CHALMERS
UNIVERSITY OF TECHNOLOGY

MARITIME DIGITALISATION: AUTOMATED DOCUMENT CLASSIFICATION

Interdisciplinary study on possible machine learning solutions for the container shipping industry

Master of Science Thesis in the Master's Degree Programme, Maritime Management

CARL BLOMSTRÖM
RODRIGO ASTORGA CASTILLO

Department of Mechanics & Maritime (M2)
Department of Electrical Engineering (E2)
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2018
Master's Thesis EX082/2018

MASTER'S THESIS EX082/2018

MARITIME DIGITALISATION: AUTOMATED DOCUMENT CLASSIFICATION

INTERDISCIPLINARY STUDY ON POSSIBLE MACHINE LEARNING SOLUTIONS FOR THE CONTAINER SHIPPING INDUSTRY

CARL BLOMSTRÖM

RODRIGO ASTORGA CASTILLO

CHALMERS
UNIVERSITY OF TECHNOLOGY

Department of Mechanics & Maritime (M2)
Department of Electrical Engineering (E2)
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2018

Maritime digitalisation: Automated document classification
Interdisciplinary study on possible machine learning solutions for the container shipping industry
CARL BLOMSTRÖM
RODRIGO ASTORGA CASTILLO

Supervisors: Henk Wymeersch, Department of Electrical Engineering (E2)
Olle Lindmark, Department of Mechanics & Maritime Sciences (M2)
Examiner: Henk Wymeersch, Department of Electrical Engineering (E2)

Master's Thesis EX082/2018
Department of Mechanics & Maritime Sciences (M2)
Department of Electrical Engineering (E2)
Chalmers University of Technology
SE-412 96 Gothenburg
Sweden
Telephone: + 46 (0)31-772 1000

Printed by Chalmers University of Technology
Gothenburg, Sweden 2018

Maritime digitalisation: Automated document classification
Interdisciplinary study on possible machine learning solutions for the container shipping industry

CARL BLOMSTRÖM
RODRIGO ASTORGA CASTILLO
Department of Mechanics & Maritime Sciences (M2)
Department of Electrical Engineering (E2)
Chalmers University of Technology

Abstract

Thesis report presents a conducted interdisciplinary study between the fields of Maritime Management and Electrical Engineering. The aim is to research how and to which degree machine learning algorithms can be applied within the area of maritime industry. The purpose is to find potential ways to innovate current burdensome administrative tasks, particularly within the freight forwarding and container shipping industry, through suggested methods of automated document classification. The study consisted of a mixed research method, with the double diamond approach as chosen design thinking model. The authors discuss possible issues identified in previous attempts to find similar solutions and research on the subject, in particular, related to organisational and cultural barriers, concluding that currently, available technology offers good improvement opportunities of many back-office activities.

Keywords: Management, Shipping, Digitalisation, Machine learning, Strategy.

Acknowledgements

The authors would like to take the opportunity in thanking everyone who has been involved in this thesis, whether through interviews or by providing guidance and motivation, without mentioning any names in particular.

Also, the authors would like to thank Chalmers Ventures and Chalmers Innovationskontor for their great feedback concerning how to formulate the identified business case.

A special thanks go out to current employers for being helpful and understanding regarding the need for time to focus on completing this master's thesis.

Finally, but not least, the authors would like to thank the supervisors Henk Wymeersch and Olle Lindmark who have provided guidance, inspiration, feedback and support throughout the entire journey.

Carl Blomström & Rodrigo Astorga Castillo, Gothenburg, June 2018

Contents

List of Figures

Abbreviations

AI	Artificial Intelligence
API	Application programming interface
BL	Bill of Lading
CNN	Convolutional Neural Network
DL	Deep Learning
FCL	Full Container Load
LCL	Less than Container Load
ML	Machine Learning
NLP	Natural Language Processing
NVOCCs	Non-Vessel Operating Common Carriers (e.g. forwarders)
OCR	Optical Character Recognition
PDF	Portable Document Format
TEU	Twenty-foot Equivalent Unit
VOCCs	Vessel Operating Common Carriers (e.g. carriers)

1. Introduction

This section will introduce the thesis by providing the background and describe the formalities of the report.

1.1 Background

Machine Learning, abbreviated as ML, has become a substantial technological trend where computers are now learning to complete tasks without being directly programmed to do so. Compared to traditional programming methods where system rules are identified and encoded by humans as if; or else; functions, ML algorithms automatically create system rules based on input data which is enabling digital applications to be more advanced than ever before. (Sebastiani, 2002)

Today, industries are at the beginning of the fourth industrial revolution, or industry 4.0, where adoption of digital solutions based on ML and AI will play a significant role for increased business efficiency (World Economic Forum, 2016). Development in previously disjointed fields, such as AI, ML and blockchain is starting to make its way into different industries where it is believed that enterprises across all sections will soon find smart systems at the heart of decision making, core business processes and automation (Gesing et al., 2018).

It is generally viewed that the maritime industry is both a conservative and reactive industry when it comes to the application of new technology, which could result in potential benefits being missed (Thakur, 2016). It was recently highlighted in a McKinsey consulting report that shipping companies have abundant opportunities to increase current performance in commercial, operation, network and fleet activities by adopting digital solutions and algorithms to make better and more effective decisions (Glave et al., 2014). According to industry representatives, ML will be crucial for future business and success, where investments in predictive analytics and data scientists and engineers to produce reliable, repeatable decisions will eventually have to be made (Kuik, personal communication, 2018-01-10).

The transportation of goods is generally characterised by requiring an awful lot of documentation. According to UNCTAD's review of maritime transport, in particular of the container shipping sector, an estimated 140 million TEU's were transported in 2016 (UNCTAD, 2017) where each of these container shipments requires a set of different types of shipping documents in order to be processed. Shipping documents for commercial trade include at least three mandatory documents; a Bill of Lading, abbreviated as BL, a commercial invoice and a packing list, where in addition to this, several other types of documents may be required depending on commodity and country of origin. This means that in addition to 140 million containers being transported every year, there are additional 420 million documents that must be handled and processed in back-office operations.

Document handling within container shipping is a heavy administrative burden which is still mainly being conducted by manual handling. Handling procedures are highly standardised throughout the entire world, which makes it a relatively simple task. Today, most companies within the maritime industry get their flexibility and capabilities from humans. However, humans have an impact on quality in both time and accuracy, as well as relatively high variable costs. There is a significant amount of ‘trapped value' in this industry, as employees across the globe are stuck with manual administrative processes and large piles of documentation related to every shipment. Therefore, in high volume low margin businesses, automation and digitalisation are of utmost importance and are key enablers for future success (Kuik, personal communication, 2018-01-10).

1.2 Aim

The overall aim of this thesis is to investigate if ML can be applied within the container shipping industry.

Another aim is to gain understanding on how ML algorithms can be utilized within the maritime industry, where this interdisciplinary thesis combines the knowledge from the different sciences.

The final, and preliminary, aim of this report is to create a prototype of a system that can automatically classify and handle documents frequently used in container shipping.

1.3 Purpose

The main purpose is to assess if the container shipping industry is mature for the technical development that is ML and industry 4.0.

Another purpose is to reduce the gap between maritime and technology, where the authors are confident that combined forces can result in beneficial synergies for all parties involved.

1.4 Thesis statement

It is hypothesised that it is more beneficial for container shipping companies to deploy ML algorithms to handle administrative tasks compared to current manual processing.

1.5 Research questions

The research conducted for this thesis seeks to answer the following:

1. Which ML algorithms and additional tools are most suitable for creating an automated document handling and classification system for container shipping?
2. What is the potential for these tools in terms of quantifiable performance metrics compared to human labour?
3. What are the barriers to implement ML solutions in the container shipping industry?

1.6 Delimitations

Specifications on requirements for a fully automated system in terms of programming and integrating such a system into already existing corporate IT infrastructure is outside of the scope of this thesis.

Classification testing will only be performed on three main documents used for container shipping; BL, commercial invoice and packing list. All other documents that may be required for special purpose shipments will be outside of the scope of this thesis.

The thesis will only focus on the container shipping industry as it is outside of the scope of this thesis to make comparisons between this and other maritime sectors.

The geographical limitation is supposed to focus on the global container market. However, due to large part of the data will be gathered and collected from within the EU, Scandinavia in particular, it is worth bearing in mind that some details could differ compared to other parts of the world.

Both the thesis and the development of the model will be performed without funding from external parties.

1.7 Thesis outline

This interdisciplinary report will research the topic of ML and investigate to which degree this technology can be applied to the container shipping sector. The investigation will focus on how ML can be used to automate a part of the document handling, mainly by assisting the process of sorting documents, i.e. document classification. The accuracy of the ML system will then be evaluated and compared regarding time and quality for manual labour, where conclusions on the potential benefits or disadvantages of such a system will be made. As the maritime industry is commonly known for its low levels of technology adoption, the topic of digitalisation will be covered, where the opportunities and threats that come with this development will be covered as well.

It will not necessarily help by just knowing programming, computer science and electrical engineering to solve real-life issues, which is why an ocean freight domain expertise and industry perspective on the

applicability of ML solutions on specific tasks is needed. As the authors are ocean freight experts currently working for the maritime industry, end-to-end process knowledge will help to enhance the outcome of this report further. The thesis will also include a qualitative section on digitalisation within the industry where current trends and barriers are analysed. By investigating how ML algorithms can assist with providing automated document handling models within the maritime industry, in particular by automated document classification, potential business improvements and cost saving can thus be identified.

2. Theory

The theory section consists of three parts. First out is the container shipping industry where the topic of maritime logistics, container shipping, current practices and processes will be covered. Second is technology, where ML and relevant knowledge related to the thesis topic will be covered. The last part consists of a business section, where diffusion of innovation and strategic aspects to digitalisation will be covered.

2.1 Industry

Shipping is the life and blood of the global economy and accounts for around 90 percent of international trade (UNCTAD, 2017). Transporting cargo by sea bring many benefits compared to other modes of transportation in terms of cost and environmental sustainability due to ships ability to utilise economy of scale (Lee and Meng, 2015). The shipping industry consists of different segments which are mainly divided into container, dry bulk and tanker (Song and Panayides, 2015).

2.1.1 Container shipping

Container shipping is a relatively new mode of shipping that originated in the 1950's with the introduction of the standardised steel shipping container and the first ever container ship (Lee and Meng, 2015). Container transport is often referred to in terms of TEUs, Twenty-foot Equivalent Units, where a standard 20-foot container is referred to as one TEU, and a 40-foot container is referred to as two TEUs. During the last few decades, container shipping has seen impressive growth from 28,7 million TEUs transported in 1990, to 140 million TEUs transported in 2016 (UNCTAD, 2017). During the same time, container throughput in ports increased from 88 million TEUs in 1990 to 686 million TEUs in 2016, indicating that a container is on average unloaded and loaded five times between the first port of loading and the last port of discharge (UNCTAD, 2017). Containers are unloaded and loaded in transit as a result of complex liner service networks with demand on global supply chains in terms of frequency, direct accessibility and transit times. The container traffic is the absolute number of containers being carried by sea, while the container throughput is the number of times a container is handled in port, e.g. discharged and loaded (Song and Panayides, 2015).

While container shipping occupies a relatively small share of the global maritime fleet, 12 percent (Song and Panayides, 2015), it is the fastest growing sector (UNCTAD, 2017) with revenues of 2,6 trillion dollars a year (The Economist, 2018). Container shipping currently concentrates more than half of the worlds trade value (Song and Panayides, 2015).

Container shipping is part of maritime logistics, which is a strategically crucial part of the global logistics system and is defined as the process of planning, implementing and managing the movement of cargo and information which is involved in ocean carriage (Song and Panayides, 2015). Since containerisation has resulted in marine transport being integrated with land transport by the provision of intermodal transports, maritime logistics has also come to include land-based logistic services as well, along with other value-adding logistic services such as stripping, stuffing, storage, warehousing, inventory management, supply chain management, distribution centers, quality control, testing, assembly, packaging, repairing and inland connections among others. Intermodal transports are defined as door-to-door operations using two or more modes of transportation in an integrated manner, where at least one leg consists of ocean transport. The primary goal of maritime logistics is to minimise cost while maximising customer satisfaction by providing ocean carriage as well as value-adding logistics services (Song and Panayides, 2015).

Container trade is as of today dominated by business-to-business models and integrated supply chains where demand for value-adding logistics services and the integration of intermodal or multimodal transports are becoming increasingly important for business stability and future revenue (Song and Panayides, 2015). Increase globalisation and the geographic separation between supply and demand has increased the expectation on transport services where shippers and consignees have become more demanding with respect to the quality of transport services. Users of container transport services expect fast and reliable services, covering a vast geographic network at a competitive cost. This has resulted in the sequential growth of carriers and freight forwarders, either organic or through mergers and acquisitions where considerable investments in new vessels and service expansions are being made (Song and Panayides, 2015).

While container shipping has increased around twice the rate of the global GDP between 1985 and 2007 (The Economist, 2018), and increased volumes from 28,7 million TEUs transported in 1990, to 140 million TEUs transported in 2016 (UNCTAD, 2017), the operational principles of the industry have not changed much since taking its current form in the 1950s and 1960s (The Economist, 2018). The container shipping industry has been slow in digitalising and adopting new technology, and widespread scepticism remains about the viability of such initiatives (Alphaliner, 2018)

2.1.2 Container transport providers

Container shipping involves many parties where carriers and freight forwarders are the primary providers of ocean transport services for customers (Song and Panayides, 2015). Carriers which are often referred to as VOCCs, Vessel Operating Common Carriers, both own and operate their vessels and provide transport and other value-adding logistic services to customers. Freight forwarders, often referred to as NVOCCs, Non-Vessel Operating Common Carriers, do not own their vessels and act as freight traders, cargo consolidators and intermediaries between the shipper and carrier (Berglund, 2017). Since freight forwarders arrange ocean freight for many customers, they are able to buy transport services from carriers at beneficial rates. Whether the cargo owner decides to arrange container transport through a carrier or a freight forwarder usually comes down to the marketing mix four P's; price, product, place and promotion (Song and Panayides, 2015). It is generally said that large organisations with high volumes are able to arrange more beneficial contracts with the carrier directly, while small to medium-sized organisations with lower volumes will get their best offer through a freight forwarder.

In addition to the carrier and the freight forwarder, there are other parties involved in transporting a container from point A to point B. Terminal operators, trucking companies, banks, insurers and customs authorities, all play essential roles in the global supply chain network where they fill functions mandatory to international ocean freight (Song and Panayides, 2015). Involved parties for a general container shipment as well as the information exchange required between parties is illustrated in Figure 1, used with permission by DHL.

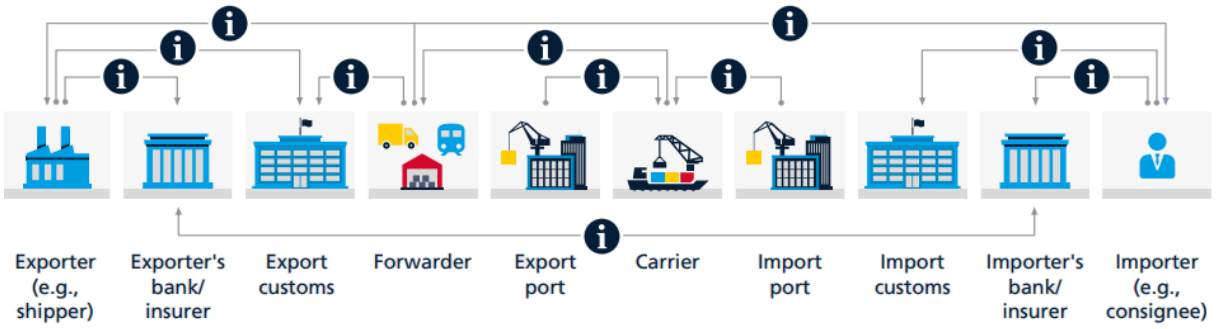

Figure 1: Transporting containers, involved parties and information flow (DHL Trend Research, 2018)

2.1.3 Import process

When a container reaches its destination, it is unloaded in a container terminal which is a controlled customs zone. Before the container may leave the terminal to be transported into the country, there are various obligations that must be fulfilled (Lee and Meng, 2015).

2.1.3.1 Bill of lading

A BL is the single most important document for all ocean shipments as it contains all relevant information regarding the shipment such as shipper, consignee, origin, destination, payment terms, freight operator along with the terms and conditions for the shipment (Lee and Meng, 2015). The BL also serve as a contract

of carriage where all the terms for the transport service are expressed (Lee and Meng, 2015). According to Black's Law Dictionary (1968), the legal definition of a BL is the following (see Appendix A.1):

> *"A bill of lading is an instrument in writing, signed by a carrier or his agent, describing the freight so as to identify it, stating the name of the consignor, the terms of the contract for carriage, and agreeing or directing that the freight be delivered to the order or assigns of a specified person at a specified place. [...] It is receipt for goods, contract for their carriage, and is documentary evidence of title to goods [.]"*

The BL in its current form dates back to the 1700's and is a document of title to the goods, giving the holder of the BL the right to claim ownership to the cargo that is stated therein (Lee and Meng, 2015). In Mason v Lickbarrow, [1794] 5 TR 693, 101 ER 380, KB, an 18th century English court case, in which the fundamental aspects of the BL, its negotiability and the legal process concerning the passing of property were determined, the following statement was made concerning entitlement to the goods (see Appendix A.2a):

> *"It cannot indeed be disputed but that, as between the assignee and the indorsee [i.e. the person to whom the negotiable document is endorsed], the indorsement of a bill of lading is a complete transfer of the property which the consignee has in it [.]"*

The BL is also a tradeable document which means that the ownership of the cargo can be sold multiple times during transport (Lee and Meng, 2015). While this is unusual in container shipping, bulk and tanker shipments are frequently sold during transport where the cargo may have had several different owners during the time of transportation (Lee and Meng, 2015). According to the opinion of Ashhurst J, ruling judge in Mason v Lickbarrow, he stated the following (see Appendix A.2b.i-ii):

> *"But, as between the vendor and third persons, the delivery of a bill of lading is a delivery of the goods themselves [...] Though the bill of lading in this case was at first indorsed in blank, it is precisely the same as if it had been originally indorsed to this person; for when it was filled up with his name, it was the same as if made to him only."*

According to the opinion of Grose J, ruling judge in Mason v Lickbarrow, he stated the following concerning the monetary value of a BL (see Appendix A.2c.i-ii):

> *"[The] bill of lading transfers the property. [...] A bill of lading carries credit with it; the consignor by his indorsement gives credit to the bill of lading, and on the faith of that money is advanced. "*

Lastly, the BL is not issued until the cargo is loaded on the vessel. Therefore, it is also a conclusive receipt and an acknowledgement that the goods have been loaded onboard the vessel (Lee and Meng, 2015). This follows from the opinion of Buller J, ruling judge in Mason v Lickbarrow, were he stated the following (see Appendix A.2d):

> *[A] bill of lading is an acknowledgment by the captain, of having received the goods on board his ship [.]"*

The Swedish regulations on BLs, translated as "konossement", are mainly stated in the 13th chapter of the Swedish Maritime Code, i.e. "Sjölag (1994:1009)", concerning "carriage of general cargo", which deals with BLs and other transport documents, in paragraphs 3, 9, 19, and 42-59. The regulations are also stated in the 14th chapter, which deals with the chartering of ships, particularly in paragraphs 2, 5, 15, 18, 27, 50, 62-63, and 71. Chapter 13, which are based on the Hague, Hague-Visby and Hamburg Rules, is to a large extent mandatory law whereas the rules in chapter 14 can (and most often is) be set aside by agreement between the parties. Finally, the regulations are also stated in the sixth section, concerning the final provisions of the code, in the nineteenth chapter, which deals with the statutes of limitation of certain claims, more specifically in paragraph 1.

The BL must be received in original before the goods can be released from the port (Lee and Meng, 2015). The carrier issues the BL as a master BL and by the freight forwarders as a house BL in the country of export where the shipper signs it before being carried to the country of destination (Lee and Meng, 2015). In the country of destination, it is received by the consignee who signs the BL before sending it to the freight forwarder or carrier for release of the cargo (Lee and Meng, 2015). Cargo may never be released without having received the original BL for the shipment, and the carrier or freight forwarder is liable for compensating the cargo owner for the full value of the cargo if it was to be released without the BL being received (Lee and Meng, 2015).

While original BLs are still being used to a wide extent, they are causing various issues in an increasingly digital business environment. Due to this fact, the maritime industry has introduced express, or telex release, BLs which may be sent and received in digital format. The main difference between the original and express BL, beside the easier administration, is that the express bill is non-negotiable and cannot be traded. However, it must still be received, stamped, signed and scanned by the included parties (Lee and Meng, 2015).

2.1.3.2 Customs clearance

The next step is to get the container customs cleared. Customs clearance must be performed by a licenced party, which is usually the carrier or the freight forwarder (Lee and Meng, 2015). For the customs clearance the BL, along with all other commercial documents, such as commercial invoice, packing list, certificates and other mandatory documents for specific origin countries, must be accounted for (Tullverket, 2018). In the process of customs clearance, the customs authority will clear the container after assessing risk based on the type of cargo and after making sure that the correct toll and tax fees have been applied (Tullverket, 2018). All documentation used for the purpose of the customs clearance, supporting documentation, must also be archived by the party who submitted the customs declaration to the customs authority, for a period of five years (Tullverket, 2018). Various other regulations, such as the Swedish Book-keeping Act, i.e. "Bokföringslag (1999:1078)", also provide regulations on how the responsible party must archive transport document and supporting documents for shipments.

2.1.3.3 Release from the terminal

When the cargo has been cleared for customs and released by the carrier or the freight forwarder, it is the terminal operator who is responsible for managing the release of the container to leave the customs controlled zone and be transported into the country (Lee and Meng, 2015). As a result of this, the terminal has an elaborate process of receiving the correct information and documentation from involved parties, along with an identification procedure for transport operators in which they must verify whom they are when arriving in port for picking up the container. The container is only released to the operator who can provide the right codes for pickup along with the correct documentation (Lee and Meng, 2015).

