Scalability of machine learning within the heavy asset industry

A multi case study for identifying barriers and possible solutions

Steffen Novak Mollestad

Master's Thesis
Master of Science, Innovation and Entrepreneurship
30 credits

The Department of Informatics
The Faculty of Mathematics and Natural Sciences
UNIVERSITY OF OSLO

May, 2019

Scalability of Machine Learning within the Heavy Asset Industry

Steffen Novak Mollestad

http://www.duo.uio.no/

Reprosentralen, University of Oslo

Abstract

The purpose of this research paper is to identify the barriers of and propose possible solutions to machine learning scalability in the heavy asset industry, as these are blocking further innovations within the industry. The research has been designed with an inductive approach resulting in explanatory and explorative multi case study with Facebook, Uber, Tesla, GE Digital, Arundo Analytics and C3 as selected cases.

The research findings highlight lack of uniform data handling within the heavy asset industry, limit for generalization of ML models due dissimilarities of industrial objects of analysis, and knowledge gap in external organizations as barriers. Further, a conceptual framework for how to approach these hurdles have been proposed.

Acknowledgement

This thesis marks the end of my M.Sc. in Innovation and Entrepreneurship. As of my background as an industrial engineer I hold the interest of both the scientific technological problems, technical industrial cases as well as innovational aspects. By diving into this research theme I have been able to cultivate technology within the scope of innovation management. This thesis hence aims to explore these two dimension accordingly, in the quest of academic and corporate contribution and personal curiosity.

This work is a result of many contributors and it's my greatest honor to have had the chance of interacting with all these intelligent and dedicated people. First off, the thesis is conducted in collaboration with Arundo Analytics AS. I would like to give them a huge thanks for letting me take part of a problem that is highly relevant during the time of writing, and for offering me resources to perform this research. Special thanks to Edoardo Jacucci and Marcus Furuholmen for setting up this opportunity. Also I'd like to thank every interview respondent for freeing up time to be interviewed. I'd also like to thank my supervisor Øystein Stavø Høvig for incredible effort of late readings, valuable input and guidance. Your input was vital for the quality of the thesis. Finally, for being unconditionally supportive I'd like to thank my wife and daughter.

My hope is that this relatively small contribution of research can be helpful within the academic and corporate field.

Drammen, May 2019
Steffen Novak Mollestad

Contents

Figures

Tables

Abbreviation

ML – Machine learning
AI – Artificial intelligence
DS – Data Science
UI – User Interface
MLaaS – Machine learning-as-a-Service
OEM – Original Equipment Manufacturer
ISV – Independent Software Vendor

1 Introduction

Heavy loads of data are being collected and stored from all types of sources nowadays and to analyze these data machine learning (ML) models are typically being applied in various applications. ML models are fed with input data and determines a output – typically a prediction – based on training from previous data. These models have proved to achieve high accuracy conditioned by the right training and engineering for a set of input data. By operationalizing these models – and at scale – tremendous value is being released. Due to potential high-quality predictions from ML it has therefore been applied and deployed into a broad range of sectors, e.g. telecom, banking, insurance, etc. The common denominator in those cases is usually that analytics are being performed handling data about people.

This is not the case for the heavy asset industry (further referred to as industrial or industry) – e.g. maritime, oil and gas, chemicals – where the application of ML models often are related to optimizing production and maintenance, and thereby analysis of industrial equipment. Cloning a ML model for one equipment to another similar type does not apply, hence applying ML models seems to be harder to scale for the industrial cases. Companies have been trying to – and still do – to solve the scalability issue within the industry, but no-one has really succeeded yet.

The hurdle of industrial scalability are blocking further innovations within the industrial sector, and are therefore relevant to solve in the context of innovation management. This is the point of departure for this thesis. As of this, the research have been aiming towards investigating what is causing these limitations. In quest of discover what could unleash innovation within the industry, it has been needed to dive into the technological sphere of ML and understand how this technology have been implemented and scaled within other sectors. This brings two dimensions to the thesis – both innovational and technological – thus will be analyzed accordingly. In addition, as academic literature also emphasizes that aspects such as organizational and business also impacts the processes of new innovations, thus will be included in the analysis.

In this case study Facebook, Uber, Tesla, GE Digital, Arundo Analytics and C3 have been examined. This research points to key barriers for scaling of industrial ML and possible strategies to overcome them.

1.1 Research Problem

The research first seeks to understand the characteristics of successfully scaled ML, further identify the barriers to scaled industrial ML, followed by outlining possible strategies and framework for dealing with those barriers. Due to the scope the thesis will seek to answer three different questions and therefore causing the analysis to be segmented accordingly into three sections.

Research question 1:
(1) *What is characterizing scaled machine learning in the selected cases and how does it contribute to business value and innovation?*

The objective is to outline the attributes of scaled ML thus understand how mass-scaling of ML models has been driving business value and innovation.

Research question 2:
(2) *What is the barriers making ML scalability challenging in the industry?*

This part of the thesis will seek to investigate specific elements which are blocking industrial ML scaling. The objective will therefore be to identify these elements and evaluate their impact within the industrial sphere, hence reveal what distinguish the industrial case from other cases.

Research question 3:
(3) *What are possible strategies to solve these scalability barriers?*

Based on the prior analysis performed, strategic guidelines for industrial scalability will be outlined. By doing so, the objective is to overcome the barriers of scaling thus guide further innovation within industrial ML.

1.2 Structure of the thesis

In the first chapter it is described the background and relevance for the subject of the thesis. In the second chapter the theoretical background and context is being explained. In the third chapter the methodology of the thesis is outlined including case selection and data collection. In the fourth chapter the analysis is presented where there is three sections, each connected directly to the corresponding research questions. In the fifth chapter the conclusion is presented alongside with its implications.

2 Theoretical background

In this section relevant theoretical framework are presented to give a fundament and solid understanding of the research problem. In particular, the first chapter focuses on context of relevant technologies and terminologies. The second chapter are focusing on the concepts of innovation management and its context. The third chapter outlines a conceptual framework and its content which will work as the baseline for the thesis analysis. Finally, the fourth chapter will summarize the theoretical background.

2.1 Technology context and terminology

In this chapter essential terminologies and context will be explained and hence works as a fundament for understanding the scope and focus of the thesis.

Context of ML. Heavy loads of data are being collected and are available from all types of sources – also commonly labeled as Big Data – and to analyze these data machine learning (ML) models are typically being applied. Previously, organizations could put together teams of statisticians, modelers and analysts in quest to explore and exploit the data manually (Fawcett & Provost, 2013). As the volume and variety of data have increased significantly this has outpaced the capacity to perform manual analysis. In addition, computers have increased in computational power, network being omnipresent and algorithms developed for connecting datasets to more various analytic cases, leading to new opportunities (Fawcett & Provost, 2013). This has led to significant use of data analysis, and then followed by the evolution of data science as profession which typically are the main users of analytical tool such as ML.

AI, ML and deep learning. In context of this, terms like AI, ML and deep learning are frequently being applied. They are sometimes used interchangeably despite being distinct in meaning. Artificial intelligence (AI) have within computer science been defined as *"the science and engineering of making intelligent machines"* (McCarthy, 2007). Machine learning (ML) is a subset of AI which specifically have *"the ability to learn without being explicitly programmed"* (Skymind.ai, 2019) thus have the ability to modify itself without human intervention by being exposed to new data. This brings distinct differences, e.g. *"symbolic logic – rules engines, expert systems and knowledge graphs – could all be described as AI, and none of them are machine learning"* (Skymind.ai, 2019). ML algorithms is also often referred to as a model, which also will be the case in this thesis. Further, deep learning is a subgroup of ML which accounts for more specific types of computer algorithms, such as *deep* artificial neural networks and *deep* reinforcement learning (Skymind.ai, 2019) whereas deep refers to a technical definition – which will not be included here. As of this, the technological aspect has to be understood on the level and in context of learning algorithms from input data.

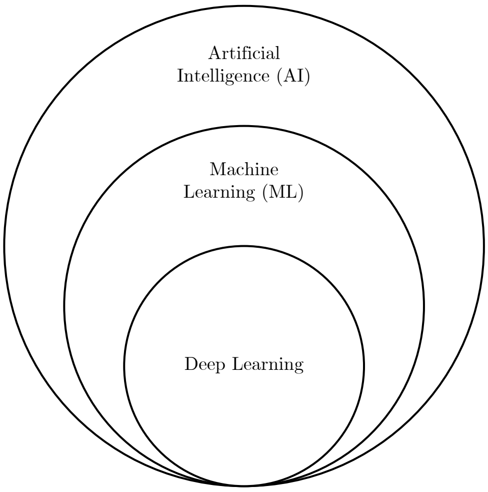


Figure 2.1. **Artificial Intelligence (AI) vs. Machine Learning vs. Deep Learning.**

Scalability. The term scalability is varying depending on context. Generally, scalability can be understood as *"the ability of a business or system to grow larger"* (Cambridge University Press, 2019). When narrowing it down to software, Gage (2018) defines software done at scale as *"program or application works for many people, in many locations, and at a reasonable speed."* Due to the purpose of the thesis and its context, *industrial machine learning scalability* will be defined as:

> *the ability for machine learning to be applied into large number of industrial equipment and devices at acceptable speed with reasonable accuracy.*

There exists a massive range of companies that are using machine learning in their products and services. This ranges across different verticals (Lungariello, 2018), but with the tech companies being at the forefront of application. This especially accounts for the FAANG-group (Facebook, Apple, Amazon, Netflix, Google) which uses it for highly personalized products and services, whereas they all have in common that they are doing analysis of people, and analytics are e.g. being utilized for targeted ads and customer recommendations. Their capabilities of serving their extensive customer base with ML services proves their tremendous scaling capabilities.

Industrial ML. The use of ML have also gained significant impact within the industry which also can be illustrated with the exponential growth in industrial-related publications of ML (Hajizadeh, 2018). Industrial ML can be applied for *"saving time, reducing costs, boosting efficiencies, and improving safety"* (Lungariello, 2018), and are specifically being commonly utilized for optimization of operations and maintenance. Anyhow, industrial companies have not been able to scale their ML solutions broadly unlike non-industrial companies despite extensive efforts in industrial ML and broad range of different industrial analytics vendors. These ranging from big established original equipment manufacturers (OEMs) like Siemens, ABB, GE Digital, and to smaller independent software vendors (ISVs) like C3, Aspentech, Arundo Analytics, Uptake and more.

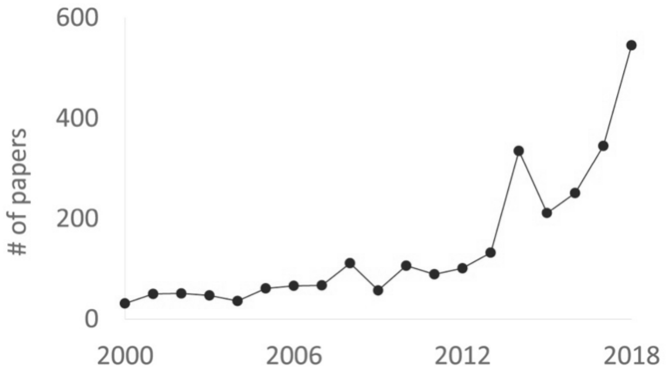


Figure 2.2. **Exponential growth of ML publications in the industry.** (Hajizadeh, 2018)

Technology trends. The compute environments of ML are typically in the cloud (centralized computing) at companies such as Amazon or Microsoft as they have tailored their cloud and datacenter solutions for ML. In the industrial setting, this implies that data have to be transferred from the industrial location to the datacenter location for computation, hence analytic insights must be transferred back to the industrial location. More recently decentralized computing – referred to as edge computing – have gained attention as this makes local computation possible, which in the industrial setting enables on-site analytics with limited latency.

In this chapter essential terminologies and context have been explained and hence works as a fundament for understanding the scope and focus of the thesis.

2.2 Innovation context

In this chapter the basis for innovation theory will be outlined and set the perspective of how scaled ML and its considerations should be viewed. It will further be used as the baseline for a conceptual framework for the analysis of the thesis.

The thesis itself is placed within the context of innovation, as the objective is to find new ways to enhance the application of machine learning (ML) models within the industrial setting thus scaled (industrial) ML potentially should be considered an innovation itself. Despite it exists various definitions on the term *innovation*, Joseph Schumpeter's definition has become recognized and defines innovation as "*new combinations of new or existing knowledge, resources, equipment and so on"* (Schumpeter, 1934). As of today, innovation is considered to be a positive contribution into the society and even "*wedded into the economic ideology*" (Godin, 2015). In contrast, innovation has historically been viewed as wicked and destroying for the order of the community, as it was "*introducing change into the established order*" (Godin, 2015).

Innovation has historically also been claimed to be associated with unexpected discoveries, and that luck and serendipity has often been the answer to these innovations (Trott, 2017). Closer investigations show that luck is rare, but instead that discoveries comes as a result of people having fascination within a specific area of science or technology in combination with extensive efforts leading to new discoveries (Trott, 2017). This substantiates the words by Louis Pasteur: *"chance favors the prepared mind"*.

It is in the latest century that innovation has gained the praise for being specifically valuable. As innovation has been given the positive association, it has also been argued for being the engine of long-term economic growth (Schumpeter, 1934), and even more lately an essential component for organizations survival, or as Christopher Freeman (1982) wrote: *"...not to innovate is to die*".

The innovation process itself are likely to be viewed as fairly abstract, but according to Trott (2017) holds some key elements. These are (i) creative individuals which is affected by scientific and technological developments leading to knowledge inputs, (ii) firms' architecture and external linkage which is affected by the social changes and market needs resulting in new marked needs and opportunities, and (iii) companies operation functions and activities resulting in development of knowledge, processes and products. As the thesis will work within the scope of an innovation process, these key elements and the following innovation theory will be a part of the foundation to understand the thesis.

2.2.1 Types of innovation

Innovations has given various impact but also hold various characteristics, and researchers therefore have classified it into sets of contrasting types (Gopalakrishnan & Damanpour, 1997). The types are usually classified based on the degree of change associated with it (*incremental versus radical*), the activities and areas affected by the innovation (*products versus process*), and innovations related to social structure and technology (*administrative versus technical*) (Gopalakrishnan & Damanpour, 1997). For the purpose of this thesis, we will only cover the first and third contrasting type.

2.2.1.1 Incremental versus radical innovation

The first contrasting innovation is the distinguishment of *incremental* and *radical* innovation, and are related to the magnitude of the innovation. Incremental innovations (also named *evolutionary, continuous*) call for marginal change in existing practices, and are related to improvements of existing methods and products (Gopalakrishnan & Damanpour, 1997). Radical innovations (also named *revolutionary, discontinuous, breakthrough*) (Crossan & Apaydin, 2010) changes the fundamental conditions for an organization or an industry resulting in departing from existing practices, and tends to provide greater improvements than demanded (Trott, 2017). As of the definitions for both incremental and radical, the innovation is related to the technical aspect. Anyhow, incremental and radical innovation are understood as two extremes on a scale of innovation, resulting in innovations to be classified as more or less radical and incremental.

The term *disruptive* innovation is often misunderstood as radical innovation. Disruptive innovation hold the characteristic of creating new markets that captures existing markets (Christensen, 2003). Hence is it not about the technology itself, but related to how the technology is used within the business context. This implicates that despite an innovation to itself being considered incremental, it could also be disruptive market-wise. Disruption theory also predicts that companies who leverage disruptive innovation have relatively higher chance of success against large incumbent companies (Christensen, 2003).

The introduction of new evolutions of iPhone models (iPhone 6, 6S, 7, 8, 8S, etc.) are more likely to be categorized as incremental innovations, as they tend to "only" improve the existing product, but the introduction of the first iPhone in June 2007 – which clearly disrupted the cell phone industry and probably cannibalized Nokia, BlackBerry and Sony Ericsson (Dediu, 2012) – are considered to be both disruptive (market-wise) and (more) radical (technology-wise).

The radical and incremental dimension also highlights various benefits related to the actors in the market. For incumbents incremental innovations are likely to be most beneficial as this allows them to leverage existing knowledge and resources, which incumbents tends to have a competitive advantage of (Trott, 2017). However, they will most likely struggle with radical innovations as they are more constrained by managers' mindset, but also have less incentive to come up with or make use of radical innovations that could cannibalize their existing products. Hence, this also explains why established industry leaders almost exclusively introduces incremental innovations (Christensen & Overdorf, 2000). This gives opportunities to new entrants as they are less constrained, and particularly regarding the limited need for change of knowledge (Trott, 2017).

2.2.1.2 Technical versus administrative innovation

Lastly, Gopalakrishnan & Damanpour (1997), points to a more general level of innovation, which distinguish between the technical and social structural aspects. "*Technical and administrative innovations are, respectively, related to the technical and administrative cores of the organization*" (Gopalakrishnan & Damanpour, 1997). Administrative innovations apply more directly to its management and indirectly to the basic work activity of the organization, such as organizational structure, human resources and administrative processes (Gopalakrishnan & Damanpour, 1997). Technical innovations are directly related to the basic work activity of an organization, which pertains to products, processes and technologies that are applied to produce products and deliver services. Within this segment, it is been argued that radical and disruptive technologies likely will lead to cheaper, more convenient, simpler and smaller products than previous (Christensen & Overdorf, 2000).

2.2.2 Innovation models

Until the 1980s, the innovation model was primarily understood as a linear model of science and innovation (Trott, 2017). As innovation research progressed, the linear model could only be proven valid for a few limited applications which led the field of innovation theory to further develop. As a result to this, *the interactive innovation model* acknowledge innovation as a continuous interactive process as a result of "*marketplace, the science base and the organization's capabilities*" (Trott, 2017), and thereby distinguish from the linear innovation model. Despite innovation models still evolve and are challenged, this model has proved to outline the key elements in innovation process, whilst it still is heavily simplified: (i) the market as a major source of innovation; (ii) firm competences enable firms to match technology with demand; and (iii) external and internal sources of innovations are important (Stefano, Gambardella, & Verona, 2012). The innovation process also can be thought of as a *"complex set of communication paths over which knowledge is transferred"* (Trott, 2017), and this model gives room for such view.

As societies has evolved into a more knowledge-based economy – meaning *"economies which are directly based on the production, distribution and use of knowledge and information"* (OECD, 1996) – Chesbrough et. al. (2006) argues this leads to a "*new mode of open systems involving a range of players distributed up and down in the supply chain"* (Trott, 2017, s. 26), and coined this concept *open innovation.* Cohen and Levinthal (1990) had previously explained two modes of R&D – internal and external – and the importance of internal R&D to utilize external technology, an ability named "*absorptive capacity*". The key difference – and claimed new paradigm – with open innovation though was argued to be that internal and external knowledge being equally important during the innovation process.

2.2.3 Summary of innovation

In the section of *innovation context*, the innovation theory have been outlined and put in the perspective of how scaled ML and its considerations should be viewed. Innovation should be understood in the setting of being a complexed interactive process with several interconnected elements.

The scope of innovation comes with many nuances, both academically and practically. To classify scaled (industrial) ML as an innovation it must according to the definition of Schumpeter fulfil the attributes of "*new combinations of new or existing knowledge, resources, equipment and so on",* which it might seems reasonable to assert. If so, it must also be seen in context of innovation management and processes and their included aspects. Anyhow, this will be a part of the research analysis.

2.3 Theoretical framework

This chapter outlines a conceptual framework, how it has been selected and its content which will be used as the base for the thesis analysis.

This thesis in itself can be understood within the scope of innovation management, and due to the purpose of the research a conceptual framework has been developed. The conceptual framework is based on innovation process (Trott, 2017), context of technical vs administrative innovation (Gopalakrishnan & Damanpour, 1997), and model of interactive (Trott, 2017; Stefano, Gambardella, & Verona, 2012) and open innovation (Chesbrough, Vanhaverbeke, & West, 2006). Specifically how the elements have been selected is explained in the upcoming paragraphs.

Aspects from the innovation process (scientific and technological developments), technical innovations (products, processes and technologies), interactive innovation model (science base), in combination with the purpose of the thesis, are compressed to the first element which is *technology.*

Based on the innovation process (firms' architecture and external linkage), interactive innovation model (marketplace), in combination with the need to compare industrial versus non-industrial cases, led to the second element which is named *business characteristics.*

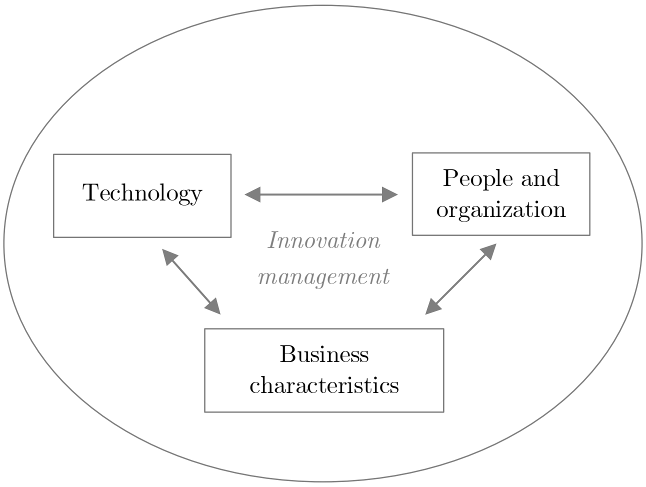


Figure 2.3. **Framework for research evaluation**.

The innovation process (individuals and knowledge inputs and firms' architecture), administrative innovation (organizational structure, human resources and administrative processes), interactive innovation model (organizational capabilities) are condensed into the final and third element which is labeled as *people and organization*.

Thereby the research will focus on industrial ML with respect to (i) technology, (ii) business characteristics and (iii) people organization and hence will be evaluated accordingly.

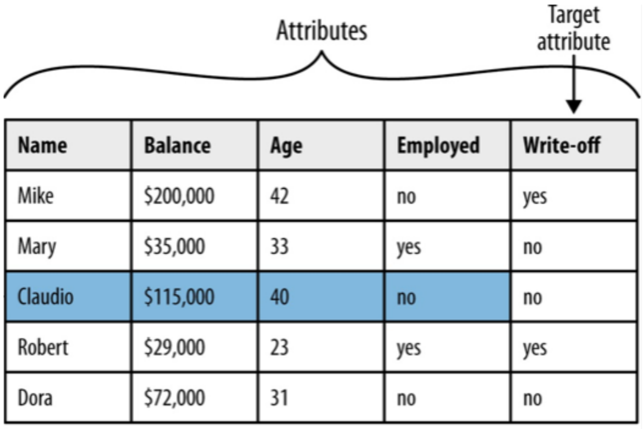
2.3.1 Technology

Technology in this thesis will be concern machine learning and software scalability. This section will explore the technological ...

2.3.1.1 Machine learning (ML)

Machine learning (ML) could generally be defined as the *"design and study of software artifacts that use past experience to make future decisions"* (Hackeling, 2014). The unique property is then that ML have the ability to not be specifically programmed from every case and condition, but can learn from a set of data.

The data is the prerequisite for learning – also referred to as training or modeling – and the data have different attributes/variables in columns, which is referred to as features.


Attributes
Target attribute

Name	Balance	Age	Employed	Write-off
Mike	$200,000	42	no	yes
Mary	$35,000	33	yes	no
Claudio	$115,000	40	no	no
Robert	$29,000	23	yes	yes
Dora	$72,000	31	no	no

Figure 2.4. **Data attributes/variables/features**. (Fawcett & Provost, 2013)

The machine learning process is based on the *Cross-Industry Standard Process for Data Mining* (CRISP-DM), which highlights that processing data is an agile process meaning that *"iteration is considered to be the rule rather than the exemption"* (Fawcett & Provost, 2013). By iterating – even just for the first time – one will also explore the data itself, leaving more insights to the data science team. The iteration results in a few core processes, as illustrated in figure 3.5, (1) business understanding, (2) data understanding, (3) data preparation, (4) modelling, (5) evaluation, and eventually (6) deployment. In the context of ML, the modelling process is where a models is learning – more commonly referred to as training. Despite the ML process being fairly the same in different cases, they still have some variations due to the specific demands and conditions.

After the model is deployed and hence put into production, the model can now take in data for prediction, which are grouped into batch data – which is accumulated data put together into a larger data set – or real-time data which consists of smaller amounts data streaming at a set frequency.

