fewer interviewees at some fields of expertise. This led to inconclusive findings twice, both at the rate of monitoring and the end user communication. As the insight gained in these measurables were insufficient, the authors deemed them inconclusive.

Has the study asked the right questions?

The authors were aware of the concern that in semi-structured interviews, there might be a different conception of different terms and concepts. The authors treated this concern by avoiding using ambiguous terms and continuously asking the interviewee if something needed to be clarified. The authors also approached the interviewees by adapting to their background and expertise, as the authors have an interdisciplinary background that makes it possible to somewhat adapt to both a technical language and a business language.

Has the study answered the research questions?

RQ1 was answered by conducting semi-structured interviews in the qualitative analysis. The analysis identified a snapshot of Yara's current state of big data technology adoption. The snapshot attempted to provide a generalized impression of the state of big data technology adoption in Yara. As a result of this, there is a limited level of detail in causes, effects and suggestions. RQ2 was answered through analyzing the findings in RQ1 in relation to theory presented in chapter 2. For the reader's convenience, a summary of the findings on RQ1 and RQ2 for Yara's performance in each value driver is presented in its own insight section below each respective value driver.

Has the study been able to quantify the results?

The scope of the study has been a limiting factor for quantification of the results. The study does not investigate specific topics in depth where it could be plausible to quantify results. It would however be more feasible for Yara to initiate efforts based on the results of the study if it had been able to present quantified results. This is the reason for the thesis to call for several student projects and further internal investigations by Yara.

Summary

Based on these reflections the authors are confident that the thesis observes, identifies and measures what it says it does (Bryman, 2012). They are also confident that due to the scope and the process of the study, the insights are generalizable across the departments and throughout the company, despite the lack of quantification.

The study does not investigate whether Yara's situation and opportunities are applicable for other companies. However, based on the literature study and the authors' relation to industrial companies throughout the master's degree, there is reason to believe that several aspects of the thesis could be generalized for companies with similarities to Yara.

6. Conclusion

The research in this thesis was based on the assumption that Yara has great opportunities to strengthen their big data capabilities and thus unlock large values. It had the objective to answer the following research questions:

RQ1: *What is the current state of big data technology adoption through digitalization in Yara?*

RQ2: *What should be emphasized in the coming years to unlock potential value from big data?*

To answer these research questions, a qualitative case study was conducted through interviews with employees with various responsibilities and backgrounds in Yara. The observations made during these interviews were then analyzed in chapter 4 in relation to theory presented in chapter 2. An overview of the analysis can be found in Figure 7. The findings on RQ1 are summarized in 6.1, and the findings on RQ2 are summarized in 6.2.

6.1 The Current State of Big Data Technology Adoption in Yara

In order to increase the measurability of the current state of big data technology adoption in Yara, this thesis investigated Yara's current performance within five big data value drivers identified by (Manyika, et al., 2011). The findings within each value driver are presented and discussed in the following chapters:

H1. Data Transparency and Accessibility

H2. Data-Driven Decision Making

H3. Internal Process Monitoring and Optimization

H4. Precisely Tailored Products and Services

H5. Enhanced Innovation Through Product Data

The findings are relatively unambiguous and indicate that Yara are at an early stage of adopting big data technologies. This thesis identifies increased data accessibility as a catalyst for other value drivers, and observes several issues related to how data is made available and accessed

(2.6.1, para. 1). These include a lack of a data lake (H1.1.1), data being stored in silos or ad-hoc (H1.1.2), a low focus on systems alignment following acquisitions (H1.3.2), and employees accessing data through key persons (H1.2.1).

Decision-making in Yara seems to be data-driven to the extent that it is reasonably practicable with Yara's current ability to make high quality data available. Employees indicated that analyses are kept at a high level (H2.2.1), and that it is hard to find the underlying data in aggregated reports (H2.3.2). Aggregating data for analyses and reports also seems to be time-consuming, decreasing the time that is made available for reflecting on, and controlling the output of these analyses (H2.1.1).

Yara does not seem to be utilizing advanced analytics or modeling for large scale process optimization (H3.2.1). Employees do, however, indicate that Yara is working on improving the degree of connectivity at facilities, logistics units, and in machines and equipment (H3.1.2), (H4.1.1). As Yara keep improving the connectivity, they will experience larger volumes of data coming from a higher variety of sources being generated at an increased velocity (H4.1.2). There seems to be limited focus on interconnecting optimization systems across departments and regions (H3.4.1).