2.1.3.4 Document requirements

The required shipping documents are issued by many of the included parties. The commercial documents such as packing list and commercial invoice for the cargo is issued by the shipper, the seller of the cargo, who performs a transaction with the consignee, the buyer. BLs are then issued by the carrier as a master BL and by the freight forwarder as house BL. Certificates, if necessary, can be issued by the seller to prove the origin of the cargo or by the terminal operator who is stuffing the container where they assure that cargo is correctly stuffed and fumigated. Insurance documents, if applicable, are issued by either an insurance company, the carrier or the freight forwarder (Lee and Meng, 2015).

All the required documentation must then be collected by the party who performs the export declaration at origin and by the party who perform the import custom declaration at the destination. This is usually the freight forwarder or the carrier (Lee and Meng, 2015).

2.1.4 Administrative work

Due to a complex system, many import containers spend more prolonged time than necessary in terminals due to missing documents (Lee and Meng, 2015). This also causes extra cost for cargo owners in form of demurrage and detention as well as problems for shipping companies due to lack of empty containers for new shipment bookings (Lee and Meng, 2015). Moving goods across borders has an estimated annual cost of 1,8 trillion USD, where documentation and administration is estimated to make up one fifth of these costs (McKevitt, 2018). While some operators apply document management systems or EDI for transferring documents between different parties involved in the supply chain, most document exchange is still conducted through email and fax where the documents are scanned before being sent between different parties as a hard copy PDF, Portable Document Format, files (The Economist, 2018)

2.2 Technology

This section will cover the technology related to the topic of this thesis and is divided into four subcategories consisting of mathematical and statistical framework, ML, Classification algorithms and related topics. A brief introduction of each section is provided under each heading.

2.2.1 Mathematical and statistical framework

As ML is based on mathematical and statistical frameworks, a short introduction on these topics are provided under this section. This section seeks to explain the underlying principles of ML and enables enhanced understanding of section 2.2.2. ML and 2.2.3 Classification algorithms.

2.2.1.1 General statistics

Statistics, as a body of knowledge within applied mathematics, usually refers to calculated quantities taken from sample data (Hayslett and Murphy, 1995). This is often selected by a stochastic process, i.e. being random, where the term 'observations' is more appropriately applied to the actual 'facts and figures' (Hayslett and Murphy, 1995). The goal of any data analysis is to extract an accurate estimation from raw information (Alexopoulos, 2010), and in the words of Dr. Wright (2003):

> *"Conducting data analysis is like drinking a fine wine. It is important to swirl and sniff the wine, to unpack the complex bouquet and to appreciate the experience. Gulping the wine doesn't work".*

It is possible to say that there are two kinds of statistics, the first being the numerical description of data, i.e. descriptive statistics, and the second being the process of reaching conclusions about the sample population, i.e. inferential statistics (Hayslett and Murphy, 1995). In other words, it is the science of making the least incorrect decision based on uncertainty and incomplete information (Hayslett and Murphy, 1995). Rather than justifying doubtful assumptions and choosing samples based on already made conclusions, it is crucial for the researcher to not fall into the dishonest practice of using statistics "*as a drunken man uses lamp-posts,* [which is] *for support rather than for illumination*", as stated by the 19th century Scottish novelist Andrew Lang (Hayslett and Murphy, 1995).

During the process of analysing, it is essential to organise the data into a meaningful form so that the researcher can get an overview more conveniently, which is commonly done by constructing frequency distributions (Manikandan, 2011). It is to structure the data into tables or other graphical representation of the sample population in categories, with the purpose of showing whether the observations are concentrated in one area or rather if they are spread out across the measurement scale (Manikandan, 2011). In other words, this will illuminate the uncertainties, thus preparing the path for the analyst to find the most probable solution to the given statistical problem, mainly by making it possible to recognise patterns in the data. Robert (2007) points out that the main purpose of statistical theory is to make inferences based on observing random phenomenon and to interpret, rather than explaining, the probability distribution underlying this phenomenon.

2.2.1.2 Probability and likelihood

Probability and likelihood tie into what is known as probability theory, which according to Bishop (2006) "provides a consistent framework for the quantification and manipulation of uncertainty", and when combined with decision theory it allows for optimal decision-making, i.e. confident predictions, even if the available information is incomplete or ambiguous. In short, it is the measuring of both the probability and likelihood of an event occurring out of a total number of trials, implicitly considering an infinite amount of trials as a limit, i.e. $N \to \infty$ (Bishop, 2006).

The main distinction between probability and likelihood briefly explained according to Bishop (2006), is that the "probability" is a numerical value that expresses how probable an observed outcome is, given a set of characteristics of the sample population distribution, i.e. parameter values. This numerical value is then regarded as the "likelihood" of the set of parameter values or characteristics, given the observed outcomes (Bishop, 2006). Another important fact stated by Bishop (2006), is that the probabilities for the events in question to occur must by definition lie in the interval [0,1], and if they are mutually exclusive as well as including all possible outcomes, they must also sum to one.

To better understand the rules of probability, it is possible to exemplify in more general terms by involving two random variables (see Appendix B), X and Y forming a two-dimensional array, i.e. a matrix. A variable, or variate, can, on the one hand, be continuous, i.e. quantities that can take any numerical value within a certain range, like in the case of heights or weights (Bali, Goyal, and Watkins, 2007). It can also be discrete, which can be both numerical and categorical, but is distinguished in the sense that it can only take particular values, e.g. the names of a group of persons or the number of objects (Bali, Goyal, and Watkins, 2007). In short, it is a quantity that is not constant as the name implies, but rather it is an assigned quantity, which means it has a fixed place of storage in the data memory (Sinclair, 1990). Variables are for the most part numbers or strings in nature and have a particular place of storage with a unique label (Sinclair, 1991).

2.2.1.3 Linear regression

In statistics, linear regression describes the relationship between one dependent outcome variable (y) and a set of independent predictor or explanatory variables, i.e. covariates or "regressors" $(x_1, x_2, \ldots, x_p)$, the letter p denoting the variable as being a predictor (Hosmer et al., 2013). This is used to find the statistical model that fits the best with the input and which output is easily interpretable (Hosmer et al., 2013). In other words, it is the study of the conditional distribution $(y)|(x)$, given the $p * 1$ vector of the independent variable (x), using continuous and discrete variables, depending on the assumptions made (Olive, 2017). The model describes an approximation of a state or process, and according to Levins (1966), unlike a scientific hypothesis, it is neither a hypothesis or theory nor is it verifiable directly by an experiment. In other words, he states that (Levins, 1966):

> *"for all models of true or false, the validation of a model is not that it is 'true' but that it generates good testable hypotheses relevant to important problems",*

and as the famous quote by Box (1979) goes:

> *"All models are wrong, but some are useful".*

As previously mentioned, inferential statistics is used by default to answer particular questions about the data, to test hypotheses, to describe specific associations, i.e. correlations, or to model the relationship within the data (Alexopoulos, 2010). The main definition of correlation is if the change in one variable affects a change in the other (Bali, Goyal, and Watkins, 2007) (see Appendix C).

There are three types of regression as stated by Yan and Gang Su (2009): firstly, the simple linear regression; secondly the multiple linear regression which involves more issues such as e.g. detection of regression outlier; and thirdly the nonlinear regression which assumes that (y) is not linear in regression parameters and is a bit more complicated than the two other models. Briefly put, the model consists of a function of regressors added with a random error (Yan and Gang Su, 2009).

The first type is expressed by $y = \beta_0 + \beta_1 * x + \varepsilon$, where ε denotes the random error ε, β_0 being the y intercept, and β_1 is the gradient or slope of the regression line (Yan and Gang Su, 2009). This expression assumes that ε is normally distributed with the expectation $E(\varepsilon) = 0$ and a constant variance $Var(\varepsilon) = \sigma^2$ (Yan and Gang Su, 2009). The second type is expressed by $y = \beta_0 + \beta_1 * x_1 +, \ldots, + \beta_p x_p + \varepsilon$, where $\beta_0, \beta_1, \ldots, \beta_p$ are regression coefficients (Yan and Gang Su, 2009). The third and final type is expressed by $y = \frac{\alpha}{1+e^{\beta t}} + \varepsilon$, where y is a nonlinear growth as a function of time t, and α and β being the model parameters (Yan and Gang Su, 2009).

As previously mentioned, the purpose of doing regression analysis is to investigate and establish a causal relationship between the response variable (y) and the set of regressors (x_p). This way, the response variable (y) is predicted based on a $p * 1$ vector of predictor variables $x = (x_1, x_2, \ldots, x_p)^T$, were x^T is the transpose of x (Olive, 2017). By screening them, it is possible to identify which are more important, relatively speaking, to explain (y), in order to determine the relationship in a more optimised manner (Yan and Gang Su, 2009). In the case of multiple linear regression, these are special cases of the linear regression model which can be jointly presented by defining the population model in terms of the sufficient predictor $SP = \beta^T * x$, and the estimated model, $ESP = \hat{\beta}^T * x$, which is in terms of the estimated sufficient predictor, i.e. ESP (Olive, 2017).

It is important to make a proper initial assessment of the scientific problem, making sure that it falls into the scope of regression analysis, as well as making distributional assumptions on the random error in the model (Yan and Gang Su, 2009). Thus, the model can use the assumptions as a basis when solving the problem and even to draw statistical conclusions (Yan and Gang Su, 2009). When collecting data, apart from verifying and testing the assumptions, there are a number of things to have in mind so that the model is applicable in the desired real-life scenario, like data cleaning, experimental design, sample size determination (Yan and Gang Su, 2009). Different diagnosis methods may be used in order to evaluate the appropriateness of the model to answer the scientific questions, its level of performance and if it is in need of refinement (Yan and Gang Su, 2009).

The sample correlation coefficient r will only show if there exists a linear relationship, but not specifically what the relationship is, in such a way where it is possible to predict values for any of the variables x and/or y (Hayslett and Murphy, 1995). In other words, to recognise a pattern. A difference between a correlation problem and a regression problem is that the x variable is not randomly picked, whereby the researcher

conducting the experiment is expected to choose the values based on the given problem that is being solved (Hayslett and Murphy, 1995).

2.2.1.4 Logistic regression

In contrast to linear regression, the outcome variable (y) in a logistic regression model is binary or dichotomous, meaning the outcome is categorically partitioned and belonging either to one of two parts (Hosmer et al., 2013). To further understand this distinction, Kleinbaum and Klein (2002) present an illustration of a multivariable problem using a typical example taken from the epidemiologic research where researchers want to evaluate the extent to which a disease is associated with a specific type of exposure.

The example in question uses the exposure variable (E), like smoking status, classified as either "yes" or "no", with a set of control variables $\{c_1, c_2, \ldots, c_n\}$ like age, sex and ethnicity, in order to describe or predict the dependent variable (D), which in this case is a disease outcome (Kleinbaum and Klein, 2002). The dichotomous disease outcome will be represented by $(y) = 0$, meaning evidence of disease is absent in the individual subject, respectively $(y) = 1$, meaning evidence is present (Hosmer et al., 2013).

Generally speaking, the logistic regression model differs to the linear regression model, in how the assumptions are made and how the model is chosen (Hosmer et al., 2013). Apart from this, the same general principles apply which are used in linear regression, thus motivating the approach to logistic regression analysis through the techniques used in linear regression (Hosmer et al., 2013). The logistic function, called $f(z)$ is given by the following expression, $f(z) = \frac{1}{1+e^{-z}}$, which has the range $0 \leq f(z) \leq 1$, and the variable z which has the range $-\infty < z < +\infty$ (Kleinbaum and Klein, 2002). Because the denominator consists of the sum of $1 + e$, i.e. the natural logarithm, to negative z, when $z = -\infty$, then $f(-\infty) = 0$, respectively when $z = \infty$, then $f(\infty) = 1$(Kleinbaum and Klein, 2002).

This is one of the main reasons for why this model is popularly used, as the probability estimates will be rather simple to understand as it will always be a particular value between 0 and 1 (Kleinbaum and Klein, 2002). The other reason for it being popularly used is the S-shape of the logistic function, which can be seen in Figure 2, which makes it easy for the experimenter to see whether the z value is positive or negative (Kleinbaum and Klein, 2002). Thus, this function is also called the "sigmoid function", as the term "sigmoid" means S-shaped (Bishop, 2006).

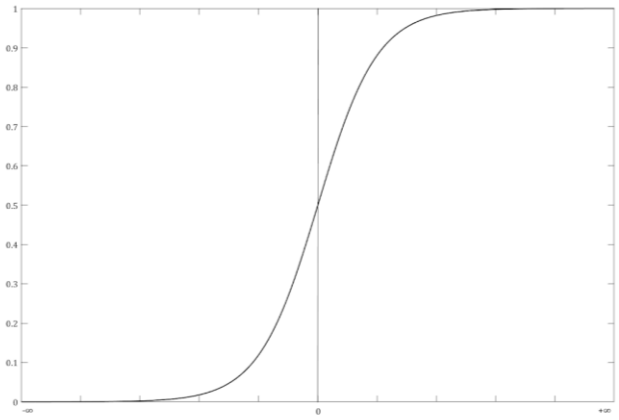

Figure 2: Graphical representation of the logistic "sigmoid" function

To further elaborate on this model, Kleinbaum and Klein (2002) continues with the example taken from epidemiologic research, in which the z variable represents an index which combines many contributions of several risk factors, i.e. multivariable, were $f(z)$ represents the actual risk for each value of z. This means that the graphical representation of the function, resembling the letter S, indicates that the level of risk for the individual in question to develop a disease, is low until reaching a certain threshold, after which it rises rapidly, until surpassing a specific range of intermediate values of z, as illustrated in Figure 2 (Kleinbaum and Klein, 2002). The value z consists of the linear sum $\alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots \beta_k x_k$, where k is the number of considered variables. The model would ultimately be expressed as $f(z) = \frac{1}{1+e^{-(\alpha+\beta_i x_i)}}$, were x_i are independent variables of interest, for $i = 1, \ldots, k$, and α and $\beta_i$being constant terms representing unknown parameters (Kleinbaum and Klein, 2002). Thus, the probability model can be defined as a nonlinear logistic

regression model and can be denoted by the conditional probability statement $P(D = 1|x_1, x_2, \ldots, x_k)$ (Kleinbaum and Klein, 2002).

2.2.2 Machine learning

This section seeks to explain the fundamentals of ML and AI. The topics of learning techniques, algorithm training and algorithm evaluation are covered before presenting recent years rise to ML 2.0 with increased usability and accessibility. Finally, the concept of ML pipelines and fully automated systems are explained.

2.2.2.1 Machine learning

ML, abbreviated as ML, is a field in computer science where existing data is used to predict or respond to future data (Paluszek and Thomas, 2017). Compared to traditional methods of giving computers detailed instructions on how to respond in a possible scenario, ML algorithms are data driven, giving them the ability to learn and adapt based on the input data (Chapmann, 2017), as illustrated in Figure 3, which was inspired by Jason Brownlee at Machinelearningmastery.com.

Figure 3: Traditional programming (coding) vs ML algorithms, inspired by Jason Brownlee

ML is becoming increasingly important for system development where it is not feasible, or sometimes even not possible, to write manual algorithms to perform a specific task (Paluszek and Thomas, 2017). ML algorithms can be used for many different purposes, to obtain insights, recognise patterns and make predictions from data and are as of today successfully applied in many different fields, where autonomous driving, finance prediction, medical diagnoses and text and speech recognition are just a few examples (Paluszek and Thomas, 2017). To understand how and where ML algorithms can be useful and applied it is important to understand the underlying principles and techniques, essentially how ML algorithms work (Paluszek and Thomas, 2017).

ML belong to the area of autonomous learning where systems learn without human intervention, and it is closely related to the fields of pattern recognition and computational statistics (Paluszek and Thomas, 2017). The term ML was originally coined in 1959 by Arthur Lee Samuel where it was defined as "*the field of study that gives computers the ability to learn without being explicitly programmed*" (Samuel, 2000). In an article for IBM Journal of Research And Development, computer scientist Arthur Samuels wrote about his studies on ML using the game of checkers, where he managed to verify that computers can be programmed so that they will learn to play a game better than the person who wrote the program, in a remarkably short period of time (Samuel, 2000). A more modern and more technical definition is provided for the term ML as well by Tom Mitchell (1997) where he states the following:

> *"[a] computer program is said to learn for experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E".*

Mohammed et al. (2017) also quote Professor Tom Mitchell, by restating an important point related to the topic of ML being an intersection of Computer Science and Statistics:

> *"Whereas Computer Science has focused primarily on how to manually program computers, Machine Learning focuses on the question of how to get computers to program themselves [...] Whereas Statistics has focused [...] on what conclusions can be inferred from data, Machine Learning incorporates additional question about [what algorithms can be most effectively used for]".*

All learning algorithms are data-driven, where often large amounts of data are required to train a learning algorithm (Paluszek and Thomas, 2017). Data sets may be collected by humans or by control systems to be used for training purposes. Learning systems are based on models, where each model provides a mathematical framework for learning (Paluszek and Thomas, 2017). Depending on the desired output of the learning system, different models are applied. Most traditional ML models are often human-derived and

based on human observation and experiences. However, more modern ML algorithms develop own models without a human-derived structure (Paluszek and Thomas, 2017).

2.2.2.2 Machine learning and Artificial Intelligence

ML and Deep Learning, abbreviated as DL, are key components in the development of artificial intelligence (Gesing et al., 2018). While artificial intelligence is often oversimplified in the popular press as one unified technology, this is not the case. Generally, AI applications consist of sensing, processing and learning components (Gesing et al., 2018), where ML represents the latter. In order to process the real world and take in information, AI systems are heavily relying on sensing components where real-world information, such as text, sound, video or images can be captured and digested in order to be processed and used by learning components (Gesing et al., 2018).

DL is a subcategory of ML where many state-of-the-art and human level performance algorithms are being developed (Géron, 2017). DL is a result of deep neural networks and has gained huge momentum since 2006, when Geoffrey Hinton et al. published a paper titled "A Fast Learning Algorithm for Deep Belief Nets", showing that deep neural networks were capable of recognizing handwritten digits with state-of-the-art accuracy over 98 percent compared to human accuracy which is roughly 95 percent (Géron, 2017). The relation between AI, ML and DL is illustrated in Figure 4, used with permission by Gesing et al.

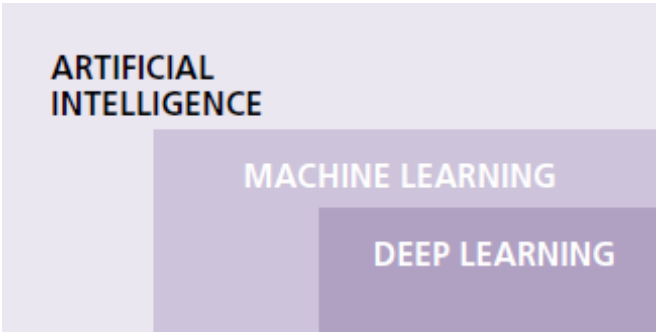

Figure 4: ML in the context of AI and DL (Gesing et al., 2018)

2.2.3 Learning techniques

ML algorithms are generally divided into one of two broad categories, supervised or unsupervised learning (Chapmann, 2017). Modern ML methods consist of many scenarios that transcend the classic but simple supervised and unsupervised learning techniques, such as reinforcement learning or semi-supervised learning (Witten et al., 2017). However, it is deemed that reinforcement learning and semi-supervised learning are outside of the scope of this thesis. A general overview of the main branches of ML and their general areas of application is illustrated in Figure 5, which was inspired by Vishakha Jha at TechLeer.com.

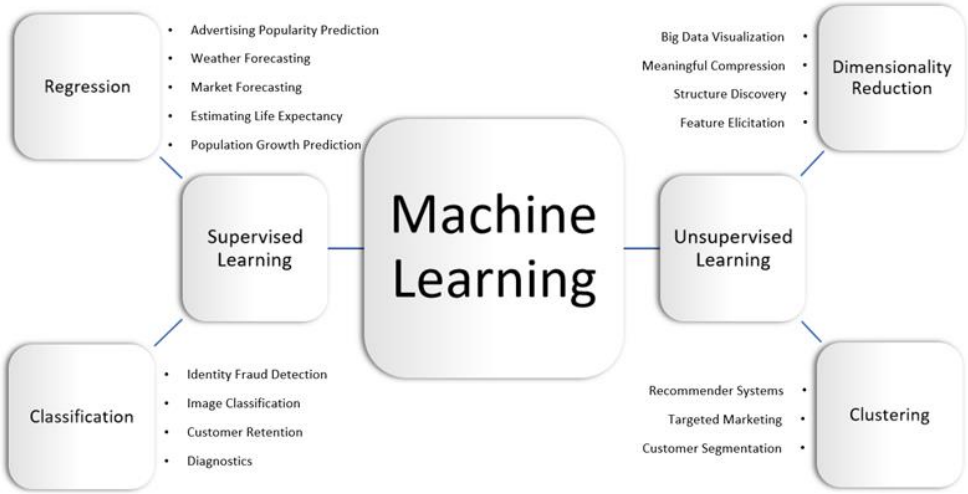

Figure 5: General overview of ML, inspired by Vishakha Jha

2.2.3.1 Supervised Learning

In supervised learning, the learning algorithm is provided with the correct output label to all input data (Ng, 2018). This gives the algorithm the ability to map a relationship between the input vector (x) and output vector (y) and develop a hypothesis as to which parameters are present when a specific input generates a specific output. The training is supervised in the sense that the training sets are human derived and labelled by a human supervisor, where the process of classifying the output for the input data is what is referred to as labelling (Paluszek and Thomas, 2017). The goal of supervised learning is for the algorithm to approximate the mapping function so well so that new input data (x) can be predicted to have the output variable (y). If enough training data is provided, the algorithm should learn how to produce correct outputs when brand new inputs are introduced (Paluszek and Thomas, 2017). Supervised learning problems are mainly divided into classification or regression problems (Ng, 2018).