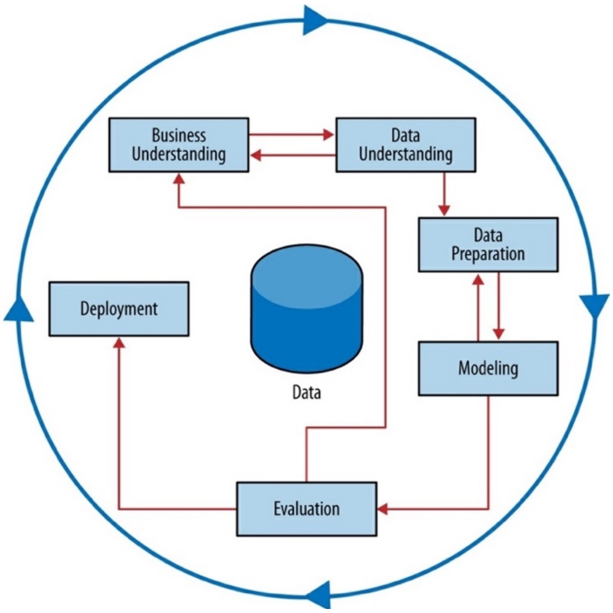


Figure 2.5. **CRISP-DM.** (Fawcett & Provost, 2013)

2.3.1.1.1. Concepts of machine learning algorithms

When it comes to how the ML systems learn they are usually categorized into learning either with or without supervision of humans. As of this, the two main types of ML are commonly called *supervised* and *unsupervised* machine learning, which occupies opposite ends of the scale.

Supervised learning is a program which *"predicts an output for an input by learning from pairs of labeled inputs and outputs; that is, the program learns from examples of the right answers"* (Hackeling, 2014). In other words, it aims to learn based on desired outcomes already known. Supervised learning is most commonly applied for regression or classification analysis.

On the other hand, unsupervised learning has not been given any labels of what is wrong or right, but instead are trying to discover patterns in the data (Hackeling, 2014). Such algorithms are often used when data is unlabeled and for the purpose of grouping data points – more commonly termed clustering – for data who are sharing similar characteristics.

As supervised and unsupervised learning are found on the ends of the spectrum, there is also possibilities of *semi-supervised learning* in the middle. An example here is *reinforced learning* which receives feedback for its decision in an environment with changing states. Reactions are then measured by a previously defined target and thus a reward value is returned to the algorithm serving as a target for optimization (Hackeling, 2014).

2.3.1.1.2. Evaluation metrics

As CRISP-DM shows, data is the sole input of the training and are therefore the prerequisites of the quality of the ML model. In quest of getting the "best" model, it would be defined by *"attributes like how interpretable, simple, accurate, fast and scalable the model is."* (Robinson, 2019). The challenge though being that these attributes typically are contrary to each other, e.g. being accurate often comes at the cost of compute-heavy and slow models. In the data science field it is therefore a big need of optimizing the models for the purpose of the analysis and the business question asked.

Generally speaking, data are usually segmented into two sets of data during the ML process; training data set and testing data set. When testing and evaluating the model, the output metrics is therefore based on the test set.

Confusion matrix. This is a table that is often used to describe the performance of a classification model on a set of test data for which the true values are known. For a classification model with binary decision problem the results will be labeled either positive and negative. As input data also are classified binary, this leaves for four different cases:

- *True positive* (TP) is the number of when the algorithm correctly classifies the output as positive.
- *True negative* (TN) is the number of when the model correctly classifies the output as negative.
- *False positive* (FP) is the number of when the model incorrectly classifies the output as positive, but really is negative.
- *False negative* (FN) is the number of when the algorithm incorrectly classifies the output as negative, but in reality is positive.

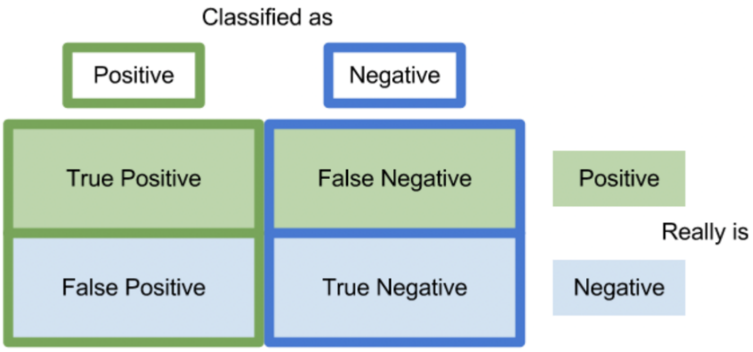


Figure 2.6. **Confusion matrix** (Fawcett & Provost, 2013)**.** In green the positive class, in blue the negative class.

These terms are very useful when computing the performance of the model. There exists several metrics for measuring the performance, but *accuracy* is considered to be one of the simplest and most straight-forward metrics. It is calculated as follows (Fawcett & Provost, 2013):

$$Accuracy = \frac{n\ correct\ predictions}{n\ total\ predictions} = \frac{TP + TN}{TP + TN + FP + FN}$$

Accuracy will return, due to the nature of the formula, a ratio output ranging from 0 to 1. The accuracy range depend on conditions, type of analysis and cases.

In addition, metrics such as recall, precision, and f-measure are common metrics, but are outside the purpose of the thesis.

2.3.1.1.3. Machine learning subgroups and concepts

As previously presented, ML falls into the group of AI, but also have different subgroups itself. As mentioned, one of the are to be deep learning. Below, concepts of ML are described briefly which will be relevant for later.

Anomaly detection. This is the identification of *"rare items, events or observations which raise suspicions by differing significantly from the majority of the data"* (Zimek & Schubert, 2017), hence called anomaly detection, but also outlier detection. Three main techniques are considered to be *unsupervised*, *supervised* and *semi-supervised* learning.

Automated machine learning – also referred to as autoML – is the case of automating processes within the ML processes. E.g. data processing and modelling would be automated, which could increase the effectiveness of data science workload (Hutter, et al., 2019)

Transfer learning. This topic is fairly new in terms of wide appliance, despite being discussed in academia for decades (Pratt & Thrun, 1997). The concept of transfer learning focuses on *"storing knowledge gained while solving one problem and applying it to a different but related problem"* (West, Ventura, & Warnick, 2007). By so with minimal retraining one could apply a very comprehensive model to new related cases. Findings in the case of industrial context show that transfer learning *"outperforms unsupervised anomaly detection in the target domain"* (Vercruyssen, Meert, & Davis, 2017) and hence shows that the approach is promising within the industrial setting.

2.3.1.2 Software scalability

As pointed out in section 2.1, scalability in this thesis is defined as *the ability for machine learning to be applied into large number of industrial equipment and devices at acceptable speed with reasonable accuracy.* According to Fisher & Abbott's book *The Art of Scalability* (2015)*,* there is three main dimensions to scaling, also framed as the Scalability Cube:

(i) Y-axis; scale by splitting different things (functional decomposing),
(ii) X-axis; scale by cloning (horizontal duplication),
(iii) Z-axis; scale by splitting similar things (data partitioning),

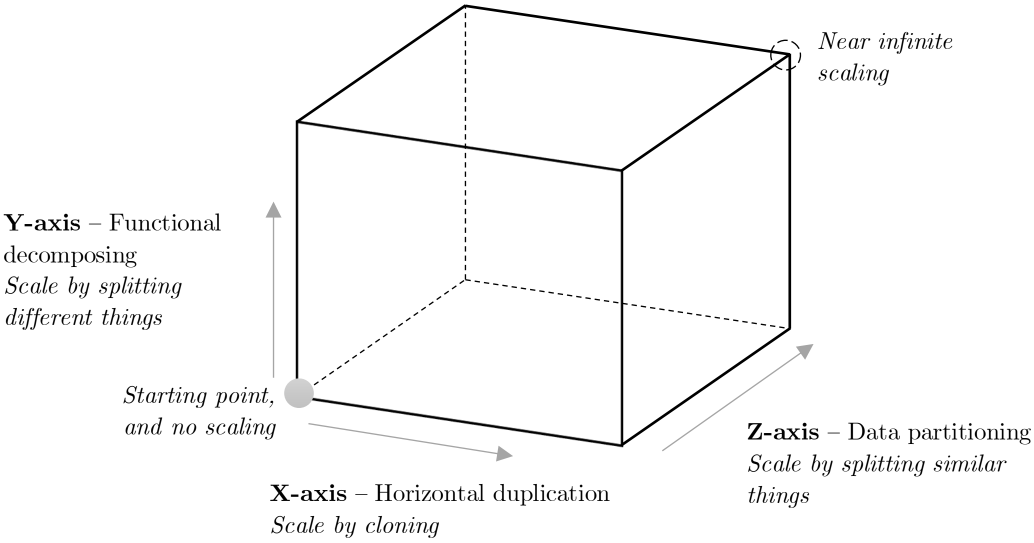

Near infinite scaling
Y-axis – Functional decomposing
Scale by splitting different things
Starting point, and no scaling
X-axis – Horizontal duplication
Scale by cloning
Z-axis – Data partitioning
Scale by splitting similar things

Figure 2.7. **The Scalability Cube.** (Fisher & Abbott, 2015)

(i) Scale by splitting different things (Y-axis). By *different things* meaning methods, functions and services. Such scaling splits a monolithic application into a set of services to increase scalability, and focuses on separating services and data along noun or verb boundaries, and are usually executed by making use of microservices. E.g. in commerce could be splitting browse, checkout, login. For each service, it should its own non-shared data to ensure fault isolation and high availability.

(ii) Scale by cloning (X-axis). This meaning that each server runs multiple identical copies of the service (if split) behind a load balancer. Hence, the X-axis will range from multiple copies to one monolithic system.

(iii) Scale by splitting similar things (Z-axis). In this respect, it's similar to X-axis scaling. The big difference is that each server is responsible for only a subset of the data. Similar things could thus apply to segment *customers* based on geography (ex. North America and Europe). Such split would also benefit in cases of local privacy laws, such as GDPR in Europe.

Despite this theory being not directly related to the more narrow field of ML the purpose is to help use this as an baseline for understanding scalability of ML within the industrial context. Hence, this makes the problem approach more tangible and will work as a guide for possible new frameworks.

2.3.2 Business characteristics

The characteristics of the businesses are important to evaluate as they (potentially) are determining the prerequisites of company's possibilities and boundaries, and so also into the context of innovation.

2.3.2.1 Business models

The term *business model* have been used inconsistently among authors ranging from *"core repeated processes, a mediating construct between technology innovation and economic value, or a set of building blocks"* (Weiblen, 2014). Anyhow, the common denominator in business model research is that it describes the logic of *value creation* and *value capturing* of a firm (Weiblen, 2014; Chesbrough, 2010).

As the digital disruption has emerged, Weill and Woerner (2015) proposed a framework for business models in the digital era. Their research findings were that enterprises who understood their end costumer well and the majority of their revenues from digital ecosystems had significant higher revenue growth and profit margins than their industry average, with 32% and 27% accordingly (Weill & Woerner, Thriving in an Increasingly Digital Ecosystem, 2015).

Business design

Knowledge of end customer

	Value chain	***Ecosystem***
Complete	**Omnichannel business** • "Owns" customer relationship • Multiproduct, multichannel customer experience to meet life events • Integrated value chain *Examples: banks, retailers*	**Ecosystem driver** • Provides a branded platform • Ensures great customer experience • Plug-and-play third-party products • Customer knowledge from all data • Matches customer needs with providers • Extracts "rents" *Example: Amazon*
Partial	**Supplier** • Sells through another company • Potential for loss of power • Skills: low-cost producer, incremental innovation *Examples: insurance via agent, mutual fund via broker*	**Modular producer** • Plug-and-play product/service • Able to adapt to any ecosystem • Constant innovation of product/service *Example: PayPal*

Table 1. **Business models in the digital era** (Weill & Woerner, Thriving in an Increasingly Digital Ecosystem, 2015)

The proposed framework are classified into four groups: (i) *omnichannel business*, (ii) *ecosystem driver*, (iii) *supplier* and (iv) *modular producer*. Each of these axis can be understood as scales, where companies can share characteristics form the different quadrants, but generally the companies tends to lean into one of the groups.

Supplier model. Typically companies operating in the value chain of a bigger company, and therefore have limited knowledge about their end customer.

Ecosystem driver. This holds a specific key characteristics being the ability to gain customer knowledge from all data, but also working as a platform connecting customers with products and providers.

Omnichannel model. Such business model gives access to their products across various channels and has an integrated value chain. Omnichannel companies knows their end-customers and profits on this knowledge.

Modular producer. Companies having a such model have products and services that can adapt to a wide variety of platforms and needs to be best in their category to survive. Most modular producers don't get to see all the customer data, but are *"limited to the data from the transactions they process"* (Weill & Woerner, Thriving in an Increasingly Digital Ecosystem, 2015). This also limits to further gain knowledge of the end customers.

2.3.2.2 General characteristics

For classification purposes, a framework for enabling analysis of businesses has to be established. These properties are determined based on general factors describing companies.

Characteristic	Description	Elements / example

Company type	Related to what core business activity, and indirectly to their offering.	(i) Original Equipment Manufacturer (OEM);
		(ii) Independent Software Vendor (ISV);

Commercial Transactions	Refers to if the customer is the end consumer or a business.	Business to business (B2B); Business to consumer (B2C)
Industrial analysis or not	This aspect will be essential due to classification of the companies.	Industrial; Non-industrial
[...]	[...]	[...]

Financials are a typical high level metric for company analysis. Due to the nature of the research problem this does not fit in, so it will be left out of the scope.

2.3.3 People and organization

As previously mentioned, people and organizational aspects has been highlighted to be a central factor within the innovative process (Trott, 2017).

Regarding people, it is being argued that the innovative process is impacted by *"creative individuals which is [...] leading to knowledge inputs"* (Trott, 2017). Further, people's ability to obtain and increase their knowledge will be considered as a valuable element of analysis, which also will be related to the theory of absorptive capacity.

In context of digital organizations, Snow et. al. (2017) emphasizes, to catch up with the increase competition and surroundings, that digital organizations needs to design for actor-oriented principles in aim for more engaged and productive members. A such organization consist of three main elements: (i) actors, (ii) commons and (iii) protocols, processes and infrastructures. These actors – being either individuals, teams of firms – *"must possess the capabilities and values to self-organize, and engage in self-management rather than hierarchical directions"* (Snow, et al., 2017). These actors also understand the overall structure and processes of the organization, and their decisions are taken in alignment with the companies goals and good. In addition, commons, infrastructures and protocols are applied to guide and facilitate actor behavior, connecting organization members with one another and supporting their activities (Snow, et al., 2017). As of this key outcomes in such organizations is organizational autonomy.

In relation Fisher and Abbott (2015) refers to agile organization which also hold the characteristics of self-organizing and autonomous teams. Despite solving technological problems, the scalability issues is based on people developing the technology, and thereby people are important to understand and process elegantly. Hence, a *"scalable solution requires the alignment of architecture, organization and process"* (Fisher & Abbott, 2015, s. 61).

Specifically, agile organizations also brings various team models depending on the activities and demand (McKinsey & Company, 2019). Self-managed teams are suited when the demand are relatively predictable, but whereas a specialist team is needed in case of variable demand. In addition end-to-end cross-functional staff are suited for facing more creative and customer related activities.

2.3.4 Summary of theoretical framework

In the chapter of theoretical framework it has been outlined a conceptual framework, how it has been selected and its content which further will be used as the base for the thesis analysis.

The technological factors will take input from the machine learning and software scalability theory. Business characteristics will consider the different properties for the selected businesses evaluated. People and organization will take into account how companies (wanting to scale) have structured themselves internally in combination with people and how they meet other actors.

2.4 Summary of theoretical background

In this section relevant theoretical framework have been presented to give a fundament and solid understanding of the context of the research problem.

The first chapter focused on context of relevant technologies and terminologies, with highlighting the relevance and nuances of ML and the context of current applications. The second chapter focused on the concepts of innovation management and its context, working as a base for the conceptual framework. It also asserted scaled ML as to be considered an innovation itself which requires validation from the analysis. The third chapter outlined the conceptual framework and its content which will work as the baseline for the thesis analysis. The technological factors will take input from the machine learning and software scalability theory. Business characteristics will consider the different properties for the selected businesses evaluated. People and organization will take into account how companies (wanting to scale) have structured themselves internally in combination with people and how they meet other actors.

3 Methodology

In this section the chosen research method and design will be described, and it will lay out reasons for choices regarding cases. In addition, validity and reliability that will be discussed.

3.1 Research design and method

In this chapter the choices of research design and methodology will be explained hence bring clarity to the approach of this research.

A research design is a tentative disposition that explains how the research will be executed (Ringdal, 2001). Complementary, Yin (2014, s. 29) defines research design as *"the logical sequence that connects the empirical data to a study's initial research questions and, ultimately, to its conclusions"*.

In research design theory, methods are segmented into two main groups: (1) Quantitative strategies which seeks to understand coherences of variables (Ringdal, 2001) which usually depend on a big amount of data such that statistical analysis is able to be performed. It is therefore argued that quantitative data – in its "purest" form – is metric data (Yin, 2014). The other group is (2) qualitative methods which seeks to understand the depth of a case or a specific problem and are therefore performed with close observations to a few objects (Ringdal, 2001). Thereby, qualitative data often expressed in the format of text.

Despite these methods contrast – both in execution and results format – they could also be made use of to complement and emphasize each other's findings, and such approach is more commonly named *mixed methods*. It is argued to address broader and potentially increasing the reliability of such, but are also considered to be more costly in execution and complexed in analysis (Yin, 2014). Despite the timespan and resources available in this research project, mixed methods have been attempted to achieve in quest of quantification of secondary data and qualitative interviews.

Due to the formulation of the research questions, the intention is describe the current situation, but in extension also investigate possible solutions. This leads to choosing a research design as both explanatory and exploratory which is considered to be best suited for such research questions. Explanatory design – also referred to as descriptive – are preferable when the field and/or object of research is highly complex and there is lack of existing theory within the this field, usually due to a recently new phenomenon (Yin, 2014). The objective of a such approach will be to describe and explain the phenomenon. On the other hand, explorative design seeks to start off from observations and explore the nuances of object of research. A such design is based on the inductive approach which starts with interpreting and analyzing data and observations which results in generating theoretical framework, whilst the deductive approach starts with a hypothesis which it will try to either approve or disprove based on the findings (Yin, 2014).

The research design chosen was *case study*, due to fulfilling the characteristics where such is preferable. These characteristics are when (1) *"how"* or *"why"* research questions is being posed, (2) the research have limited control over the events, and lastly (3) the focus is on a contemporary phenomenon (Yin, 2014). In addition, the objective of the case study is to look in *"depth at one, or a small number of, organizations, events or individuals"* (Easterby-Smith, Thorpe, & Jackson, 2015). As of this, case study are considered to be the preferred choice of research design. Specifically, case study research holds five components that are considered to be key (2014, s. 29):

1. a case study's questions
2. its propositions, if any
3. its unit(s) of analysis
4. the logic linking the data to the propositions
5. the criteria for interpreting the findings.

The case study questions is based on the field *scalability of machine learning models within the heavy industry*, and the research methods are determined by *what* and *how* the questions are asked. Further, propositions are not significantly relevant in particular cases of explorative designs as the exploration is the field of importance, and not predefined propositions. Instead it should be stated a purpose such that the research has a clear direction. Unit(s) of analysis are different cases of ML scalability, specifically being companies, hence the research is designed as a multi case study. Data are empirically analyzed with respect to the theoretical framework.

The thesis has been performed qualitatively as explorative and exploratory research with an inductive approach, meaning it will seek to understand a case resulting in forming theory.

The research progress first started with massive collecting of information from documents, scientific papers and interviews regarding the topic of machine learning scalability. In second phase, the various cases of scalability have been examined and analyzed resulting in a set of identified characteristics. In third phase, based on the characteristics the various factors for successful scaling have been identified. Fourth phase, characteristics which are similar and distinct between the non-industrial and industrial context have been classified. Hence, in the fifth phase, the largest barriers to scaling of industrial ML have been examined. Finally, in sixth phase, possible solutions and framework have been proposed to overcome the barriers.

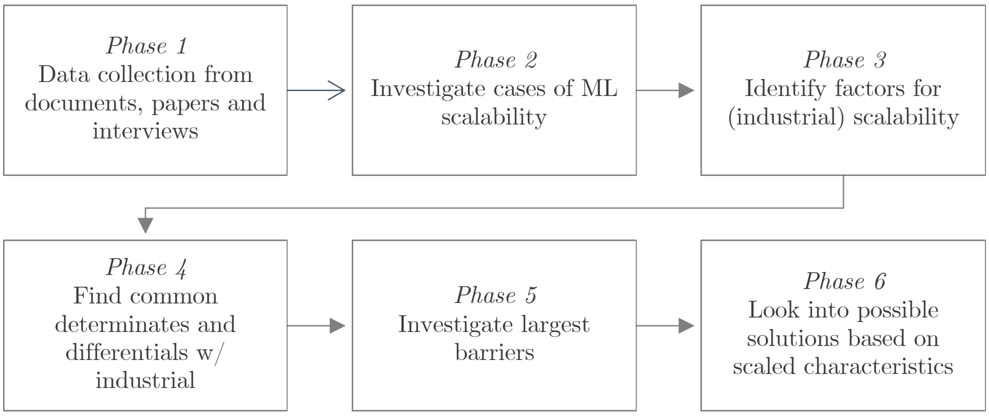

Phase 1
Data collection from documents, papers and interviews
Phase 2
Investigate cases of ML scalability
Phase 3
Identify factors for (industrial) scalability
Phase 4
Find common determinates and differentials w/ industrial
Phase 5
Investigate largest barriers
Phase 6
Look into possible solutions based on scaled characteristics

Figure 3.1. **Research design and method: Progress scheme**

In this chapter the choices of research design and methodology have been explained in aim to bring clarity to the approach of this research.

3.2 Case selection

In this chapter it be laid out the criteria for the choices of selected cases, followed by a brief introduction of each case.

In this thesis, theoretical sampling has been applied (Gobo, 2008), where cases have been selected based on perceived insights they provide for the research topic and their relevance for theory development.

The objective with the case selection has been to find valid cases which have been able to scale ML successfully, and on the other hand industrial cases which should have the base – prerequisites of resources, capabilities and technology – for being able to scale ML models within the industry. Hence, the cases have been primarily based on the following criteria:

- Non-industrial software companies with successful ML scaling initiatives
- Industrial software companies with ML scaling initiatives

Selection of non-industrial companies. There exists a massive range of companies that are using machine learning in their products and services. This ranges across different verticals and horizontals, but with the tech companies being at the forefront of application. For the purpose of this thesis, the selection have also been weighted based on companies with various objects of analytics and officially published material regarding their ML initiatives. Regarding published data this seems to be more typical among the largest tech companies like the FAANG-group (Facebook, Amazon, Apple, Netflix, Google).

For various objects of analysis, the selection of tech companies are typically very similar related to analysis of people. Anyhow there is some exceptions. Uber does analytics of people and driving patterns – and Tesla Inc – which does analytics for enabling self-driving cars. Uber has been publishing detailed on their blog regarding how they have been developing their software, including their ML challenges and how they have solved those. Tesla have been more reticent with sharing their ML progress, but did recently host an event where they described in-depth their ML progress and current capabilities. As of this, both Uber and Tesla have been selected as cases.

Selection of industrial companies. Within the industrial selection of companies there a broad range of different industrial analytics vendors. These ranging from big established original equipment manufacturers (OEMs) like Siemens, ABB, GE Digital, and to smaller independent software vendors (ISVs) like C3, Aspentech, Arundo Analytics, Uptake and more. The selection for industrial companies have been weighted based on their prerequisites (resources, capabilities, technology), but also market position and available data material.

Hence, GE Digital there exists good amount of data due to their long-term market position as they originally are an OEM. For C3, which hold other business characteristics (ISV) but are also working on the same challenge. For Arundo it has been possible to do interviews of respondents who have unique knowledge in the field of this research, but also have market experience from other companies such as C3 and GE Digital. Hence Arundo was also chosen as they could provide information of the other cases.

The selection thereby consist of:

- Facebook Inc. (non-industrial)
- Uber Inc. (non-industrial)
- Tesla Inc. (non-industrial)
- GE Digital (industrial)
- Arundo Analytics (industrial)
- C3 (industrial)

The following selection could consequently be put on two axis, (1) successful scaling of machine learning models and (2) industrial vs. non-industrial application.

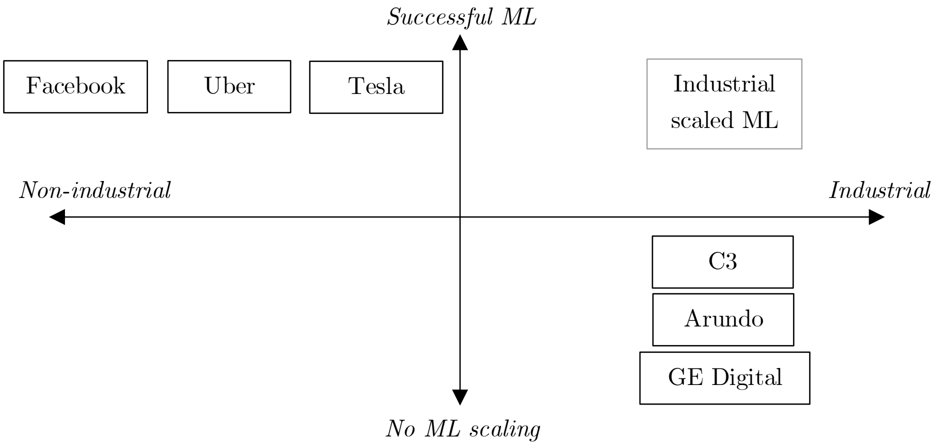


Figure 3.2. **Axis: Successfully scaled ML versus industrial and non-industrial.**

3.2.1 Facebook Inc.

Facebook is an online social media and social networking service company based in California, US, founded in 2004. It's being one of the largest companies in the world by market capitalization (Picardo, 2019), and with their social network platform serving 1.5 billion active users daily (Statista, 2019).