Yara uses advanced models to give agricultural recommendations. These appear to be tweaked manually, mainly relying on the expertise of agronomists (H4.2.1). Yara also seem to be focusing on improving their external data collection capability (H4.1.1). Depending on their analytical capability and their generative capability, Yara could see themselves rapidly increasing their micro-segmentation capability.

Yara does not seem to have a proper feedback loop for end user feedback or product data (H5.1.1). This affects the knowledge Yara has on the performance of their products, and on how customers use current products, possibly inhibiting the insight Yara is able to base future innovation on.

The thesis finds that, throughout the organization, data seem to be used mainly for descriptive purposes today (H2.1.2). This means that data is used to describe past or current events (2.5). Employees seem to have a hard time accessing underlying details of aggregated reports and

analyses (H2.3.2), making it hard to diagnose what's causing observed events. This inhibits Yara's ability to address issues with correct actions.

6.2 Suggestions for What Should be Emphasized in Coming Years

Through the literature study, the thesis finds that making data more easily accessible to relevant stakeholders can create value. Allowing data to be accessible across departments and systems can reduce search and processing time, decreasing time spent aggregating data for reports and analyses. Making data available throughout the organization enhances cooperation and alignment between departments and regions. Increased internal transparency through facilitating for a higher degree of traceability and accessibility of data also allows for better measurability within the organization. This leads to increased managerial knowledge, boosting decision-making, resource planning and process optimization (2.6.1, para. 1).

A data lake should be a sound investment for Yara as it is an effective way of increasing data accessibility and transparency (2.6.1, para. 2). For a data lake to be successful, Yara must first have a clear vision on what the use-cases of a data lake would be. Yara also need to have a strong focus on improving data governance and computer security as these are important supporting factors of a data lake.

Making data accessible is the prerequisite for being able to extract value from it (2.6.1, para. 1). The next step is being able to use the data (2.4, para. 2). Cognizant identifies the value of analysis to scale with the volume of data, and the analytical capability (Figure 4). Hence, Yara should invest in increased analytical capability as increasing volumes of data are made accessible.

For decision makers to become increasingly data-driven, they need to embrace the output of analyses as evidence (2.6.2, para. 1). To increase the degree of trust decision-makers have in data, input data needs to hold a high level of integrity (2.6.2, para. 1). Sometimes it is beneficial to investigate underlying causes of events observed in analyses. Hence, facilitating for traceability of data, i.e. making it easier to trace aggregated reports and analyses back to detailed data, is important to increase Yara's ability to diagnose what is causing observed events (2.6.1, para. 4), (2.5). Data governance and data cleaning are both activities that enhances data integrity

and facilitate for data to be traceable (2.3.1) (2.6.1, para. 3). A high level of data governance can prevent the need for data cleaning, hence the authors suggest that Yara invest in high standards for data governance of future data and focus on cleaning historical data that are deemed useful for future appliances.

By focusing on further increasing the degree of connectivity in internal processes, Yara should be able to improve their ability to monitor and optimize their processes (Insight H3). Due to Yara's immense scale (1.2.1, para. 3), even marginal increases in efficiency can lead to large cost reductions. A high degree of interconnectivity would allow process control systems to communicate with each other, further increasing Yara's ability to increase internal efficiency (Insight H3).

As Yara's data collection capability from external sources keep increasing, Yara should be able to increase their performance in offering specifically tailored products and services (Insight H4). They should also be able to gain valuable feedback from customers, enhancing innovation. This would increase customer value by delivering products and services that better suit customer needs (2.6.4, para. 1). Yara should strive to develop proof of concepts for advanced algorithms modelling crop yield and product appliance (H4.2.1, suggestions).

A higher level of big data technology adoption should allow Yara to utilize their data for increased value creation in all the five value drivers. Although the investigation of the current state of big data adoption in Yara indicates that Yara is at an early stage of big data technology adoption, the authors are confident that Yara could increase their utilization of data quite rapidly during the coming years. By investing in data infrastructure, data governance, and planning how systems can be interconnected, Yara would ensure that new systems are able to communicate, and that data is more accessible. New data would also have higher integrity, and thus be more fit for advanced analytics. In the realm of data, prevention is better than cleaning.