Classification problems are when an algorithm is given the task of predicting a class for a specific input data. Compared to regression where the output is a continuous value, classification algorithms provide a discrete value output which may be a simple as true or false (Ng, 2018). While the simplest classification tasks may result in a binary classification, multiclass classification algorithms are able to predict several different outcomes of k classes (Paluszek and Thomas, 2017). Classification algorithms may be used for many different purposes where a decision must be made based on available data. As previously mentioned, a common classification example exists within healthcare where patient data could be used to decide if a patient needs further treatment or not (Ng, 2018). A multiclass classification problem example is document classification, where the learning algorithm could classify a set of documents into k different classes.

The goal of regression algorithms is to map input variables to a continuous function so that a continuous output (y) can be presented for a new input data (x) (Ng, 2018). Regression problems mainly apply linear or polynomial regression functions rather than logistic regression as in the case of classification (Ng, 2018). Typical examples for regression problems are found within real estate where a house or apartment is used as input data. The algorithm may then predict a market value for the estate based on a chosen value of input parameters (Ng, 2018).

There are many different supervised ML algorithms that each have own specific areas of application, strength and weaknesses. Common examples of supervised learning algorithms are linear regression, logistic regression, naïve Bayes, support vector machines, decision trees, neural networks (Ng, 2018).

2.2.3.2 Unsupervised Learning

In unsupervised learning, the learning algorithm is provided input data for which there is no correct output (Paluszek and Thomas, 2017). Unsupervised learning problems can be classified into clustering and association problems where it is often used to find patterns in data or segment a large amount of data into different groups (Chapmann, 2017). In this sense, unsupervised learning does not apply training and validation sets as in supervised learning, and the input data does not come with the label Y (Paluszek and Thomas, 2017). The main goal of unsupervised learning is to model the underlying structure or distribution of the data. With enough training data, the unsupervised learning model can be developed into a strategic tool which can find the correct result of any problem (Chapmann, 2017).

Patterns in data may be useful for many different reasons where the main advantage of unsupervised learning is that algorithms may provide insights that might not have been known in advance (Chapmann, 2017). Unsupervised learning can also reveal hidden patterns in data and detect anomalies as the algorithms are able to process large amounts of data in a relatively short period of time (Paluszek and Thomas, 2017). Unsupervised learning has proved very useful in many different scenarios such as character, voice and face recognition where it is very time-consuming to label all the outputs manually (Chapmann, 2017). In character recognition, an unsupervised learning algorithm may be fed with a massive data set consisting of different handwritten characters, and rather than labelling the (y) parameter for each data input, the learning algorithm performs clustering of all input that shares similar features and cluster all the same words or letters into one single category. Unsupervised learning is also useful for anomaly detection, where a learning algorithm can provide information if a certain input data deviates from the rest (Chapmann, 2017).

There are many different unsupervised ML algorithms that each have their specific areas of application, strength and weaknesses. A few common examples of unsupervised learning algorithms are cluster analysis, k-means, and neural networks (Chapmann, 2017).

2.2.4 Algorithms

The algorithm is a sequence of computational steps that transform a set of values, i.e. the input, into some value or another set of values (Cormen et al., 2009). It is said to be correct if it solves the given computational problem, meaning it halts for every input sequence, i.e. input instance, in the correct output (Cormen et al., 2009). As described by Cormen et al. (2009), it could be described in the following way:

> *"An algorithm can be specified [...] as a computer program, or even as a hardware design. The only requirement is that the specification must provide a precise description of the computational procedure to be followed".*

2.2.4.1 Algorithm training

A sorting problem is formally defined as per below (Cormen et al., 2009):

Input: A sequence of n numbers $< a_1, a_2, \ldots, a_n >$.

Output: A permutation (reordering) $< a_1', a_2', \ldots, a_n' >$ of the input sequence such

that $a_1' \leq a_2' \leq a_n'$.

The reason for why it is vital to possess certain knowledge regarding algorithms and how they are devised is that different algorithms differ in how efficiently they perform the computations (Cormen et al., 2009). In an example regarding two algorithms for the task of sorting, Cormen et al. (2009) describe how a slower computer B can outperform a faster computer A by running more than 17 times faster, by choosing one algorithm instead of another.

ML algorithms are data-driven and need to be trained on data to generate good results (Paluszek and Thomas, 2017). A general rule within ML is that more training data often result in better learning parameters (Ng, 2018). While this is not always true, famous quotes such as one stated by Banko and Brill (2001), remain cited to this day:

> *"It's not who has the best algorithm that wins, it's who has the most data"*

Problems may arise when the algorithm is not provided with enough training data or if the training data is not sufficiently diverse since small training sets make it more challenging for the algorithm to discover the patterns and rules (Paluszek and Thomas, 2017).

Gathering and preparing the input data for ML algorithms is a time-consuming effort (Witten et al., 2017). Many instances show that real data is often low in quality and needs to be cleaned before it can be applied as input to a learning algorithm (Witten et al., 2017). This has resulted in the term data cleaning which is the process of cleaning data before it is used in the learning algorithm and is often performed by humans who carefully check the input data for anomalies (Witten et al., 2017). Data cleaning is an important step for generating a well-performing ML system (Witten et al., 2017).

Once the input data has been gathered and cleaned, a large part of the job is already done (Witten et al., 2017). For training purposes, it is essential to split the data into different sets. This is done because a learning algorithm would perform unrealistically well if it were to be evaluated on using the same data as it has been trained on (Witten et al., 2017). The original data is split into two different sets, where one set is used for training, and the other is used for evaluation (Witten et al., 2017). The training data set may then again be divided into two different sets, one for training and one for validation, as presented in Figure 6, inspired by Mostafa Eissa at CodeProject.com.

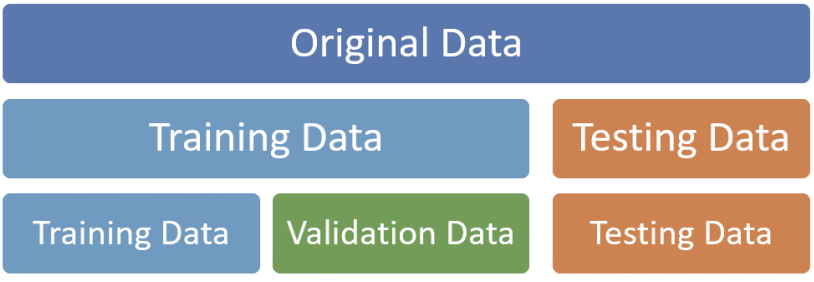

Figure 6: Training, validation and testing data split, inspired by Mostafa Eissa

The validation set, sometimes also referred to as the cross-validation set, plays an important role (Witten et al., 2017). Rather than training a model on training data and then evaluating its performance on the test data, the model is trained on training data and evaluated on the valuation data. This is known as the "holdout method", where the testing data is held out from the ML algorithm until it has been trained, tested and tweaked for optimal performance (Witten et al., 2017). If the same test data is repeatedly reused for evaluation during algorithm tuning and algorithm selection, it eventually becomes part of the training data. A crucial step when training ML algorithms are therefore to evaluate performance on data which the model has not seen before (Witten et al., 2017). According to Witten et al., (2017), it is only then, when providing brand new input data, that it is reasonable to evaluate if the learning parameters are good or bad as the objective is to make sure that the model performs well on new, previously unseen, data (Witten et al., 2017).

How much of the original data that goes into each set may vary, but a usual split among ML practitioners is 70/30 where 70 percent of the data is used for training, and 30 percent of the data is used for testing (Witten et al., 2017). It is important to keep the training set as big as possible since more data generally generate better learning parameters and improve the performance of the learning algorithm (Chapmann, 2017).

When splitting data sets, it is also necessary to maintain any class proportions that may be present (Witten et al., 2017). This is done so that one class is not underrepresented in the training data when that same class is overrepresented in the test data, as this may cause problems with performance (Witten et al., 2017). A method to prevent biased data sets is called stratification, where the data is randomly sampled in a way that allows each class to be adequately represented in the different sets. While stratification is generally worth doing, it only provides a low safeguard against biased representation (Witten et al., 2017).

2.2.4.2 Algorithm evaluation

Algorithm evaluation is a critical step for making progress in ML (Witten et al., 2017). Just because an algorithm is generating outstanding results on the test set, does not mean that the algorithm has developed a hypothesis that works well for generalising new input data. Algorithm evaluation provides a systematic way to evaluate how well different algorithms work and also allows for comparison between different learning algorithms (Witten et al., 2017).

All possible outcomes of a classification algorithm are presented in Figure 7, where the green outcomes are correct and the red outcomes incorrect. If the algorithm predicts $(y) = Pos$ correctly for input data, a true positive is generated. If the algorithm makes the same prediction for input data where the actual $(y) = Neg$, a false positive is generated. The same model applies for true and false negatives (Ng, 2018).

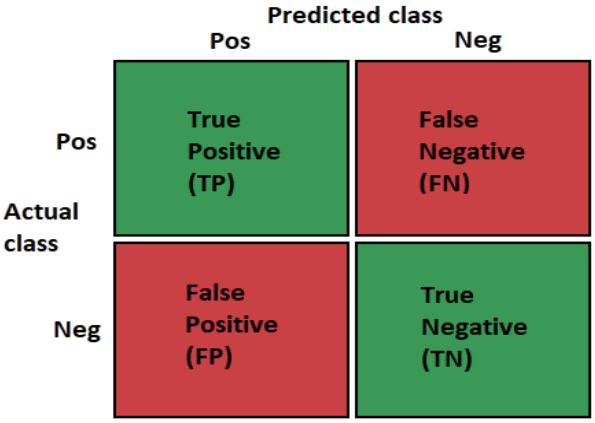

Figure 7: Prediction outcome matrix

Classification algorithms mainly measure performance in terms of accuracy (Witten et al., 2017). Accuracy is the proportion of successful classifications compared to errors made over a whole data set. Some authors prefer the term error rate, where instead the proportion of errors is compared to the successes over the whole set (Witten et al., 2017). Accuracy is calculated as following:

$$\text{Accuracy} = \frac{tp+tn}{tp+tn+fp+fn}$$

The simplest way to calculate algorithm accuracy is to train the algorithm on the training set and measure accuracy on the validation and test set, where the error rate on the test set will provide the most accurate estimation of the real algorithm performance (Witten et al., 2017).

However, accuracy can be calculated in different ways. A popular method to calculate the accuracy of a learning algorithm is to use stratified ten-fold cross-validation (Witten et al., 2017). Many tests have shown that ten-fold cross-validation is a good way of calculating with high accuracy (Witten et al., 2017). When performing ten-fold cross-validation, the data is then randomly divided into ten parts. For the method to work, it is crucial that the data be stratified, so that class representations remain similar to the total data set. The algorithm is then trained on nine out of ten sets, and the tenth set is used for calculating accuracy. The procedure is then repeated ten times so that each of the ten data sets serves as a test set on one occasion. After the accuracy has been then calculated for each holdout set, the accuracy is averaged to generate an overall error estimate for the learning algorithm (Witten et al., 2017).

Even though accuracy is a popular feature to calculate for classification problems, it does not give insight as to why the errors are occurring (Ng, 2018). The next natural step is error analysis, where the reason for failure is investigated. The most common problem for a learning algorithm is over- or underfitting where the learning curve is suffering from a high bias or a high variance problem (Ng, 2018). A conventional approach to investigate if an algorithm is suffering from any of these issues is visualisation, where the learning curve may be studied in relation to the given input data (Ng, 2018). Underfitting, also known as a high bias problem, is when the learning curve fail to generalize new examples due to the function being too simple to capture the variance in the data, while overfitting, also known as a high variance problem, is when the algorithm is forcefully fitting the learning curve to the training data where the learning curve is often too good to be true (Ng, 2018). Another common way to perform error analysis is to manually examine the examples that the algorithm made errors on and see if there are any systematic trends in these examples that may cause the error (Ng, 2018).

If a data set is very biased, e.g. in the case of determining if a patient has cancer or not, the number of patients who actually have cancer may be very low. The learning algorithm could then predict the same outcome for all input data, $(y) = 0$, and yield a success rate of 98 percent, or error rate of two percent, if only two percent of the samples actually have cancer. Hence, in this case, accuracy is not a very good performance measure (Ng, 2018). Due to this reason, there are other important metrics to use in algorithm evaluation (Ng, 2018). Precision and recall are two commonly used performance metrics. Precision and recall are interrelated, meaning that trade-offs have to be made where one must be prioritised over the other (Ng, 2018). Whether precision or recall is the prioritised metric depends on their relevance for the task at hand (Ng, 2018).

Precision also referred to as specificity, measure the fraction of selected items that are relevant (Ng, 2018). Given a specific output class, e.g. $(y) = 1$, not all data labelled by the algorithm is going to be correct. If ten samples are labelled as $(y) = 1$, while only eight of these are correct, the precision for this scenario is 80 percent. Precision is an important metric in scenarios where the output must have high confidence, where higher precision leads to reduced recall (Ng, 2018). Precision is calculated as following:

$$\text{Precision} = \frac{\text{tp}}{\text{tp+fp}}$$

Recall also referred to as sensitivity, measure the fraction of relevant items that are selected (Ng, 2018). Given a specific output class, e.g. $(y) = 1$, some of the data labelled as other classes are going to belong to that outcome. If eight samples are positively labelled, when the total data set consist of ten positive inputs, the recall for this scenario is 80 percent. Recall is an important metric in scenarios where it is important not to miss any instances, where higher recall leads to reduced precision (Ng, 2018). Recall is calculated as following:

$$\text{Recall} = \frac{\text{tp}}{\text{tp+fn}}$$

Precision and recall come together as an F-score (Ng, 2018). The F-score is the harmonic average of the precision and recall where it is used to assess the performance of a classifier. The F-score reaches its best value at one and worst at zero, where an F-score value of one indicates an optimal combination of precision and recall (Ng, 2018).

Another important factor is cost. In most practical ML situations, the cost of a misclassification error depends on the type of error it is—whether, e.g., a positive example was erroneously classified as negative or vice versa (Ng, 2018). When doing ML, and evaluating its performance, it is often essential to take these costs into account (Ng, 2018).

2.2.4.3 Machine Learning 2.0

In an article by Schreck, Kanter, Veeramachaneni, Vohra and Prasad (2018) published in the Harvard Business Review, a streamlined way of creating a ML model is presented. In particular, how a four-person team managed to develop within eight weeks the proof-of-concept and deploy the necessary models (Schreck et al., 2018).

ML has for a long time only been available to a minimal group of people within the field of applied mathematics. Despite many ML discoveries from the academic world with research papers showing the capability of ML models and large amounts of data available for algorithm training, businesses outside of the technology sector are having trouble deploying ML solutions to solve real-life issues in their day to day operation (Schreck et al., 2018).

Studies such as "Machine learning 2.0 Engineering data-driven AI products" have shown that it is not about applying the latest and most advanced ML techniques that matter most for generating business value, rather, it is about actually applying something, where even traditional and more robust ML algorithms have the potential to generate tangible business value (Schreck et al., 2018).

ML has seen many technical developments during recent years which has taken performance up to new levels. However, another even more important development within ML is the reduced barriers of entry, where today, even those without data science experience are able to access and train ML algorithms for different purposes. What is referred to as "Machine learning 2.0" is the paradigm of reduced barriers of entry for ML practitioners where many ML APIs and libraries, such as Scikit-Learn and TensorFlow, are available online as open-source and free of charge. Even more advanced ML algorithms are available online for a minor fee through cloud computing services provided by tech giants such as Windows, Google and Amazon, where algorithms can be trained using the customers own data in order to evaluate the performance of different ML algorithms (Schreck et al., 2018).

Another indirect result of this development is a considerable decrease in time for deploying ML systems. Twenty years ago, a ML model development could span for months or years and require several engineers, where today it is possible to create a ML model from raw data that is put to use within weeks with little or no background knowledge of the technical aspects. The reduced barriers of entry and speedy process also allows for domain experts to be more involved in ML projects where they can come in at the right level to ensure that the ML system is doing something useful and, more importantly, that can generate value in their process or organization (Schreck et al., 2018).

2.2.4.5 Pipeline

A linear sequence of data processing components is called a pipeline and are commonly used in ML systems, mainly due to requirements of data manipulation and transformation before the input can be integrating into a ML algorithm (Géron, 2017). The pipeline allows for an automated workflow where input and output can be generated without having to apply each component individually (Géron, 2017). Pipelines work by allowing a linear sequence of data transformations to be chained together culminating in a fully working system (Géron, 2017).

An important step of designing a well-functioning ML system is to design the ML pipeline where components must be carefully selected for the task at hand (Géron, 2017). While different components are linked together, they usually run asynchronously where each component is fairly self-contained (Géron, 2017). Pipelines also make the ML system easier to understand from an outside perspective as it provides a data flow graph (Géron, 2017).

2.2.5 Classification algorithms

Due to an increased number of documents being available in digital form the interest in texts classification and automatic document handling has increased (Lyfenko, 2014). Automated document handling using ML principles falls under the category of back-office AI (Gesing et al., 2018), where such solutions have gained huge industry attention during recent years (Gesing et al., 2018). One of the primary goals of automated document handling is to classify input data, i.e. a particular document, to a predefined class. Such problems are already being solved with the help of different systems, where one example would be the Automatic

Document Classifier system, abbreviated as ADC, as presented below by Lyfenko (2014) in the paper on a conceptual scheme for text classification systems.

However, there is a large variety of ML methods that can be used for both text and image classification, and in turn an automated document classification. As the given task can be dealt with in many different ways, there is no one right answer on how to solve the issue at hand. The most popular algorithms for document classification purposes are neural networks, naïve Bayes classifiers, Rocchio classifiers and support vector machines (Lyfenko, 2014).

This section will cover two classification algorithms that are identified by the authors as being suitable for the task at hand, i.e. neural networks and Naïve Bayes. In relation to the topic of investigation, the following models will be described briefly as they provide practical benefits and tools for the particular questions this study is aiming to answer and could potentially serve the intended purpose.

2.2.5.1 Neural networks

Neural networks, or artificial neural networks, were originally designed with the goal of having machines that can mimic the human brain (Ng, 2018). While neural networks are very advanced, they could be simplified as that each neuron in the network has a mathematical model for determining output from input (Ng, 2018). Neural networks could be defined as an advanced form of pattern recognition, where they just like any other ML algorithm must be trained using sample data. Neural networks are defined by three main properties; activation function, network topology and training algorithm (Géron, 2017).

The activation function transforms the input and forwards the output throughout the network (Géron, 2017). While there are many different activation functions, such as, a commonly used activation function is the previously described sigmoid function which generates a 0 or 1 output (Géron, 2017). The neurons could then be said to be activated if the output of the model in a specific neuron is 1, and not activate when the output is 0 (Géron, 2017).

A neural network is formed by inputs and outputs with hidden layers in between as illustrated in Figure 8, inspired by blog.webkid.io. The neural network topology is an integral part of the neural network and is specified in terms of how many layers, units in each layer and outputs units or classes exist in the network (Géron, 2017). Neural networks can model very complex problems as they may contain hundreds of artificial neurons, where each neuron has a number of weighted variables as input (Géron, 2017).

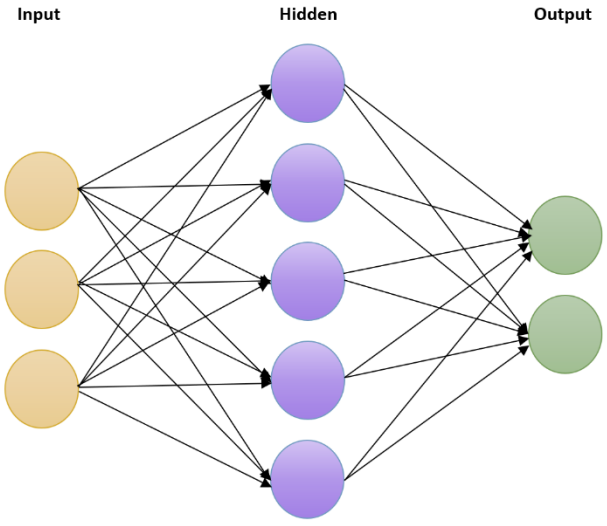

Figure 8: Neural network with one hidden layer, inspired by Webkid

Neural networks may act through different training algorithms, were a commonly used algorithm is the backpropagation algorithm (Géron, 2017). Backpropagation is a neural network terminology and aims to minimise the cost function and to further improve the algorithm performance (Géron, 2017). This means that the neural network is first calculated from input to output before being calculated again from output to input, i.e. backpropagated (Géron, 2017).

2.2.5.2 Deep neural networks

A deep neural network is a neural network with multiple hidden layers between the input and output (Géron, 2017). While ordinary neural networks usually apply one or two hidden layers, deep neural

networks include many more which allows the network to map more complex functions (Géron, 2017). Studies have shown that DL algorithms scale better with data compared to traditional ML algorithms, where deeper networks generally produce greater results (Ng, 2018). Deep neural networks gave rise to DL, which as of today is a popular area of research where many state-of-the-art algorithms with human-level performance are being developed (Géron, 2017).

A popular deep neural network is the convolutional neural network, abbreviated CNN (Bhattacharjee, 2016). The CNN is a deep, feedforward neural network which is commonly applied to tasks related to image recognition, video analysis and Natural Language Processing (abbreviated as NLP) (Bhattacharjee, 2016). A CNN consists of several hidden layers where the hidden layers usually consist of convolutional layers, pooling layers, fully connected layers and normalisation layers (Bhattacharjee, 2016). The CNN is a popular classification algorithm that requires little pre-processing compared to other image classification algorithms. One of the main benefits of the CNN is that it has the ability to characterise and classify raw sequence data, thus removing the need for manual feature generation (Bhattacharjee, 2016).