It is designed as a multi-sided platform – meaning it offers its services to two or more groups of users with different value propositions – where its user groups are people wanting to connect with other people and organizations wanting to advertise. With their goal of deliver relevant and interesting content to all of their users, they cannot be dependent on manual selection of content by people, but instead need to use advanced algorithms which can deliver this automatically in the moment. Here, machine learning models finds its purpose and are used to delivering significant value to the products and hence the users (Hazelwood, et al., 2018).

At the core of this is data. Data are collected from people's digital devices (smart phones, computers, etc.), giving Facebook the ability to gain massive amounts of data at first-hand. Being able to serve 1.5 billion users daily, it's fairly obvious to claim that Facebook has indeed been able to scale machine learning. Facebook also have their own research department publishing articles frequently inclusive the topic of machine learning, giving the opportunity to get insights from the source itself. As of this, the thesis will use Facebook as an example of scalability being doable, and will investigate Facebook's approach to scaling and extract relevant information.

3.2.2 Uber

Uber is a transportation network company based in California, US, founded in 2009. Their core value proposition started with peer-to-peer ridesharing, which they distribute through their mobile app, and as of 2018 Uber counted 95 million rides (Statista, 2018), 15 million trips a day (Business of Apps, 2019) and 3 million drivers (Uber, 2019). Since startup they have developed a large range of different products and services, e.g. Uber Eats for delivery of food.

Uber has applied ML to a broad range of its services to ensure accurate information to its users, and with their extensive customer base they need to depend on non-human calculations and decision to deliver a proper service. Uber has also published several articles on how they are using machine learning in their platform. As of published articles and available information, Uber are interesting to investigate as they hold several characteristics related to this thesis.

3.2.3 Tesla (Autopilot)

Tesla is an automotive and energy company based in California, US, founded in 2003. It has gained massive attention for its innovative products and services – including their luxury high-tech cars – and is considered to be the 4th most innovative company in the world (Forbes, 2018).

As a part of their portfolio they released Autopilot in their cars in October 2014 which is an advanced driver-assistance system, as of today has features like lane centering, adaptive cruise control, self-parking and more. After the Enhanced Autopilot release in October 2016, they also claim all their new cars to be able of full autonomy driving on long-term due to the hardware upgrade (named *hardware version 2*) but conditioned by legal approval and a well-trained system (The Verge, 2016; Fortune, 2015). The CEO, Elon Musk, claimed in 2015, that: "*The whole Tesla fleet operates as a network. When one car learns something, they all learn it*" (Fortune, 2015). As Tesla makes such claims, this follows that they have been able to scale machine learning into their fleet of cars. And as of the quote "*When one car learns something, they all learn it*", this is also referred to as *fleet learning.* This can more formally be understood as *network effect based machine learning* (Strobl, 2017), and has been followed up later with proofs of the Tesla cars improving their driving behavior as time pass (Electrek, 2018).

As Tesla has been able to show proof of fleet learning and thereby implicate scalability of ML, in addition to handle data of physical observations, the Tesla Autopilot results in being a highly relevant case to study and research.

3.2.4 GE Digital

GE – formally named General Electric – is an American multinational conglomerate with headquarters in Boston, incorporated in New York in 1892 (Wikipedia, 2019). As of today, GE's segments ranges from aviation, health care, power, oil and gas, finance, manufacturing, and more. In 2015, GE founded their digital subsidiary GE Digital with focus on software and analytics for the industry, more specifically industrial internet of things (IIoT). This in continuation of their release of the IIoT software Predix Platform in 2013. The precursor, GE Software, was founded in 2011.

GE have had high ambitions for their industrial software (The Street, 2015), and are still making interesting claims regarding *pre-built industrial analytics* and *self-learning analytics.* Given the context and their offerings, it's interesting to include GE Digital – and more specifically Predix – into the scope of the cases.

3.2.5 Arundo Analytics Inc.

Arundo Analytics Inc. – further only referred to as Arundo – is an international software company providing advanced industrial analytics the heavy industry. Founded in 2015 with offices in Houston, TX, San Francisco, CA, and Oslo, Norway. They currently serve various industrial cases with advanced analytics for large corporations.

They are currently serving analytical solutions on in several environments, e.g. DNV GL's and ABB's digital platform. With their insights and experience in the sphere of the industrial analytics case, in combination with broad knowledge to the industry, makes them highly relevant to investigate during this research. Hence they've been chosen as one of the cases of analysis.

3.2.6 C3

C3 is a software company providing an digital platform for advanced analytics and software applications serving various industries. Founded in 2009 and based in San Francisco, CA.

C3 claim to have been able to scale ML in the industry and have gained significant impact in the market with their AI suite consisting of analytical tools and their specific software applications, e.g. Predictive Maintenance and Sensor Health. They also hold characteristics of being an independent software company working on industrial cases. These considerations makes them interesting to investigate in this research.

3.2.7 Case selection summary

In table 2 the general characters of the selected cases are classified and will be the baseline for the analysis, but business characteristics in particular.

Characteristics	**Facebook**	**Uber**	**Tesla**	**GE Digital**	**Arundo**	**C3**
B2B vs. B2C	Both	B2C	B2C	B2B	B2B	B2B
Industrial	No	No	No	Yes	Yes	Yes
Company type	ISV	ISV	ISV/OEM	OEM	ISV	ISV
Core business	Social network w/ ads	Ridesharing	Automaker	Digitizing industrial	Industrial analytics	Industrial analytics
Proven successful scale of ML	Yes	Yes	Yes	Not yet	Not yet	Not yet

Table 2. **General characteristics of selected cases**.

In the chapter for *case selection* it has been laid out the criteria for the choices of selected cases, followed by a brief introduction of each case, in addition to a presentation of the general characteristics of the various cases.

3.3 Data collection

In this chapter the approach of how data have been retrieved will be explained, specifically regarding document collection and qualitative interviews.

3.3.1 Types of data

As previously mentioned, there is a distinction between quantitative data and qualitative data and can be understood as one dimension of data. Quantitative is at its best expressed as numbers, whereas qualitative data typically are expressed as text.

Another dimension to data is primary and secondary data. *Primary data* are where a researcher collects and structures the data and have the ability to bring very in-depth and relevant data which could ensure high validity of the data, but have the disadvantages of being less reliable as the researcher could impact – both consciously and unconsciously – the output of data basis. As for *secondary data* is when a researcher uses data collected by others and when finding secondary data the data itself will can only be limited affected by the researcher, hence will obtain its reliability. Anyhow, it is not certain the data are directly relevant for the unit of analysis, consequently the validity has to be ensured by carefully picking the right secondary data but also make proper searches which secures skipping relevant material.

As for this thesis, primary and secondary data have been coherent to understand the characteristics of scaled ML, but with slightly different purposes. Secondary data have been used as the base for the broader understanding of scaled ML and consequently sketching the research landscape, and in some cases also been used for in-depth investigation in cases where it has been possible. The qualitative interviews have been applied for in-depth investigations of the industrial aspects but also of specific considerations during the research project. For secondary data, document collection approach have been utilized, whereas for primary data the qualitative interviews have been performed.

3.3.2 Secondary data: Document collection

Every case study should contain this type of data collection as it is stable, specific and serves a broad scope (Yin, 2014). Anyhow it has pitfalls when it comes to the selection of documents, which could lead to research bias. To ensure maximum validity official company published material have been preferred, and substantiated and supported by commentary material. The types of documents being collected includes academic research papers, blog posts (both official and commentary), news articles, videos (official) and internal company documents.

Facebook. Official research material is considered to be fairly extensive as Facebook are established their own R&D department – named Facebook Research – where they have been publishing academic papers frequently plus more educational ML videos, which suited the research theme of the thesis. Especially, two papers where particularly important which were the "*Applied Machine Learning at Facebook: A Datacenter Infrastructure Perspective*" (Hazelwood, et al., 2018) and *"Machine Learning at Facebook: Understanding Inference at the Edge"* (Hazelwood, et al., 2019).

Uber. At Uber they have an official blog – called Uber Engineering – specified for the technological and engineering discussion of the services at Uber containing detailed articles. This made it relatively easy to find valid content. Particularly the article *"Scaling Machine Learning at Uber with Michelangelo"* (Hermann & Balso, 2018) have been significantly important as it contained an extensive insight into how Uber are working and thinking regarding scaling ML. In addition, commentary material with external viewpoints have been used.

Tesla. Previously, Tesla have been fairly reticent regarding publishing technical material of their ML solutions. As of this there has been many news articles commenting Tesla's progress and development working as an indication of their development. Anyhow, in May 2019 they held the *Tesla Autonomy Day* (Tesla Inc., 2019) event which they streamed and published online. In this session they gave a detailed presentation of how they are building their Autopilot service, including how they are thinking in regards of scaling ML.

GE Digital, Arundo and C3. As for the industrial cases available secondary material were limited. This led to obtaining marketing material (whitepapers and product offerings), press releases and some blog posts as an indication. As marketing material tends to bend the truth they have been given the least impact. As for Arundo it has been given access to a sample of internal company documents. Anyhow, for these cases primary data have been considered the most valid data.

3.3.3 Primary data: Qualitative interviews

Interviews is being considered as one of the most important sources to evidence for case studies (Yin, 2014). This due to its characteristics of being focused and targeted on the research objective, and therefore have the property of being highly insightful. However, it also includes the risk of biases, such as response bias where the informant withholds valuable information or poorly formulated interview questions which effects the answers in return (Yin, 2014). Interviews are segmented into (1) unstructured, (2) semi structured and (3) depth interviews. When performing interviews it being specifically pointed out *"the ability to pose and ask good questions is [...] a prerequisite for case study investigators"* (Yin, 2014), hence making sure the questions are right chosen, formulated and presented impacts the results and have been considered cautiously throughout the interviews. The interviews in this case study have been performed semi structured, due to the explorative research approach and the need for flexibility to adapt ad hoc during the interviews. Consequently, the interview guide have been developed accordingly and can be found in the appendix.

Interviews have primarily been performed by approaching employees internally from Arundo, as they offered themselves to be available for interviews during the whole research period, in addition to having extensive industry experience and insights – including GE Digital and C3 – which emphasizes why they have been selected as a case. Consequently, the interviews have been used as the main source for the industrial cases.

During the interviews it has been preferred to use an audio recorder ensuring getting the details – prior consent by the respondents – alongside with notes highlighting key elements in the interviews. Postprocessing of the interviews have been decoded into key elements and further been integrated into the table of characteristics and qualitative (text) analysis.

3.3.3.1 Respondents

As the respondents are anonymized they have been assigned with an interview ID. Their expert domain and experience is described below including the date of when the interview was performed.

Int. ID	Domain	Experience	Date
1	Business	Management Consulting, Technology companies, (C3)	01.28.19; 02.14.19; 03.14.19; 05.10.19
2	Data Science	Academia	02.04.19
3	Data Science	Oil & Gas	02.05.19
4	Data Science	Academia	02.05.19
5	Business	Academia	02.20.19
6	Data Science	Academia	02.22.19
7	Business	Oil & Gas, (GE Digital, C3)	03.04.19
8	Business	Oil & Gas, Maritime	03.14.19
9	Business	ISVs, (GE Digital)	04.02.19
10	Software	Academia	04.02.19
11	Software	SW Consulting	04.12.19
12	Business/Software	Maritime	05.03.19
13	Software	Utilities (Hydropower)	05.08.19

Table 3. **Respondents.**

In the chapter of *data collection* the approach of how data have been retrieved have been explained, specifically regarding document collection and qualitative interviews.

3.4 Reliability and validity

In this chapter the considerations regarding reliability and validity will be explained, highlighting the awareness of this aspect.

When performing research, one must ensure to reach for as high validity and reliability as possible. This is always related to the methodology and is therefore important to be considered thereby.

Reliability. This is, for case studies, to "*demonstrating that the operations of a study - such as the data collection procedures can be repeated, with the same results"* (Yin, 2014). Consequently, secondary data are typically more easily available hence retrievable. Therefore the data considered in the non-industrial cases have higher reliability than the data retrieved in the industrial cases based on the previously mentioned considerations. To ensure the highest degree of reliability, the respondents' answers have been analyzed to both other respondents' answers and to the secondary data, but also – vice versa – the secondary data have been compared to the respondents' answers, which should help increasing the external reliability of this thesis.

This is also connected to the philosophic sphere of epistemology which is the study of theory of knowledge, also formulated as the study of *"how we know what we know"* (Easterby-Smith, Thorpe, & Jackson, 2015). This can be understood as a scale of two contrasting types; positivism – which seeks to ensure truth by being totally independent of the unit of analysis and hence only focused on quantitative data and by so seeks to explain a phenomenon – and social constructivism – which seeks to ensure truth by being a part of what is being observed collecting rich qualitative data which can be applied for understanding the unit of analysis (Easterby-Smith, Thorpe, & Jackson, 2015). As for this research study, it has been vital to ensure proper understanding of scaling of industrial ML but also the ML scalability itself. Further one could argue the non-industrial cases they have been performed more positivistic than the industrial cases. The methodology of characterizing the cases in tables have been an attempt to quantify the study as this should amplify possible causal explanations – hence towards more positivistic approach. Anyhow, taking a more social constructivism approach for the industrial cases have been vital in performing the thesis with sufficient validity.

Validity. This is the consideration of whereas a study researched what it indented to do, hence how relevant the study is (Easterby-Smith, Thorpe, & Jackson, 2015). As the research is primarily exploratory this brings internal validity not be considered as relevant (Yin, 2014). For external validity this is considered in the context whether or not the findings can be generalized. Generalizations in science are usually based on a *"multiple set of experience that have replicated the same phenomenon under different conditions*" (Yin, 2013:21). As this thesis is exploratory in nature executed with a smaller selection of cases the findings in this thesis should accordingly be understood as one experience of many which should be challenged and tested for its validity. It is therefore argued that this should be seen in context of other similar studies before claiming that these findings are generalizable.

3.5 Ethics and privacy

In this chapter it will be highlighted general considerations impacting the ethics and privacy in this thesis.

As there will be collected various types of data, this also leads to different treatment accordingly. For secondary data collected there will be limited to none concerns regarding privacy, as this is already taken care of by those who did the collection. Anyhow, there will still be needing to treat the secondary data with respect in terms of treating them as others property, and during the analysis point out what results are dependent to others work.

Regarding primary data it is of big importance to take the privacy of contributors seriously, both for ethical reasons but also especially due to the new EU regulation termed General Data Protection Regulation (GDPR). This means that any data that can identify or be related to an identifiable living individual is considered to be personal data (European Commission, 2016), but also gives the ownership of the data to the contributor and not the collector which is a big turnover. Personal data therefore has to be handled with great care, and during the research period therefore has to be anonymized and not traceable, both directly and indirectly. This will be highly relevant when the research will enter its observation phase, but also before and after observations when in contact with contributors, e.g. when asking people to join as contributors or sending them transcripts of conversations. To collect personal sensitive data, it has been needed to report to the Norwegian Centre of Research Data (NSD) on beforehand with reference-ID 186502.

Another ethical aspect of observation is in the case of where respondents say that some facts are “off the record” or in case where I as the observer catches something that is sensitive and not intended for me, this will require ethical handling as its not data approved for recording. As an observer I ought to leave this out of any reflections, though this potentially could to some extent effect the viewpoints.

4 Analysis

The analysis section is based on the outlined research questions thus will first focus in chapter 4.1 on exploring and understanding the phenomenon of scaled ML, followed in chapter 4.2 by an analysis of the barriers based on the previous findings, and finally in chapter 4.3 present possible solutions to these barriers.

4.1 What is characterizing scaled machine learning in the selected cases and how does it contribute to business value and innovation?

This chapter will lay out the different cases and their associated characteristics. During the analysis the characteristics of the successfully scaled cases will be outlined, and hence be compared to the other cases. This should consequently reveal the value of scaled ML.

4.1.1 Non-industrial case 1: Facebook

4.1.1.1 Technology

Most of Facebook's services and products are leveraging ML (Hazelwood, et al., 2018). This includes services and products like the *News Feed*, *Facer* (face detection and recognition), *Lumos* (extracts high-level attributes and embeddings from an image and its content, enabling algorithms to automatically understand it), *Search* function, *Language Translation*, *Sigma* (general classification and anomaly detection framework), and *Speech Recognition.* How these ML models are viewed by Facebook and for what services they are used, see Appendix 7.2.

To enable machine learning at scale Facebook have developed an internal infrastructure to handle the massive demand for data analytics. This infrastructure "*includes 'ML-as-a-Service' flows, open-source machine learning* *frameworks and distributed training systems"*, and in addition, this includes running ten datacenter locations as of 2018 (Hazelwood, et al., 2018).

Facebook's *ML-as-a-Service* (MLaaS) flow has distinct similarities with CRISP-DM model, but what is referred to as deployment is in Facebook's case running the ML models in production. Despite training typically being executed offline – disconnected from real-time data – there exists some cases, *"particularly for recommendation systems, additional training is also performed online in a continuous manner"*. This indicates that Facebook have developed a system for online training of ML models.

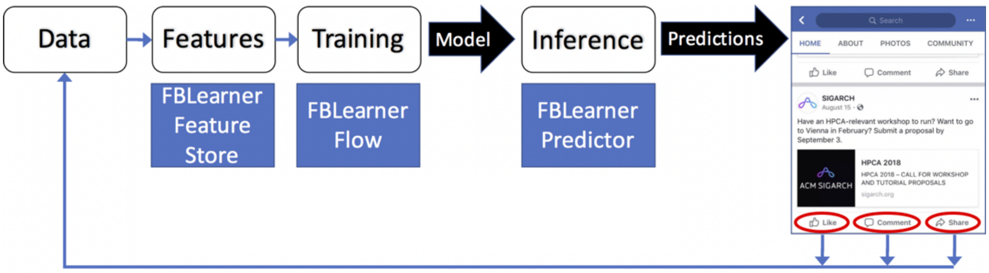


Figure 4.1. **Facebook's Machine Learning Flow and Infrastructure**. (Hazelwood, et al., 2018)

In the MLaaS of Facebook, they have clearly done functional decomposing – aka. *scaling by splitting different things* (Fisher & Abbott, 2015) – by making use of microservices that increases the efficiency and scale of the ML.

The ML flow starts off with the *FBLearner Feature Store*, which essentially is *"a catalog of several feature generator"* and functions as a marketplace where teams can share and discover data (Hazelwood, et al., 2018). Obvious implication of this is that helps the teams with data availability and accessibility to get the relevant data.

Next stage is the *FBLearner Flow* is the platform for ML training and so-called pipeline management system executing a *"workflow describing the steps to train and/or evaluate a model and the resources required to do so"* (Hazelwood, et al., 2018) (see Appendix for screenshots). Mentioned in mobile context only, Facebook has also developed *FBLearner AutoML* designed for *"optimal configurations for experiments"* (Hazelwood, et al., 2019).

After modelling is completed, the models are run in *FBLearner Predictor* which is the production engine. Predictor can be used in two ways; (i) as a multi-tenancy service, or (ii) as a library that can be integrated in product-specific backend services (Hazelwood, et al., 2018).

When it comes to feed ML models with data, Facebook have several techniques to do this efficiently, such as "*decoupling of data feed and training, data/compute co-location, and networking optimizations*" (Hazelwood, et al., 2018). The amount of data being leveraged at a training task for Ads and Feed Ranking is more than hundreds of terabytes, which also emphasizes the vast amount of data that has to be processed. In addition, it's also being highlighted that success is predicated on the availability of extensive, high-quality data (Hazelwood, et al., 2018).

For mobile phone (edge) usage of ML, Facebook has implemented some less energy-intensive ML models which causes it to have less accuracy. This explicit refers to optimize models for running in an environment constrained by performance and memory, and as an example this is specifically done by "*reduce[ing] the precision of a large multi-GB embedding table from 32-bit single precision float to 8-bit integers*" (Hazelwood, et al., 2019). Hence, this eliminates data details but simplifies computations. Thus this enables for real-time predictions on the phone itself with marginal latency, and thereby are able to increase the user experience.

Understanding the entity model of an Facebook ML model would also be preferable, but such information as seemed hard to retrieve. Anyhow, for an Ads model a very limited set of features are public, but some of the most essential ones are (Facebook, 2018): *Impression ID* – the unique ID of the event when the user views an ad; *User ID* – the ID of the Facebook user; *Post ID* – the ID of the ad post; *Clicks* [True/False] – output whenever the user clicked on the ad or not; *Post Country* – nationality of the post; *Post Category* – type of group the ad item is belonging to (e.g. shoes, hats, glasses, etc.)

As of one can see, these features are combination of constant (historical) variables and variables generated from user interaction. It therefore reasonable to assume that these features can relatively easily be applied to every user and hence generalized.

Specifically for anomaly detection, Facebook points out in a paper (Laptev, 2018) that working with anomaly detection is hard due to lack of labeled and realistic time-series data. Hence, they have worked on a process of generating realistic time-series data with anomalies, and their Deep Anomaly Generator approach has shown superior performance compared to more common approaches like pure synthetic data.

Facebook: Technology	
Object of analysis (OA)	People
n of (similar) OA	1.5 B daily users
ML microservice(s)	Internal MLaaS, including feature store

IT infrastructure	Self-operated datacenters
Real-time edge analytics	Yes
Training	Online
Data vendor	Self
Data setup environment	Standardized Application Store, Web client
Kind of AI/analytics	Anomaly detection (Sigma), classification (Sigma, News Feed), computer vision (Lumos, Facer), natural language processing (Speech Recognition) [...]

4.1.1.2 Business characteristics

In the case of Facebook, as they have enormous amounts of data, it will be reasonable to assume that one of their biggest challenges to scaling has been the ability to analyze the data sufficiently in terms of speed and accuracy, and hence having the needed infrastructure to do so.

The massive data basis will unarguably come from their end-to-end dataflow where Facebook is the one generating the data, and by so are the vendor of data themselves. This could be explained by that they are delivering their product and service in touch with the people, thus also are their *object of analysis.* As of this, this also gives Facebook the ability to set the standard for and control the data flow and quality first at hand.

Given that they serve 1.5 billion active users daily which also indicates they're having 1.5 billion of the same objects of analysis (OA). Despite people being different based on age, gender, nationality, cultural differences and so on, it is still reasonable to assume that finding patterns and personal characteristics is highly doable due to the amount of data, and thereby abstraction and generalization of the OA. Hence, Facebook obviously have large data sets to both train, test and validate their models on, and thereby achieve high precision in their predictions.

The basis for the large amount of data can be understood partially from their business model being an ecosystem, as they enable an platform aligned with their vision statement which is *"to give people the power to build community and bring the world closer together"* (Facebook, 2019), and fulfilling the majority of the characteristics of an ecosystem business model (Weill & Woerner, Thriving in an Increasingly Digital Ecosystem, 2015).

Facebook: Business characteristics	
Business model	Ecosystem driver
Company type	Independent Software Vendor (ISV)
Commercial Transactions	B2C and B2B
Company age	15 years
Analytics	Non-industrial

4.1.1.3 People and organization

When it comes to people and organizational contents regarding scaled ML at Facebook this seems to be hard to retrieve, so such considerations is hard to evaluate. Nevertheless, their internal MLaaS implicit outlines an organizational collaboration of software engineers and data science staff, since MLaaS requires both such skills.

At Facebook the decision makers – what to do with the data basis and hence what analytics to be run – sits internally in the organization. In terms of common understanding, but likely also culture, it is likely that Facebook has good understanding of how to utilize their data thus the conditions and limitations.

Facebook: People and organization	
Team organization	N/A
Analytic decision maker	Internal

4.1.1.4 Implications and value of scaled ML

"Looking forward, Facebook expects rapid growth in machine learning across existing and new services" (Hazelwood, et al., 2018). The quote is not surprising, given the significant benefits ML analytics results in, by delivering tailored content to 1.5 billion individuals daily. Unarguably one could claim that this has led to increased customer experience and value, but also given them an competitive advantage. So to say, there is obvious advantages – both for the end user and the company itself – of having large-scale ML analytics running.

Implications of such though is the larger the analytics have been scaled, the larger is the dependency. Given they still expect this to grow even further, the dependency will also probably increase accordingly. As these ML algorithms help analyze individual cases rapidly, it also becomes a tool with great impact.