Investments in technology, however, is not enough. Yara also need to digitally mature through digital transformation (2.1.1, para. 1). Strategy, not technology, drives digital transformation (C. Kane, Palmer, Nguyen Phillips, Kiron, & Buckley, 2015). Hence, it is crucial for Yara to develop and continually update a digital strategy encompassing the entire organization. This thesis recommends utilizing the framework presented in 2.1.2 to develop a detailed digital roadmap that is reiterated upon as needed. This roadmap should include all of Yara's digital

endeavors and be aware of which changes digital implementations will lead to at the process level, organization level, business domain level and society level.

Bibliography

- Acaps. (2017). *Data Cleaning*. Retrieved from Acaps - See the crisis, change the outcome: https://www.acaps.org/sites/acaps/files/resources/files/acaps_technical_brief_data_cleaning_april_2016_0.pdf
- Aaltonen, A., & Tempini, N. (2014). Everything Counts in Large Amounts: A Critical Realist Case Study on Data-based Production. *Journal of Information Technology*, 97-110.
- Amatriain, X. (2016, January 14). *Medium*. Retrieved May 9, 2018, from What's the relationship between machine learning and data mining?: <https://medium.com/@xamat/what-s-the-relationship-between-machine-learning-and-data-mining-8c8675966615>
- Ariker, M., Heller, J., Diaz, A., & Perrey, J. (2015, November 23). *How Marketers Can Personalize at Scale*. Retrieved June 7, 2018, from Harvard Business Review: <https://hbr.org/2015/11/how-marketers-can-personalize-at-scale>
- Bobriakov, I. (2017, May 9). *Top 15 Python Libraries for Data Science in 2017*. Retrieved June 7, 2018, from Medium: <https://medium.com/activewizards-machine-learning-company/top-15-python-libraries-for-data-science-in-in-2017-ab61b4f9b4a7>
- Bryman, A. (2012). *Social Research Methods 4th ed*. New York: Oxford University Press.
- Carson, D., Gilmore, A., Perry, C., & Gronhaug, K. (2001). *Qualitative Marketing Research*. United Kingdom: Sage Publications Ltd.
- C. Kane, G., Palmer, D., Nguyen Phillips, A., Kiron, D., & Buckley, N. (2015). *Strategy, not Technology, Drives Digital Transformation*. MIT Sloan Management Review. Massachusetts: MIT Sloan Management Review. Retrieved from MITSloan Management Review.
- Cognizant. (2016, March). Driving Value Through Data Analytics: The Path from Raw Data to Informational Wisdom. *Cognizant 20-20 Insights*.
- Corporate Finance Institute. (n.d.). *Types of Synergies*. Retrieved from Corporate Finance Institute: <https://corporatefinanceinstitute.com/resources/knowledge/valuation/types-of-synergies/>
- Datafloq. (2016, July 13). *4 Examples of Big Data Implementation in Customer Micro-Segmentation*. Retrieved from DATAFLOQ: <https://datafloq.com/read/4-examples-big-data-implementation-customers/2171>

- dataPARC. (2017, February 1). *Static, Live or Dynamic - Which Report is Best?* Retrieved May 27, 2018, from dataPARC: <http://blog.dataparcsolutions.com/static-live-dynamic-report-best>
- Diebold, F. X. (2012). *A Personal Perspective on the Origin(s) and Development of "Big Data": The Phenomenon, the Term, and the Discipline**. University of Pennsylvania. Philadelphia, PA: Penn Institute for Economic Research.
- Drew, R. (2017, February 6). *Scalable Personalization: Challenging the Mass-Market Mindset*. Retrieved from emedia: <http://www.emedia.com/scalable-personalization-challenging-mass-market-mindset/>
- Eleftheriadis, R. J., & Myklebust, O. (2016). A guideline of quality steps towards Zero Defect Manufacturing in Industry. *Proceedings of the 2016 International Conference on Industrial Engineering and Operations Management* (pp. 332-340). Detroit: IEOM Society International.
- EMC, IDC & Cyclone Interactive. (2014, April). The Digital Universe of Opportunities.
- Fleck, M. M., Forsyth, D. A., & Bregler, C. (1996). Finding naked people. (pp. 593-602). Berlin, Heidelberg: Springer Berlin Heidelberg.
- Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*(35), pp. 137-144.
- Günther, A. W., Mehrizi, H. R., Huysman, M., & Feldberg, F. (2017). Debating big data: A literature review on realizing value from big data. *The Journal of Strategic Information Systems*, 26(3), 191-209.
- Gillon, K., Aral, S., Ching-Yung, L., Mithas, S., & Zozulia, M. (2014). Business Analytics: Radical Shift or Incremental Change? *Communications of the Association for Information Systems*, 34, 287-296.
- Gung. (2016, April 27). *Cross Validated*. Retrieved May 9, 2018, from What is the difference between data mining, statistics, machine learning and AI?: <https://stats.stackexchange.com/questions/5026/what-is-the-difference-between-data-mining-statistics-machine-learning-and-ai>
- Gupta, S. (2014, August 10). *Garbage In - Garbage Out: Your analysis can be as good as the data you use*. Retrieved May 29, 2018, from LinkedIn: <https://www.linkedin.com/pulse/20140810110847-44264733-garbage-in-garbage-out-your-analysis-can-be-as-good-as-the-data-you-use/>