2.2.5.3 Naïve Bayes

According to Hristea (2013), naïve Bayes classifier is a model which has been used widely, both due to its level of efficiency as well as it being able to process evidence from a large number of features combined. In the context of Word Sense Disambiguation, the Bayesian classifier works in such manner that it looks at the words around an ambiguous word, taking into consideration a relatively large context window (Hristea, 2013). The classifier does not select any particular features, on the contrary, it processes the particular information with certain potential usefulness, contributed by each content word, as well as considering in which sense the word is most likely to be used, based on its ambiguity (Hristea, 2013). The reason for why it is called naïve is because the made assumption is that the describing attributes for the words are all conditionally independent (Hristea, 2013). In the context of Bayesian statistics, in order to determine $P(Y|X = x_k)$, for any new instance x_k, one way to learn $P(Y|X)$ as presented by Wang and Tseng (2015), is to use the training data to estimate $P(X|Y)$ and $P(Y)$, together with the sum and product rules (see Appendix B).

Lyfenko (2014) used this model when investigating automatic document classification (ADC) systems for documents in a natural language, as the main objective is to automatically sort the documents based on their textual contents, into the most appropriate classes, thus the name text classification. The basic algorithm of the ADC system consists of three stages: data acquisition, data processing, and analysis of the results as illustrated by Figure 9, which is used with the permission Lyfenko (2014). As also mentioned by Hristea (2013), the method is based on a model which has a strong, i.e. naïve, assumption of independence of vector components, and propose to calculate the probability that a document, denoted by d, belongs to a particular class, denoted by c, through the Bayes' theorem (see Appendix D), which Lyfenko (2014) expresses in the following way:

$$P(c|d) = P(c) \prod_{1 \le k \le n_d} P(t_k|c)$$

$P(t_k|c)$ denotes the conditional probability that the term, or feature, t_k, will be included in a document belonging to class c (Lyfenko, 2014). $P(c)$ is the prior probability that the document belongs to c, and the lexical units of d, consists of so-called tokens, which total number is denoted as n_d, and which sequence is denoted as $\{t_1, t_2, \dots, t_k\}$ (Lyfenko, 2014). In this naïve Bayesian classification method, Lyfenko (2014) states that the class which is the most probable to be chosen, denoted as c_{map}, where "map" denotes the *maximum a posteriori*, i.e. maximum prior probability. The method is expressed as $c_{map} = \text{argmax} \hat{P}(c|d)$ which are the points in the probability function for c, given d, at which the values are maximised, and which formula is expressed in the following way:

$$c_{map} = \text{argmax}\, P(c|d) = P(c) \prod_{1 \le k \le n_d} P(t_k|c)$$

According to Lyfenko (2014), this method is recommended for classification models in cases with documents having tight restrictions. The method itself provides practical benefits in terms of it having a simple mathematical model, as well as being relatively rapid to operate with a high enough performance in classification (Lyfenko, 2014).

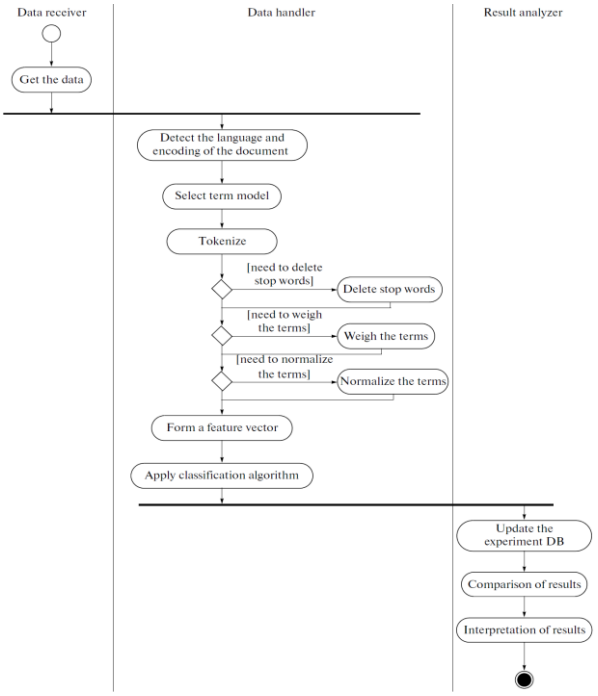

Figure 9: Conceptual model of the ADC system (Lyfenko, 2014)

2.2.6 Related topics

This section seeks to explain the topics of OCR and Artificial data synthesis for the purpose of document classification and testing. As OCR is a key enabler for paperless trade and information extraction, it is deemed necessary to include for the purpose of this thesis.

2.2.6.1 OCR

Optical Character Recognition, abbreviated OCR, is one of many machine recognition techniques which aim to identify, collect and enter data into a computer system without human interaction (Chaudhuri et al., 2017). Speech recognition, radio frequency identification, magnetic stripes and barcodes belong to the family of machine recognition techniques where each type has its own relevant area of application. OCR is the process of identifying alphanumeric data from images, e.g. letters, numbers, words or sentences, that can then be translated into machine-encoded text that may be edited, searched or used for other applications, such as data mining or ML (Chaudhuri et al., 2017). As a scanned document is just an output image of the original paper document, where text contained is just a set of pixels as far as the computer is concerned, text cannot be extracted without the help of an OCR tool. Because of this, OCR is one of the most important technologies to achieve paperless trade (Tarnopol, 2017). If, however, the document has remained digital throughout the supply chain, OCR is not required, and text can easily be extracted from the native electronic file (Tarnopol, 2017).

OCR is achieved through segmentation, feature extraction and classification where the outcome is measured in terms of confidence and accuracy (Holley, 2009). An OCR software does not know if an identified character is correct or not. Rather it calculates the confidence level for each identified character where the letter or number with the highest confidence is generated. Based on this logic, an OCR software does not know if the presented outcome is correct or not, it can only be confident or not confident that the answer is correct. Whether a character is correct or not can as of today only be determined by a human interpreter (Holley, 2009).

Accuracy is measured in terms of successfully identified characters where a good accuracy rate for machine written text is between 98 and 99 percent (Holley, 2009). The performance of an OCR system is directly dependent upon the quality of input documents, where more constrained input results in better performance (Chaudhuri et al., 2017). Because of this, data cleaning and pre-processing are important parts of OCR as a high accuracy rate of an OCR application says more about the input data, rather than the OCR technology (Tarnopol, 2017). To generate optimal results from OCR applications high-resolution images and clear contrast between letters and background are by far the most important factors (Tarnopol, 2017). Even though OCR system has become one of the most successful applications of technology in pattern

recognition and artificial intelligence fields, many commercial systems are still not able to compete with human reading capabilities with desired accuracy levels (Chaudhuri et al., 2017).

Development in OCR has gained considerable momentum during the last decade, where increased computing power, higher quality input and increasing areas of application for the information converted from images and documents all play essential roles (Chaudhuri et al., 2017). OCR applications are today successfully combined with autonomous learning methodologies, such as neural networks and hidden Markov models, fuzzy set reasoning and NLP, which result in very high accuracy compared to traditional methods (Chaudhuri et al., 2017). OCR systems can be divided into two different groups, special purpose machines and PC based scanners. Special purpose machines are commonly used in commercial settings for specific applications where high throughput is important. Such systems are still related to high cost, and an effective special purpose machine may cost several million dollars (Chaudhuri et al., 2017).

Recent year advancements in computer technology have made OCR tools available to the public where it possible to fully implement recognition part of OCR in software packages to work on personal computers (Chaudhuri et al., 2017). The first OCR online service available to the public was launched in 2000, and as of 2015, Google offers OCR tools to scan any file in over 200 languages for free. PC based OCR software still has some limitations when it comes to speed, and character sets read (Chaudhuri et al., 2017).

2.2.4.2 Artificial data synthesis

One of the most reliable ways to get a high-performance ML system is to train a learning algorithm on a large data set (Ng, 2018). Authentic and large data sets are however hard to come by since they usually contain a lot of sensitive or classified information (Koperniak, 2017). Artificial data synthesis is a way to generate new training examples, either from scratch or already existing data (Ng, 2018).

Recent studies have shown that it is even possible to apply ML systems to authentic data sets which are then able to create fully synthetic artificial data sets, based on the input from the authentic data (Koperniak, 2017). Artificial data is a valuable tool for education and training since real data is sometimes too sensitive to work with. Research that compares the impact of using artificial data rather than real data on learning algorithms show that the difference in outcome is relatively small (Koperniak, 2017), where such result suggests that artificial data can successfully replace real data in a software writing and testing environment.

The first way to use artificial data synthesis is to create new training examples from scratch (Ng, 2018). This is a very time-consuming process where training examples are generated through manual labour, where, e.g. letters, words or documents may be altered or generated in various ways. By providing the same original document in different versions with changed fonts, words or sentences, the learning algorithm can use different versions of the original data as individual training examples (Ng, 2018).

The second way to use artificial data synthesis is to introduce different distortions on an already available data set (Ng, 2018). Entire documents can be distorted in different ways allowing the learning algorithm to train on different versions of the same original document where, e.g. noise, warping and scanner imperfections have been added. Many scanned documents contain some kind of noise due to printer quality and manual handling. Therefore, adding noise to training data could prove valuable. When adding distortions, it is important that they represent the kind of distortions that would actually occur in a real example (Ng, 2018).

2.3 Business

This section seeks to highlight the barriers to the implementation of new technologies and innovations as well as present both the relevancy and the strategic benefits of digitalisation within a business context.

2.3.1 Diffusion of innovations

The theory that seeks to explain how, why and at what rate innovations spread on a large scale is called diffusion of innovations (Rogers, 2003). While adoption is an individual process, diffusion signifies a group phenomenon, where diffusion is the means by which innovations are adopted on a large scale and translated into social and economic benefits (Tidd, 2010). Understanding diffusion of innovations is critical since there is a wide chasm between the development and successful adoption of a new service or product where a majority of all innovations never reach their intended markets (Tidd, 2010). Therefore, a better understanding of why and how innovations are adopted, or not, can assist in developing more realistic business plans. While diffusion of innovation often focuses on a specific product or technology, an innovation also includes ideas, beliefs and practices, as, e.g. management practices such as lean production or six sigma (Tidd, 2010).

Conventional marketing approaches work well for many products and services but do not work as well for innovations. This is because there are many barriers to widespread adoption of innovations where economic, behavioural, organisational and structural obstacles all have the potential to make or break an innovative product or service (Tidd, 2010).

- **Economic barriers** include cost, access to information and insufficient incentives.
- **Behavioural barriers** include priorities, motivations, rationality and change.
- **Organizational barriers** include goals, routines, power, influence, culture and stakeholders.
- **Structural barriers** include infrastructure, sunk costs and governance.

Along with the barriers, there are also specific characteristics that influence the rate of adoption (Tidd, 2010). These characteristics are grouped into three categories which contain characteristics of the innovation, characteristics of the individual or organisation and characteristics of the environment. Characteristics of the innovation are arguably the most important one since diffusion rates of different innovations are highly variable. However, contextual and environmental factors are also relevant since the diffusion rates for the same innovation differ between organisations and different industries (Tidd, 2010).

Adoption of innovations usually generates significant benefits, where technological innovations are the source of productivity and quality improvements. Most innovations take several years before they reach their full potential, and therefore, most innovations fail to reach the stage of diffusion and widespread adoption which results in a limited impact (Tidd, 2010).

Several authors have tried to visualise the diffusion of innovation process where the model presented by Rogers (1962) remain the most cited and commonly used (Tidd, 2010). Rogers model, presented in Figure 10, is often referred to as the idealised diffusion curve, which consists of five types of adopters. The five types, or populations, as proposed by Rogers (1962) consist of innovators, early adopters, early majority, late majority and laggards and are all said to represent a specific proportion of the entire population. The figure consists of two curves where the blue is a bell-shaped curve representing adopting among different categories and the yellow a cumulative logistic function representing full adaptation (Rogers, 1962).

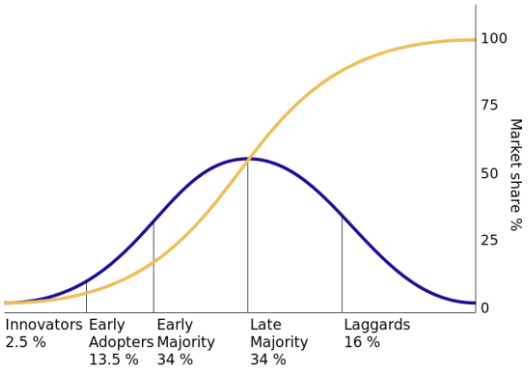

Figure 10: Rogers diffusion of innovation model (Rogers, 1962)

2.3.2 Automation of manual tasks

In a news release by Accenture published in March of 2018, they present recent developments and tests within blockchain technology which provides confirmation on the large potential for digital solutions to significantly reduce the number of resources needed to both share and distribute relevant data within the freight and logistics industry. This has led to container companies such as APL and freight forwarders such as Kuehne + Nagel to almost eliminate their need for handling shipping documents in paper form, thus optimising their operations enormously in terms of operations costs (Accenture, 2018). A significant portion of the data included in the documentation can be replicated and by making it possible for it to be digitally entered, this has both simplified many parts of the shipping processes as well as provided several benefits for the entire logistics chain (Accenture, 2018).

This follows along with a report published by the European Economic and Social Committee in 2017 regarding the "Impact of digitalisation and the on-demand economy on labour markets and the consequences for employment and industrial relations". In the report, the committee states that digitalisation will have major implications for the labour market, due to an increasing part of the many labour tasks being able to be performed automatically with practically no human intervention, particularly in the case for simple and repetitive tasks, e.g. within administration (EESC, 2017). Thus, digitalisation will

probably reduce the number of work roles in many different areas, while at the same time, create both more service-oriented low-skilled jobs as well as new forms of employment (EESC, 2017). These jobs will require higher qualifications, which in turn will enable both organisations and workers to focus on more high-skilled tasks (EESC, 2017). The organisational interaction between the human workforce is projected to be increasingly concentrated to online platforms, thus making it possible for more traditional business and industries to potentially use these platforms to outsource activities and even to allocate and schedule the work more fittingly (EESC, 2017).

According to an article in the Wall Street Journal from May 2017 (Shumsky, 2017), the spending for the US market only for management-consulting services reached $58.7 billion during the year of 2016. Most of the spending is to cover administrative task performed by managers which most of the time result in data analysis and presentations (Libert and Beck, 2017). Considering recent advancements in different AI and ML applications, these tasks have become more suitable for the machines to take care of. Libert and Beck (2017) believe these developments will change many jobs, e.g. accountants, consultants and lawyers, who will be focusing more on judgment rather than the analysis. This due to "Quant Consultants" and "Robo Advisers" being more efficient and accurate compared to human specialists.

The EESC report states that it is currently not possible to draw any conclusions about how specific processes and groups will be affected by digitalisation, given the high level of ambiguity the effect new technologies have on already existing services and products (EESC, 2017). However, as has been shown by history like in the case of the introduction of the industrial steam engine in the 19th century, and the large-scale usage of electricity in the 20th century, this type of changes are likely to increase productivity (EESC, 2017). The increase concerns the reduced need for human labour as well as increased demand due to reduced production costs and in turn even prices towards consumers (EESC, 2017). In a study made by Arntz et al. (2016), they show that more than 70 percent of the activities in 9 percent of jobs could be automated, and approximately 30 percent of the activities of 60 percent of the current jobs within the OECD countries are automatable based on technologies which effectiveness have already been proven (Dawson et al., 2016).

2.3.3 Labour costs

The EESC study (2017) further states that new automation techniques will not be adopted from one day to another. Rather the transition will take relatively long periods of time, all depending on the level of difficulty to automate the particular activities, the cost for the actual development and implementation of said changes, as well as the costs and overall availability of manual labour on the global market.

Other aspects which will be considered are the economic benefits for the different industries, the level of willingness of the societies for these changes to occur and also if current legislation even allows for changes in this direction or if any amendments will be required (EESC, 2017). It is, however, fair to say that by continuing the path towards increased digitalisation, this will potentially lower market entry barriers by both creating new markets to enter as well as open up already existing market for current firms and industry suppliers/providers to offer their respective services (EESC, 2017).

Regarding the usage of tools based on Information and Communications Technology (ICT), these have become both more adapted to companies' particular needs, as well as more affordable for different organizations to implement, thus pushing the trend of 'platformisation' further, having been enabled by the more widespread usage of algorithms (EESC, 2017). This has, in turn, made it possible for more conventional companies to experience a transformation concerning how they manage both their processes and assets, resulting in more innovative solutions, which have led to them in many cases changing their business strategies completely (EESC, 2017).

2.3.4 Projected future of administration

In a study by Kolbjørnsrud, Amico and Thomas (2016) regarding the potential of AI technologies to automate many of the administrative tasks which are often both time and resource consuming, they surveyed 1770 managers from 14 different countries and interviewed 37 executives responsible for the digital transformation within their organisations. In their article, they present five different practices which they identified to be needed by managers to be successful. These were the following: Leave Administration to AI, Focus on Judgment Work, Treat Intelligent Machines as "Colleagues", Work Like a Designer, and to Develop Social skills and Networks (Kolbjørnsrud, Amico and Thomas, 2016). The final recommendation stated in the article is to adopt AI in order to automate the administrative tasks, and instead of replacing human judgment rather this technology will enhance it (Kolbjørnsrud, Amico and Thomas, 2016). They also conclude that for leaders to prepare both themselves and their organizations for the constant technological improvement, they must take the following steps: Explore early and experiment with AI; Adopt new KPI's

like learning and decision-making effectiveness in order to drive adoption; and to Develop training and recruitment strategies for creativity, collaboration, empathy and judgment skills (Kolbjørnsrud, Amico and Thomas, 2016).

2.3.5 Strategy

Strategy theory emerged as a subset of management and organisation theory in the 1960s in response to needs for research into bigger questions (Strannegård and Styhre, 2013). Strategy has many definitions where one of the most commonly used is that business strategy is an organisation's plan for achieving its vision, prioritising objectives, competing successfully and optimising financial performance (Strannegård and Styhre, 2013). A well formulated and executed strategy is important for any organisation, as it establishes a foundation where success can be monitored, measured and created (Strannegård and Styhre, 2013). Since its introduction in business, strategy has been getting increasingly diverse, where strategy now includes the management of both internal and external factors, analysis, formulation, implementation, administration, leadership and much more (Strannegård and Styhre, 2013).

Recent years change in business environment has changed strategic thinking (Strannegård and Styhre, 2013). Where just ten or fifteen years ago a strategy could be presented as a five or ten-year plan, changing market environments and disruptive forces does not allow for such long-term planning anymore (Strannegård and Styhre, 2013). Digitalisation has played a key role in this development, where the pace of change in the new digital and now global arena as faster than ever before (Strannegård and Styhre, 2013).

2.3.6 Digital strategy

Going forward, digital strategies will become an essential topic for almost every company in every industry (Bughin et al., 2018). According to a recent McKinsey study, only 8 percent of companies think their current business model will survive digitalisation in its current form and pace (Bughin et al., 2018). While business disruption is nothing new, the speed and magnitude of disruptions in a digital setting are unprecedented (Bughin et al., 2018).

Digital strategy is more than a set of technologies to be bought. It is the strategy around the abilities those technologies create (McDonald, 2015). The key question in digital strategy is how a business can gain using information technology to raise human performance (McDonald, 2015). Ever since the breakthrough of information technologies in organisations, it has been understood that higher level of performance may be achieved if firms react quickly to innovative ideas or stakeholder demands regarding information technology (Strannegård and Styhre, 2013).

Digital strategy consists of two paths, extending digitalisation and transforming activity (McDonald, 2015). The extending digitalisation strategy consist of applying new technology to existing activities or by repeating previous digital strategies to cover new functions and processes (McDonald, 2015). While this is the most common and the easier digital strategy to succeed in, it is not always the most sustainable as digital breakthroughs may sometimes call for an entire remodelling of the day to day operation and processes to gain maximum benefits in an organisational environment (McDonald, 2015). The transformation activity strategy revolves around transforming the business processes to be more compatible with the new technology. While this strategy may yield superior results, it is less common as it involves a lot of change and transformation in organisations which often prove difficult (McDonald, 2015).

2.3.7 Why digital strategies fail

While strategy is important, strategies sometimes fail (Strannegård and Styhre, 2013). There are many different reasons for overall strategies to fail, such as organisational culture or lack of top management commitment (Strannegård and Styhre, 2013). A recent McKinsey study reveals the top reasons why digital reasons fail and claim that there is a surprisingly large number of organisations who underestimate the increasing momentum of digitalisation, the behavioural changes and technology driving it (Bughin et al., 2018). Below are the four top reasons listed, which will be elaborated further on in this section.

1. Fuzzy definitions
2. Misunderstanding of the economics
3. Overlooking of ecosystems
4. Over-indexing the usual suspects

The first pitfall of digital strategies is fuzzy definitions, where very few organisational leaders have a broad, holistic understanding of what digital really means (Bughin et al., 2018). Lacking a clear definition of digital result in companies struggling to connect their digital strategy to their business (Bughin et al., 2018).