Anyhow, the output of such power tool will be determined by the actor utilizing it. The highly targeted Ads with good intentions becomes an extremely helpful tool, but also for bad intentions could lead to malicious results. Despite malicious consequences previously has been mainly limited to potential outcomes, the case of Cambridge Analytica in early 2018 illustrated explicit how powerful tools can influence and be harmful, and thus shake democratic western values and awake the privacy concern (Wired, 2019). Anyhow, the case of Cambridge Analytica will not be investigated further, but is worth noticing as an instance for tools with great implications.

Facebook is using ML in nearly all of their services, providing key capabilities in almost all aspects of user experience. This includes ranking posts for News Feed, speech and text translations, advertisements and their search engines, but with the News Feed dominating the total compute load (Hazelwood, et al., 2018). Even though Facebook has built architecture themselves for handling large scale analytics, an option to handle this demand would be a cloud computation solution. Hence, this has enabled massive development and adoption of ML resulting in innovative solutions for delivering individualized content to 1.5 billions users daily. Unarguably, ML has increased service and product value for the users.

4.1.2 Non-industrial case 2: Uber

Today Uber have an internal platform named Michelangelo which they developed since 2015 and still are doing. It consists of three major strategic pillars: (1) organization, (2) process and (3) technology.

Aspects of interest within the technological pillar will be outlined below, and pillars of organization and process will be outlined in organizational aspects as they are interconnected.

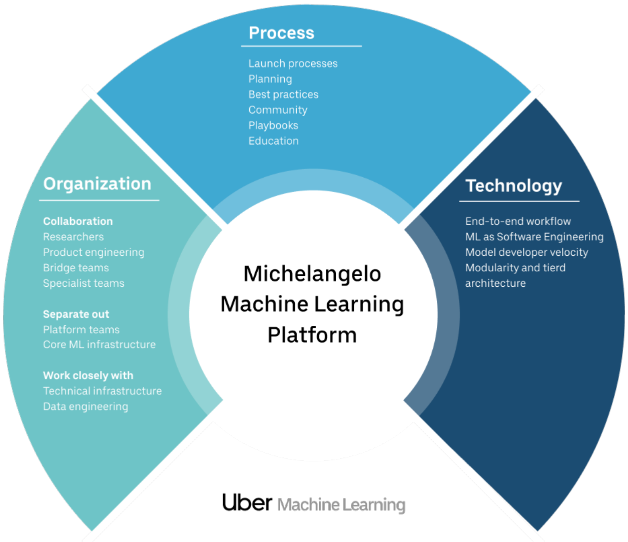


Figure 4.2. **Uber's ML platform.** (Hermann & Balso, 2018)

4.1.2.1 Technology

As of today, Uber is using ML in a wide range of their services due to enabling the company itself to scale its services. These services are *estimated times of arrival* (ETAs), *marketplace forecasting*, *customer support*, *ride check* (alert and help if anomalies during ride), *one-click chat*, and also their *self-driving car project*.

In 2015, ML was not a widespread tool being utilized, in contrast of what is per 2019, but Uber argued *"it was obvious that there was opportunity for ML to have a transformational impact"* (Hermann & Balso, 2018).

The technology pillar of Uber holds four elements which Uber claims to be the most essential ones; (1) End-to-end workflow, (2) ML as Software Engineering, (3) Model developer velocity, and lastly (4) modularity and tiered architecture (Hermann & Balso, 2018).

End-to-end workflow. Uber highlights that ML is way more than training models, but has to do with support of the whole cycle of data modeling (CRISP-DM). Thus, they argue with the importance of having a set of integrated tools for all the steps of the ML workflow. Hence, a brief walkthrough of Ubers end-to-end workflow will be outlined. When it comes to managing data, Uber has developed a centralized *feature store* which *"allows teams to share high-quality features and easily manage the offline and online pipelines"* (Hermann & Balso, 2018). Regarding training of models, this is done using Uber's Data Science Workbench (DSW). As the fundation is the *DSW Management Service* which – by 2017 – does session, file, package and job management, in addition to empower collaboration, displaying dashboard and handling quota enforcement (Joshi & Geracioti, 2017). On top of this users can do simple model training in their preferred notebook or web user interface (UI), but also *"compose complex transformation pipelines, ensembles and stacked models"* (Hermann & Balso, 2018).

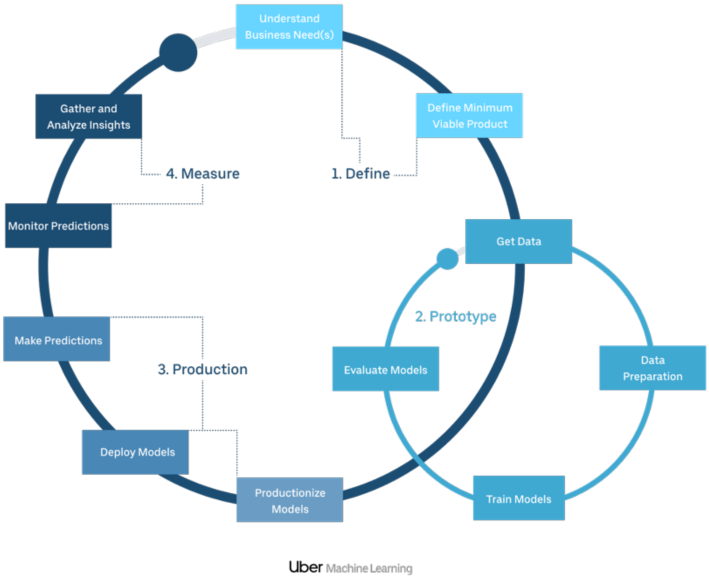


Figure 4.3. **The workflow of a machine learning project.**

Due to the highly iterative process of data and model alignment, Uber specifically developed a model evaluation and comparison tool enabling better visualization and hence understanding of model performance. When deploying the models, this can be done either through the web UI or by using the API, and for both online and offline models, *"the system automatically sets up the pipelines for data from the feature store"* (Hermann & Balso, 2018). After deployment, monitoring the ML models with their *data quality monitoring tools* are done by the two following approaches. The first – and most accurate approach – is to log model predictions made in production and combine this with the actual outcome, as outcome is collected from the data pipeline. The second approach applies to the cases where either the outcome is hard to collect or the outcomes and predictions is hard to join together. As of this, the approach is to monitor distributions of the features and predictions over time. This element however boils down to creating a platform consisting of microservices.

ML as software engineering. The second area is the analogy of ML as software engineering, which means to *"apply patterns from software development tools and methodologies"* (Hermann & Balso, 2018). This explicit means when creating and training a model it is important to keep track of the assets and configuration giving the opportunity to reproduce and/or improve the models. Example being *"in case of transfer learning in deep learning models, we track the entire lineage so that every model can be retrained, if needed"* (Hermann & Balso, 2018). This approach originates from cases where it's been hard to reproduce the due to the data and/or training configuration has been lost.

Model developer velocity. Uber highlights several principles which has proven the DS teams to work more effective: (1) *Solve the data problem so data scientists don't have to.* This is solved by using the Michelangelo's feature store and feature pipelines, as this solves a range data processing problems. (2) *Automate or provide powerful tools to speed up common flows.* Specifically, Uber has developed an internal tool named *AutoTune* which is general purpose optimization-as-a-service tool at Uber, designed with the objective *"to more efficiently search for an optimal set of hyperparameters"* (Hermann & Balso, 2018). (3) *Make the deployment process fast and magical*, which could be to enable single click deployment and hiding of the unnecessary details. (4) *Let the user use the tools they love with minimal cruft.* Michelangelo allows interactive development in Python, notebooks, CLIs, and includes UIs for managing production systems and records. (5) *Enable collaboration and reuse.* This boils down to Michelangelo's feature store which Uber again highlights as critical, as it is enabling teams to reuse important predictive features already identified and built by other teams. (6) *Guide the user through a structured workflow.* This helps the user to better understand how to work with the modeling.

Modularity and tiered architecture. The fourth and last technological area is to have simple components that can be assembled in targeted ways for greater flexibility for cases which are less common and/or more specialized. E.g. has been the development of prebuilt workflows, and interactive learning and labeling tool in this case specifically for computer vision.

Key lesson learned at Uber. Throughout Uber's scaling process, they also highlights some key lessons learned. First, when it comes to ML, data is both the hardest and the most important thing to get right, and here the Uber feature store is critical, as it enables sharing of high-quality features, automated deployment and monitoring. Broken data is the most common cause of such problems. Secondly, real-time ML is also highlighted as challenging as there is few proper tools which solves hybrid online/offline capabilities as most existing tools are built for Extract, Transform, Load (ETL) or online streaming. Uber emphasizes this to be a big part of their focus. Thirdly, making use of open source and commercial components at scale is challenging. Fourth, letting developers use their preferred tools and lastly, develop iteratively based on user feedback.

The findings from Uber's technological scaling is primarily directly related to the Y-axis in the scalability cube, by providing different solutions to different problems during the ML workflow. This including structured general-purpose tools in addition to more customized simple tools for special cases.

Uber: Technology	
Object of analysis (OA)	People, driving patterns
n of (similar) OA	95M users monthly[1], 15M Uber trips each day[2]
ML microservice(s)	Internal MLaaS *(Michelangelo)* including feature store
IT infrastructure	Cloud and own datacenters (Hermann & Balso, 2018)
Real-time edge analytics	N/A
Training	Online, centralized
Data vendor	Self
Data setup environment	Standardized application store (App Store)
Kind of AI/analytics	Stacked models

4.1.2.2 Business Characteristics

As Uber are having 95 million users monthly (Statista, 2018) and 15 million trips a day (Business of Apps, 2019) this unarguably results in large amounts of data. This gives ground to assume processing and analyzing data sufficiently would be

[1] (Statista, 2018)

[2] (Business of Apps, 2019)

a critical challenge. From the investigation, one could argue that this could be one of the reasons for their development of the MLaaS named Michelangelo.

The data collection has its basis in the mobile application setup, hence when people – both drivers and customers – are downloading the Uber app. The setup is convenient and hassle free, probably due to automation on the backend.

Uber characteristics	
Business model	Ecosystem driver
Company type	Independent Software Vendor
Commercial Transactions	B2C
Company age	10 years
Analytics	Non-industrial

Due to their business model sharing the characteristics with ecosystem, and hence end-to-end dataflow, this also results in advantages of data collection and availability, but also in the format data is collection, giving opportunities to streamline the data collection process. Interestingly enough, Uber also points out that broken data is the most common problem, despite they controlling the dataflow.

Uber's OA is most likely to mainly focused around people and car driving patterns, which by it selves and interconnected will return highly valuable insights. Abstraction and generalization of such OA would probably also be relatively doable considering the amount of data available.

4.1.2.3 People and organization

As mentioned, Uber last two pillars are *organization* and *processes*, and will further be discussed in this section.

Uber emphasizes that the challenge is "*in allocating scarce expert resources and amplifying their impact across many different ML problems"* (Hermann & Balso, 2018) as the requirements for the problems is varying. Hence this also impacts the need for optimal organizational structure. As of this Uber has been iterating for finding a suitable structure.

Firstly, there is product (engineering) teams which builds, deploys and owns the ML put into production, and their focal point is narrowed into a specific products only. Such teams would also then have the ability to detect custom needs and create suitable tools accordingly. Secondly is the specialist team which have deep expertise across many domains, and acts like a expert resource pool used whenever for product teams as specific problems occur. As products are put into production they make sure to fill the expertise gaps will full-time experts freeing up the specialists. Thirdly, there is research teams – also named Uber AI

Labs – which collaborate on problems and guide future research, but also develop new techniques and tools which can be made available to product teams. Lastly, there is platform teams who builds and operates *"a general purpose ML workflow and toolset that is used directly by the product engineering teams to build, deploy, and operate machine learning solutions"* (Hermann & Balso, 2018).

For processes Uber emphasizes two main processes which is to share best practices of ML, and instituting more structured processes. This is caused by that *"ML systems are particularly vulnerable to unintended behaviors, tricky edge cases, and complicated legal/ethical/privacy problems"* (Hermann & Balso, 2018). This has led to the expiration of processes regarding model launching, coordinated planning across ML teams, connected and collaborating community, and education processes.

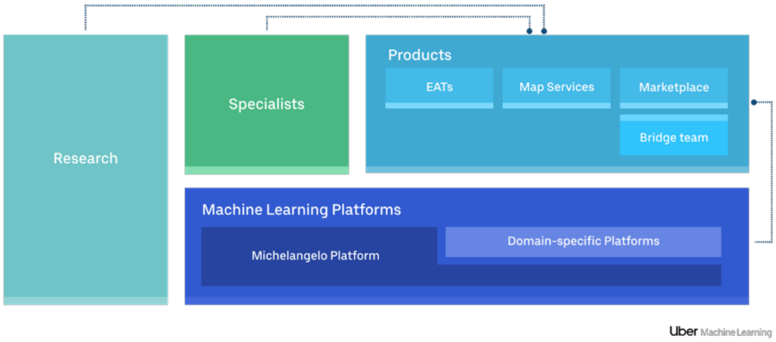


Figure 4.4. **Organizational interactions of different teams in Uber's ML ecosystem**. (Hermann & Balso, 2018)

One could also argue that the principles of Uber's team organization share a majority of its characteristics with the agile organization (Fisher & Abbott, 2015) and the actor-oriented principles (Snow, et al., 2017). E.g. the flexibility and dynamic between product teams, specialists and researches really show – in theory – how the organization adapt to (sudden) needs and demands, and hence optimize knowledge and expertise accordingly. Hence, arguably Uber is able to adapt to the customer demand and needed scalability, thus increase their competitive advantage in the long run.

At Uber the analytics decision makers sits internally in the organization as well as the data collection team and the data science team, thus it is likely that they understand the analytics field.

Uber: People and organization	
Team organization	Flexible organization: Balance of specialized and focused teams;
Analytic decision maker	Internal

4.1.2.4 Implications and value of scaled ML

The scaled ML at Uber gives basis for a broad range of services which otherwise would have been very hard, if not impossible, to deliver. E.g. would be the ETAs which solely is a prediction and hence based on pure ML. When using Uber, this is highly valuable to the consumer as such information improves certainty of what time one could be at an event or not, which surely impacts the decisions of planning (rest of) the day for the consumer. As this particular service is scaled and applied to all 15 million trips a day, the impact is unquestionable big.

Anyhow, as quoted from Uber, ML systems are tricky in terms of ethical, legal and privacy, and hence probably scale ML as well. Such cases has also been publicly questioned and discussed. From the analysis performed Uber is likely to understand *"your place of work, favorite eating joints or shopping destinations, how often you travel, your residence and much more"* (Bajpai, 2018). This however is not unusual – and is also the case with Facebook – but rather the issue is how this data is being utilized. It's being argued that such information is being used by companies to generate additional revenue outside of the core business by e.g. selling the data (Bajpai, 2018) or by promoting their own services by highlighting individual people without consent (Frizell, 2014).

Having scaled ML internally, this surely gives incremental improvement – and hence incremental innovations – to Uber's line of products. As their product(s) have improved, their customer base has also increased. So to speak, one could argue that their service has been disruptive (innovation) and that ML has been one of the enablers to making their service available to the extensive size of customers.

4.1.3 Non-industrial case 3: Tesla

"*Autopilot advanced safety and convenience features are designed to assist you with the most burdensome parts of driving"* (Tesla Inc., 2019). This highlights the value proposition of Autopilot – safer, more convenient, less boring. Elon Musk – CEO of Tesla – has also mentioned that the system will never be perfect, but it's likely to *"reduce accidents by a factor of 10"* (CBS Interactive Inc., 2018) which anyway must be considered to be of significant value.

4.1.3.1 Technology

The Autopilot system heavily relies on machine learning models for doing analytics in real-time and with their deployment of fleet learning, Tesla claims the ML system is improving for every new distance driven. As mentioned, Tesla has the ability to set the terms for both the software and hardware which determines the conditions for the data and ML workflow.

For hardware this have impacted their selection of sensors on the cars and led to custom-built components optimized for data processing and ML analytics. By April 2019, Tesla shipped new cars with their third generation of hardware – called Hardware Version 3 – consisting of two custom-built data processors for high performance of analytics, one forward radar, three forward cameras (narrow, main, wide), four side looking cameras (two forward and two rearward) and twelve ultrasonic sensors for with narrow range. All these sensors are then being utilized for different purposes of the analytic process, but with the cameras as being the primary source of their analytics as they primarily rely on computer vision with neutral network algorithms (Tesla Inc., 2019). As Tesla has these sensors on they are able to gather content-rich data resulting in a massive data basis. Anyhow, Tesla highlights that it is not the large content itself that is critical but the variety of the data which represents real cases, which tends to originate from large data set. They argue that is all about *"the long tail"* which covers for less frequent special events – hence varied data – which is argued to be critical for increasing the ML accuracy of the computer vision algorithm. Tesla has access to *"car's speed, acceleration, braking, battery use, and [...] "short video clips" during accidents"*, according to the company's privacy policy (The Verge, 2018). By so they also then have information of how their fleet of cars are handling in terms of steering and therefore get highly accurate information regarding how drivers operate during special events and further how cars in Autopilot mode should or should not be steering. Tesla therefore emphasize the importance on what they compile down to *"large, varied and real"* data.

Even when the cars are not set in Autopilot mode the cars are running in what is named *Shadow Mode* which registers what actions the Autopilot would have been taking if it was activated. By so, the software collects data from events where human behavior or events and ML predictions differs – in other words mispredictions in their ML models – which causes a data trigger of sending data to Tesla centrally and then highlights relevant events for further investigation. An event can consist of half a million of images regarding this special events and will work as the data basis in Tesla's training process. They are also able to automatically label the events the cars are into – due to their computer vision algorithms which gives great efficiency gains in terms of data structuring and labeling. This data trigger is executed locally in the cars and is the basis for Tesla's "data engine" and enabler for their "fleet learning" approach. Figure 4.5 shows how the *data engine* – in other words the ML workflow process – is executed at Tesla and works as a feature store. Since the models are being updated and pushed into the fleet continuously, the Tesla cars also tends to become better as time passes.

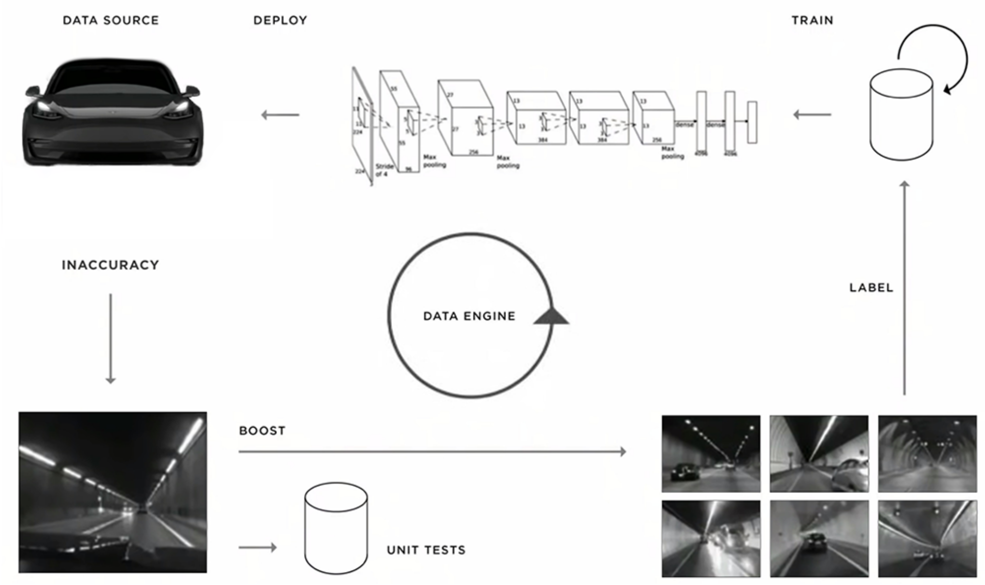


Figure 4.5. **Tesla Data Engine.** (Tesla Inc., 2019)

Tesla: Technology	
Object of analysis (OA)	Surroundings, driver behavior
n of (similar) OA	500,000 vehicles on the road globally (Muller, 2019); 500,000 images from special events
ML microservice(s)	Data engine (feature store concept)
IT infrastructure	N/A
Real-time edge analytics	Yes
Training	Centralized, pilot and validation test with Shadow Mode
Data vendor	Self
Data setup environment	Pre-installed from manufacturing
Kind of AI/analytics	Primarily neutral networks, (anomaly detection)

4.1.3.2 Business characteristics

Tesla have chosen a pure omnichannel business model in their B2C, in contrast to the normal approach of being a supplier through car retailers. By such Tesla has been able to gain direct customer relationship combined with their data collecting capabilities from their sold cars, thus increasing their knowledge of the end customer. As Tesla is considered to be both an OEM and ISV themselves, they also set the terms for the end-to-end data flow, both regarding hardware and software. Their analytics are focused on non-industrial objects, hence they're analytics are classified as non-industrial and will be elaborated in the next section.

Tesla: Business characteristics	
Business model	Omnichannel
Company type	OEM/ISV
Commercial Transactions	B2C
Company age	16 years
Analytics	Non-industrial

4.1.3.3 People and organization

At Tesla there exists dedicated teams for different products from various backgrounds, like data engineers, electronic and mechanical engineers, but also data scientists doing data processing and analytics. Specifically material regarding their team dynamics has been hard to retrieve. Due to lack of valid material this will be hard to investigate.

Anyhow, as for the cases for Facebook and Uber, the analytics decision makers sits internally in the organization as well as the end-to-end dataflow is covered in Tesla's chain. It is therefore considered likely that they understand the scope of the analytics field.

Uber: People and organization	
Team organization	N/A
Analytic decision maker	Internal

4.1.3.4 Implications and value of scaled ML

From the wide intake of data and specialized ML modelling, Tesla has been able to scale its capabilities and deliver self-driving features to 500,000 vehicles and are enabled by the infrastructure built-in into the vehicles. It is therefore reasonable to assume that the scale are directly related to the data collection ("large, varied and real"), the label accuracy and hence ML model accuracy. As the ML models continue to increase in performance as they prove to be doing currently, the massive application of these models brings safety and convenience gains beyond what has been able to achieve previously.

It is also being argued that self-driving cars brings ethical difficulties in terms of special cases where the ML algorithms has to choose between two evils, in such that the algorithms have to choose which action does the smallest harm. Based on the material it is hard to argue that this not might be the case, but one could also argue that due to the extensive amount of data from the various sensors that such events would be very limited and not as of a big concern as some critics claim. Given that the system are able to increase safety by a factor of 10, the net reward would reasonably therefore be increasingly better than not having the scaled ML system.

The Tesla Autopilot feature is at the moment not a fully self-driving car, but the capabilities currently show good evidence of what the Autopilot – and by so the ML models – can achieve by being utilized in the proper way. Reaching to become a fully autonomous car is unquestionably a radical innovation but also disruptive. Also, the cars are becoming incrementally better as time goes – including Autopilot – which is considered to be unique in the market place and hence disruptive market-wise. Anyhow, by the time fully autonomous cars are present, it is likely it will be considered an innovative achievement. And at the moment ML models seems to be very vital as an enabler.

4.1.4 Industrial case 4: GE Digital

GE emphasizes the elements of business, technology and people as central aspect in the Digital Ecosystem (GE, 2016), hence also support the analysis framework of the thesis.

4.1.4.1 Technology

Despite GE Digital being only four years in existing, it seems to be based on older software. In 2015 they claimed to do become an industrial software company with significant impact, expecting *“Predix Software to Do for Factories What Apple's iOS Did for Cell Phones”* (The Street, 2015). GE’s approach has been to deliver analytics with the formula of combining advanced data science, physic-based models and applied engineering knowledge, but also emphasizes the need for computing on the edge rather in the cloud (GE Digital, 2016).

Per 2019, the Predix software is an industrial IoT platform, connecting industrial assets and data to the platform – which can be done by an edge solution – delivering solutions for monitoring, analytics and more, in addition to specific industrial applications. Anyhow, there is several claims that this software is not so good as it hoped for. *“The platform [Predix] is not developer friendly, takes ten-times as long to complete normal tasks compared to best-in-class IoT platforms, and does not have a modern IT architecture”* (Sdx Central, 2018). Such considerations was also confirmed from interview sessions. One could therefore argue that they are lacking modern infrastructure and software capabilities, probably due to their IT legacy systems of previously developed products and solutions. This might root in their more conventional Asset Performance Monitoring (APM) product which helps operational engineers and administrators overview their physical assets, hence industrial equipment, but when it comes to advanced analytics these are fairly limited.

GE Digital also deliver an edge solution – named Predix Edge – which enables streaming data from devices in the field in addition to run some analytics.

Based on interviews, GE Digital's advanced analytics are not specifically tailored for the various industrial equipment – despite the marketing material claims so – but deliver relatively simple generalizable rule-based metrics, hence are lacking ML intelligence. They also deliver some ML but based on findings, they have not been sufficient in performance.