- Heckler, B., & Gates, D. (2017, 06 09). *What i4.0 Means for Supply Chains*. Retrieved from IndustryWeek: <http://www.industryweek.com/supply-chain/what-i40-means-supply-chains>
- Hurwitz, J. S., Nugent, A. F., Halper, D., & Kaufman, M. A. (2013). *Big Data For Dummies®*. Hoboken: John Wiley & Sons, Inc.
- IBM: The Big Data & Analytics Hub. (n.d.). *IBM: Infographics & Animations*. Retrieved from The Four V's of Big Data: <http://www.ibmbigdatahub.com/infographic/four-vs-big-data>
- Infosys Limited. (2017). *Effective Data Governance*. Retrieved May 28, 2018, from Infosys: <https://www.infosys.com/data-analytics/insights/Documents/effective-data-governance.pdf>
- Kerravala, Z., & Miller, L. C. (2017). *Digital Transformation For Dummies®, Mitel Special Edition*. 111 River St.: John Wiley & Sons, Inc.
- Laney, D. (2001, February 6). 3D Data Management: Controlling Data Volume, Velocity, and Variety. *Application Delivery Strategies By META Group Inc*.
- Linstedt, D., & Inmon, W. (2014). *Data Architecture: A Primer for the Data Scientist: Big Data, Data Warehouse and Data Vault*. Morgan Kaufmann.
- Loebbecke, C., & Picot, A. (2015). Reflections on societal and business model transformation arising from digitization and big data analytics: A research agenda. *The Journal of Strategic Information Systems*, 24(3), 149-157.
- Maini, V. (2017, August 19). *Machine Learning for Humans*. Retrieved May 8, 2018, from Machine Learning for Humans, Part 3: Unsupervised Learning: <https://medium.com/machine-learning-for-humans/unsupervised-learning-f45587588294>
- Maini, V. (2017, August 19). *Machine Learning for Humans, Part 2.1: Supervised Learning*. Retrieved May 07, 2018, from Machine Learning for Humans: <https://medium.com/machine-learning-for-humans/supervised-learning-740383a2feab>
- Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Hung Byers, A. (2011). *Big data: The next frontier for innovation, competition, and productivity*. McKinsey Global Institute.
- McAfee, A., Brynjolfsson, E., Davenport, T. H., Patil, D., & Barton, D. (2012). Big data: the management revolution. *Harvard business review*, 90(10), 60-68.
- Megahed, F. M., & Jones-Farmer, A. (2013). A Statistical Process Monitoring Perspective on Big Data.