The second pitfall is a misunderstanding of the economics of digital, where the core economic principles from the past no longer hold true in the new realities of digital competition (Bughin et al., 2018). The digital economy provides consumers with unlimited choice and price transparency, shifting the power dynamics of competition and resulting in winner-takes-it-all economics where only a few market-leading companies will remain profitable and where first movers and some superfast followers will have a significant advantage. (Bughin et al., 2018)

The third pitfall is overlooking ecosystems, where digital platforms and ecosystem economics upend the fundamentals of supply and demand (Bughin et al., 2018). As industries slowly transform into ecosystems, where platforms allow digital companies to move fast and easily across industries, it is essential to understand the new economies rules to move ahead (Bughin et al., 2018). In an ecosystems environment, today's competitor may turn out to be a partner, where failure to grasp the ecosystem business approach may lead to missed opportunities and failing strategies. (Bughin et al., 2018)

The fourth pitfall is over-indexing the usual suspects, where the digital era allows other than established corporations to pose a threat to an entire industry (Bughin et al., 2018). Excessive focus on the ordinary competitors is dangerous, as new entrants and disruptive forces may now appear from any industry. (Bughin et al., 2018)

2.3.8 Changing forces

What are the deeper forces behind significant changes in the market and how do these forces contribute in the actual digital transformation of them? These questions need to be answered when developing future business strategies in an increasingly digitalised world. Dawson, Hirt, and Scanlan discuss this matter further in an article published by McKinsey in 2016, in which they highlight the importance of understanding the nature of the disruptions faced by companies in terms of simple market dynamic, as well as supply and demand. Additionally, the article mentions the necessity for companies to monitor market competition and the process of launching digital initiatives in order to meet these changes and simultaneously remain competitive (Dawson et al., 2016). The two primary sources of digital transformation and disruption are firstly the making of new markets related to unconstrained supply, e.g. by increasing capacity availability, and secondly undistorted demand, e.g. through tailoring, and how these two can be connected in more optimised ways (Dawson et al., 2016).

Suggestions provided by McKinsey (2016) on how to create disruptions before the competitors are to turn product-portfolio into services, and to create more streamlined order processes, e.g. by automating and speeding up transactions, thus making them more user-friendly. This can even result in actually removing the middlemen from the supply chain (Dawson et al., 2016). Secondly, it is the combination of how to reimagine business systems, e.g. by digitalising and automating key supply processes, and the actual creation of new value propositions, e.g. by performing tasks which were previously being done by the customers themselves (Dawson et al., 2016). Suggestions on how to create disruptions in the second source, e.g. to improve connectivity of physical devices and to extend product and service portfolio with digital features (Dawson et al., 2016). Thus, companies can 'hyperscale' their platforms, making it possible for them to take advantage of current customer relationships and stored information about, e.g. consumption patterns (Dawson et al., 2016).

2.4 Summary

The theory section described the container shipping industry, technology and business aspects of the research topic, being Maritime Management and the application of Electrical Engineering to the identified issue within back-office activities, in particular, ML and DL solutions.

The container shipping subsection described how and why the container shipping industry is a vital part of the global economy. The subsection is followed up by presenting the main shipping service providers, the container import process, documentation requirements and finally what the administrative work and document handling process generally entails.

The technology subsection presents the reader with a general mathematical and statistical framework needed to understand the fundamentals of the engineering aspect of the topic. The parts included are essentially about ML and different learning techniques, which algorithms are used, in particular, for the task of automated document classification. The subsection is finalized by mentioning two highly regarded algorithms for the topic at hand being the naïve Bayes model and neural networks. The neural network section goes further into describing the rise of deep neural networks and deep Learning, where DL algorithms have surpassed human-level performance with accuracy levels of above 95 percent. The topic of

OCR is covered where state-of-the-art OCR tools have reached accuracy levels of 99 percent. The topic of Artificial data synthesis is covered due to their relevance to the topic at hand.

The business subsection described the diffusion of innovations model, covering economic, behavioural, organisational and structural barriers to widespread adoption of innovations. The topic of automation of manual tasks is covered where recent studies show that more than 70 percent of the activities in 9 percent of jobs could be automated, and approximately 30 percent of the activities of 60 percent of the current jobs within the OECD countries are automatable based on technologies which effectiveness have already been proven. In future administration, it is suggested to leave administration to AI, focus on judgment work, treat intelligent machines as "colleagues", work like a designer and to develop social skills and networks. In order to succeed with digitalization efforts and improvement measures, a clear strategy is needed where the topics of strategy and digital strategy are further elaborated. To further enhance the understanding, the topic of why digital strategies fail and changing forces are elaborated.

3. Method

The following section explains the research methodology used in this study. As this thesis aims to answer both qualitative and quantitative questions, a mixed method research approach is chosen and explained. Second follows a description of how design thinking is applied and how the chosen design thinking model, double diamond, works. This is followed up by a gaining access section, where the collection of qualitative and quantitative data is explained along with ethical considerations.

3.1 Mixed method research

For this thesis, a mixed method research approach has been chosen. Mixed method research is commonly applied within the business and management field as quantitative and qualitative methods can be fruitfully combined to create synergy effects that enable a more rounded and complete picture to be drawn (Bryman and Bell, 2011). As quantitative and qualitative methods have strengths and weaknesses in different areas, the response of applying mixed method research creates a strategy that allows for the various strengths to be capitalised upon and the weaknesses to be somewhat offset (Bryman and Bell, 2011). While mixed method research is a relatively new approach, it has become increasingly used and accepted in social sciences as it may help overcome practical constraints and limitations of only applying one of the traditional methods (Bryman and Bell, 2011).

Mixed method research is the combination of quantitative and qualitative methods within a single project (Bryman and Bell, 2011) where a quantitative approach has been chosen for the technical aspects, and a qualitative approach has been chosen for the business and management aspects. The quantitative research may then facilitate the qualitative research and vice versa, where gaps can be filled by using the mixed method approach (Bryman and Bell, 2011). As this thesis aims to explore the application of already existing ideas and concepts of ML for business operations within container shipping, it will also be categorised as an empirical study. The concept of using ML for automating administrative work in the proposed field, along with proposed research questions, form the conceptual model and hypothesis for which the empirical evidence will be collected (Bryman and Bell, 2011). For mixed method research to be successfully applied, it is crucial to be aware of the strengths and weaknesses, as well as to recognise that mixed method research is not superior to research that employs only a single research strategy (Bryman and Bell, 2011). As mix method research is a mix of quantitative and qualitative research, both these topics will be briefly covered in below.

3.1.1 Quantitative research

Quantitative research methods consist of systematic observation, recording and collection of numerical data where the applied statistical techniques and calculations are used for hypothesis testing as well as drawing additional conclusions (Habib et al., 2014). The goal of the quantitative method is to determine whether the predictive generalisations of the proposed research questions hold true (Habib et al., 2014). The quantitative approach is applied to the technical aspects of the thesis as it is most suitable for evaluating quantifiable performance metrics. To generate valid results, measurements must be objective, quantitative, and statistically valid where precautions must be undertaken in order to assure such results (Habib et al., 2014).

3.1.2 Qualitative research

A qualitative research approach is typically associated with an inductive way of linking data and theory (Bryman and Bell, 2011). A qualitative approach is chosen for the business and management aspects of the

thesis as it is the better way of achieving a deep understanding of a subject according to Bryman and Bell (2011). The deeper understanding is generated by enabling the study to be more exploratory while using more flexible and open research methods. By allowing the management and business section of the thesis to be more exploratory and flexible, it will increase the possibility of gathering unexpected data that may prove to be important for the overall success of a project as suggested in this thesis. To further aid into the analysis, the data will be iteratively processed, meaning that the researchers go through the data repeatedly, backtracking several times to tie empirical findings closer to theory and refining the thesis as a whole (Bryman and Bell, 2011). This process will also help the researchers make more sense out of the collected data to come up with better concepts and reveal new themes (Bryman and Bell, 2011).

3.2 Design thinking approach

For the purpose of this thesis, a design thinking process will be applied. Design thinking is a valuable method when solving problems, as it provides a clear framework on how to reach a solution from an initial idea or inspiration (UK Design Council, 2007). The double diamond approach, developed by the British design council in 2005, provides a good framework of natural steps for designing new solutions and can be applied in any industry (UK Design Council, 2007).

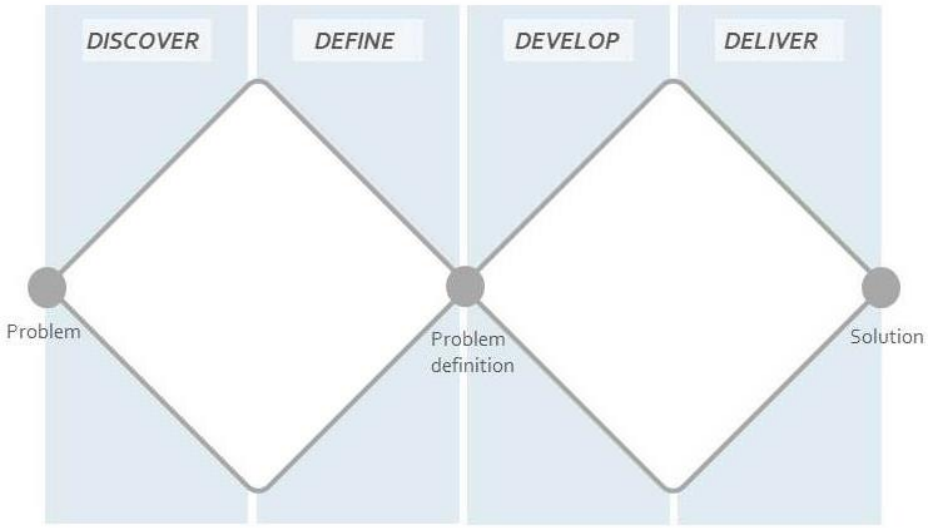

Figure 11: Double diamond (UK Design Council, 2007)

The double diamond is a tool which helps set up, frame, organise, structure and execute design challenges and projects (Nessler, 2016). Much like the project management 4D approach, the double diamond framework consists of several steps which provide structure throughout the project from problem definition to solution. The different steps are illustrated in Figure 11, which is used with permission by the UK Design Council. The double diamond also includes a diverge and converge element where phases one and three are diverging, and phases two and four are converging, thus forming the shape of two diamonds. During the diverging phases, it is the researchers task to open up as much as possible without limitation to possible solutions, whereas the converging phases focuses on narrowing down the findings and ideas gathered in the previous phase (Nessler, 2016).

- Discover
  - The first phase begins with an idea or inspiration to solve a specific problem. This phase is also referred to as the research phase where insight on the problem is gained and where market and user needs are identified (UK Design Council, 2007). This phase consists of a diverging element where it is important to approach the problem with an open mind set free from biases that may limit the potential outcome (Nessler, 2016).
- Define
  - The second phase consist of interpretation and alignment of the previously identified user and business needs (UK Design Council, 2007). Automated solutions are a hot topic within container shipping, as it involved many reparative and time consuming administrative tasks. Decisions are made on how to move forward by narrowing down identified technical options from phase one, thus converging.

- Develop
    - The third phase, Develop, is also referred to as the ideation phase where potential solutions or concepts are created, tested and iterated (UK Design Council, 2007). The process of trial and error by testing different learning algorithms in MATLAB, along with evaluating additional features such as information extraction in the development phase will assist the authors with insight on how to improve and refine their solution.
- Deliver
    - The fourth and last phase of the double diamond is Deliver. This is also referred to as the implementation phase where solutions developed in the previous phase are narrowed down, evaluated and finalized (UK Design Council, 2007).

3.3 Literature review

A literature review is conducted in order to gain knowledge on the topic of ML, digitalisation, strategy and maritime logistics. Related topics to ML, such as NLP, and Optical Character Recognition were investigated due to their relevance to the specific topic. The main purpose of the literature review is to analyse and understand which state-of-the-art tools and algorithms are available and which of these that are most suitable for the problem at hand.

3.4 Gaining access

Gaining access to real data and people from the container shipping industry was no problem due to the background of the researchers. Both researchers are currently working within the maritime industry and have a background from the maritime management program at Chalmers, resulting in a large contact network. Contacts within different freight forwarding and shipping companies helped us locate people with the most relevant knowledge and expertise to participate in interviews. However, according to Bryman and Bell (2011), just because access has been gained to an organisation, this does not mean that people are willing to respond to the researchers' inquiries, as some topic may be considered sensitive. The next objective for the researchers was therefore to establish good relationships with the made contacts for them to better understand the particular needs and expectations the researchers had for the interviews.

3.5 Interviews

According to Bryman and Bell (2011), there are many challenges when performing qualitative research, where one of the main challenges for researchers is to collect own data.

Interviews are the most widely employed method in qualitative research (Bryman and Bell, 2011). In relation to the qualitative approach to business and management aspects, qualitative semi-structured interviews were performed to gather primary data. According to Bryman and Bell (2011), semi-structured interviews are good when fairly specific topics need to be covered, as it provides the interviewer and the interviewee with a degree of freedom where new follow up questions can be added if deemed necessary.

Another purpose of the interviews was to gain an understanding of the gap between the academic, shipping and IT industry where possible common grounds and solutions could be identified. The last and perhaps most significant purpose of the interviews was to identify the need for technical solutions within the container shipping industry where the author's personal biases had to be set aside.

The respondents from the academic side are listed in Appendix E. The questionnaires used for the interviews can be found in Appendix G and Appendix H.

3.6 Data set for training purposes

Original shipping document sets, including BLs, commercial invoices and packing lists for training purposes have been provided by, and with the expressed permission of, a global NVOCC service provider. All sensitive data, such as shipper, consignee, cargo, cargo value, container numbers and shipment references were altered as per the guidelines in 2.2.4.2 so that the original shipment cannot be traced. The total data set consists of 455 documents with an even distribution of BLs, invoices and packing lists. The shipping documents represent a wide spread of import shipments made by large and small Swedish companies covering both LCL, Less than Container Load, and FCL, Full Container Load, shipments.

3.7 Preparation of data

In order to use the data for training and testing purposes, the data must be pre-processed to match the expected algorithm input. The original documents are mainly received in image format, e.g. PDF and JPG format, where such input cannot be directly applied to most ML algorithms, mainly because of too large images which would need to be resized. Before processing the data, it is decided that classification would be performed on documents as images rather than text. Image classification resembles human behaviour better as the practice of sorting the documents into correct classes is done visually based on the framework and structure of the document, rather than reading through its content. Data preparation and pre-processing will be performed in MATLAB to fit the requirements of different image classification learning algorithms.

3.8 Development

The experimentation will be performed with the CNN for classification purposes. The task of the CNN is to classify the input data into three different categories, BLs, invoices and packing lists.

CNN's provide many benefits to the user as it allows for raw data input in .jpg format and has several pre-trained networks, such as GoogLeNet, ResNet and AlexNet available. This process is known as transfer learning where a network has been previously trained on millions of images to build good learning parameters. The same network is then retrained on the authors own data with the new categories.

Due to superior performance, the AlexNet is chosen which contains eight layers with weights. Briefly explained, it consists of five convolutional layers and three fully-connected layers which output is fed to a final 1000-way softmax (Krizhevsky, Sutskever and Hinton, 2012). The layers included in the model, i.e. the neural network topology, are listed below:

```
1   'data'    Image Input                   227x227x3 images with 'zerocenter' normalization
2   'conv1'   Convolution                   96 11x11x3 convolutions with stride [4  4] and padding [0  0  0  0]
3   'relu1'   ReLU                          ReLU
4   'norm1'   Cross Channel                 Normalization   cross channel normalization with 5 channels per element
5   'pool1'   Max Pooling                   3x3 max pooling with stride [2  2] and padding [0  0  0  0]
6   'conv2'   Convolution                   256 5x5x48 convolutions with stride [1  1] and padding [2  2  2  2]
7   'relu2'   ReLU                          ReLU
8   'norm2'   Cross Channel                 Normalization   cross channel normalization with 5 channels per element
9   'pool2'   Max Pooling                   3x3 max pooling with stride [2  2] and padding [0  0  0  0]
10  'conv3'   Convolution                   384 3x3x256 convolutions with stride [1  1] and padding [1  1  1  1]
11  'relu3'   ReLU                          ReLU
12  'conv4'   Convolution                   384 3x3x192 convolutions with stride [1  1] and padding [1  1  1  1]
13  'relu4'   ReLU                          ReLU
14  'conv5'   Convolution                   256 3x3x192 convolutions with stride [1  1] and padding [1  1  1  1]
15  'relu5'   ReLU                          ReLU
16  'pool5'   Max Pooling                   3x3 max pooling with stride [2  2] and padding [0  0  0  0]
17  'fc6'     Fully Connected               4096 fully connected layer
18  'relu6'   ReLU                          ReLU
19  'drop6'   Dropout                       50% dropout
20  'fc7'     Fully Connected               4096 fully connected layer
21  'relu7'   ReLU                          ReLU
22  'drop7'   Dropout                       50% dropout
23  'fc8'     Fully Connected               1000 fully connected layer
24  'prob'    Softmax                       softmax
25  'output'  Classification Output         crossentropyex with 'tench', 'goldfish', and 998 other classes
```

3.9 Evaluation

When validating the algorithm, the data is split as per 2.2.4.1, where 70 percent of the data will be used for training purposes. The data used for training the algorithm is randomised to ensure that instances from each file are used to not introduce any biases. The remaining 30 percent will be used for testing and validation of how well the algorithm performs.

The measurements that will be used to evaluate the algorithm are accuracy, calculated as per section 2.2.4.2 on Algorithm evaluation, and speed. Because the data set consist of an equal distribution of the three different classes the baseline for classification is calculated accordingly. As the CNN is a very capable algorithm, results above 0.8 are expected. The closer the result is to $Y = 1$, the better the algorithm is. While this is not very realistic, a minimum requirement for the classifier to be valuable is set at $Y = 0.8$.

The algorithm will be run ten times on a single GPU, allowing for very fast training sessions. For each run, the accuracy and loss function is calculated and plotted against the number of iterations to allow for visual representation. The values are recorded for each run, and the mean of the ten measurements is presented as the most accurate result as per 2.2.4.2. The need for several runs is caused by the randomisation of the data, where different accuracy and learning outcomes will be generated in every attempt.

3.10 SWOT Analysis

Within project management, a commonly used tool for strategic planning is the SWOT analysis, where SWOT stands for Strengths, Weaknesses, Opportunities and Threats (Kerzner, 2015). The purpose is to find solutions to complex problems, in addition to identifying the external and internal factors related to the problems in questions, within a project or a business venture (Kerzner, 2015). Strengths and weaknesses are the internal resources and assess the capability to solve the problem, while opportunities and threats are the external results that may or may not occur depending on if the analysis produces a potential solution to the problem (Kerzner, 2015).

3.11 Ethical considerations

Ethical questions are critical to discuss as they can disturb the integrity of the whole piece of research (Bryman and Bell, 2011). These questions come in many forms and shapes, and the researchers have been extra careful when setting up the research strategy and design, as well as when collecting primary data not to conduct any ethical violations.

Collecting primary data from the industry, where both authors are currently active, from own organisations and competitors may pose various ethical dilemmas. Some information on internal systems, processes, developments and internal projects were shared off the record and is therefore knowingly excluded from the thesis. The respondents from the shipping and IT industry also requested to be anonymous due to the sensitivity of some of the information shared. The companies from the shipping and IT industry have been made anonymous as well to prevent any type of repercussions to anyone and to allow for as much information as possible to be shared without compromising the integrity of any of the respondents. The authors have taken into consideration the legal requirements related to the data protection and privacy for the engaged individuals, and to hinder the export of any personal data outside the EU and EEA, in accordance with the new GDPR regulation.

The authors have throughout all interviews been very clear that they are both students and employees at their respective organisations so that respondents are aware of whom they are sharing their information with. Thus, the results of the qualitative research method can become more trustworthy by ensuring high integrity.

3.12 Summary

A mixed methods research is conducted to cover qualitative and quantitative topics of the interdisciplinary study. The goal of the quantitative method is to determine whether the predictive generalisations of the proposed research questions hold true and the goal of the qualitative approach is to achieve a deeper understanding of the subject. A design thinking model is applied as it provides a clear framework on how to solve an initial idea or inspiration. A literature review is conducted to gain knowledge on the topic of ML, digitalisation, strategy and maritime logistics.

Semi-structured interviews are conducted with people from tech and maritime industry to provide input to the qualitative part of the thesis. A wide range of people is interviewed in order to get a good understanding of the current situation and issues going forward.

Data set is provided from a global NVOCC for training and testing purposes. The data is altered through Artificial data synthesis due to the sensitive nature of the commercial documents. The data is also pre-processed for the input requirements of the convolutional neural network. A CNN is used for training and testing purposes. The outcome is evaluated in the performance measures of accuracy and speed.

The researchers have been extra careful when setting up the research strategy and design, as well as when collecting primary data and performing algorithm testing not to conduct any ethical violations.

4. Results

In this section, the results are presented. The first part consists of the quantitative results from document classification testing and the qualitative results from academic, tech and container shipping enquiries.

4.1 Document classification

Data preparation, development, training and testing was performed as per section 3.6 through 3.9

The CNN managed to build good learning parameters for classifying shipping documents as by reading them as images, where high training and validation accuracy was achieved in the developed model. For computational efficiency different number of epochs were tested to investigate how much training was required for the model to develop good learning parameters. One epoch is one complete presentation of the data set for the learning algorithm, where three epochs were in the majority of attempts enough for the model to build good learning parameters.

Accuracy and time were measured and compared between three and four epochs. Training is performed ten times for each setting. The validation accuracy as per the ten-fold cross-validation method and elapsed time for training are presented in Figure 12.

Runs	Results for three epochs		Results for four epochs	
	Validation accuracy (%)	Elapsed time (s)	Validation accuracy (%)	Elapsed time (s)
1	99,25	12	99,25	15
2	95,49	12	100	15
3	100	12	100	15
4	100	13	98,5	15
5	100	11	100	16
6	100	12	100	15
7	100	11	100	15
8	98,5	12	96,24	15
9	93,23	11	99,25	15
10	98,5	11	96,24	15
Average	**98,497**	**11,7**	**98,948**	**15,1**

Figure 12: Shipping document classification results

As can be seen, the average accuracy does not differ much between three and four epochs where only marginal improvements of 0,451 percent are generated through additional training. Measured on individual attempts, however, the gap between validation accuracy in the worst performing instances is 93,23 vs 96,24 percent, resulting in a difference of 3,01 percent. For some individual instances, it is believed that additional training would be beneficial.

The time required for training when increasing from three to four epochs increased by 29 percent. As the authors' data set is rather small, a time difference in a few seconds does not have much impact. However, if deployed in real scale, the increase in time would be considerable.