GE Digital: Technology	
Object of analysis (OA)	Industrial equipment
n of (similar) OA	>1000 (estimate)
ML microservice(s)	MLaaS (Predix)
IT infrastructure	Cloud and Edge
Real-time edge analytics	Yes
Training	N/A
Data vendor	Self (when as a service)
Data setup environment	Edge (stream); System integration (batch)
Kind of AI/analytics	Rule-based metrics, anomaly detection,

4.1.4.2 Business characteristics

GE has historically been an OEM and hence their business design has been aligned as an omnichannel. More recently, they have been trying to move over to an ecosystem driver (Weill & Woerner, What's Your Digital Business Model?, 2018), which substantiates with their platform approach.

As GE is an OEM, they also have the opportunity to sell their machines in the format of a *uptime-as-a-service*, meaning the customer pays for the runtime – thus the value creation – of the equipment is generating, unlike a fixed one-time purchase of the asset itself. In such cases GE is owning the asset and so the data, and hence they are able to collect and manage the data itself. As of this, this should hypothetically enable new opportunities for their GE Digital division. It is reasonable to assume they are having top of the line engineering knowledge and equipment data. In combination of their uptime-as-a-service delivery, it is also basis for arguing they have the updated data on their equipment performance.

GE Digital serves, as rest of GE, in the B2B market. Since they deliver the industrial equipment itself, gaining data from it probably can be narrowed to only equipment data and by so not reveal other sensitive business data from their customers. This could hence be seem as an competitive advantage in GE Digital's favor. Anyhow, there has been several bad news regarding their digital initiative, and GE in general has been suffering significantly as reflected in the stock price the latest 3 years.

GE Digital: Business characteristics	
Business model	Omnichannel (developing towards ecosystem)
Company type	OEM
Commercial Transactions	B2B
Company age	4 years (GE Software: 8 years), legacy of GE
Analytics	Industrial

4.1.4.3 People and organization

When it comes to team organization contents regarding scaled ML at GE Digital this seems to be hard to retrieve, so such considerations is hard to evaluate.

Anyhow, in the case of GE Digital, the analytical products are to be delivered for customers which are typically operating the industrial equipment. The customers are hence the analytic decision maker and the stakeholder which determines how the analytics are supposed to be utilized. As of this the analytic decision maker are classified as external in this case.

GE Digital: People and organization	
Team organization	N/A
Analytic decision maker	External

4.1.4.4 Implications and value of scaled ML

Given that GE Digital will be able to scale ML in the applications and services where they currently deliver, they hold a great business opportunity to deliver insights into enormous number of companies. From the data basis retrieved in this research there still seems to be a long way to go for GE Digital, both in terms of technology and business.

4.1.5 Industrial case 5: Arundo Analytics

As Arundo is the one addressing the problem of scalability, they hold marginal content of successful scaling. Hence this section will applied for understanding the characteristics of Arundo.

4.1.5.1 Technology

As Arundo specializes on industrial analytics in the heavy-asset industry, their OA is the industrial equipment, which is an umbrella of various groups of equipment. This includes equipment's like compressors, heat exchangers, turbines, pumps, and so on. The groups but also subgroups itself varies broadly in characteristics and results in a low number of similar OA. This makes it hard to generalize and make proper abstractions of the OA. Anyhow, Arundo has shown some promising results in making equipment specific ML templates which is ML models which has been pre-trained. Despite this could be applied to plural instances, this usually comes at the cost of lower accuracy, but as this is still under development it is likely it will improve by time.

As Arundo also rely on other companies for historical data of industrial equipment. Since the historical data (usually) is the prerequisite for ML model performance, the case of where the other companies set the terms of the data content brings challenges. Anyhow, the Arundo Edge product comes in from a new angle, setting new terms for data collection and then only collects data which are relevant for data analysis.

Arundo has developed products intentionally designed for MLaaS. So to speak their service serves approximately the same purpose as for those who are built for internal usage, but in addition can be used by the customers. Below will be a walkthrough of the products.

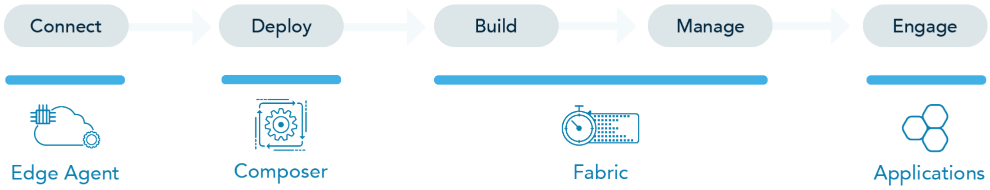


Figure 4.6. **Arundo Products in the ML process** (Arundo Analytics, 2019)

The products is built along the machine learning process, hence CRISP-DM. The first product, the Edge Agent software, enables ingestion and collection of data from industrial equipment by being installed in the presence of the industrial utilities. Further, this enables analytics in environments which is disconnected and hence do not have access to internet and cloud. Secondly, the Composer software enables data scientists deploying with few clicks desktop-models into the Arundo cloud environment, named Fabric. Thirdly, the Fabric software is a cloud service which manages models, data streams and data pipelines. Lastly, they also have applications to work on top of the ML analytics, e.g. the

Equipment Condition & Performance Monitoring (CPM) which is a dashboard displaying KPIs and other analytical metrics.

As the industrial ML process seems to differ from the more conventional process, Arundo has sketched what they believe to suitable. This takes the equipment installation phase into account, in addition to visualization of the insights of numerous individual equipment as their solutions ultimately shall present – hence help – the business with better business decisions.

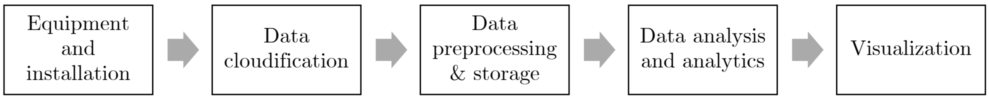

Equipment and installation
Data cloudification
Data preprocessing & storage
Data analysis and analytics
Visualization

Figure 4.7. **Overview of the ML process, specific within the industrial context.** (Arundo interviews, 2019)

When the data scientist at Arundo are working on ML analytical cases they rely on the industrial data gathered by engineering operators working in the field with the industrial equipment. What kind of data being gathered are determined by procedures and motives of the operators field of expertise, hence does not directly apply to the needs of a data scientist. In addition, it is also been noticed that industrial companies have programs which does not record operational data if the values are approximately equal. This also results in datasets with missing data, making analytics harder as this removes valuable information also regarding data distributions.

Arundo's primary field of ML analytics is currently related to advanced anomaly detection algorithms. This is most likely due to the business use cases they are working on.

Arundo: Technology	
Object of analysis (OA)	Industrial equipment
n of (similar) OA	<20
ML microservice(s)	MLaaS (product suite)
IT infrastructure	Cloud and Edge
Real-time edge analytics	Yes
Training	Offline
Data vendor	Heavy-asset industrial companies (typically operating engineers) especially for historical data; (Edge product for edge collection)
Data setup environment	Edge Agent (stream); System ingestion (batch and stream)
Kind of AI/analytics	Anomaly detection

4.1.5.2 Business characteristics

Arundo's products can be used standalone in the Arundo environment but are primarily currently being utilized on top of other services and platforms, like Microsoft's ML cloud platform – Azure – but also on industrial-specific analytical platforms like DNV GL's Veracity and ABB's Ability. Such approach shares mostly its business model characteristics with the cross section of *modular producer* – as of the capability of adapting to any industrial ecosystem and the plug-and-play product – and the *supplier* – as they deliver their products/service through other companies.

Arundo are an independent software vendor in the B2B sphere, meaning that they are dependent on data from the companies approaching them. As they are a recently founded company, it is likely that they are able to take advantage of the latest and most optimal software tools for these particular business cases, despite that it takes great effort and skills building sophisticated software.

Arundo: Business characteristics	
Business model	Modular producer & supplier
Company type	Independent Software Vendor (ISV)
Commercial Transactions	B2B
Company age	4 years
Analytics	Industrial

4.1.5.3 People and organization

The technology-related people in Arundo are generally organized either from a product or project perspective, but flexible in adapting to new demands. By so they are able to achieve focus and momentum for each direction. Specifically for the product team located in Oslo, this meant that they were given freedom and autonomy in cases where there was a clear objective. If the team seemed to lack in its productiveness, the manager would balance the team with a more rigid approach, and hence if the manager saw the team being too constrained by structure and rules he would balance it by giving them more autonomy. The manager hence emphasized the importance of balancing the team for the right purpose and situation, and not solely go for an "agile" approach without further do.

In the case of Arundo, the analytical products are to be delivered for customers which are typically operating the industrial equipment. The customers are hence the analytic decision maker which is labeled as external. The findings in the thesis shows that there seems to be a common understanding that these external decision makers do not possess the needed knowledge of what the possibilities and limitations there exists within the analytics field, hence what is considered to be feasible in terms of industrial analytics. As of this there seems to exists a knowledge gap between analytics and business value.

Arundo: People and organization	
Team organization	Flexible teams balancing on demands
Analytic decision maker	External

4.1.5.4 Implications and value of scaled ML

Scalability has not been achieved – yet. Given that ML can be applied successfully at scale – serving a broad range of industrial use cases – there is great belief of that this will be very valuable for industrial companies optimizing their operations. This is grounded on small-scale analytics deliveries. At scale comes also insights and there is concerns form various industrial companies regarding sharing intimate operational data.

Great business value has been proved related to ML analytics of industrial equipment, and is the reasons for why Arundo – among many others – have been targeting this industry segment the last few years.

4.1.6 Industrial case 6: C3

4.1.6.1 Technology

As of C3 ecosystem approach they do deliver a platform tailored for handling data – which is their core technology – and providing a basic set of applications including some general anomaly detection applications.

Their ML capabilities starts with the AI Suite which delivers microservices for every step in the ML workflow, with comprehensive data integration followed by their data management service, time-series visualization and ML model management service. The data management *"include data federation, management of and interaction with multiple databases, and persistence data in the appropriate data store"* (C3, 2019). It seems to be mainly focused on data integration, but does not include a feature store. C3 has also built Integrated Development Service which they call a low-/no-code environment for developing and operationalize ML during the whole ML process, which consists of a broad range of microservices.

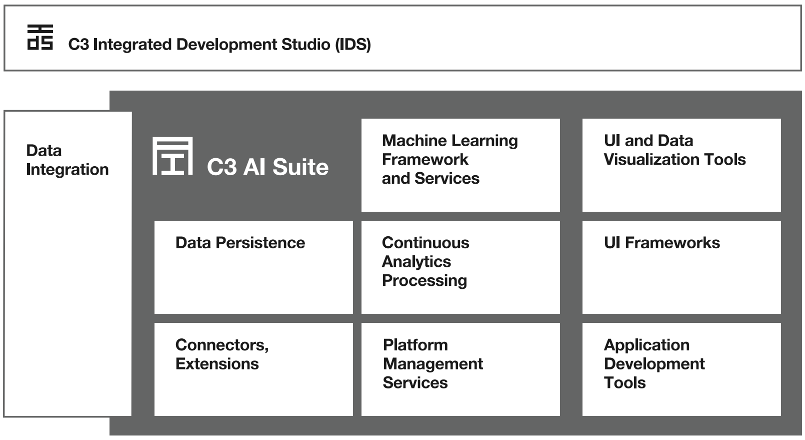

C3 Integrated Development Studio (IDS)
Data Integration
C3 AI Suite
Machine Learning Framework and Services
UI and Data Visualization Tools
Data Persistence
Continuous Analytics Processing
UI Frameworks
Connectors, Extensions
Platform Management Services
Application Development Tools

Figure 4.8. **C3 AI Suite.**

As of these findings there is clearly connection to the *functional decomposing* (Y-axis) in the scalability cube as of the microservices which is suited for the whole ML process.

It is also reasonable to assume that C3 does not invest heavily in data science knowledge targeted for industrial equipment, as they aim for delivering the platform to customers and third-party companies building the applications.

C3 seems to be highly cloud-based and hence do not deliver solutions for edge computing, despite that their latest articles are referring to the topic. Therefore on can argue only that they explicit use cloud as their IT infrastructure.

C3: Technology	
Object of analysis (OA)	Industrial equipment
n of (similar) OA	N/A
ML microservice(s)	MLaaS (C3 AI Suite)
IT infrastructure	Cloud
Real-time edge analytics	No
Online training	N/A
Data vendor	Heavy-asset industrial companies;
Data setup environment	System ingestion (batch and stream)
Kind of AI/analytics	Anomaly detection

4.1.6.2 Business characteristics

C3 have built a platform where (industrial) companies can build their analytic solutions and make use of various applications, but also are applying their solution on top of other ML cloud platforms such as Amazon Cloud and Microsoft Azure. Hence, they hold attributes primarily as an ecosystem but with slippage a modular producer. Based on information from the market, C3 is positioning them self for moving into the ecosystem business model as they want corporative collaborations to make use of their platform. By so they will probably gain more data from the industrial space.

C3 are an independent software vendor in the B2B space, meaning that they are dependent on data from the companies approaching them. As of their age there should be reasonable to assume that there is little to nothing old software architecture, which are confirmed by their use of modern approaches in their market material.

Their applications and solutions are primarily suited for industrial cases – like C3 Predictive Maintenance and C3 Energy Management – but also deliver more general services which can be used for fraud detection.

C3 has structured it's technology such that third-parties are the ones who build their applications and hence advanced ML solutions, and by so letting third-parties keeping their own IP. This approach could potentially help C3 with limiting their need for sensitive insights of other companies.

C3: Business characteristics	
Business model	Ecosystem
Company type	ISV
Commercial Transactions	B2B
Company age	10 years
Analytics	Industrial

4.1.6.3 People and organization

Due to the business of C3 have specialized on their platform development. Based on interviews C3 has therefore limited data scientists working on ML but are therefore more specialized on the platform development with their data engineers and software developers.

Regarding the analytic decision maker C3 are delivering analytical platform and products to customers which are typically operating the industrial equipment. The customers are hence the analytic decision maker which is classified as external.

C3: People and organization	
Team organization	N/A
Analytic decision maker	External

4.1.6.4 Implications and value of scaled ML

C3 have already delivered results to several industrial companies, such as the energy company Shell which has chosen C3 as their AI platform. There is a underlying belief that this platform could work as an enabler for further digitalization and by so greater insights by the use of ML.

4.1.7 Characteristics of cases

Below is a table summarizing various attributes of the different cases, followed by a high level discussion of the characteristics.

	Facebook	**Uber**	**Tesla**	**GE Digital**	**Arundo**	**C3**
Scaled ML	Yes	Yes	Yes	Not yet	Not yet	Not yet
Business model	Ecosystem driver	Ecosystem driver	Omnichannel	Omnichannel (→ecosystem)	Modular producer & supplier	Ecosystem
Company type	ISV	ISV	OEM/ISV	OEM	ISV	ISV
Commercial Transactions	B2C & B2B	B2C	B2C	B2B	B2B	B2B
Company age	15 years	10 years	16 years	4 years, but legacy of GE	4 years	10 years
Analytics	Non-industrial	Non-industrial	Non-industrial	Industrial	Industrial	Industrial
Object of analysis (OA)	People	People, driving patterns	Surroundings, driver behavior	Industrial equipment	Industrial equipment	Industrial equipment
n of (similar) OA	1.5 B daily users	95M users monthly, 15M Uber trips each day	500,000 vehicles; 500,000 images/event	>1000 (estimate)	<20	N/A
ML microservice(s)	Internal MLaaS	Internal MLaaS *(Michelangelo)*	Internal feature store (*Data Engine*)	MLaaS (*Predix*)	MLaaS (product suite)	MLaaS (*C3 AI Suite*)
IT infrastructure	Self-operated datacenters	Cloud and own datacenters	N/A	Cloud and Edge	Cloud and Edge	Cloud
Real-time edge analytics	Yes	N/A	Yes	Yes	Yes	No
ML Training	Online	Online, centralized	Centralized, pilot and validation test with Shadow Mode	N/A	Offline	Offline
Data vendor	Self	Self	Self	Self (when as a service or Edge solution)	Industrial companies; Edge collection	Industrial companies;
Data setup environment	App Store, Web client	Standardized application store (App Store)	Pre-installed from manufacturing	Edge (stream); System integration (batch)	Edge Agent (stream); System ingestion (batch and stream)	System ingestion (batch and stream)
Kind of AI/analytics	Anomaly detection, classification, computer vision, natural language processing [...]	Stacked models	Primarily neutral networks (for computer vision), anomaly detection	Rule-based metrics, (anomaly detection)	Anomaly detection	Anomaly detection
Team organization	N/A	Flexible, focused	N/A	N/A	Flexible, balancing demands	N/A
Analytic decision maker	Internal	Internal	Internal	External	External	External

Table 4. **Characteristics of cases**.

As from the cases investigated there is some distinct differences. All of the scaled ML cases are found in the B2C market which may indicate that consumers are more comfortable to share private information than businesses. This might be seen in connection with the number for similar OA's which clearly have an extensive difference in figures, ranging from 9 to 6 figures versus 4 to 2 figures. Whether or not this is rooted in the chosen OA for the industrial cases is difficult to determine, but clearly show that there is different prerequisites for scaling ML in the industrial context than for the more conventional B2C but also for the analytics decision maker.

Anyhow, due to these characteristics will work as the basis for analysis and for identifying the barriers of industrial scalability of ML they will be further analyzed in the next section of the analysis.

4.1.8 The impact and value of machine learning scalability

From the described cases examples, it can be clearly stated that machine learning is been heavily applied into various core activities, and are playing a key role into driving new product and service propositions.

Despite machine learning as a subject not being a particular novelty of this decade, the scale and magnitude of ML should be considered to be of significant innovative value and impact. One could then argue that the innovations in machine learning – like deep learning – is considered as incremental technically, but has led to great new innovations in their magnitude of business cases, hence hold more disruptive character. This also has given opportunities for Tesla, as their advance in ML – contrary to other established automotive companies – has given them a competitive advantage. Such characteristics is also what the disruptive theory supports, and gives room for classifying the Tesla Autopilot as more disruptive (Christensen, 2003). Anyway, there will be good reason to assert that fully autonomous vehicles are both highly radical and disruptive innovation.

There also seems to be an connection with the extensive application of ML at scale in combination of an ecosystem business model resulting in tremendous competitive advantages as it ultimately improves the customer experiences. This probably also contributes to the increasing *"winner takes it all"* analogy of the big tech companies (Barwise, 2018) which commonly have the ecosystem business model – hence Facebook and Uber. If so, this also implies the importance for companies to solve the large-scale ML case in need of staying ahead in the competition.

With great power comes great responsibility, and the risks of misuse must also be taken into account. E.g. the enormous scale of Facebook also lead to the opportunity of misuse in terms of Cambridge Analytica, who used the platform to mislead and misinform their targeted audience. As of such, this illustrates the importance to be aware of the magnitude of large-scale ML thus its implications. Anyhow, ML originates from data of observations and hence knowledge of the observations. This emphasizes the relevance of the old quote *"Knowledge itself is power",* often credited Sir Francis Bacon (Wikipedia, 2019). Thus companies should be aware and take action regarding responsibility for ethical use of data analytics and data insights, and hence integrate such considerations into the company and product/service strategy.

The increasing amount of data gathered show it could improve the ability to gain insights of phenomena's of a new level, and so also advance the *absorptive capacity* of organizations. Organizations with the ability to scale ML should therefore chase it accordingly due to the aspect of possible innovational benefits and hence competitive advantages.

4.1.9 Conclusion on characteristics

Chapter 4.1 have laid out the different cases and their associated characteristics. During the analysis the characteristics of the successfully scaled cases have been outlined and be compared to the other cases. By so it have also been revealing the value of scaled ML.

The discussion has shown that ML at scale has generally proved to be achievable and have proved to be highly valuable and beneficial, both to the end-customer and for the business. As ML scales it also comes with great magnitude which unarguably brings new dimensions to responsibility issues, thus scaled ML should be considered both radical technology-wise and disruptive market-wise. Anyhow, the successful cases hold certain characteristics which distinguish from the industrial cases, and hence clearly also show that there is different prerequisites for scaling ML in the industry than for B2C. These characteristics will hence be investigated more in-depth in the next chapter.

4.2 What is the barriers making ML scalability challenging in the industry?

In this chapter an analysis of what distinguish the cases of successful cases and the industrial cases will be performed, followed by identification of which elements that are likely causing to be barriers of industrial ML model scalability.

4.2.1 Characteristics of scaled ML

The next chapters will analyze the various characteristics in context of technology, business, and people and organization.

4.2.1.1 Technology

In this section the industrial ML process from Arundo will be used as a baseline (fig 5.7).

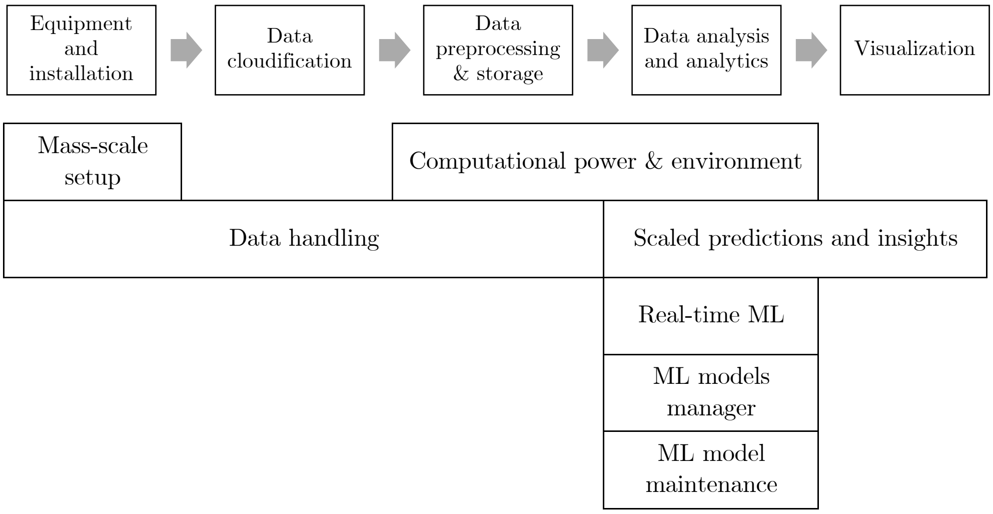

Equipment and installation
Data cloudification
Data preprocessing & storage
Data analysis and analytics
Visualization
Mass-scale setup
Computational power & environment
Data handling
Scaled predictions and insights
Real-time ML
ML models manager
ML model maintenance

Figure 4.9. **Technology characteristics**.

Mass-scale setup. Within the successful cases, like Facebook and Uber, the end-user is setting up the software themselves through a plug-and-play approach, typically from an standardized format, ex. app platform (App Store) or web-browser. The setup takes only a few minute. In the case of GE, this is more uncertain, but seems to require software engineering skills and hence is not very efficient. In the case of Arundo their Edge Agent is setup by running an installer on the computer connected to a part of the industrial production systems gathering data. Within this Edge Agent all data streams has to be mapped manually. Hence when trying to connect all of the whole equipment suite on a production facility this is a comprehensive action.

Data handling. At Facebook and Uber it is natural to assume they have continuous data streams given their opportunity to design for an end-to-end data flow. This leads to consistent data flows which results in data processing with relatively few exceptions and rare conditions. Commonly for the successful cases they have created a centralized feature store for historical data. This ensuring high-quality features which can be easily managed, shared and extracted into data and ML workflows.

In the industrial cases consistent data streams does not seem to be the case. Insights from performed interviews show that there could come batches of data from days before and are missing data labels, causing data structure difficulties. Time labeling in the different systems also tends to vary, which makes time stamping a challenge, and the data collected are limited to be values of a generic range making it hard to identify the source, e.g. what sensor is sending the data. Further, in many industrial systems data are only being recorded – hence collected – if the sensor data changes. From a data science perspective, this causes broad range of valuable data to be missing from the data sets. In addition, IT in the industrial setting are built on old technology and infrastructure which makes it hard to retrieve all the needed and wanted data. So to speak there is a technology laggard which blocks for large data accessibility.

Computational power. For companies having massive amounts of data one barrier of scaling is related to computational power. As the cases for Uber and Facebook shows, there seems not to be technical constrains regarding having sufficient power of compute, either for a pure cloud based approach or hybrid cloud and datacenter approach. Edge approach serves its own purpose as it can provide instant analytics in remote environments and make less compute intensive operations and then eventually transmit data to the cloud. Such were also the case for Facebook in their News feed service. In addition, the edge approach solves a security and policy concern for many industrial companies that don't want their (raw) data to be transmitted into the cloud.

For the case of Arundo, they are relying on cloud approach in combination with edge solutions in remote environment. The approach to GE Digital seems to be quite similar, as they are both offering cloud and edge solutions. C3 on the other hand seems to be the industrial case where edge computing is not present. Hence, computational power has proved to be easily solved by cloud approach of purchasing large amounts of computational power and having the sufficient work force to process the data. Hence, these two activities boils down to capital. When the models are being set to production they are typically being deployed into the cloud environment, which usually is done by using an third party vendor, like Microsoft Azure or Amazon Cloud. In this environment the models can run and vary their demand for computational power depending on the needs.