- Megahed, F. M., & Jones-Farmer, L. A. (2015). Statistical perspectives on “big data”. In *Frontiers in Statistical Quality Control 11* (pp. 29-47). Springer.
- Ohara, D. (2012, April). *The big machine: creating value out of machine-driven big data*. Retrieved April 9, 2018, from Splunk White Papers: https://www.splunk.com/web_assets/pdfs/secure/The_big_machine_creating_value_out_of_machine_driven_big_data.pdf
- Oracle. (2017, January). *Digital Twins for IoT Applications*. Retrieved from Oracle Fusion Applications: <http://www.oracle.com/us/solutions/internetofthings/digital-twins-for-iot-apps-wp-3491953.pdf>
- Parviainen, P., Tihinen, M., Kääriäinen, J., & Teppola, S. (2017). Tackling the digitalization challenge: how to benefit from digitalization in practice. *International Journal of Information Systems and Project Management*, 5(1), 63-77.
- Pettey, C. (2017, September 18). *Prepare for the Impact of Digital Twins*. Retrieved from Smarter With Gartner: <https://www.gartner.com/smarterwithgartner/prepare-for-the-impact-of-digital-twins/>
- Qin, S. J. (2014). Process data analytics in the era of big data. *AIChE Journal*, 60(9), 3092-3100.
- Rice, B. (2017, 4 17). *What is Big Data's role in process optimization?* . Retrieved 3 14, 2018, from Plant Services: <https://www.plantservices.com/blogs/the-feedback-loop/what-is-big-datas-role-in-process-optimization/>
- Russell, S., & Norvig, P. (2009). *Artificial Intelligence: A Modern Approach third edition* . Upper Saddle River, New Jersey: PRENTICE HALL SERIES IN ARTIFICIAL INTELLIGENCE.
- Russom, P. (2011). *TDWI Best Practices Report: Big Data Analytics*. TDWI Research.
- Schmarzo, B. (2017, April 7). *Economic Value of Data (EvD) Challenges* . Retrieved April 4, 2018, from Dell EMC Infocus: https://infocus.dell EMC.com/william_schmarzo/economic-value-data-challenges/
- Schmarzo, B., & Sidaoui, M. (n.d). *APPLYING ECONOMIC CONCEPTS TO BIG DATA TO DETERMINE THE FINANCIAL VALUE OF THE ORGANIZATION'S DATA AND ANALYTICS, AND UNDERSTANDING THE RAMIFICATIONS ON THE ORGANIZATIONS' FINANCIAL STATEMENTS AND IT OPERATIONS AND BUSINESS STRATEGIES*. Hopkinton: DELL EMC.
- Seeliger, J., Awalegaonkar, K., Lampiris, C., & Bellomo, G. (2004). *So You Want to Get Lean Kaizen or Kaikaku?* Retrieved June 1, 2018, from oliverwyman.de:

- <http://www.oliverwyman.de/content/dam/oliver-wyman/global/en/files/archive/2004/OPT06-LeanKaiKaiku.pdf>
- Sentance, B. (2016, October 19). *Data Lineage and Traceability: Using Data to Improve your Data*. Retrieved May 27, 2018, from Data Management Review: <https://datamanagementreview.com/enterprise-analytics/blog-entry/data-lineage-and-traceability-using-data-improve-your-data>
- Tikhonov, M., Little, S. C., & Gregor, T. (2015). Only accessible information is useful: insights from gradient-mediated patterning. *Royal Society Open Science*, 2(11), 150486.
- Tupper, C. (2011). *Data Architecture : From Zen to Reality*. Elsevier Science & Technology.
- Wamba, F. S., Akter, S., Edwards, A., Chopin, G., & Gnanzou, D. (2015). How 'big data' can make big impact: Findings from a systematic review and a longitudinal case study. *International Journal of Production Economics*, 165, 234-246.
- Wamba, F. S., Gunasekaran, A., Akter, S., Ren, J.-f. S., Dubey, R., & Childe, J. S. (2017). Big data analytics and firm performance: Effects of dynamic capabilities. *Journal of Business Research*, 70, 356-365.
- Witten, I. H., & Frank, E. (2005). *Data Mining - Practical Machine Learning Tools and Techniques, Second Edition*. San Francisco: Elsevier.
- Wu, X., Zhu, X., Wu, G.-Q., & Ding, W. (2014). Data mining with big data. *IEEE transactions on knowledge and data engineering*, 26(1), 97-107.
- Yara International ASA. (n.d.). *Crop nutrition solutions: Digital farming*. Retrieved from Yara International ASA: <https://www.yara.com/crop-nutrition/digital-farming/>
- Yara International ASA. (n.d.). *Crop nutrition solutions: Digital hubs*. Retrieved from Yara International ASA: <https://www.yara.com/crop-nutrition/digital-farming/our-hubs/>
- Yara International ASA. (n.d.). *Mission, vision and values*. Retrieved from Yara International ASA: <https://www.yara.com/this-is-yara/mission-vision-and-values/>
- Yara International ASA. (n.d.). *Yara Homepage, 1900-1905* . Retrieved 01 28, 2018, from History 1900-1905: <http://yara.com/about/history/1900-1905/>
- Yara International ASA. (n.d.). *Yara Homepage, History*. Retrieved 01 28, 2018, from The History of Yara: <http://yara.com/about/history/>
- Yara International ASA. (n.d.). *Yara Homepage, What We Do*. Retrieved 01 29, 2018, from What We Do: http://yara.com/about/what_we_do/

Zhou, K., Zhuo, L., Geng, Z., Zhang, J., & Li, X. G. (2016). Convolutional neural networks based pornographic image classification. *2016 IEEE Second International Conference on Multimedia Big Data (BigMM)* (pp. 206-209). Beijing: IEEE.