To visualise, the learning curve is plotted as accuracy vs nr of iterations, with scores distributed over three and four epochs. Training and validation accuracy is chosen as relevant measures, where smoothed training accuracy is included for a smoother graph. The plotted measures are represented as per Figure 13.

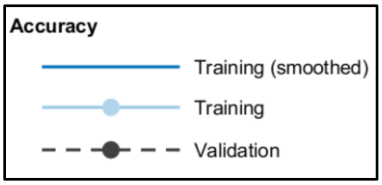

Figure 13: Legend

The learning curves from the highest and lowest accuracy from three and four epochs are presented in below and elaborated further. While ten training sessions were performed for both three and four epochs, the authors decide not to elaborate on every attempt as most share similar characteristics and patterns. Visualization is an important step for being able to draw conclusions and to further improve the model.

4.1.1 Runs with three epochs

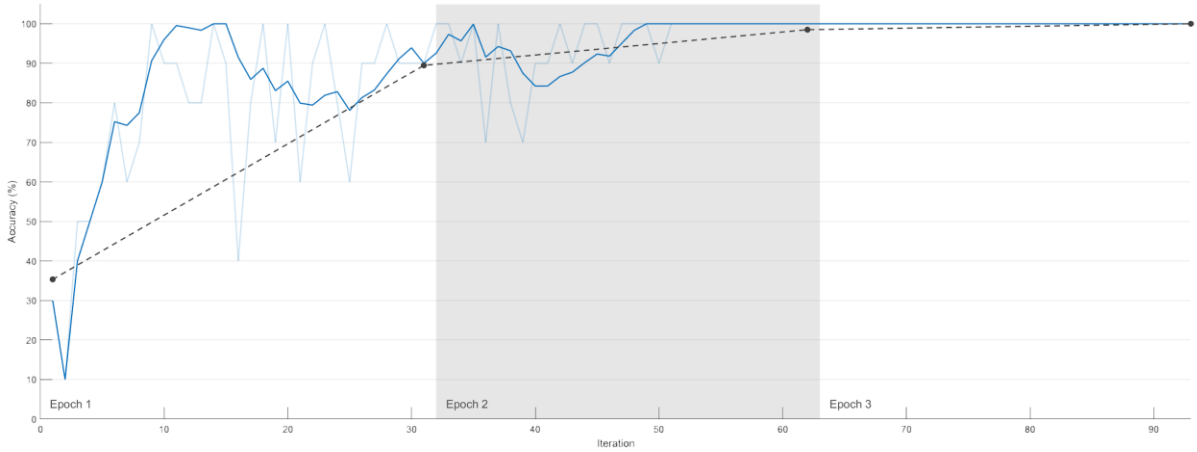

Figure 14: Learning visualisation run five, MATLAB graph

Figure 14 shows how the learning parameters are developed for the input data in training attempt five, three epochs, where training and validation accuracy reached 100 percent. It is seen that the model manages to create good learning parameters in an early stage where training and validation accuracy reach levels of 80 percent at the end of the first epoch. Further improvements are made during epoch two where the training and validation accuracy stabilises. Finally, the model reaches full training and validation accuracy in the last quarter of epoch three.

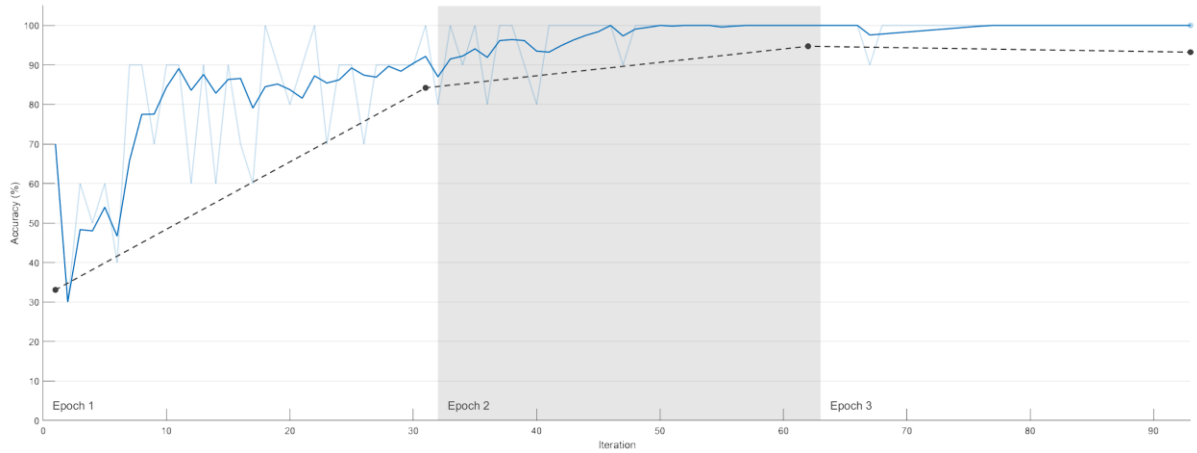

Figure 15: Learning visualisation run nine, MATLAB graph

Figure 15 shows how the learning parameters are developed for the input data in training attempt nine, three epochs. In this attempt, training accuracy reached levels of 100 percent while validation accuracy only

achieved 93,23 percent. In this attempt, it is believed that the model overfitted the data and as a result did not manage to generalise new examples equally well, thus higher training accuracy than validation accuracy. While 93,23 percent accuracy is still good, the learning parameters developed in this run are the worst for all ten attempts with three epochs. The overfitting is believed to be a result of the data randomisation where additional epochs would be needed to enhance the accuracy further.

4.1.2 Runs with four epochs

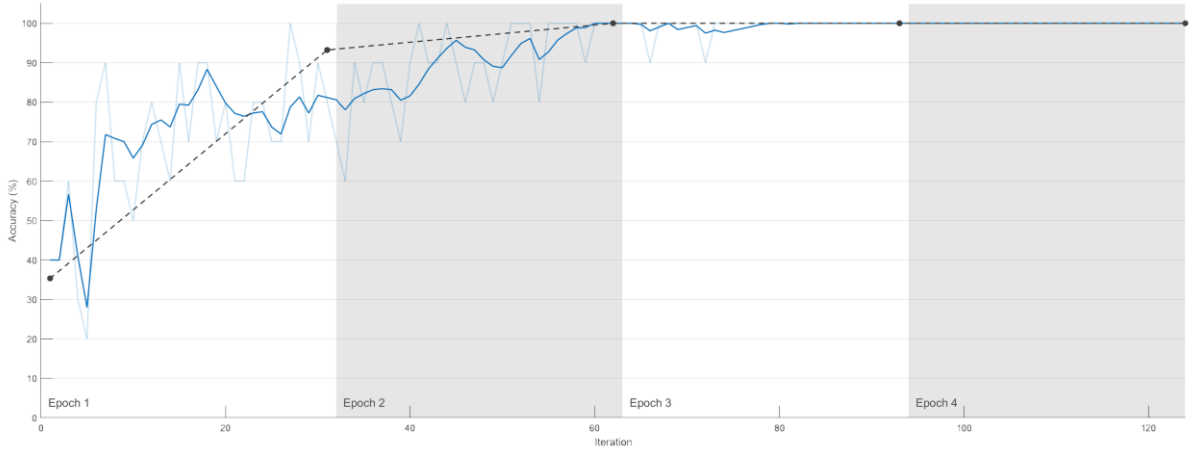

Figure 16: Learning visualisation run seven, MATLAB graph

Figure 16 shows how the learning parameters are developed for the input data in training attempt seven, four epochs, where training and validation accuracy reached 100 percent. It is seen that the model manages to create good learning parameters in an early stage where training and validation accuracy reach levels of 80 and 90 percent respectively at the end of the first epoch. Further improvements are made during epoch two where the training and validation accuracy stabilises. Finally, the model reaches full training and validation accuracy in the second half of epoch three where it remains until the end of epoch four.

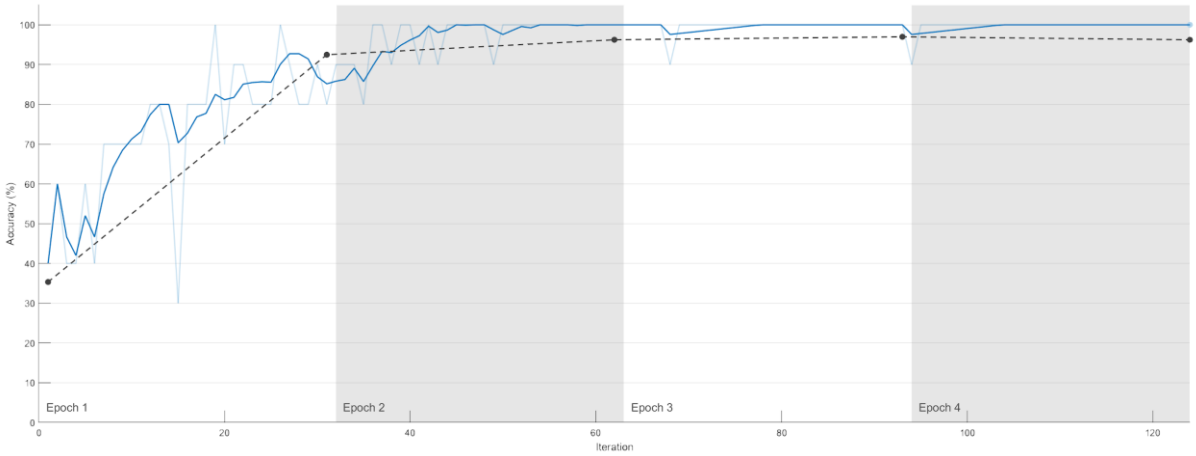

Figure 17: Learning visualisation run eight, MATLAB graph

Figure 17 shows how the learning parameters are developed for the input data in training attempt eight, four epochs. In this attempt, training accuracy reached levels of 100 percent while validation accuracy only achieved 96,24 percent. In this attempts, it is believed that the model overfitted the data and as a result did not manage to generalise new examples equally well, thus higher training accuracy than validation accuracy. While 96,24 percent accuracy is still good, the learning parameters developed in this run are the worst for all ten attempts with four epochs.

As an additional feature and for the authors to be able to validate the accuracy, the model is requested to display three sample validation images with their predicted labels. An example can be seen in Figure 18 where the model has successfully managed to classify a BL, an invoice and a packing list with the correct output label for all ten runs with both three and four epochs. For this reason, only one figure is used to represent the outcome. The label is stated above the displayed image of the input document.

Figure 18: Labelled output

The CNN has proven to be a successful algorithm for classification of the authors' data set. Further comments on document classification with advantages and disadvantages is covered in section 5, discussion.

4.2 Enquiries

This section provides a synthesized text of the collected responses on the questionnaires (Appendix G and Appendix H) provided by both the academic and industry respondents. The quotes provided can be found in Appendix F.

4.2.1 Academic enquiry

In order to efficiently gather a relatively large collection of questions, reflections and statements from this sector, which could, in turn, contribute to the topic of this thesis, two important conferences (see Appendix E) were attended in which many relevant speakers were invited. To synthesise and insert the contributions from the seminars appropriately, the gathered material will be paraphrased to fit the context of the research inquiries, of course without any alterations of the actual intent and meaning of the expressions. All seminars are uploaded and available at Chalmers and GAIAs own YouTube channels, which makes the gathered material easily verifiable.

4.2.1.1 Digitalisation Conference

On the topic of AI, Dr Fersman mentioned that even if many current market actors who are active within this field, express that they are dealing with AI, in fact, they are dealing with something called ‘digital assistance', like in the case of Microsoft's Cortana, Apple's Siri and IBM's Watson. She further stated that the accepted definition of ML by Ericsson is both ML and AI, being on the one hand driven by massive amounts of data managed by data science, and simultaneously being driven by facts and knowledge managed by logic. Because the value is in the data, by having enough of it regarding different issues, then the problem solving is rather simple she states, and if the stored data in a ‘data lake' is sorted and filtered, then a large part of the data processing is practically done. The main objective is to actually turn this raw data as an unpolished diamond and polish it in order to make it shine, as she phrases it.

Dr Fersman went on to say that rather than telling the computer what to do, from the perspective of cognitive science it is more reasonable to use the technology by asking questions on what it has done. The usage of ML on data should be applied with reasoning based on the target business or industry, and regarding key findings she has made in her own research, it is crucial to ensure safety and trust as well as to have the right data at right time and place. The applied context is to be connected with the business objectives of the respective company, and the potential of data applications is in its interoperability across domains. In turn, this will drive new value and enable completely automated operations, and management, which she states will be necessary in the future. Significant challenges for data applications is to achieve high availability of connectivity as well as heterogeneous data streams and modelling formalisms.

Dr Kahl spoke on the topic of ‘Computer Vision after Deep Learning’, in particular with regards to the task of detecting and classifying images, as well as instance segmentation of identified objects contained in images. He mentioned that even if there still exists some challenges within this field, the technological development has been very rapid. At the same time as there are “*no more low-hanging fruit*”, there is still a need for getting output which is “*something more semantically rather than just a simple yes or no response*”. He further stated that later developments within DL has been achieved by combining geometry and semantics, and are more application-driven. However, he stated the fact that new challenges require good knowledge in mathematics which goes against this type of development, as the general perception is that “*mathematical equations are not needed anymore*”.

Continuing on the topic of semantics and linguistics, Dr Lazaridou stated that language is interactive and a multi-party agent process, rather than consisting of self-interested agents. Due to this, she stated that it is possible to express the inference of this matter that Language equals the sum of Structure, consisting of pattern learning, in particular, NLP techniques (about 95 percent), and a particular Function. It is not necessary to reinvent the wheel, she stated, rather the aim should be to profit from previously gathered knowledge. During her presentation, she cited an article about sequence-2-sequence by Sutskever et al. (2014) in which the researchers studied the relationship of context, i.e. previous sentences in a text, and a predicted reply to, e.g. a question or a random statement.

Dr Johansson gave the following definition of NLP in his presentation:

> *"the engineering discipline where we design systems that process text automatically."*

He continued by stating that ML became dominant in NLP techniques in the mid-1990's, and since then, e.g. automatic translation has become increasingly useful. Recent developments in information extraction have gone so far that there currently exists software which can automatically create summaries of entire documents. However, he went on to ask the question regarding if NLP is considered to be an applied science, what particular science is the one actually applied? The reason for the question is mainly because the aim for any researcher is to conduct an analysis in a controlled and structured manner by using models which are explainable. He further stated that the researcher wants to be reliable in terms of being able to give guarantees, as well as to know in forehand how to avoid spurious errors including which systems are most probable to fail. This is to finally be able to justify the way in which the analysis has actually been conducted.

In his presentation, Dr Johansson provided a brief illustration on the way in which this type of analysis was conducted in 2008 compared to in 2018, in which previous feature engineering and intermediate analysis of relevant parameters have been proven to be both time-consuming and useless. He cited two articles, the first by Johansson and Nugues titled 'Dependency-based Semantic Role Labeling of PropBank', published in 2008, and the second by Zhou and Xu titled 'End-to-end Learning of Semantic Role Labeling Using Recurrent Neural Networks', published in 2015, with which he concluded that the analysis often works better when inserting the data into a "*big messy neural network*", as he phrased it. Dr Johansson went on to restate questions asked by other PhD colleagues of his, e.g. Devdatt Dubhashi and Asad Sayeed. The questions were the following:

> *"Should we think on how to build data sets; how does language emerge in a cooperative task; how can we design reliable domain-specific NLP systems; how can we supervise ML for NLP using semantics?"*

4.2.1.2 GAIA conference

Mr Kågebäck started off the conference by asking a highly relevant question on the topic of 'Learning to Communicate using Deep Reinforcement Learning', in which he summarised the questions asked during the seminars from the previous digitalisation conference:

> *"Rather than finding out what is the most probable response to a problem, how should I formulate the problem to reach my intended goal?"*

Dr Gillblad continued in this line by asking the question on the topic of AI Beyond ML, regarding how efficient and actionable are current ML techniques. In the case of DL being an ML technique, "*it is like the new steam engine of our days*", as it is both powerful and able to perform work which is fundamental to create a revolution, "*but it is neither efficient nor is it deployable*". He stated further that ML in general still lacked in efficiency in terms of computation, development time and terms of data, as learning from small data is essential. He went on to state that ML lacked severely in interpretability and usability of results, in which it is possible to draw a parallel on the previous question asked by Mr Kågebäck. The main areas in which it lacks are explainable models, timely results, correlation vs causation and finally uncertainty measures. Because humans are generally bad to determine causation compared to computers, it is crucial to create explainable models, in the sense that the black-box solutions need to be broken up.

Further on in the seminar, Dr Gillblad continued on this line by asking how it would be possible to move beyond DL, even ML, and actually perform computations on abstract concepts? Many fundamentals remain unsolved, mainly regarding why DL actually work and how engineers can achieve robust engineering. On the subtopic of expanding the environment for ML, he mentioned that real environments have constraints, e.g. with data sources in regard to access, ownership, security and privacy, which very much relates to the topic of this thesis. Another thing he mentioned which related equally much to the thesis topic, is the

essentiality of multimodality in the application of ML techniques, e.g. with the tasks of simultaneously looking at a text and a picture together.

On the topic of 'Scoring Data', Dr Simmons discussed the process of how to build a quality model intended to be used for performing quality measures and evaluations as well as for specifying certain requirements. The example presented during her seminar included risk scoring, which in this case was more of a categorical measure related to a particular type of threat. An important aspect is that, generally, the performance metrics need to be useful for the end-user. E.g. if the score or number provided is consistently too high or too low, in order for it to be useful, "*it doesn't matter if it's technically correct*", she stated. As previously mentioned in the theory part of this thesis, precision, or specificity, is an essential metric in cases where the output must have high confidence. Recall, or sensitivity, on the other hand, is more relevant in cases where it is important not to miss any instances, in which one of these two metrics needs to be prioritised. A rule-based model with sub scores would allow for customisation, and to this point, she mentioned that black box models can be somewhat opaque and do not offer as much utility for end users. Although the latter type of models performs extraordinarily well, this statement is in line with what was previously mentioned by Dr Gillblad.

Dr Simmons further stated that a clear strategy is crucial in the building process for it to result in a useful model for the end users. This relates to one of the questions asked by Dr Johansson in the previous conference. She suggests that samples should be gathered for testing, with automated classification, and also to make results transparent to the users. The process could take a Top-Down approach, which consists of looking at some samples manually and try to understand the problem, which would be good for generating ideas but not for the case of automation. A reversed approach, being Bottom-Up, would make the model more scalable compared to the previously mentioned approach, but it can only work provided there is a strategy to base the work on. She further stated that to build a model successfully, it is critical to find subject matter experts from which expert opinions can be received. Also, categorical judgements make the feedback more useful, e.g. simple 'yes/no' verdict, as well as with clearly defined feedback, e.g. 'Would you take x action based on y data'.

4.2.2 Tech industry enquiry

Two software development companies were interviewed to get an understanding of their challenges when providing solutions to customers. Both companies are actively providing data-driven and AI-powered solutions to their customers where one of the companies are working exclusively within the shipping industry. The companies have been provided fictional names. The quotes in this subsection are translated from Swedish, and can be found in Appendix F. The questionnaire used for this enquiry can be found in Appendix G.

4.2.2.1 Company Alpha

Alpha is a software development company that operates exclusively in the shipping industry. Interviews were carried out with a software developer and the CEO. The key findings and notable quotes are presented in below.

Alpha experiences that it is quite hard to build and sell software to the shipping industry, where several reasons are presented. The first reason is the IT infrastructural barriers, where most shipping companies already operate a variety of different systems. Some actors run as many as fifteen or twenty systems for their day-to-day operations, where new systems must be compatible and able to integrate with already existing infrastructure that has more than often been provided by several different suppliers. Customers rarely want to implement new systems that do just one thing. It must be all-in-one solutions where one system can handle a wide range of tasks. As most shipping companies are already operating several different systems, it is hard to motivate why additional systems are needed.

The performance requirement for new software is very high where many actors are sceptic about implementing brand new software that has not yet been commercially proven. As shipping is a low margin industry, the cost from one error could take ten or twenty shipments to make up for. Therefore, IT systems have to work perfectly, every day and every time as it can get costly when things go wrong.

As shipping is not a high-tech industry, customers are having a hard time specifying what they need. One of the key enablers for Alpha when providing solutions to the industry is that most employees have a past within the maritime industry where domain expertise and contact network allows them to provide and explain solutions in a way that customers can understand. Without this expertise, Alpha believes that they would not be able to operate towards the shipping segment. It was further elaborated;

> *"The shipping industry is not interested in software, and it is very unsexy coming from an IT company trying to provide them with software solutions when they don't even know what they need as they do not have the right competence." [C.1]*

The shipping industry has a lot of senior people in leading positions who do not see the benefits of adopting new technology. Looking at CEO's and boards of directors for shipping companies, there is a clear overrepresentation of senior staff members who often have spent their entire careers within the shipping industry. It was stated;

> *"The possibilities we see today in terms of technical solutions simply did not exist when these people were running the day-to-day operations. This makes it hard for them to see why such systems would be needed." [C.2]*

It is believed that a generational shift is needed for shipping companies to realise the true potential of smart software solutions truly.