Scaled predictions and insights. In the successful cases of scaling, they ultimately hold the characteristics of ML models which can be applied to a large number of instances with sufficient accuracy. As for Facebook and Uber, their object of analysis (OA) is people and with their extensive data basis it is very likely that they are able to generalize their data thus the requirement of scaling ML models. With a such large quantity of data – ranging from 1.5B-100M users – it is likely that it covers significant number of events of rare instances.

In the case of industrial analytics the typical OA is industrial equipment. Within this group there exists a large number of subgroups – e.g. compressors, heat exchangers, pumps, generators, etc. - with distinct differences themselves. Even for the group of compressors there is several new subgroups with different attributes, and results in narrow dataset relatively to the successful cases. In the case of Arundo this means 20 or less similar OA. For GE being an provider of industrial equipment they are likely to possess a large database of similar OA. In the use case of predictive maintenance, generally speaking, industrial equipment do fail rarely, causing the data of failures to be very limited. Given the constrains of similar OAs, this results in a challenging analytical conditions. Whereas a solution is scaled for 1000s of equipment, this also brings challenges into effective and intuitive visualization of the operational status.

The *horizontal duplication* (X-axis) in the scalability cube could relate to generalization of the ML models, in such that one model could be scaled and be valid for *n* industrial assets.

ML in production. The successful cases show that it is important to have rigid and low-latency systems that can deal with real-time predictions resulting in instantly. As for the industrial cases, they also show proven real-time analytics.

Both Facebook and Uber have systems for monitoring and maintaining their real-time ML model performance. As of this models that drift in performance will be more easily detected. This approach also includes having observation ML models which detects the performance of models in production or models which adapts due to new data (adaptive learning), making sure the models have sufficient level of performance.

ML models manager. Incorporated into this is the approach of having an service where one can easily manage the models and its associated dependencies like data streams and premises. This is the case across all cases; Facebook, Uber, C3, GE Digital and Arundo.

Generally, there is a broad practice of having MLaaS in all of the cases, bringing microservices as an essential and feasible part of scaling, hence substantiate the *functional decomposing* (Y-axis) in the scalability cube.

4.2.1.2 Business characteristics

Data vendor. As for Facebook and Uber, both control the data flow end-to-end, giving them rich data on their own terms. Anyhow, this is also the case for GE Digital as they sell their industrial equipment provided as a service. For companies like Arundo and C3 this is different as they are dependent on data from the operators, which is the case when being an independent software vendor (ISV).

For the case of Facebook and Uber their data flow control are rooted in their ecosystem driver design and hence give significant benefits. Yet, from the data basis GE has not been able to scale ML for these industrial assets. Such considerations could therefore argue that having complete control of data and hence more data is a great advantage, but not necessary the catalysator for industrial scaling. Becoming an ecosystem hence seems to be highly preferable in the case of scaling as this gives more data insights and hence better data basis to perform analytics.

For B2C, consumers they seems to be more comfortable with giving away personal – and thus sensitive – information. Operating in B2B might seem to be harder as industrial companies are more reluctant to share sensitive data thus resulting in small data samples for external data analysis. This could hence also be a factor which impact data accessibility.

Company age. A common denominator is that the companies are relativity young in terms of age, specifically meaning about 10 years or less. Despite not the age itself being vital in the case of scalability, it anyhow could be an indication of a modern and hence software intensive company. As of such this leads to great focus on software development on modern approaches which all seems to be the case. The case of GE Digital is a special case, despite that the digital division first was initiated 8 years ago, research show that they are based on old software infrastructure and hence suffer from these constrains. This is argued to be related to their former IT infrastructure before they established their software division. Companies starting with clean sheets when building their IT infrastructure they have the opportunity to choose more flexible and adaptive solutions.

4.2.1.3 People and organization

Organizational structure. From the case findings it clearly seems that scaling ML requires efforts from different disciplines, particularly from software engineers and data scientists. As argued by Fisher and Abbott (2015) the "*alignment of architecture, organization and process*" is vital for scaling a solution, which became very clear in the case of Uber which has distinct teams for different purposes and work with distinct overall goals giving employees autonomy. For the other cases this is more subtle, but one can argue by their products and deliveries that they have organized in ways for rapid and dedicated product and solution development. As of this one could then argue that such findings support both the theory of *agile organizations* by Fisher and Abbott, and the *organizational autonomy* by Snow et. al.

Analytics and OA knowledge. In the successful cases of ML scaling the required knowledge for applying analytics to the specific OA is assumed to be limited to data science and psychology as the OA is people. In the case of industrial equipment this is different. Engineering know-how brings ML modelling complexity as it results in the need of deep engineering knowledge in combination of deep data science competence and business understanding. This challenge results in a *knowledge gap* as it is difficult to find people with holistic competences.

In addition where industrial ML solutions are being integrated, it also causes some understanding difficulties as they "don't talk the same language". This also then account for the cases of knowledge gap for analytics decision makers and brings execution and initiative challenges. Hence, this count for cases where the decision makers are external.

4.2.1.4 Non-industrial vs. industrial case

Table summarizing characteristics non-industrial vs. industrial:

Non-industrial	Industrial
- Objects of analysis are similar and have same type of sensors, resulting in unified and similar data. Results in extensive data basis, ranging from millions to billions in size of similar OA.	- Data are significantly different, due to dissimilarities of equipment and even just for the same equipment model. Results in limited data basis, as similar OA is constrained by the sum of 10-1000.
- Often relatively consistent and uniform data (streams).	- Data can be ingested into the system relatively inconsistently, resulting in data handling difficulties.
- Tend to control the data flow end-to-end, often due to delivering B2C products that are dominant in the market, resulting in large, varied and real data sets.	- Industrial analytic companies are dependent on other companies data, typically being operators. Data sets are real, but due to limited data size the sufficient variety is hard to retrieve.
- Typically have an ecosystem business model resulting in improved data accessibility.	- Data vendors are industrial operators, resulting in ISV to not being ecosystems, which can lead to less data accessibility.
- Analytic decision maker are typically sitting internally in the organization.	- Applying industrial ML requires engineering and specific industrial equipment knowledge, in addition to data science. Brings difficulties as this relates to analytics decision makers who typically are externally positioned.

4.2.2 Barriers to industrial ML scaling

Based on the qualitative analysis performed there is reasons to assume that the following aspects are considered to be the largest barriers of scaling.

Technology. Handling industrial data properly is hard and brings difficulties in terms of data tangibility and size. This seems partially to be rooted in old infrastructure which lacks modern software needs causing in massive manual processing. Specifically this is also related to inconsistent data streams, correct time stamping and mapping of data streams. Hence, the data cannot be handled universally, which also impacts data size and accessibility. As seen from the successful cases data should preferably be large, varied and real. For the industrial cases they are rarely large, nor varied, but real. Anyhow, large datasets are not always needed for solving industrial cases as relevant data is what solves problems, but the case seems to be that within large and varied data sets typically relevant data are also embedded. Retrieving sufficient relevant data and being able to structure and handle it elegantly is blocking application, hence scaling. The lack of universally data handling is therefore considered to be a significant hence the first barrier.

In the industrial case, the object of analysis (OA) are industrial equipment. This OA brings challenges itself as it differs, even for identical models of equipment. Even in some rare cases it leads to applying different ML model for the same model of equipment. Applying ML models which can run for plural equipment with satisfactory performance are therefore probably only doable to a smaller set which share essential common features. Hence, optimizing ML models for the right equipment and number of equipment is important, and thus limited application and generalization of the same ML models are considered the second barrier.

Business characteristics. Aspects within this group does not seem to be barriers in themselves, but are impacting the size of the other barriers. Specifically this counts for commercial transactions of B2C and ecosystem business model which seems to unleash more access to data. In contrast, being an independent B2B company without an ecosystem model could increase the scaling hurdles, but are not considered to be significant. Company age itself is not considered to be a barrier, but could to some extent represent company software legacy and modernity, and hence laggard of technology infrastructure.

People and organization. Applying ML models at scale requires knowledge from both the field of analytics and the field where the analytics should be applied to, which in this case applies to industrial equipment. From a data scientists viewpoint this typically account for lack of engineering knowledge, but as for decision makers this applies to lack of data science and analytics knowledge. As a result decision makers tends to be not enough focused on the relevant data that can answer business questions. This seems to account for external analytic decision makers in particular, and one could may assert that this is related to the companies and individuals absorptive capacity. As of this, knowledge gap in external organizations is considered to be the third and final significant barrier.

4.2.3 Conclusion on barriers

In this chapter an analysis of what distinguish the cases of successful cases and the industrial cases have been performed, followed by identification of which elements that are likely causing to be barriers of industrial ML model scalability.

As outlined, the following aspects are considered to be barriers of ML scalability within the industrial setting: (i) lack of uniform data handling, (ii) limit for generalization of ML models due to dissimilarities of OA, and (iii) knowledge gap in external organizations.

4.3 What are possible strategies to solve these scalability barriers?

In this chapter the barriers of industrial scalability of ML will be discussed, and further based on material found possible solutions to these barriers will be proposed. As outlined in the previous chapter, these are considered to be (i) lack of uniform data handling, (ii) limit for generalization of ML models, and (iii) knowledge gap in organizations.

4.3.1 Lack of uniform data handling

This barrier, as explained in the previous chapter, it is rooted in the various old IT technologies and systems in combination with inconsistent data streams. In sum, this leads to a more fragmented and chaotic data basis and then affects the majority of the ML process workflow, directly equipment onboarding and data setup, data cloudification and processing, thus indirectly data analytics. A solution to this barrier must therefore be able to handle the various software technologies, deal with inconsistent data streams, correct time stamping and mapping of data streams.

Various software technologies. The recent trend of edge computing, which enables local connection to the actual source, can work mostly independent of the old IT and software systems. By so the data streams can be set on the analytical terms and hence be tailored for more optimal data basis, but more importantly only collect the data which is relevant. This also takes care of a majority the cases where industrial data have been erased due to unchanging values.

There seems to be an emerging interest in edge computing as more and more vendors are developing such tools, but still seems to be in an early technological phase. It is therefore being proposed applying and developing edge capabilities is important to build the foundation for and accelerate industrial scaling of ML models.

Time stamping. Due to the different systems the internal system clocks are not in sync. In the case of industrial analytics they are heavily based on time series analysis, hence syncing the various events are crucial for aligning the data. Based on the input for interviews this seems to be solvable during the edge setup, but requires mainly manual processing based on specific system specifications.

Hence, given automatic detection of this system clocks based on automatically detected specifications – e.g. from a classification ML model – this process could hence be enhanced and scaled up.

Inconsistent data streams. Data streams can be infrequently, e.g. 2 days old dataset, which forces re-calculation of the ML models and other critical metrics. Due to this there is a need of having a flexible database which can handle such data batches and re-calculations elegantly.

Based on interviews, there does not currently exists any database system which has such abilities. It is therefore suggested that industrial software companies working with analytics should aim for development of a such solution as this will enable a fundament for uniform data handling further on in the ML process. The downside will probably be that this is very expensive and hence may require collaborative forces – e.g. partnerships or joint ventures – depending on business resources.

Mapping of various data streams. The data streams from different sources are commonly containing values in the same range which makes it hard to map the actual source – hence what sensor or signal is being sent – of the data streams. This requires manual mapping with industrial documents.

A conceptual solution to a such case could be to train a ML model which could propose the likelihood of classification the various sources cross-checking with related industrial documents. It will be reasonable to assume that at first the ML model at first will not be having sufficient performance, but given the "long tail" analogy – which increases model accuracy over time due to more relevant data – it is proposed as being a possible solution with significant gains if accomplished.

Data handling in general. The majority of the proposed solutions are grounded on automation, primarily enabled by ML models. By so it seems reasonable to assume this would work as a data unifier enhancing data structuring and hence increase the data accessibility and tangibility.

4.3.2 Limited generalization of ML models

The typically small amount of specific OA related data and highly differentiated OAs makes it hard to scale ML models for a suite of industrial equipment. Hence, this limits the application of data analytics within in the industry itself blocking scaled insights. To overcome this barrier there has to be found solutions and guidelines in regards of data tangibility and model optimality. The two most prominent aspects related to limited generalization of ML models is therefore outlined.

The amount of specific OA data. As the amount of data for specific equipment is limited – e.g. in the case of Arundo limited to less than 20 instances – the relevant industrial data basis is disproportionate different from the large successful scaled cases.

Given the important role the feature store has been highlighted in the successful cases, it is reasonable to argue it may also be of great importance in the industrial setting due to the scarce amount of relevant data. An industrial feature store could work as a standardized data basis with preprocessed features ready to be applied to industrial ML models. This could help data scientist working in the industrial context more utilized with relevant data. As the feature store grows in size and instances, more data can be utilized for cross-validation and the argument of the long tail could also lead to great value in terms of model training thus performance, thus possibly model generalization across various equipment.

Due to the scarce amount of data it is also likely that extra simulated data of sufficient quality are likely to be preferred. Findings from deep anomaly generator (Laptev, 2018) showed promising results and is hence proposed as an assistance in increasing the data basis and hence ML model quality.

Differentiated industrial OAs. Training a ML model for one specific equipment will increase in performance as the model gets more and more tailored training, but since industrial equipment are varying significantly applying the same ML models broadly – which can be considered as *horizontal duplication* (scale by cloning) in the scalability cube – does not give satisfactory results. Anyhow, there could be a way to segment the industrial assets and make very specific pre-trained ML models which has seen promising results in terms of being applicable but typically comes at the cost of model accuracy.

This leads to a conceptual graph for accuracy vs. ML model scalability which is illustrated in figure 4.10.

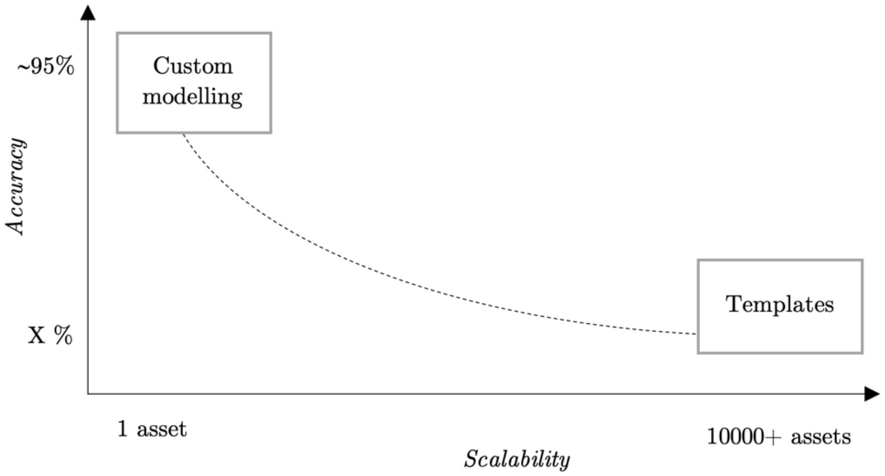


Figure 4.10. **Accuracy vs. Duplication Scalability: Practical implications in terms of model training.**

Hence, this could bring practical implications in application of ML models, as pretrained models with less accuracy brings value in cases of less critical equipment in need of basic monitoring thus can be applied for larger volumes of equipment. In contrast, critical equipment needs higher accuracy and therefore more customization which will only be applicable for individual cases. This also implicate that accuracy is directly correlated with the cost of building the ML models. Further, this could potentially also mean that based on the number of similar OAs ML models should be applied to it could be calculated what accuracy is likely to be achieved but also at what cost.

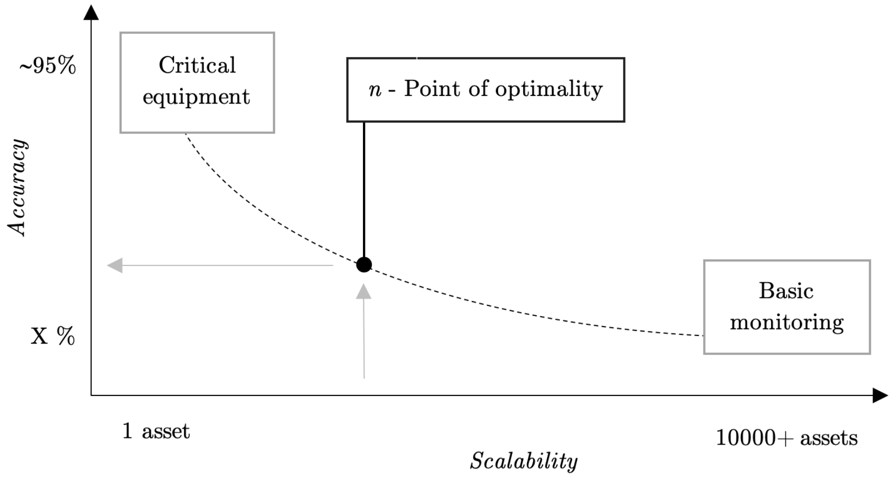


Figure 4.11. **Accuracy vs. Duplication Scalability: Practical implications in terms of application.**

Scientific papers also showed promising evidence of transfer learning for industrial cases (Vercruyssen, Meert, & Davis, 2017), hence this approach is argued to be worth further research. Transfer learning models can then first trained for templates followed by minimal training for specific cases, thus increase accuracy for greater number of OAs, which is illustrated in figure 4.12. As of this it is considered that abstraction and generalization of ML models are more likely to be feasible, first within the subgroups of each equipment type – e.g. for positive displacement compressors and dynamic compressors – followed by abstraction of the equipment groups themselves such as compressors, pumps, etc. Practical implications are also likely to be decreasing of the cost to build industrial ML models thus cost to scale.

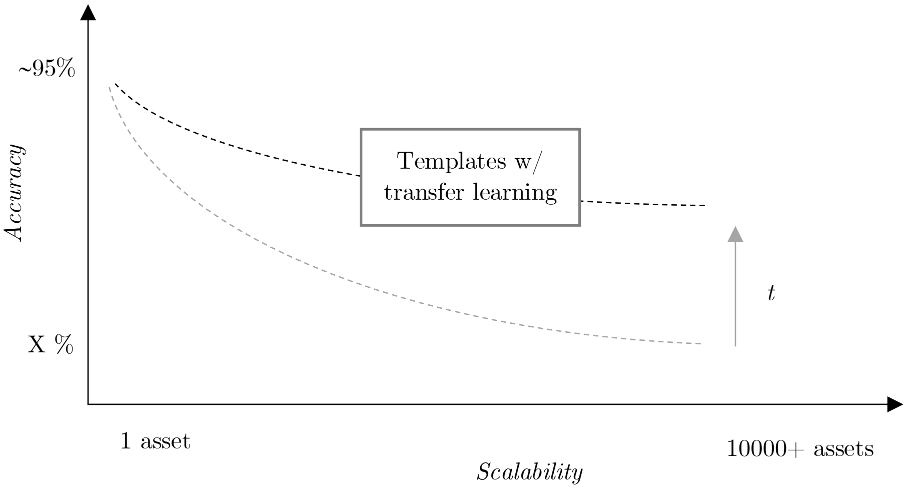


Figure 4.12. **Accuracy vs. Duplication Scalability: Transfer learning**

Limited generalization of ML models. Two aspects of limited generalization of ML models have been highlighted, whereas the differentiated industrial OAs seems to be most impactful. The proposed framework brings a basis to ML models in terms of model training and application for approximately same and various type of industrial OAs, which should be a considered valuable in the quest of generalizing ML models.

4.3.3 Knowledge gap in external organizations

As highlighted, external analytic decision makers tends not to possess enough knowledge about the capabilities and limitations of analytics. By so the decision makers tends to be too much focused on the data they have stored rather than the relevant data that can answer specific business questions, which also exemplifies the need for more analytics knowledge. Closing this knowledge gap should therefore be considered.

One of the reasons the knowledge gap exists in the industry is probably also related to the still lack presence of the applied ML technology but also due to that ML technology is considered being relatively unmatured. As the technology will be more common, mature and more applied it is reasonable to assume the general knowledge of analytics will be increased and implicit also the prerequisites for analytics. It is likely this will take time which implies that other actions has to be considered for solving the barrier at the time of research.

It is therefore argued to contribute nuancing the industrial analytics field in the areas of influence. This particularly could relate to customer meetings where starting off with determine what business problems are in need of being solved rather than what data have been stored, followed by identifying what problems contain relevant data to be utilized. Also in the same discussion highlight what are the prerequisites for good ML and analytic projects. One could also argue that this could be an ingredient into the marketing and public presence for companies serving analytical products.

4.3.4 The industrial ML scaling cube

Based on the various considerations in the analysis, it is possible to align and understand the case of industrial ML scalability in the context the scalability cube (Fisher & Abbott, 2015). As of this there is proposed and illustrated a new cube (figure 5.17) in quest of making it tangible for the industrial setting.

Y-axis (functional decomposing). Originally intended for microservices, the ML microservices and infrastructure have been placed, also more commonly referred to as MLaaS. Every case showed a proof of a such approach, despite it seemed to be that the successful ones were a bit more advanced in terms of data and feature store. The more tools and microservices that will be developed for a more rapid development cycle, the more likely it is that it will be easier to broadly apply and scale ML models.

X-axis (horizontal scaling or scaling by cloning). This axis have been suited for applying – hence cloning – the same ML model for the same industrial equipment. The more assets of the same group of equipment the model can apply to, the greater the scale of the model is.

Z-axis (data partitioning). With respect to the case of industrial ML, this could apply to the variety of industrial equipment that ML models has to be able to apply to for the instance of being able to scale.

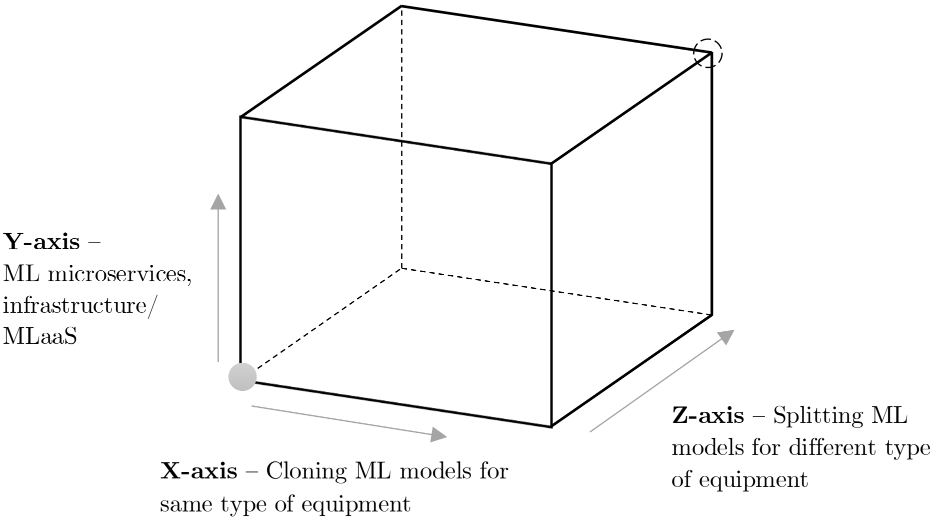

Y-axis –
ML microservices,
infrastructure/
MLaaS
X-axis – Cloning ML models for
same type of equipment
Z-axis – Splitting ML
models for different type
of equipment

Figure 4.13. **Industrial scaling of ML models.**

The intersection of cloning ML models for *n* equipment (X-axis) and splitting ML models for *n* equipment (Z-axis) brings questions of optimization. The less ML models are operating in production, the easier it is to monitor, update and maintain the models. Contrary, the more ML models are put into production the higher performance the models are likely to achieve. Hence, it is therefore being proposed that the following question should be asked for industrial analytic companies: *for what industrial equipment, on what level and for what number do you split you ML models?*

4.3.5 Conclusion on possible strategies

In this chapter the barriers of industrial scalability of ML have been discussed and possible solutions proposed.

For the first barrier – *lack of uniform data handling* – several strategies for unifying the data have been proposed including establishing a data feature store, highly impacted by automation enabled by ML models. For the second barrier - *limit for generalization of ML models* – it has been outlined a framework for how to deal with the barrier with a deliberated approach in addition to proposing further R&D into the technology of transfer learning. For the third barrier – *knowledge gap in external organizations* – it is emphasized as critical the importance of aligning expectations and common understanding with external organizations and external analytic decision maker in particular.

5 Conclusion and implications

In this section the findings from the analysis will be described in addition to the implications which follows the findings. Finally the limiting aspects of this thesis will be highlighted.