Appendix 1. Interview Guide

Introduction

Hi! We are two engineering students writing a master's thesis in cooperation with Yara's digital farming division. The topic of the master's thesis is big data, and we are currently running diagnostics on all divisions of Yara to identify the current state of big data technology adoption through digitalization in Yara. We do this because we believe Yara has a huge potential for further value generation through big data technologies. Our thesis answers the following research questions:

RQ1: *What is the current state of big data technology adoption in Yara?*

RQ2: *What should be emphasized in the coming years to unlock potential value from big data?*

To answer these research questions, we are conducting interviews with Yara employees. The insight we get from the interviews will be included in the thesis anonymously. We have an NDA in place and will not disclose any information that is sensitive to Yara. We will only use the interviews for internal diagnostics in Yara. The more accurate findings we get, the more accurate recommendations we can provide. We therefore ask your honest subjective opinion while answering the interview. We wish to discuss the following five topics:

H1: Internal data transparency and accessibility

H2: Data usage in decision making

H3: Process control and optimization

H4: Micro-segmentation and highly specific customer knowledge

H5: Product development and data collection from current products

We are interviewing a lot of different Yara employees spread throughout the organization, so if there are somethings that seem unclear, just ask us to elaborate. We're asking for your honest subjective opinion. We would like you to answer what you think is true, or what you feel is true from your perspective. Does that sound okay? Great, let's start!

Questions

Table 4: Semi-structured interview guide for employees in Yara. Questions tailored to specific employees and their field of expertise is not included.

| Number | Questions | Relevance/Value Driver |
|--------|---|------------------------|
| 1 | <p>Could you please start by explaining a bit about yourself and your role in Yara?</p> | General |
| 2 | <p>If you are going to use data, it's important that you can access the data – or find it somewhere. In a general term, how easy is it to access the data/information you need in Yara?</p> <ul style="list-style-type: none"> a) Where do you find this data? b) How do you access it? c) What kind of information is it important for you to access? d) Is that data mainly from internal or external sources? e) What about data from other departments? | H1 |
| 3 | <p>I want to address the topic of decision making. During work you are probably faced with decisions of varying magnitude. Could you talk a little bit about how you use your experience and how do you use statistics while making decisions?</p> <ul style="list-style-type: none"> a) We are trying to dig into how Yara use data to improve their decision making. Do you believe Yara could improve their decision making by using more data? b) How? c) Analytics? Visualization? d) How is the data presented to decision makers? e) Do you feel like data you use are precise enough? f) Do you trust it? | H2 |
| 4 | <p>We would like to talk about improved process optimization with big data.</p> | H3 |

| | | |
|----------|---|-----------|
| | <ul style="list-style-type: none"> a) What kind of processes are you collecting data from? b) What data is that typically? c) Do you use that data for process improvement or mainly for monitoring? d) In production – do you typically save snapshots (samples) or do you save continuous data? e) How is that data stored? <p>Additional topics to address if relevant: Management planning, HR planning (ERP systems), production processes and logistics</p> | |
| 5 | <p>We would like to talk a bit about segmentation of products and services with big data. Does Yara collect customer data?</p> <ul style="list-style-type: none"> a) What kind of data are you (planning on) collecting? b) From where? c) Does Yara use customer data to give individual recommendations or offers? d) Does Yara use customer data or other data to predict future needs of the customer? e) Have you heard about GDPR? Great! So - GDPR is approaching and will be enforced May 25th. Do you know if Yara are taking actions to prepare for this new legislation? | H4 |
| 6 | <p>We would like to talk a bit about improvement of products and services from big data. Do you collect data from products? For instance, N-adapt sensor, apps etc.</p> <ul style="list-style-type: none"> a) What kind of data do you collect? b) Is this data used to drive product development/Improvement of services etc.? c) Do you know if Yara is taking any actions to collect more data from their products? | H5 |

| | | |
|--|---|--|
| | d) Does Yara measure how customers actually use different products? For instance, which features or services they value the most. | |
|--|---|--|