Another problem is that everyone says they want to do AI, while very few companies are willing to go from idea to reality. In the shipping industry especially, there is also a lack of first movers as most companies apply a 'wait and see' attitude. AI has seen a very slow adoption rate within the industry but is believed to have a large long-term impact. The few companies who actually want to be first movers lack a clear understanding of how AI works, and some are even requesting AI solutions without specifying what they want them to do. There are very few companies within the shipping industry that understands AI and the advantages of applying such solutions still needs to be clearly illustrated. While the following quote might provoke a certain chuckle, it carries an underlying truth;

> *"AI in shipping is like sex in high school, everyone is talking about it, but no one knows what it is." [C.3]*

An important matter when trying to solve problems with software for customers in the shipping industry is to investigate who the problem belongs to. It was highlighted that every shipping company have more or less the same problems. When everyone is having the same problem, it is hard to justify investments into solving that particular problem. It was further stated;

> *"Many problems in the industry are as of today everyone's problem, where everyone's problem is no one's problem. As long as everyone have the same problems, everyone is on the same level." [C.4]*

A clear benefit when providing solutions to the shipping industry is that other industries have already done this journey. It was from a software development perspective elaborated that;

> *"There are many other industries that are way ahead when it comes to software solutions and automation, which is good, we can just look at what they are doing and do the same thing." [C.5]*

It was concluded that that the shipping industry is a perfect candidate for automation and smart software solutions as there are a lot of high paid employees who do a relatively simple and routine-based job. The administrative cost for handling cargo is rather high, where automation can help bring down these costs. Alpha further states that the shipping industry is not an easy market to penetrate and that domain knowledge is required from the supplier.

4.2.2.2 IT company Beta

Beta is a software development company that provide solutions mainly within the telecom industry. Interviews were carried out with an AI expert, a software developer and the CEO. The key findings and notable quotes are presented in below.

Beta has seen a huge rise in customers who want to do AI during recent years. Beta claims that there is an ongoing AI revolution within the software development industry which have provided several new programming languages, algorithms and frameworks to solve customers' problems using data.

When it comes to providing AI solutions, technology is rarely a problem. Rather than using the latest algorithms, it is about generating value for the customer. Beta claims that simpler models are usually better for creating customer value. While some of the latest and more complex algorithms can map complex functions, there is always the risk that the model makes connections that do not exist. The problem with more advanced algorithms, especially within deep learning, is that it is hard to explain what is going on.

While many customers are requesting smart and high-tech solutions, rather than being high-tech, it is more important to understand the customer needs and to generate value. Many times, it turns out that the customer wants AI just because it is AI, when in fact what they need is something way more straightforward. It was further elaborated;

> *"Some customers are requesting AI without even knowing what it is. Someone has read an article once and thought this sounds good" [C.6]*

When it comes to solution selling, the most important part for the customer is that the solution works. Even if many customers are requesting AI, there are very few who actually care about the particular technology behind a specific solution. It was further elaborated that customers often get confused when a software developer goes into to specific details and that AI solutions are often requested for marketing and PR purposes. Most customers are satisfied with black box solutions.

Beta also provide consulting services within AI where they have been involved in assessing a wide range of projects. Beta was presented with the author's thesis project, suggested solution and initial idea for a type of software intended for the container shipping industry, in order to assess its potential. Beta were surprised when they heard about current processes within the container shipping industry where it was concluded that current processes are 'relics from the past'. The CEO elaborated;

> *"This is more or less what other industries looked like ten or fifteen years ago. We have been developing similar solutions to other industries for years." [C.7]*

Beta claimed that a prototype software for automated document handling using AI and information extraction could be developed based on the authors' data set. Technology for developing such software would not be a problem and combining off-the-shelf solutions could probably generate a software with an overall accuracy between 80 to 90 percent as documents are machine written and follow similar frameworks. Even the timeframe to develop would take Beta approximately two weeks.

4.2.3 Shipping industry enquiry

Ocean freight experts, managers and operational workers from five different companies within container shipping were interviewed to get an understanding of current systems, challenges and digitalisation strategies. All companies are among the top six largest carriers or freight forwarders in the world. The key findings and notable quotes are presented below, with two particular companies used for illustration reasons. The companies have been provided fictional names. The questionnaire used for this enquiry can be found in Appendix H.

4.2.3.1 Current systems

There is an apparent gap when it comes to back and front-office solutions. While most front-office solutions that can be seen and interacted with by the customer are modern and up to date, allowing, e.g. data-driven tracking, customer service and forecasting, back-office solutions used for the internal day-to-day operations are in most organisations severely inefficient. To further elaborate, back-office solutions refer to internal systems used by the employees who handle the day-to-day operations and front-office solutions refer to systems and platforms available for customers to for example book, manage and supervise their shipments.

It is motivated that the main effort is being put on front-office solutions as it can be seen by customers and stakeholders and generates an instant value. The container shipping industry is very competitive in terms of IT solutions to customers where being able to provide instant tracking, forecasting, quoting services and supply chain management platforms have the ability to decide whether a new account is won or lost. As the industry generally competes with price, these value-added services have a significant impact on the value proposition presented by forwarders and carriers.

The strategic focus on the front-office has resulted in back-office solutions being put aside where some organisations still apply internal systems originating from the 1980s. While this is not the case for every organisation, current internal systems are mainly being updated with additional features rather than replaced entirely. Many interviewees believe there is a limit to what current systems can achieve based solely on additional plugin features.

Attempts have been made by two of the interviewed organisations to improve their internal system through collaborations with global tech companies during recent years. While one of the interviewed companies claims to have succeeded and managed to get ahead of their competitors though custom-built systems, the other company's IT project failed and resulted in write-down costs of 345 million euros. It was stated that the main reason for the project failure was due to a lack of understanding of the shipping industry from the

IT provider. Due to this, most container shipping companies that are not first-movers turn to dedicated software providers to procure off-the-shelf solutions that have been commercially proven within the industry. The overall results of this approach are that everyone applies more or less the same systems and are experiencing the same issues.

Further, it was identified that front-office systems had created a skewed image of digitalisation within the industry. A few companies that appear more digitalised and high-tech through their front-office solutions and are labelled as best-in-class among carriers and forwarders apply the same or sometimes even worse back-office systems as their competitors.

While AI is a hot topic in container shipping and some organizations have already initiated collaborations with consulting and tech companies to investigate the potential of such technology, large organizations as interviewed for this thesis handle this question on a global level where little or no information on current developments and projects is available on local level in Gothenburg, Sweden.

4.2.3.2 Challenges

One of the main challenges when moving into a more automated and digitalised environment for shipping companies is the culture. Throughout the interviews, many interviewees from the operational level expressed their concern for disruptive forces and the potential change that AI and increased digitalisation could lead to. While agreeing that current internal systems are indeed in need of improvement, many interviewees are overall satisfied with the current state. One employee stated;

> *"We are already digital. When I started working here, we didn't even have email, and we were faxing and sending documents by bike courier. Now we can do everything ourselves in the computer" [C.8]*

There is a general feeling of scepticism against digitalisation and automated software solutions where many examples were brought up by both operational workers and managers when technology had not worked as intended. The shipping industry is also a fairly small circle where the word on failures, especially regarding IT, gets around. To further elaborate the downside of increased automation it was stated;

> *"Just look at what happened when X implemented automated invoicing, there wasn't a damn invoice that was correct for months" [C.9]*

> *"Look at Y when the industry was hit with malware last year, everyone else was doing just fine while they couldn't do anything as they rely too much on technology" [C.10]*

Another challenge is the knowledge gap in terms of technology where there seem to be a general misunderstanding and fear of what digitalisation means. Many interviewees believe that rather than assisting, further digitalisation and the implementation of smart systems will make many employees obsolete due to automation of various tasks and processes.

Many challenges are organisational where people from both operational level and management experienced problems with the size, hierarchy and bureaucracy of their organisation. A local Head of IT estimated that it would take between three and five years to implement a new software on a local level as it has to go through the bureaucracy of the organisation all the way to the headquarter, which is not located in Sweden, to be approved before being considered. It was further elaborated that the noise in the communication channels often results in the message being lost somewhere along the way.

Another organisational challenge is the silo mentality that arises in large organisations. It was expressed from the management perspective that the same organisation may look very different in terms of goals and culture comparing between different countries and regions. A general feeling of frustration was displayed that headquarters often do not see and understand what is happening on a local level where it was further elaborated by a local manager;

> *"Sometimes our local goals do not match with the organisation's global goals, what they see from HQ is often something completely different from what we see in day-to-day operations, we say we need this, they say we need something else" [C.11]*

4.2.3.3 Digitalisation strategies

The topic of digital strategies was hard for most interviewees to answer. While every organisation have digital strategies to improve and update their current systems and processes, the question whether this includes data-driven solutions such as ML was not confirmed. Some interviewees stated that their

organisation had started small-scale projects to investigate the potential of such solutions, where no further comments could be given due to the sensitivity of the subject.

During the spring of 2018, container shipping companies have published reports on AI in logistics, where in collaboration with consulting companies such as Accenture and technology giants like IBM, benefits of applying AI in logistics have been presented. While these reports look promising, they still include several limitations of AI and ML which needs to be overcome.

As these reports are published from the corporate headquarters, little knowledge on the topic is available on a local level where most interviewees are unaware of this development. A local manager stated;

"Lucky for them (HQ) we are generating the money so that they can play with this stuff" [C.12]

4.2.3.4 Shipping company Delta

Delta is one of the largest carriers in the world. To further display the situation in container shipping the opinions expressed by employees at Delta are elaborated in below.

Delta wants to focus on their core business which is ocean transport. Initiatives have been taken from headquarters to increase value-adding logistics services such as door to door delivery, cargo insurance, customs clearance among others. There is currently too much work to do with the core business which results in opportunities being lost where employees do not have time to provide value-adding services to customers. It is estimated that additional value-adding logistic services are only provided for 20 percent of the total shipments.

Delta are experiencing problems from customers when trying to become more digital in front-office applications. Many customers do not want to use platforms or software as they prefer email correspondence being "the way things have always been done". As a result, Delta still mainly communicates by email, where different email boxes are available for different purposes.

Delta has a global service center located abroad which is responsible for many administrative tasks. The service center has a documentation department which is responsible for handling all incoming emails with attached documentation from customers. The service center handles all cases manually where staff are transferring information from paper to software systems by hand.

Current back-office systems are considered to be outdated and have only been updated by additional plugin features since the 1980s. The central operating system is considered to be fragile, as it is not particularly secure, neither capable enough to empower the staff using the systems, exposing them to the risk of human error and to make mistakes. Employees believe that there is great potential for improvement with new IT solutions which could save both time and resources. Administrative work is claimed to be very time consuming with highly standardised processes.

4.2.3.5 Shipping company Epsilon

Epsilon is one of the largest carriers in the world. To further display the situation in container shipping the opinions expressed by employees at Epsilon are elaborated in below.

Like in many other shipping companies, documentation related to shipments is currently manually handled and is provided by clients through email, being collected in a common mail inbox. There are no current EDI or other internal systems for document handling, as everything is handled through email.

Most of the inquiries from customers are related to bookings, information about their shipments and invoices. At the same time, most of the employees' daily tasks, approximately 80 percent, are administrative and not particularly focused on serving the customer by, e.g. responding to their inquiries mentioned above. This creates immense frustration from the customer's side as they want to be updated with status on their respective shipments. An indication that this might be the case for many other industry actors is found in one example of an Epsilon client which switched to another shipping line due to them being able to offer a lower price. The client eventually returned due to better customer service at Epsilon, even if this was the main reason for them leaving in the first place.

The currently used IT systems are developed in-house and are considered underdeveloped and outdated in terms of design and user-friendliness, compared to other IT-systems used individually by staff members outside of their work environment. The general feeling is that people could, on the one hand, lose their jobs by too much digitalisation. On the other hand, this change might entail more advantages than disadvantages for the staff. This is because the digitalisation would instead relieve staff members and put them on more strategically focused tasks, thus increasing excitement and engagement for the work.

The supply chain department of Epsilon is far ahead when it comes to using IT systems and technological application within their current work processes, compared to the ocean freight department. However, Epsilon does not consider themselves to be first movers as they choose to wait and see what other companies do. Both Headquarter and service center is located in the home country of Epsilon, i.e. not being located in Sweden.

All decision-making concerning everything from rates to amendation of BLs must be confirmed by headquarters, making Epsilon a very centralised and hierarchical organisation in terms of power which make internal processes time consuming and inefficient. In addition to this, there is also a substantial time difference as the headquarter is not located in Europe. Most amendments can therefore not be made until the next day when a reply has been received.

4.3 Summary

This section consists of the quantitative results of the document classification testing and the qualitative result of both the academic, technology and container shipping enquiries.

The quantitative part consisted of testing and training runs with three and four epochs. The output data consists of three classes, being BL, invoice and packing list as per section 3.6. The highest validation accuracy for both test runs was 100 percent, and the lowest was 93,23 percent for three epochs, and 96,24 percent for four epochs respectively. The average improvement was 0,451 percent which for this case could be considered as almost insignificant. However, the time required for the training runs differed by 29 percent between the epochs, which suggest that if deployed in real scale, the increase in time would be more substantial. Thus, it is concluded that it suffices with only three epochs for the model build good learning parameters.

The qualitative part being the outcome of the enquires, resulted in statements that affirmed both usages of digital solutions for back-office activities. Additionally, DL and Neural Networks algorithms were confirmed as being suitable technologies for automatic document classification tasks like the one being researched in this thesis. Apart from having better performance levels, they meet the requirements of ensuring privacy for the data owners and also increases security as it reduces the number of people involved in the handling process. Another critical aspect mentioned was that a clear strategy is crucial in the building process of the model for it be useful for the end users. Most IT-service providers have already developed highly advanced solutions which are often unknown to many potential customers, in particular, the shipping industry. One conclusion inferred from the results was that the strategy for the developed models was not aimed towards the shipping industry.

One of the main reasons why other organisations within the shipping industry did not implement existing solutions was their lack of skills and knowledge in these areas. Also due to the shipping industry being conservative and rather hard to change in general. Moreover, IT-service providers encounter difficulties when presenting these potential solutions to the industry, mainly due to their lack of understanding of the core business and how to identify particular daily issues experienced by the industry workforce.

5. Discussion

This section contains the analysis and discussion of the findings in relation to the theoretical framework.

5.1 Document classification

The classification testing performed with a CNN on BLs, commercial invoices and packing lists provide very good indications that this part of the process of an automated document handling pipeline may be done successfully. As the achieved training accuracy is very high, it was initially assumed that the model was overfitting the input data. Because the accuracy is equally high on the validation set, it is concluded by the authors that this is not the case and that the model has built good parameters for generalising new input data.

The accuracy that was achieved after tweaking features and hyperparameters may, however, be somewhat optimistic, as it is simply too good to be true. It was initially assumed that the accuracy would be somewhere at eighty percent, were in the initial test and train attempts, this was true. The algorithm performed poorly at first, where after trying out different approaches continuous improvements were made. The performance of ML systems, however, often says more about the input data than the algorithm itself, causing the authors to believe that the input data may not be sufficiently diverse to represent a real-life scenario. To further validate that classification of shipping documents can reach high accuracy levels, a much more large and diverse training and test set should be obtained.

One of the clear benefits in shipping and logistics is that all documents provided by a shipper, carrier or forwarder, share similar characteristics of the framework. The only thing that differs is the content related to particular shipments such as the invoice or BL number, cargo, origin or destination. This makes a BL, invoice or packing list from a specific shipper, carrier or forwarder when only seen a few times before, somewhat easy to classify for a neural network.

Another benefit in shipping and logistics is that shipments follow the Pareto rule, where a high percent of shipments for a carrier or forwarder is accounted for by few large customers who use the same suppliers for most of their shipments. This distribution promotes the use of ML tools as many documents are frequently recurring where a ML classification tool has good chances to learn the patterns of incoming documents.

While, as previously mentioned, many shipments do follow the exact same setup in regard to routing, shipper, consignee, cargo, and supplier there are always exceptions to this where special or ad-hoc shipments that stand out from the crowd do occur. While the classification algorithm managed to build good parameters to classify the frequently occurring shipments as used for input data, it may prove more challenging to classify shipments that are less frequently occurring. There is, however, a significant benefit in being able to classify the standard or common shipment documents as these represent a majority of the total trade.

If document classification would perform similarly in a real-life scenario, considerable time and cost savings on back-office operations could be made by automating the document handling process. In a container shipping scenario, where many of the large customers have more or less the same shipments every time, several times a week or month with volumes of hundreds or thousands of TEU a year from the same supplier or suppliers, a good enough system that is able to handle standard and frequent shipments could prove to be immensely beneficial.

5.2 Automated document handling

The administrative work that goes into document handling for container shipping is a time-consuming task where most organisations still apply manual labour handling methods for receiving, sorting and filing documents. Every container requires several different documents before it can be processed and customs cleared in order to leave the port, where these documents include as already mentioned a BL, commercial invoice, packing list as the minimum requirement.

The cost related to document handling from a carrier or freight forwarder perspective consists of direct costs for manual labour and indirect costs for the penalties that occur if document handling is not completed on time. As container shipping is a high volume low margin industry, small improvement measures and automation of repetitive administrative tasks may result in tremendous benefits in a long-term perspective.

While fully automated document handling requires many different functions from input to output, as well as experienced developers to integrate a new system into already existing infrastructure, document classification on shipping documents as investigated in this thesis do achieve good results.

A drawback of the CNN is however that input data require a specific size in order to be processed. The original documents have to be reduced in size to fit the input layer of the network where a reduction in size comes with the price of loss of information. If a CNN would be used as a document classifier in an automated document handling pipeline, it is crucial that information extraction is performed before the classification, as text information such as shipment or BL number is nearly impossible to extract after having reduced the document size as per the input requirements.

Throughout the course of this thesis work, several different end-to-end document handling tools have been investigated where different technical solutions have been applied. While ML has been the core idea for document classification in this thesis, conventional rule-based systems may work just as well in the cases of standard shipments. A combination of the two is also possible and would probably prove most beneficial, where rule-based systems can be used for easier tasks and state of the art ML algorithms can be used for challenging problems.

For automating document handling in container shipping, another key feature that has been identified is information extraction. While information extraction is outside of the scope of this thesis, it is perhaps the most important feature for an automated document handling system as it is a key element for extracting shipment number, BL number or container number from the documentation, in order to link the received documents to a specific shipment.

5.3 SWOT analysis

To further elaborate the pros and cons of automated document handling and the automation of additional administrative tasks, the topic of automation will be discussed within a SWOT analysis framework. The purpose of applying the SWOT analysis is to highlight that there are both strengths and weaknesses, as well as opportunities and threats in the development towards a more digitalised container shipping industry.

5.3.1 Strength

Forwarders and carriers are as of today very slimmed organisations where employees work more reactive than proactive. Automating time-consuming administrative tasks would increase productivity, decrease the risk of non-compliance and free a lot of human capital. It would also enable more proactive, cost-efficient and high-quality logistic services. Automation increases process control and reduces human error, where greater control and consistency also results in a higher product quality.

All of the strengths mentioned above are of course based on the hypothesis that systems work as intended, where the conducted enquiry from the shipping, academic and IT industry, as well as the authors' conducted model testing, provide good indications that this can already be done successfully. Automated systems are not intended to replace human labour. The main purpose is instead to assist employees to become more efficient and assist in the day to day operational tasks.

Automating administrative tasks also reduces risk related to the transportation of high-value cargo. As many people are involved in the supply chain where different operators have access to documentation over cargo and value, it occurs relatively frequently that high-value shipments get targeted by criminal organisations due to leaks from the supply chain administration. It is believed that reducing the "number of

eyes" on the documents, through automated document handling, will also decrease the risk of transporting cargo.

The suggested solution of automating administrative tasks also contributes to good health and well-being by reducing current burden of administrative work. Document handling is as previously mentioned a very non-stimulating task but has to be done due to legal and customer requirements.

5.3.2 Weaknesses

Automation creates a system dependency that is not experienced by many actors in the industry today. While every carrier and forwarder rely on their computers and internal systems to work to process shipments, organisations are staffed accordingly for the requirements of the day to day operations to be fulfilled. Automating administrative tasks will most likely result in staff reductions and take system dependency to the next level, where operational disruption and cost will occur if systems were to malfunction or fail.

Automation does to some degree result in a decreased understanding of the industry as employees are no longer required to handle all the administrative tasks from origin to destination. However, as the industry progresses, there is knowledge that is very relevant today that may not be as relevant in the future. During a transition phase towards a more digital and automated container shipping, manual processes and handling still need to be in the companies' knowledge base in order to handle tasks where the automation has failed.

Another weakness of automated systems powered by AI is explainability. Within DL that produces the best achieving learning algorithms today, it is even hard for developers to explain what the model is doing and how results are achieved. Even if the correct results are achieved, the question arises whether or not such a model should be applied and trusted when it cannot be explained how the correct result is produced.

5.3.3 Opportunities

The freeing of human capital allows employees to work more proactive, focus on more value adding and customer-focused activities which generate increased customer satisfaction. Automation and the freeing of human capital also allow for companies to handle higher volumes with existing staff or reduce cost by allowing the same work to be conducted with fewer employees. An increased customer focus and customer satisfaction result in a lower customer churn rates and increased long-term profitability.

The freeing of human capital allows for improvement and expansion of current business models. A current and somewhat widespread issue within the container shipping industry is that employees rarely have time to look up from what they are doing and think about improvement measures. Allowing for such a work environment is believed to generate a lot of additional benefits in terms of continuous improvements and business expansion into new areas and services.

While this thesis has aimed to investigate a very specific area of application for a learning algorithm, the container shipping industry provides many back and front-office processes where smart systems could be applied. As many administrative tasks can be specified in terms of specific processes, domain knowledge from the shipping industry combined with IT experts could allow for various other areas to be automated and improved.

Automated document handling tools may, e.g., serve as innovative solutions to the shipping industry, where traditional and conservative methods of manual handling still dominate. Increased digitalisation will assist the shipping industry in being more compatible with other industries who have already come a long way in their digitalisation efforts and ensure future competitive advantages and resilience from industry disruptors.

The container shipping industry generates a lot of data. While this data is used for a few applications and services today, shipping companies know very little about their customers compared to tech-based companies who track customers every click of a button. Smart systems could be used to make good use of the data generated by shipping companies. As ML systems are data-driven, many fields of application are possible once the industry learns to apply such technology.