The thesis aimed to understand the characteristics of successfully scaled ML, further identify the barriers to scaled industrial ML, followed by outlining possible strategies and framework for dealing with those barriers. To be able to achieve the following, the research needed to be performed with respect to two dimensions – innovational and technological – thus have been analyzed accordingly. In addition, as academic literature also emphasizes that aspects such as organizational and business also impacts the processes of new innovations, it have also been included in the analysis. Due to this three research questions was developed accordingly and evaluated in the context of technology, business characteristics, and people and organization, which have led to the following findings:

Research question (1) *What is characterizing scaled machine learning in the selected cases and how does it contribute to business value and innovation?*
There is certain distinct characteristics which distinguish the successfully scaled cases unlike the industrial ones. This specifically means they are all having B2C transactions making end-to-end dataflow more feasible and the analytic decision makers are positioned internally in the organization. Hence it is reasonable to claim the prerequisites are evidently different for the industrial B2B cases than for the successfully scaled ML.

Further, there is great reasons to assert that despite ML itself have been an incremental innovation, scaled ML brings analytical capabilities and insights efficiency which is both significant in technological change and in market offerings. Hence, scaled ML is claimed to be both radical and disruptive as innovation.

Research question (2) *What is the barriers making ML scalability challenging in the industry?*
The barriers are according to the finings considered to be lack of uniform data handling within the industry, limit for generalization of ML models due dissimilarities of industrial objects of analysis, and knowledge gap in external organizations are blocking industrial scalability.

Research question (3) *What are possible strategies to solve these scalability barriers?*
The proposed framework brings a set of possible solutions to overcome the barriers. As for *lack of uniform data handling* this should be approached by automation by ML models as well as a uniform data store. For *limit of ML model generalization* a framework has been proposed for ML models in terms of model training and application for industrial OAs. In terms of *knowledge gap in external organizations* the importance of aligning expectations and common understanding with external organizations, and external analytic decision maker in particular is emphasized as critical.

5.1 Academic implications

The research findings gives basis – based on the enhanced value and change by scaling – for asserting that innovation also should be considered in *the way a product or service is being enabled to be applied in large number of instances*, hence scaled. As there is not found academic literature which can validate such concept and approach, the term *scaling innovation* is being coined and would more specifically account for by what systems, infrastructure and processes it is needed for a product or service to be delivered and utilized for a great number of instances. This could be primarily related to the technological improvements despite – as argued with scaled ML – it comes with disruptive elements. Further, it is therefore being proposed that this term should be challenged to bring increase its external validity hence ability to be generalized.

The research also substantiates prior research (Snow, et al., 2017; Fawcett & Provost, 2013; McKinsey & Company, 2019) how autonomy and flexibility within organizations are needed for being able to align solutions with the demands in the market, hence deliver relevant products and services in a highly competitive market.

Further, it being asserted that lack of knowledge is vital for not blocking innovation, which confirms the theory of absorptive capacity (Cohen & Levinthal, 1990). It is proposed that this is critical also on the level of individuals, particularly for organizational decision makers.

5.2 Practical implications

This research brings practical implications for all of the proposed strategies, including highlighting the importance of building an solid unified data basis of relevant data. Further, the framework related to generalization of industrial ML models is argued to have a practical approach which can be transferred directly into industrial business management. This also account for the industrial ML scaling cube which brings specific focus to the concept and development of industrial ML at scale.

5.3 Limitations of the research

As for any scientific research there are potential observational errors, including this research. The research duration lasted for 4 months only, which is considered to be relatively short for a research project. Is therefore likely to argue that a longer research period could have enhanced the quality of the findings. The researcher's background is not in computer science or software engineering, which might can have impacted the interpretation of the field of ML.

The type of data available differed, as it seemed that technology and business data were more accessible than organizational data, particularly for team organization, and consequently can have impacted the findings. For qualitative interviews, as my native language is not English this could have brought language challenges which have affected the primary data. Due to the fact that approximately all interviews were performed on Arundo employees, this is unquestionable colored by Arundo's view on the industrial aspects. Anyhow the consideration was that these interviews were the most relevant and valid respondents given the constrained resources and time of the research project. Another possible limitation could also be that – despite that it has been assumed people have acted honestly in the interviews – it exists instances of where people are having their own agenda which could have impacted the answers. As for secondary data, possible limitations or sources of error could be outdated material, but also material which is biased by serving an hidden agenda (ex. for marketing material).

6 Bibliography

Accenture. (2018, January). *Getting insights from data.* Retrieved from Accenture.com: https://www.accenture.com/us-en/blogs/blog-ruehle-detwiler-procurement-data-insights

Arundo Analytics. (2019, April). *Arundo Products.* Retrieved from Products: https://www.arundo.com/products

Bajpai, P. (2018, October). *How Uber is Selling all Your Ride Data.* Retrieved from Investopedia: https://www.investopedia.com/articles/investing/030916/how-uber-uses-its-data-bank.asp

Barwise, P. (2018, July). *Nine reasons why tech markets are winner-take-all.* Retrieved from London Business School: https://www.london.edu/lbsr/nine-reasons-why-tech-markets-are-winner-take-all

Bondi, A. (2000). Characteristics of scalability and their impact on performance. *Proceedings of the second international workshop*, p. 195.

Business of Apps. (2019, February). *Uber Revenue and Usage Statistics (2018).* Retrieved from Business of Apps: http://www.businessofapps.com/data/uber-statistics/

Cambridge University Press. (2019). *Scalability | Cambridge Dictionary.* Retrieved from Cambridge Dictionary: https://dictionary.cambridge.org/dictionary/english/scalability

C3. (2019, Mai). *C3 AI Suite Services and Capabilities.* Retrieved from C3.ai: https://c3.ai/products/c3-ai-suite/c3-ai-suite-services-and-capabilities/

CBS Interactive Inc. (2018, April). *Tesla CEO Elon Musk addresses autopilot system safety concerns: "We'll never be perfect".* Retrieved from CBS News: https://www.cbsnews.com/news/tesla-ceo-elon-musk-addresses-autopilot-safety-concerns/

Chesbrough. (2010). Business Model Innovation: Opportunities and Barriers. *Long Range Planning 43*, pp. 354-363.

Chesbrough, H. W., Vanhaverbeke, W., & West, J. (2006). *Open Innovation; Researching a New Paradigm.* Boston, Massachusetts: Harvard Business School Press.

Christensen. (2003). *The Innovator's Solution.* Harvard Business Press.

Christensen, C., & Overdorf, M. (2000, March). Meeting the Challenge of Disruptive Change. *Harvard Business Review.*

Cohen, & Levinthal. (1990). Absorptive Capacity: A New Perspective on Learning and Innovation. *Administrative Science Quarterly.*

Crossan, M., & Apaydin, M. (2010, September 10). A Multi-Dimensional Framework of Organizational Innovation: A Systematic Review of the Literature. *Journal of Management Studies.*

Dediu, H. (2012, May). *The phone market in 2012: a tale of two disruptions.* Retrieved from asymco.com: http://www.asymco.com/2012/05/03/the-phone-market-in-2012-a-tale-of-two-disruptions/

Easterby-Smith, Thorpe, & Jackson. (2015). *Management and Business Research.* Thousand Oaks, CA: Sage.

Electrek. (2018, August). *Watch Tesla Autopilot's latest update driving on winding roads – showing improvements.* Retrieved from Electrek: https://electrek.co/2018/08/29/tesla-autopilot-improve-driving-winding-roads/

European Commission. (2016). *What is personal data?* (European Commision) Retrieved November 2018, from europa.eu: https://ec.europa.eu/info/law/law-topic/data-protection/reform/what-personal-data_en

Facebook. (2018, May). *Field Guide to Machine Learning, Lesson 2: Data.* Retrieved from Facebook Research: https://research.fb.com/videos/field-guide-to-machine-learning-lesson-2-data/

Facebook. (2019, April). *Facebook - Investor Relations - FAQ.* Retrieved from Facebook Investor: https://investor.fb.com/resources/default.aspx

Fawcett, T., & Provost, F. (2013). *Data Science for Business.* O'Reilly.

Fisher, & Abbott. (2015). *The Art of Scalability: Scalable Web Architecture, Processes, and Organizations for the Modern Enterprise (Second Edition).* New Jersey, 07675: Pearson Education, Inc.

Forbes. (2018). *The World's Most Innovative Companies.* Retrieved from Forbes: https://www.forbes.com/innovative-companies/list/

Forbes. (2019, February). *A Robocar Specialist Reviews The Tesla Autopilot.* Retrieved from Forbes: https://www.forbes.com/sites/bradtempleton/2019/02/27/a-robocar-specialist-reviews-the-tesla-autopilot/#78273cf542ae

Fortune. (2015, Oct). *How Tesla is ushering in the age of the learning car.* Retrieved from Fortune: http://fortune.com/2015/10/16/how-tesla-autopilot-learns/

Freeman, C. (1982). *The Economics of Industrial Innovation.* University of Illinois.

Frizell, S. (2014, November). *What Is Uber Really Doing With Your Data?* Retrieved from Time Magazine: http://time.com/3595025/uber-data/

Gage, J. (2018, May). *Deploying Machine Learning at Scale.* Retrieved from Algorithmia: https://blog.algorithmia.com/deploying-machine-learning-at-scale/

GE. (2016). *Powering Everyone.* Retrieved from GE.com: https://www.ge.com/digital/sites/default/files/download_assets/Powering-Everyone-Analytics-Ecosystem-white-paper.pdf

GE Digital. (2016, June). *GE Digital.* Retrieved from GE.com: https://www.ge.com/sites/default/files/ge_webcast_presentation_06232016_1.pdf

Gobo, G. (2008). Re-conceptualizing generalization: Old issues in a new frame. In N. K. Denzin, & Y. S. Lincoln. The SAGE Handbook of Social Research Methods, London.

Godin. (2015). *Innovation: A Conceptual History of an Anonymous Concept.*

Gopalakrishnan, & Damanpour. (1997). *A Review of Innovation Research in Economics, Sociology and Technology Management.* Elsevier Science Ltd.

Hackeling, G. (2014). *Mastering Machine Learning With Scikit-Learn.* Birmingham B3 2PB, UK: Packt Publishing.

Hajizadeh, Y. (2018, May). Machine learning in oil and gas; a SWOT analysis approach. *Journal of Petroleum Science and Engineering*, pp. 661-663.

Hazelwood, K., Bird, S., Brooks, D., Chintala, S., Diril, U., Dzhulgakov, D., . . . Wang, X. (2018, May). *Applied Machine Learning at Facebook: A Datacenter Infrastructure Perspective.* Retrieved from Facebook Research:

https://research.fb.com/publications/applied-machine-learning-at-facebook-a-datacenter-infrastructure-perspective/

Hazelwood, K., Wu, C.-J., Brooks, D., Chen, K., Chen, D., Choudhury, S., . . . Eldad Isaac, Y. J. (2019, February). *Machine Learning at Facebook: Understanding Inference at the Edge.* Retrieved from Facebook Research: https://research.fb.com/publications/machine-learning-at-facebook-understanding-inference-at-the-edge/

Hermann, J., & Balso, M. D. (2018, November). *Scaling Machine Learning at Uber with Michelangelo.* Retrieved from Uber Engineering: https://eng.uber.com/scaling-michelangelo/

Hill, & Jones. (n.d.). *Strategic Management Theory: An Integrated Approach.*

Hutter, Caruana, Bardenet, Bilenko, Guyon, & Kegl, L. (2019, March). *AutoML 2014 @ ICML.* Retrieved from AutoML 2014 @ ICML: https://sites.google.com/site/automlwsicml14/

Joshi, N., & Geracioti, I. (2017, October). *Turbocharging Analytics at Uber with our Data Science Workbench.* Retrieved from Uber Engineering: https://eng.uber.com/dsw/

Laptev, N. (2018, November). *AnoGen: Deep Anomaly Generator.* Retrieved from Facebook Research: https://research.fb.com/wp-content/uploads/2018/11/AnoGen-Deep-Anomaly-Generator.pdf

Lichtenthaler, U., & Lichtenthaler, E. (2009). A Capability-Based Framework for Open Innovation: Complementing Absorptive Capacity. *Journal of Management Studies.*

Lungariello, R. (2018, July). *7 Industries Leveraging Machine Learning.* Retrieved from New Horizon Computer Learning Centers: https://blog.nhlearningsolutions.com/blog/7-industries-leveraging-machine-learning

McCarthy, J. (2007). *WHAT IS ARTIFICIAL INTELLIGENCE? Basic Questions.* Retrieved from stanford.edu: http://www-formal.stanford.edu/jmc/whatisai/node1.html

McKinsey & Company. (2019, May 7). *Giants can dance: Agile organizations in asset-heavy industries.* Retrieved from McKinsey: https://www.mckinsey.com/industries/oil-and-gas/our-insights/giants-can-dance-agile-organizations-in-asset-heavy-industries

Muller, J. (2019, March). *What Tesla knows about you.* Retrieved from Axios: https://www.axios.com/what-tesla-knows-about-you-1f21d287-a204-4a6e-8b4a-0786b0afac45.html

OECD. (1996). *The Knowledge-based Economy.* Retrieved from www.oecd.org: http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=OCDE/GD%2896%29102&docLanguage=En

Picardo, E. (2019, February 3). *Eight of the World's Top Companies Are American.* Retrieved from Investopedia: https://www.investopedia.com/articles/active-trading/111115/why-all-worlds-top-10-companies-are-american.asp

Poole, D., Mackworth, A., & Goebel, R. (1998). *Computational Intelligence: A Logical Approach.* New York: Oxford University Press.

Pratt, L. Y., & Thrun, S. (1997, July). Machine Learning - Special Issue on Inductive Transfer. *Springer.*

Ringdal, K. (2001). *Enhet og mangfold. Samfunnsvitenskapelig forskning og kvantitativ metode.* Bergen: Fagbokforlaget.

Robinson, J. (2019, February 3). *How Facebook Scales Machine Learning.* Retrieved from Medium.com: https://medium.com/@jamal.robinson/how-facebook-scales-artificial-intelligence-machine-learning-693706ae296f?mkt_tok=eyJpIjoiWWpZell6VXpOR1k0TlRGaSIsInQiOiJZcWpNVDUxdjhEd1ZjK2tTcTFaODl6cDdlejRSR0E4dldWbE9zUXpYd3BcL002bitBdWdrMWcxUEt6cURYdWNQNjdaZGtDdFFhUk

Schumpeter, J. (1934). *The Theory of Economic Development.* Harvard Economic Studies.

Sdx Central. (2018). *GE Wants to Ditch Its Digital Assets.* Retrieved from sdx central: https://www.sdxcentral.com/articles/news/ge-wants-to-ditch-its-digital-assets/2018/07/

Shubham, J. (2018, July). *Ensemble Learning—Bagging and Boosting.* Retrieved from becominghuman.ai: https://becominghuman.ai/ensemble-learning-bagging-and-boosting-d20f38be9b1e

Skymind.ai. (2019, Mai). *Artificial Intelligence (AI) vs. Machine Learning vs. Deep Learning.* Retrieved from Skymind.ai: https://skymind.ai/wiki/ai-vs-machine-learning-vs-deep-learning

Snow, C., Fjeldstad, Ø. D., Langer, A., -, -, & -. (2017). Designing the digital organization. *Journal of Organization Design*, p. 6:7.

Statista. (2018). *Monthly number of Uber's active users worldwide from 2016 to 2019 (in millions).* Retrieved from Statista: https://www.statista.com/statistics/833743/us-users-ride-sharing-services/

Statista. (2019). *Number of daily active Facebook users worldwide as of 4th quarter 2018 (in millions).* Retrieved from Statista.com: https://www.statista.com/statistics/346167/facebook-global-dau/

Stefano, G. D., Gambardella, A., & Verona, G. (2012). *Technology push and demand pull perspectives in innovation studies: Current findings and future research directions.* Elsevier.

Strobl, C. G. (2017, May). *Update: Tesla's Fleet Learning.* Retrieved from Hackerbay Blog: https://blog.hackerbay.com/update-teslas-fleet-learning-8e34c3cd6ab4

Taylor, S., & Letham, B. (2017, February). *Prophet: forecasting at scale.* Retrieved from Facebook Research: https://research.fb.com/prophet-forecasting-at-scale/

Tesla Inc. (2019, February). *Autopilot.* Retrieved from Tesla: https://www.tesla.com/en_GB/autopilot

Tesla Inc. (2019, April). *Tesla Autonomy Day.* Retrieved from Youtube.com: https://youtu.be/Ucp0TTmvqOE

Tesla Inc. (2019, February). *Tesla Model 3.* Retrieved from Tesla Inc.: https://www.tesla.com/model3

The Street. (2015, June 3). *GE Expects Predix Software to Do for Factories What Apple's iOS Did for Cell Phones.* Retrieved from theStreet.com: https://www.thestreet.com/story/13174112/1/ge-expects-predix-software-to-do-for-factories-what-apples-ios-did-for-cell-phones.html

The Verge. (2016, October 19). All new Tesla cars now have hardware for 'full self-driving capabilities'. *The Verge.* Retrieved from The Verge: https://www.theverge.com/2016/10/19/13340938/tesla-autopilot-update-model-3-elon-musk-update

The Verge. (2016, Oct 19). *Tesla's new Autopilot will run in 'shadow mode' to prove that it's safer than human driving.* Retrieved from The Verge: https://www.theverge.com/2016/10/19/13341194/tesla-autopilot-shadow-mode-autonomous-regulations

The Verge. (2018, April 19). *How Tesla and Waymo are tackling a major problem for self-driving cars: Data.* Retrieved from The Verge: https://www.theverge.com/transportation/2018/4/19/17204044/tesla-waymo-self-driving-car-data-simulation

Tibbetts, M. (2017). *Machine Learning with Industrial Data.* Retrieved from Sintef: https://www.sintef.no/contentassets/4817fcaabf034bcfacadd625ac89498c/tibbetts_geilotalk.pdf

Trott, P. (2017). *Innovation Management and New Product Development (Sixth edition).* Harlow CM20 2JE, United Kingdom: Pearson Education Limited.

Uber. (2019). *Company Info.* Retrieved from Uber: https://www.uber.com/en-PK/newsroom/company-info/

Vercruyssen, V., Meert, W., & Davis, J. (2017). Transfer learning for time series anomaly detection.

Weiblen, T. (2014, July). The Open Business Model - Understanding an Emerging Concept . *Journal of Multi Business Model Innovation and Technology*, pp. 35-66.

Weill, P., & Woerner, S. (2015). Thriving in an Increasingly Digital Ecosystem. *MIT Sloan Management Review.*

Weill, P., & Woerner, S. (2018). *What's Your Digital Business Model?* Harvard Business Press.

West, J., Ventura, D., & Warnick, S. (2007). *Spring Research Presentation: A Theoretical Foundation for Inductive Transfer.* Brigham Young University, College of Physical and Mathematical Sciences.

Wikipedia. (2019). *J. P. Morgan.* Retrieved from Wikipedia: https://en.wikipedia.org/wiki/J._P._Morgan

Wikipedia. (2019, April). *Scientia potentia est.* Retrieved from Wikipedia: https://en.wikipedia.org/wiki/Scientia_potentia_est

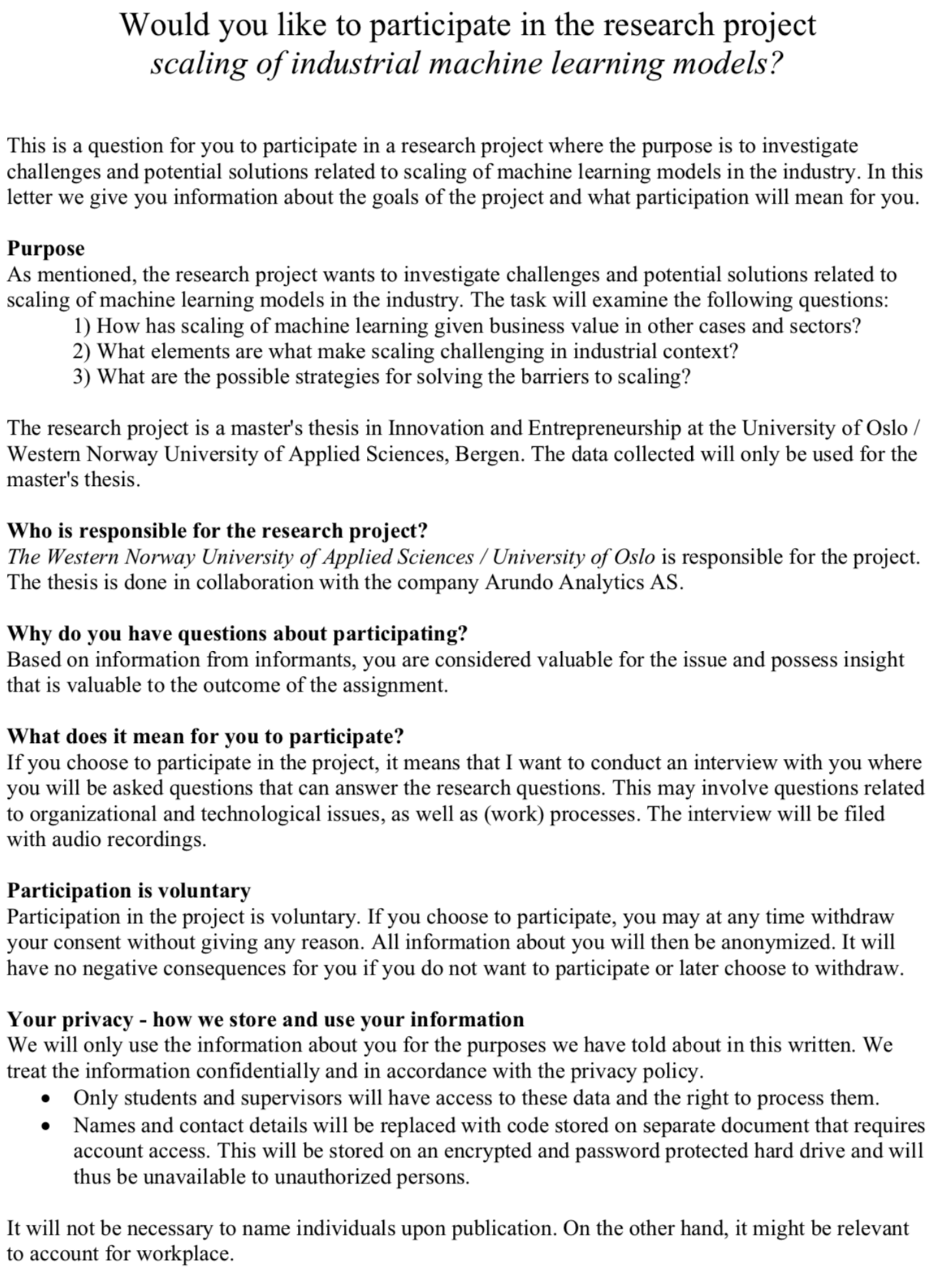
Wired. (2019, March). *HOW CAMBRIDGE ANALYTICA SPARKED THE GREAT PRIVACY AWAKENING.* Retrieved from Wired: https://www.wired.com/story/cambridge-analytica-facebook-privacy-awakening/

Yin, R. (2014). *Case Study Research Design and Methods.* Thousand Oaks, CA: Sage.

Zimek, A., & Schubert, E. (2017, September). Outlier Detection. *Encyclopedia of Database Systems*, pp. 1-5.

7 Appendix

Appendix 1. Consent statement



Would you like to participate in the research project *scaling of industrial machine learning models?*

This is a question for you to participate in a research project where the purpose is to investigate challenges and potential solutions related to scaling of machine learning models in the industry. In this letter we give you information about the goals of the project and what participation will mean for you.

Purpose
As mentioned, the research project wants to investigate challenges and potential solutions related to scaling of machine learning models in the industry. The task will examine the following questions:

1) How has scaling of machine learning given business value in other cases and sectors?
2) What elements are what make scaling challenging in industrial context?
3) What are the possible strategies for solving the barriers to scaling?

The research project is a master's thesis in Innovation and Entrepreneurship at the University of Oslo / Western Norway University of Applied Sciences, Bergen. The data collected will only be used for the master's thesis.

Who is responsible for the research project?
The Western Norway University of Applied Sciences / University of Oslo is responsible for the project. The thesis is done in collaboration with the company Arundo Analytics AS.

Why do you have questions about participating?
Based on information from informants, you are considered valuable for the issue and possess insight that is valuable to the outcome of the assignment.

What does it mean for you to participate?
If you choose to participate in the project, it means that I want to conduct an interview with you where you will be asked questions that can answer the research questions. This may involve questions related to organizational and technological issues, as well as (work) processes. The interview will be filed with audio recordings.

Participation is voluntary
Participation in the project is voluntary. If you choose to participate, you may at any time withdraw your consent without giving any reason. All information about you will then be anonymized. It will have no negative consequences for you if you do not want to participate or later choose to withdraw.

Your privacy - how we store and use your information
We will only use the information about you for the purposes we have told about in this written. We treat the information confidentially and in accordance with the privacy policy.

- Only students and supervisors will have access to these data and the right to process them.
- Names and contact details will be replaced with code stored on separate document that requires account access. This will be stored on an encrypted and password protected hard drive and will thus be unavailable to unauthorized persons.

It will not be necessary to name individuals upon publication. On the other hand, it might be relevant to account for workplace.

What happens to your information when we close the research project?
The project is scheduled to end on June 15. After this, the data will be destroyed and unavailable.

Your rights
As long as you can be identified in the data material, you are entitled to:
- insight into which personal information is registered about you,
- getting personal information about you,
- delete your personal information
- get a copy of your personal data (data portability), and
- to send a complaint to the Data Protection Officer or the Data Inspectorate about the processing of your personal data.

What gives us the right to process personal information about you?
We process information about you based on your consent.

On behalf of the *University of Oslo* and the *Western Norway University of Applied Sciences* (Dept. Bergen), NSD - The Norwegian Center for Research Data AS has considered that the processing of personal data in this project is in accordance with the privacy policy.

Where can I find out more?
If you have any questions about the study or would like to exercise your rights, please contact:
• Øystein Stavø Høvig (Oystein.Stavo.Hovig@hvl.no / 909 24 936) from Western Norway University of Applied Sciences
• Our privacy officer (personvernombud@hvl.no)
• NSD - Norwegian Center for Research Data AS, by email (personverntjenester@nsd.no) or telephone: 55 58 21 17.

With best regards

Øystein Stavø Høvig	*Steffen Novak Mollestad*
Project Manager (Researcher/tutor)	Student

Consent Statement

I have received and understood information about the project *scaling of industrial machine learning models*, and have been given the opportunity to ask questions. I agree to:

- to participate in an interview with sound recorder

I agree that my information is processed until the project is completed, approx. June 15

____
(Signed by participant, date)

Appendix 2. Interview guide

The interviews will largely be informal, and will therefore be characterized by a small common structure. This is because the research object is not persons, but a technical field that will uncover new discoveries along the way and the interviews will therefore have to be adapted and shaped along the way. Nevertheless, there will be certain things that will be consistent. This is described below.

Total interview time is estimated at approx. 30-45 min.
Phase 1: Loose talk (1 min)

- Informal chat to reassure the situation and set the framework

Phase 2: Frames for the interview (1 min)

- Inform about consent statement and purpose with study.
- Define the Master's thesis;
  - The concept of scaling: Industrial mass application of machine learning models
- Inform about why I want to interview the informant
- Allow room for questions

Phase 3: Personalization and competence (3 min)

- Name of informant
- Position / function
- Expertise in subjects

Phase 4: Professional questioning (20-35 min)
(This section depends on where in the project one is)

- From your perspective, what are the biggest obstacles to being able to scale up users of ML in industrial context?
  - Specific for; Technological, people and organizational, business characteristics?
- Which processes are used most time?
  - How do you see that it is possible to improve these processes?
- Technology: To what extent is data a problem for scaling from your perspective?
- *(Based on findings along the way in the project, these findings have also been discussed with the respondents)*

Phase 5: Summary (5 min)

- Summarize and validate answers
  - Have I understood you correctly?
- Anything else you want to add?

Appendix 3. Case material

Machine learning at Facebook

Specific regarding types of ML models Facebook sees the different types of models as such:

Specific ML models at Facebook	**Description**
Logistic Regression (LR)	efficient to train and use for prediction
Support Vector Machines (SVM)	
Gradient Boosted Decision Trees (GBDT)	can improve accuracy at the expense of additional computing resources
Deep Neural Networks (DNN)	the most expressive, potentially providing the most accuracy, but utilizing the most resources

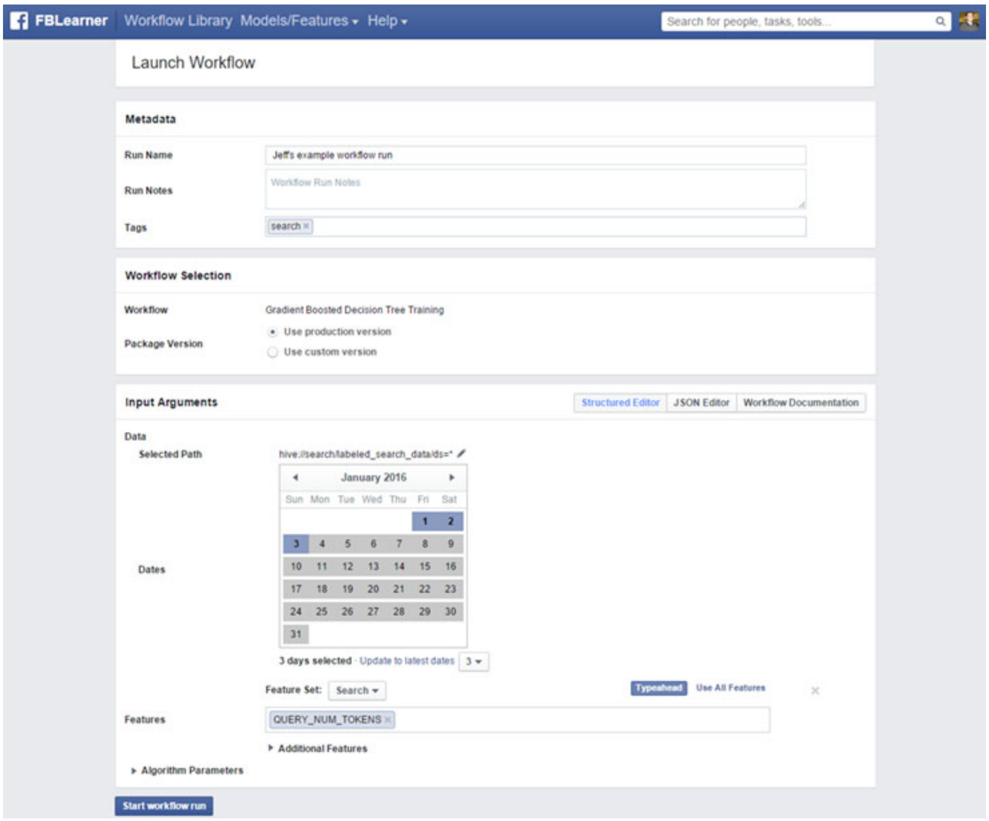
Below is also a list of how the various ML models are applied into various cases.

Models	***Services***
Support Vector Machines (SVM)	*Facer (User Matching)*
Gradient Boosted Decision Trees (GBDT)	*Sigma*
Multi-Layer Perceptron (MLP)	*Ads, News Feed, Search, Sigma*
Convolutional Neural Networks (CNN)	*Lumos, Facer (Feature Extraction)*
Recurrent Neural Networks (RNN)	*Text Understanding, Translation, Speech Recognition*

Frequency, duration, and resources used by offline training for various workloads (Hazelwood, et al., 2018):

Service	**Resource**	**Training Frequency**	**Training Duration**
News Feed	Dual-Socket CPUs	Daily	Many Hours
Facer	GPUs + Single-Socket CPUs	Every N Photos	Few Seconds
Lumos	GPUs	Multi-Monthly	Many Hours
Search	Vertical Dependent	Hourly	Few Hours
Language Translation	GPUs	Weekly	Days
Sigma	Dual-Socket CPUs	Sub-Daily	Few Hours
Speech Recognition	GPUs	Weekly	Many Hours

Screenshots of Facebook's MLaaS



Launch Workflow

Metadata

Run Name: Jeff's example workflow run

Run Notes: Workflow Run Notes

Tags: search ×

Workflow Selection

Workflow: Gradient Boosted Decision Tree Training

Package Version: ● Use production version
○ Use custom version

Input Arguments

Data

Selected Path: hive://search/labeled_search_data/ds=*

Dates:

◂			January 2016			▸
Sun	Mon	Tue	Wed	Thu	Fri	Sat
					1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30
31						

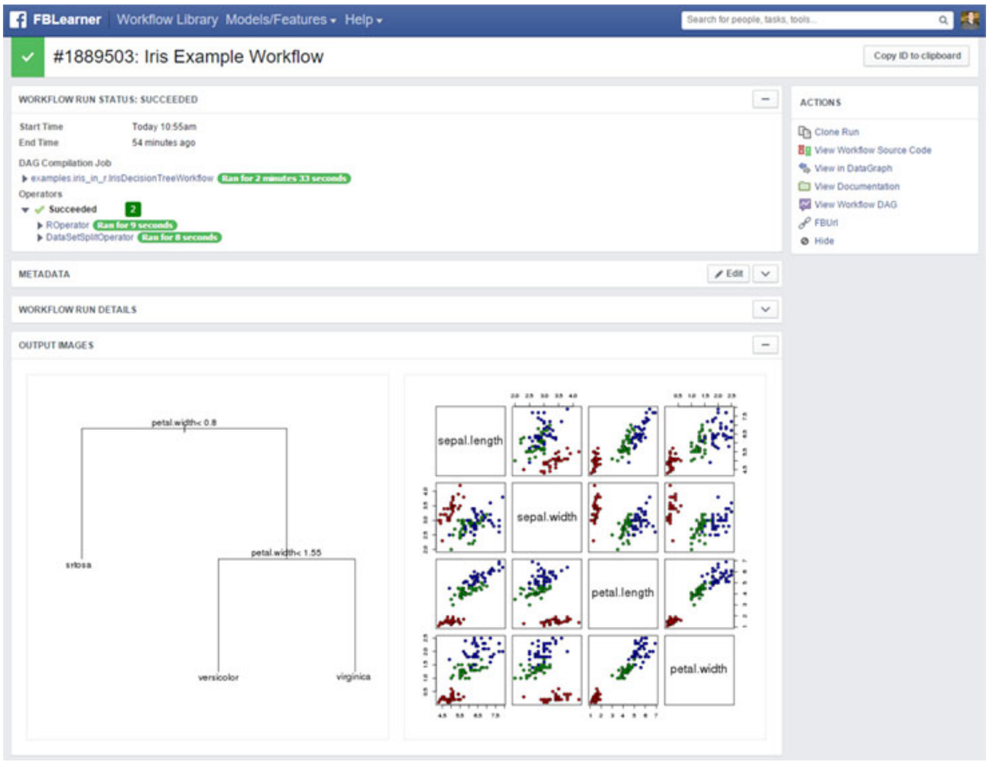
3 days selected · Update to latest dates 3 ▾

Feature Set: Search ▾

Features: QUERY_NUM_TOKENS ×

▸ Additional Features

▸ Algorithm Parameters



✓ #1889503: Iris Example Workflow

WORKFLOW RUN STATUS: SUCCEEDED

Start Time: Today 10:55am
End Time: 54 minutes ago

DAG Compilation Job
▸ examples.iris_in_r.IrisDecisionTreeWorkflow Ran for 2 minutes 33 seconds

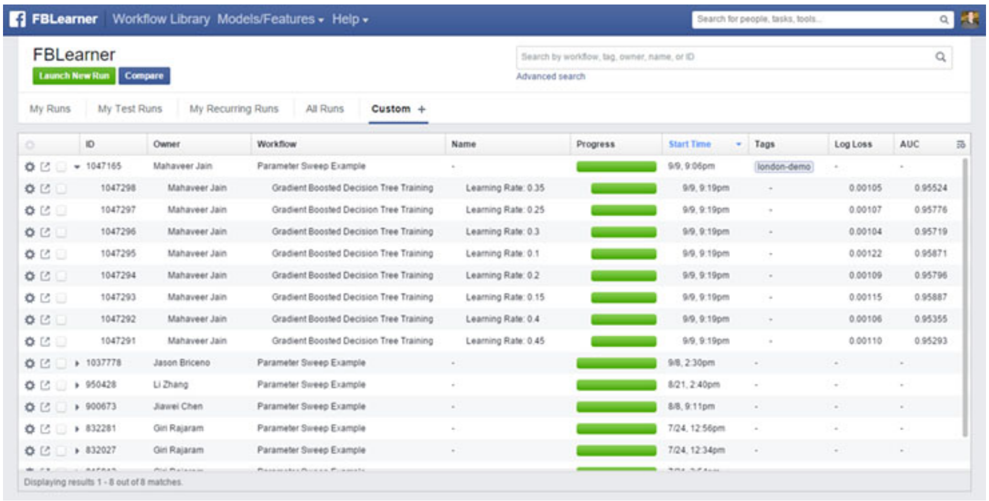
Operators
▾ ✓ Succeeded 2
▸ ROperator Ran for 9 seconds
▸ DataSetSplitOperator Ran for 8 seconds

ACTIONS

METADATA

WORKFLOW RUN DETAILS

OUTPUT IMAGES



ID	Owner	Workflow	Name	Progress	Start Time	Tags	Log Loss	AUC
1047165	Mahaveer Jain	Parameter Sweep Example	-	🟩	9/9, 9:06pm	london-demo	-	-
1047298	Mahaveer Jain	Gradient Boosted Decision Tree Training	Learning Rate: 0.35	🟩	9/9, 9:19pm	-	0.00105	0.95524
1047297	Mahaveer Jain	Gradient Boosted Decision Tree Training	Learning Rate: 0.25	🟩	9/9, 9:19pm	-	0.00107	0.95776
1047296	Mahaveer Jain	Gradient Boosted Decision Tree Training	Learning Rate: 0.3	🟩	9/9, 9:19pm	-	0.00104	0.95719
1047295	Mahaveer Jain	Gradient Boosted Decision Tree Training	Learning Rate: 0.1	🟩	9/9, 9:19pm	-	0.00122	0.95871
1047294	Mahaveer Jain	Gradient Boosted Decision Tree Training	Learning Rate: 0.2	🟩	9/9, 9:19pm	-	0.00109	0.95796
1047293	Mahaveer Jain	Gradient Boosted Decision Tree Training	Learning Rate: 0.15	🟩	9/9, 9:19pm	-	0.00115	0.95887
1047292	Mahaveer Jain	Gradient Boosted Decision Tree Training	Learning Rate: 0.4	🟩	9/9, 9:19pm	-	0.00106	0.95355
1047291	Mahaveer Jain	Gradient Boosted Decision Tree Training	Learning Rate: 0.45	🟩	9/9, 9:19pm	-	0.00110	0.95293
1037778	Jason Briceno	Parameter Sweep Example	-	🟩	9/8, 2:30pm	-	-	-
950428	Li Zhang	Parameter Sweep Example	-	🟩	8/21, 2:40pm	-	-	-
900673	Jiawei Chen	Parameter Sweep Example	-	🟩	8/8, 9:11pm	-	-	-
832281	Giri Rajaram	Parameter Sweep Example	-	🟩	7/24, 12:56pm	-	-	-
832027	Giri Rajaram	Parameter Sweep Example	-	🟩	7/24, 12:34pm	-	-	-

Displaying results 1 - 8 out of 8 matches.

https://code.fb.com/core-data/introducing-fblearner-flow-facebook-s-ai-backbone/ (May, 2016)

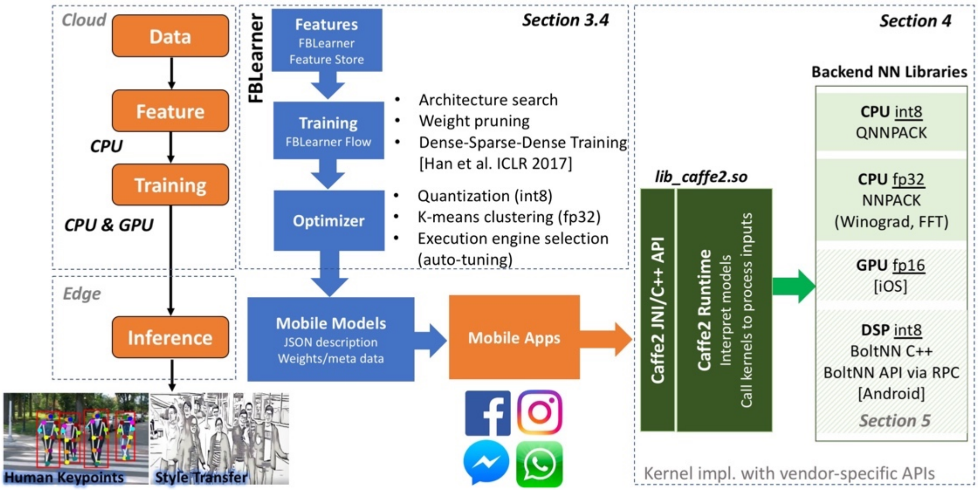
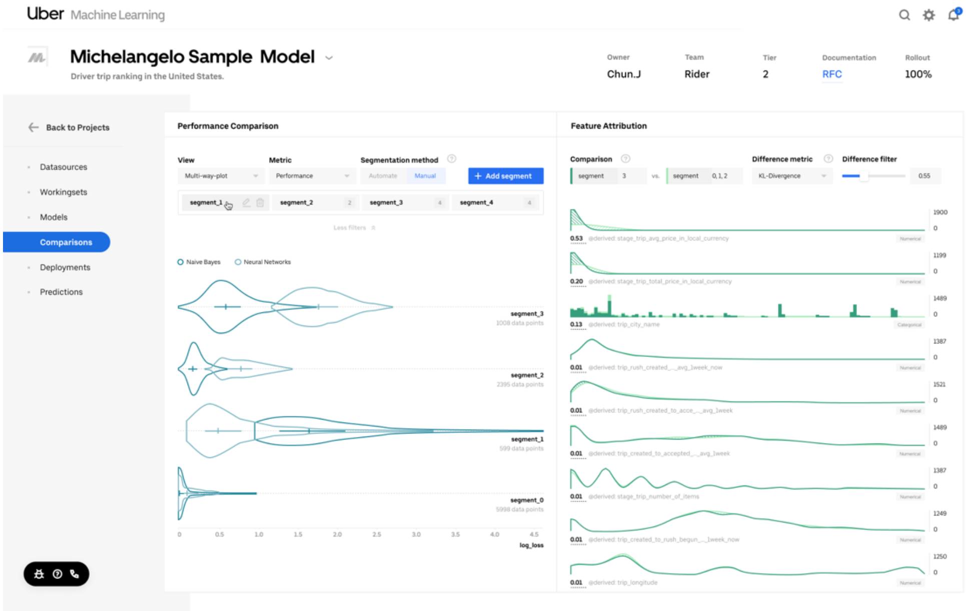


Figure 7.1. **Execution flow of Facebook's machine learning for mobile inference**. (Hazelwood, et al., 2019)

Uber ML platform



Owner: Chun.J

Team: Rider

Tier: 2

Documentation: RFC

Rollout: 100%

View: Multi-way-plot

Metric: Performance

Segmentation method: Automate Manual

segment_1 segment_2 2 segment_3 4 segment_4 4

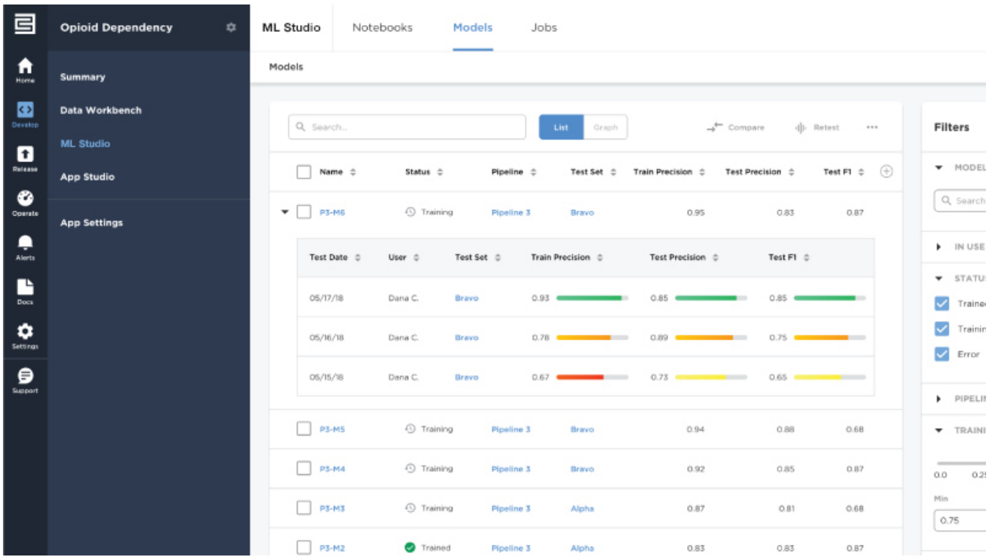
Comparison: segment 3 vs. segment 0,1,2

Difference metric: KL-Divergence

Difference filter: 0.55

Figure 7.2. **Michelangelo's model comparison page** showing a comparison of two models' behavior across different segments and features. (Hermann & Balso, 2018)

C3 ML platform



☑ Traine

☑ Trainin

☑ Error

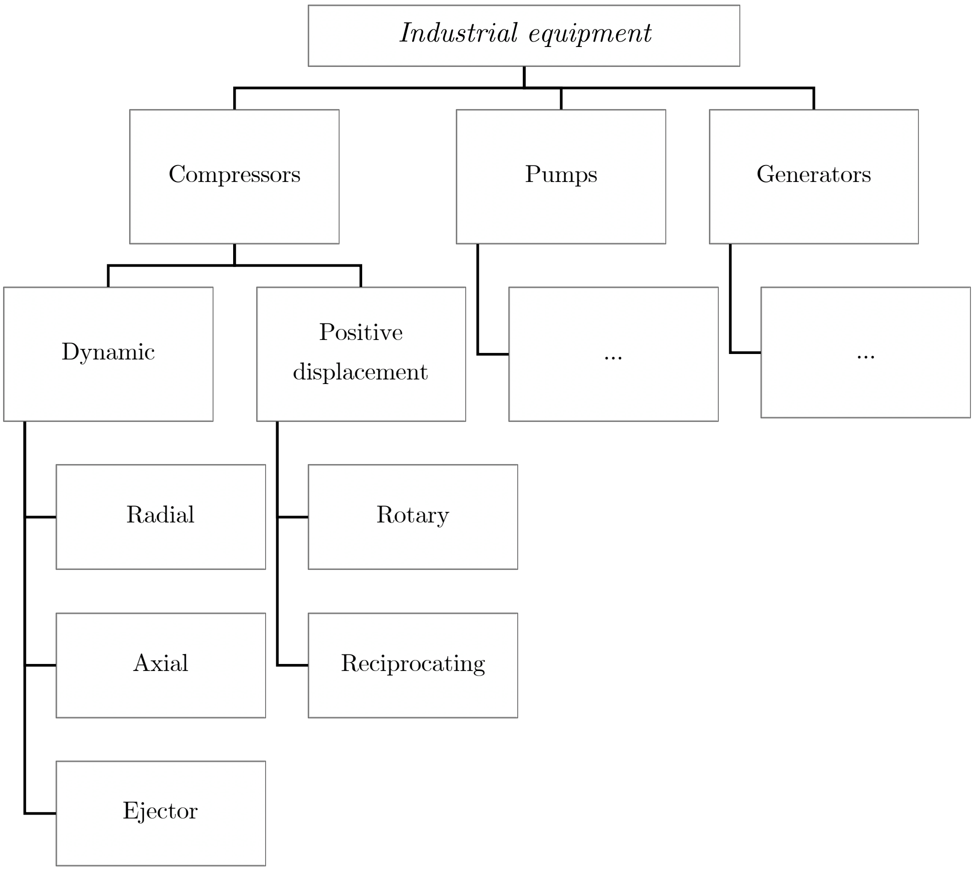
Min: 0.75

□ Name	Status	Pipeline	Test Set	Train Precision	Test Precision	Test F1
□ P3-M6	Training	Pipeline 3	Bravo	0.95	0.83	0.87
□ P3-M5	Training	Pipeline 3	Bravo	0.94	0.88	0.68
□ P3-M4	Training	Pipeline 3	Bravo	0.92	0.85	0.87
□ P3-M3	Training	Pipeline 3	Alpha	0.87	0.81	0.68
□ P3-M2	✓ Trained	Pipeline 3	Alpha	0.83	0.83	0.87

Test Date	User	Test Set	Train Precision	Test Precision	Test F1
05/17/18	Dana C.	Bravo	0.93	0.85	0.85
05/16/18	Dana C.	Bravo	0.78	0.89	0.75
05/15/18	Dana C.	Bravo	0.67	0.73	0.65

Appendix 4. Dissimilarities of industrial assets

The characteristics of an industrial equipment varies greatly, even just for the equipment groups, such as illustrated for compressors.



Appendix 5. Machine learning in oil and gas

Hajizadeh (2018) – with the oil and gas industry as reference – describes several conditions (using SWOT analysis) for the industry to apply ML. *Strengths* being the amount of raw data and expert knowledge available, and *opportunities* being hardware acceleration, transfer learning, Continuous Integration/Continuous Deployment, automated machine learning, and IoT and Edge analytics. As *weakness* Hajizadeh argues with technology laggard, waterfall model, lack of industry-wide collaborations, and availability of labeled and high-quality data. *Threats* being lack of ML strategy, oil price swings, resistance to change, safety and security, hire and retain ML talent, and technology stack.