5.3.4 Threats

Increased automation and digitalisation open up for more cyber threats. While container shipping has been somewhat of a low priority for cyber-attacks in the past, recent years have seen a rise in cyber-attacks against both vessels, carriers and forwarders. Going forward, increased security measures will have to be

taken to ensure that systems are resilient to attacks and disruptions. As can be seen in recent cases of cyber-attacks, this can result in damages to entire infrastructures, costing more than what many market actors can withstand.

Recent years discussion about the employment impact of the disruptive change ahead that is automation powered by smart systems is polarised between those who foresee limitless opportunities in new emerging job categories and those who foresee a massive labour substitution and displacement of jobs. While both are possible, it cannot be denied that current disruptive changes to business models will have a profound impact on the employment landscape over the coming years. As automation often goes hand in hand with worker displacement, a huge threat is the social impacts that this may result in. While this could be the case, it should be stressed that in container shipping many organisations have already outsourced their simple administrative tasks to global service centers in low-income countries, where these are most likely the first to disappear before affecting employees at local levels.

Increased automation makes the industry more vulnerable to disruptors as systems are easier to replicate than employees. A major discussion within the freight forwarding industry, who do not have any hard assets as compared to carriers who own and operate their vessels, is how the freight forwarding model will stand against more digital competitors when the industry enters the digital arena.

Freight forwarding has during recent years seen a rise of digital freight forwarders with automated processes which are trying to gain market shares. The greatest challenge faced by the digital forwarder is, however, the over-reliance on technology. Whereas automation has a definite appeal, the shipping industry is a people industry where the moment something goes wrong, technology goes out the window and customers, carriers and forwarders revert to the most analogue of activities, speaking directly to another human.

5.4 Digitalisation in the container shipping industry

Digitalisation, AI and blockchain are currently hot topics in many industries, where container shipping is no exception. Recent years have seen a large increase in projects and collaborations between industry leaders from the shipping and IT industry, where during the spring of 2018 alone, several new projects within AI and blockchain have been announced publicly.

While this is the case, the shipping industry is far from digital in its current back-office practices compared to many other industries. The problem for the shipping industry is not as much technical as it is organisational and strategic as many of the technical solutions that are requested from the industry for back and front-office applications and automation are already existing.

The shipping industry is known for being very conservative in its ways and there is currently a huge difference in how companies within the industry are applying and investigating new technology. While some companies are already claiming to have fully automated internal processes, others are openly admitting that they still apply traditional manual handling methods, using at least twenty-year-old software. Among carriers and freight forwarders, industry leaders such as Maersk Line and DHL Global Forwarding are currently investing heavily in new technology, while most competitors apply a wait and see attitude. While this has been the standard approach to many other issues within the shipping industry before, it may not work as well for competitors to copy-paste successful solutions to the same extent in a digital environment.

Even if AI will most likely play a significant role in the future of logistics, the authors recognise that it is not a simple task to shift current logistics operating models into an AI-powered reality. The transition towards using smart technologies to become more proactive, personalised and autonomous is a long journey that will require effective leadership, a sound understanding of business value drivers, a clear strategy, AI skills and talents in-house developers with shipping industry experience and an organisational culture that support the development of AI-driven business.

AI is believed to become a more prominent and inherent part of the day-to-day business of every carrier and freight forwarder where it will allow the industry to be more proactive, predictive, automated and personalised. While the current belief is that this will be achieved just within a few years, the AI hype will most likely fall under Amara's law, stating that the short-term effects of will be overestimated and the long-term effects will be underestimated. Even though AI and ML have come a long way during recent years, there are still obstacles to overcome in terms of scalability, explainability and usability.

AI has seen big hype cycles before, both during the 1960s and 1980s, where it received heavy funding and attention just like today. It was even announced in the 1960s that "machines will be capable, within twenty years, of doing any work a man can do", where almost 60 years later, this truth remains to be seen. While both the 1960s and 1980s AI-hypes died out, the current hype seems a lot more promising than those prior where previous technical limitations such as data availability and computational scale are supposedly said to be overcome. Whether or not the current development of AI is hype or hope, remains to be seen. The fact remains that today's current technology, business and societal conditions favour the paradigm shift to digital and smart solutions more than any other previous time in history.

5.5 Back-office vs front-office priorities

Back-office improvements through digitalisation and automation powered by AI is a hot topic for logistics in general, where the container shipping industry is an ideal candidate as it involves a lot of detail-oriented and repetitive administrative tasks, following more or less the same procedures for every import and export container. While back-office solutions are not being prioritised to the same extent as front-office solutions, the authors believe that there will be a shift in priority within a near future. In an increasingly complex and competitive business environment, companies that operate global supply chains are under very high pressure to deliver high service levels at low or even no cost where back-office automation powered by smart systems could prove a valuable asset to save both time and resources while increasing productivity and accuracy.

When it comes to digitalisation strategies and priorities, however, back-office improvements are more or less neglected for the benefit of front-office improvements. While back-office automation has the ability to generate long-term value in terms of cost savings and increased productivity and accuracy, front-office applications are the primary focus as it is desired by customers and stakeholders in order to create an instant value.

As the container shipping industry is moving into the digital arena, front-office applications such as forecasting or tracking tools can determine whether or not a customer is won or lost, while back-office operations of a carrier or forwarder are none of the customers' concern.

5.6 Diffusion of innovation for AI in container shipping

While AI is a hot topic in logistics, the shipping industry has barriers to overcome before true potential can be achieved. To get a better understanding of the barriers to AI in container shipping, the topic is presented through a diffusion of innovation framework.

5.6.1 Economic barriers

The economic barriers to AI and ML in container shipping are low and decreasing as technological progress is being made. Development in fields such as big data, connectivity, processing power along with algorithmic development have made performance, accessibility and cost of these technologies more favourable than ever before.

Access to information has increased as the topic is now being investigated from many different directions and applied in real life scenarios. This has resulted in an increased understanding of how this technology can be applied to in a variety of different scenarios.

Compared to a few years ago, ML models can now be developed, trained and tested within a matter of months. This is a direct result of new technological trends such as open source and transfer learning where many off-the-shelf solutions are already available for proof of concept testing.

5.6.2 Behavioural barriers

The behavioural barriers to AI and ML in container shipping are high but decreasing as the disruptive forces are being presented. The shipping industry, along with many other industries, cannot ignore the potentials and disruptive forces being presented by this technology. Many industry actors understand that future investments will eventually have to be made in order to remain competitive.

While the behavioural barrier poses a problem for conservative industries such as shipping, one popular statement made by Stena Line, which is one of the first Swedish shipping companies to invest heavily into AI, is that these kind of changes are ultimately about survival. If carriers and freight forwarders are to stay

relevant during the next twenty years, they need to get better and faster at everything they do. For this purpose, change is required. While change management is an issue within every industry, the shipping industry may find this barrier particularly hard to overcome.

5.6.3 Organizational barriers

The organisational barriers to AI and ML in container shipping are high and will be challenging to overcome. While there are many small to mid-sized carriers and forwarders, large organisations will find it hard to change their goals, routines and culture on a global and local level. An overall industry problem is that there are many large global organisations where implementation of new processes and software are both laborious and time-consuming to execute.

Problems arise from hierarchical structures, bureaucracy and silo mentality where one branch of the same organisation may have completely different goals and priorities compared to their colleagues in another country. To overcome the organisational barriers, a clear strategy is needed where power, influence and stakeholders have to align for this barrier to be overcome.

5.7 Limitations of study

ML, and even more sophisticated algorithms such as DL, has seen a massive increase in popularity during the last decade where many researchers are still debating whether ML is hype or hope (Brooks, 2017). Even though ML algorithms have been successfully applied in different applications, modern AI systems and ML algorithms are incredibly narrow in what they can do and suffer from problems with scalability.

Although the thesis presented both CNN and Naïve Bayes algorithms, it was not within the scope of the research to elaborate further on both branches of ML. The primary reason was due to time constraints, but also to avoid losing the focus of the research which was mainly to investigate the potential for ML solutions to be implemented successfully within administrative processes in the container shipping industry. Therefore, the authors choose to only pick the CNN algorithm for the identified business case, being how to complement the document handling process specifically with document sorting and classification.

Although the authors could have elaborated further in respect of the other maritime sectors, regarding their main characteristics compared to container shipping, which documents are involved and in turn their respective administrative processes, the authors perceive it to be out of the scope of this thesis.

5.8 Sustainability

Sustainability is an important topic in shipping logistics as there are many environmental challenges with transporting cargo around the world. While the authors presented topic has little to do with environmental sustainability, other sustainability aspects such as social and economic are highly related to the investigated topic. While many social and well-being sustainability aspects have already been presented under section 5, a summary and additional comments are presented in below.

5.8.1 Social

The social sustainability aspects of digitalisation within container shipping is a double-edged sword. While one side of digitalisation and automation assist employees in their day to day work, relieving them from burdensome administrative tasks and allowing employees to focus on value-adding activities and use their true potential, the other side of digitalisation and automation is job displacement where the industry will no longer require as many people to handle the day-to-day operations. As stated by Frank Apple, the CEO of the largest freight forwarding company in the world "Automation may mean it will soon take a third fewer people to deliver a given volume of goods" (The Economist, 2018).

The World Economic Forum predicts in their report “The future of jobs” from 2016 that about 7,1 million jobs will be gone due to automation by 2020, while only 2,02 million jobs will be generated during that same time (WEF, 2016). While this development is problematic and put stress on social sustainability aspects, history has shown before that society has always found a way to adapt to both the beneficial and the challenging effects of technological change.

5.8.2 Economic

The economic sustainability aspect of digitalisation within the container shipping industry is beneficial as companies are gaining new means to improve their businesses. Having companies that are economically

sustainable is a vital part of a functional society and economy where companies provide to the society and the local community in many different ways.

Reduced cost and increased productivity through automation and digitalisation will in a long-term perspective allow companies to flourish and hopefully invest in business improvements. For shipping companies, it is crucial that their model be economically sustainable as efforts must be put into research and development to further reduce the environmental emissions generated from transporting cargo, where a direct effect of this is improved social and environmental sustainability as well.

5.8.3 Environmental

The direct environmental sustainability aspects of digitalisation within the container shipping industry can mainly be related to the optimisation of processes and reduction of human error. Improved processes can lead to higher quality work, where unnecessary transports and accidents can be avoided, and routing and degree of filling in containers and on vessels can be improved, thus reducing the burden on the environment.

5.9 Method discussion

In this section, the authors discuss the usage of previously presented methodology regarding the reasoning behind the decision making throughout the thesis work.

5.9.1 Credibility

For the conducted research to be as trustworthy and believable as possible, the authors have used a sufficient number of sources in different formats. Sources were academic literature in the form of published books, articles and course material; academic and industry inquiries, by mail correspondence, conducting interviews and also by attending seminars and conferences on the research topic. The sample data was gathered from real market actors from which actual invoices, packing lists and copies of BLs have been provided, for the research to be as realistic as possible. The conclusions, having been based on previously formulated research questions, was therefore supported by the findings to the greatest extent possible. Both the defined questions and the aim has remained unchanged throughout the entire research, for which the authors have tried to be as consistent and systematic as possible in terms of both the chosen sources and usage of appropriate methodology.

Because the authors are currently active as ocean freight experts with end-to-end process knowledge on container shipping, with a certain level of confidence regarding common market knowledge by the majority of the industry participants, it would have been relatively easy and comfortable to merely state what is considered to be general knowledge, without necessarily making the research particularly bias. Although this might have been the case, the authors aimed to be as objective and ethically sound as possible, thus making sure that any information with a significant level of importance, and which could potentially have an effect on the final results of this research, was thoroughly checked and confirmed by several industry participants from a broad spectrum of organizations. Therefore, it is fair to state that the evidence submitted in this research, is clear, relevant, authentic and convincing, thus making it possible to be considered as credible facts.

5.9.2 Validity

The research was designed to be an interdisciplinary study within the fields of Maritime Management and Electrical Engineering. This design intended to address identified issues with technical solutions in the focus area of this research being the container shipping industry. The suggested solution for the issues was going to be derived from different parts of the fields as mentioned above. The main purpose was to enhance collaboration and establish a communication bridge for increased interaction between the fields, making it possible for more ideas to be exchanged between them, and in turn, answering the formulated research questions at hand.

The reason for this design was that early on in the study, a gap and communication barrier was identified, which the authors believe made it difficult for previous innovators and researchers to find common ground for confident collaboration and sharing of domain knowledge, without exposing themselves to the risk of losing out on potential benefits and advantages derived from suggested solutions. Thanks to the design and framework of this research, the authors were able to gather high-quality input from both sides, due to having a foot in each field, conduct experiments and performance tests of potential solutions, and finally to synthesise the information in the report of the study. This lead to the authors being capable of submitting a proposed solution to the identified issues and in turn, answer the formulated research questions.

5.9.3 Reliability

Due to the sample data being real-world documentation provided by industry actors and their respective customers, the measurement from the experimentations can be considered to be in tune with reality, thus making the data reliable. The procedure from which the suggested model was designed upon, was the same as the actual procedures in currently active organisations who deliver ocean freight services, in particular within freight forwarding and container shipping agents.

Therefore, if future research is going to be conducted, based on current or historical industry procedures, the research is believed to be consistently repeatable, as these procedures have not changed significantly during the past decade. Also, because the research is based upon internal back-office activities of global actors which do not make any particular distinctions between the offices different geographical locations, due to the nature of the global shipping industry.

5.9.4 Transferability

The way the model is purposely designed to contribute in the task of classifying different types of shipping documentation, and the actual administration tasks and procedures having been emulated from mainly shipping industry actors, are in essence relatively similar to many other industry contexts. The authors thus suggest that the presented solution applies to other areas of transportation and logistics, e.g. air, road and rail transportation, as well as inbound supply chain logistics, even for other industry contexts were the back-office activities are hugely burdensome. Regarding the model being applied to solve other more general classification problems, the authors do not see any particular difficulty, as it is mainly a matter of how well the actual procedures the model is intended to emulate are formulated. This lastly mentioned aspect is fundamental before this research findings could be externally validated, or in other confirming its transferability.

5.9.5 Triangulation

This section provides a brief discussion of the two methods used throughout this study to provide the reader with more insight into the reasoning behind the decision making of the model used.

5.9.5.1 Qualitative

An industry enquiry was carried out to get around the authors own biases regarding the need for automated document handling. Several of the largest carriers and forwarders confirmed that document handling is indeed a time consuming administrative task in need of improvement. As the authors do not have a background in programming, electrical engineering or computer science, insight into possible solutions based on machine learning and basic knowledge within the field had to be obtained.

The topics of digitalisation, strategy and maritime logistics were investigated to create a context into where the ML solution is to be applied and what the potential barriers to such a solution would be.

Finding literature and including it in the study was done mainly using a snowballing approach, whereas the researchers initiated the search with general literature and later adapting and reducing it towards more focused literature of the subject in question. From there, related literature was continuously found which would add value to the study. Mainly electronic databases were used to find literature as it allows for a faster process of finding interesting pieces and to deselect other pieces of less relevance. Physical library resources were also used in this study as some relevant literature was no available electronically through the Chalmers library.

Databases such as Google Scholar and Chalmers Library were searched for key phrases such as "Machine Learning", "Natural Language Processing", "Document classification" and books and articles within the field are digested to get an initial understanding of current strengths and weaknesses. Courses on ML and programming in MATLAB are conducted on Coursera and MathWorks. The purpose of the courses was not the become specialist in the field of ML. Rather it was to gain a broad understanding of the technology and how it works, thus incorporating both the organisational and technical aspects.

The researchers did not experience any particular challenges related to gaining access. Primary data on ML, the container shipping industry and digitalisation were gathered through interviews with people from the container shipping industry and IT providers and secondary data on the same topics were gathered from consulting reports, conferences, books and articles.

Interviews were conducted with people from academia, IT and container shipping to gain a good understanding of the topic at hand, the current potential and limitations of technical solutions and the

barriers for implementing such solutions. The main questions formulated for the interviews (as in Appendix G and Appendix H), proved to be good conversation starters which managed to branch out into relevant related areas.

5.9.5.2 Quantitative

As per the literature review and industry inquiry it was decided by the authors to proceed with the CNN for classification purposes, rather than using the NLP approach. As explained earlier, the reason for this is mainly due to the nature of the documents in question, and how the particular industry handling procedure is for these types of documents. Several of the contributors from both the conducted interviews and the attended conferences mentioned that Neural Networks perform with higher accuracy and also that the time taken for the models to learn the parameters and features of input data was considerably less compared to other models. This was also mentioned and even exemplified in the Coursera ML course taken during the initial phase of the master thesis.

As in many other cases, humans try to identify a set of routine practices for specific tasks, and simply follow the procedure, rather than using all "neurons" for the same task each time. This is both for efficiency reasons in terms of not wanting to overthink each step and to just perform the task, but also to avoid making errors when deviating from what has already proven to work. In the case of document handling within container shipping, the documents in question are relatively simple to learn after seeing hundreds of these each day. This has, in turn, made the handling clerks experts in distinguishing one document from another. Thus, for the evaluation part, the desired validation accuracy from the industry enquiry is $Y = 1$.

After initial testing of different learning algorithms available in MATLAB, and to satisfy these requirements, the authors identified CNN to be the model who resembles humans the most, both in accuracy and also in the way to determine which class each document belonged to. The only aspect in which this model really outperforms any human is in terms of time taken to perform the task successfully.

6. Conclusion

This section provides brief conclusions to the most central parts of the report, i.e. the thesis statement, the research questions and finally what the authors suggests as future research.

6.1 Thesis statement

> *It is hypothesised that it is more beneficial for container shipping companies to deploy ML algorithms to handle administrative tasks compared to current manual processing.*

This research concludes that above statement is proved to be valid.

6.2 Research questions

1. *Which ML algorithms and additional tools are most suitable for creating an automated document handling and classification system for container shipping?*

CNN algorithms for classification of standardised documents which mainly require relatively simple visual imagery analysis.

Naïve Bayes algorithms for classification on more particular documents in a natural language which mainly require text and content extraction.

The rise of deep learning has made many traditional machine learning algorithms obsolete. Deep learning algorithms are in some areas of application surpassing human-level performance where this development is a valuable asset for many industries. The convolutional neural network is a deep learning algorithm which manages to classify shipping documents as images with very high accuracy. Document classification was decided to be performed as image recognition rather than text recognition as this approach for deciding whether a document is a BL, invoice or packing list resembles human logic. Deep learning algorithms, such as the CNN, combined with OCR for information extraction of valuable information from documentation are found to be key components for automating the document handling process within container shipping.

2. *What is the potential for these tools in terms of quantifiable performance metrics compared to human labour?*

The convolutional neural net managed to train 319 and classify 136 input documents with an average accuracy of 98,49 percent in an elapsed average time of 11,7 seconds. While classification accuracy would most likely be the in the same region if done manually, the time required for the algorithm to sort and classify a total of 455 documents is far lesser, allowing automation to save a lot of time and manual labour.

3. *What are the barriers to implementing ML solutions in the container shipping industry?*

The main barriers to implementing ML solutions and increased digitalisation are organisational and behavioural. Trends within container shipping with mergers and acquisitions are resulting in larger and larger organisations where change on global level poses a problem for implementation of new processes and systems. The shipping industry is also known for its conservatism where behaviour barriers are clearly seen and deeply rooted in the organisational culture.

While technical barriers exist, they are mainly related to currently existing IT infrastructure and a low level of understanding of the maritime industry from an IT industry perspective, how to identify issues in current handling processes and even how to adapt already existing technical solutions on these processes.

The economic barriers are considered low and decreasing as technological progress is being made. Development in fields such as big data, connectivity, processing power along with algorithmic development have made performance, accessibility and cost of these technologies more favourable than ever before.

While barriers still exist, it is concluded that the disruptive forces posed by AI are too large to be overseen. In order to overcome the barriers and succeed, close collaboration between the software and logistics industry is required to bridge the existing knowledge gap and increase the understanding of strengths, weaknesses, opportunities and threats from both sides.

6.3 Further research

As the need for education within electrical engineering and computer science will increase in society, not only related to maritime studies, this research concludes that more emphasis needs to be put in higher educations for maritime students to grasp the technical aspects and learn how to use software tools, such as MATLAB, for different purposes. The objective should not be to turn students into programming experts, but at least provide them with a high enough level of both skills and knowledge, thus enabling them to work with data and in turn contribute more extensively in solving computational issues within their specific field of expertise.

The two main reasons are first that IT experts will not be able to identify specific issues with current processes in all specific industry sectors, as they will not experience them on an individual level. Secondly, for programmers and computer engineers to suggest different solutions, the end-users, in this case the maritime sector, needs to be more involved and possess a certain knowledge of available solutions for a good result.

Future research should also investigate in how to map out the maritime industry further and begin illustrating different business cases in more concrete terms before it can expect other industries to increase and even develop current collaborations. Like in many other aspects of life, it is easy to become blind to the commonly occurring things, which makes it even more vital to provide partners with clear overviews of the industry fundamentals.

In particular, future research is suggested to continue where this thesis left off, and identify different non-value adding back-office activities in order to investigate the potential for automation. The objective should be to find solutions which complement human labour, rather than aiming to replace them, as the human qualities related to flexibility, creativity, teamwork, empathy and judgment skills, play key roles in reaching business excellence and continuous improvements.

Reference list

Accenture (2018) Industry Consortium Successfully Test Blockchain Solution Developed by Accenture That Could Revolutionize Ocean Shipping, *Accenture*. [Electronic] 14th Mar. Available at: https://newsroom.accenture.com/news/industry-consortium-successfully-tests-blockchain-solution-developed-by-accenture-that-could-revolutionize-ocean-shipping.htm Accessed: 2018-04-20

Alexopoulos, E.C. (2010) Introduction to multivariate regression analysis. *Hippokratia.* [Electronic] vol. 14, no. Suppl 1, pp. 23. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3049417/ Accessed: 2018-04-08

Alphaliner (2018) Alphaliner Weekly Newsletter, *Alphaliner*. [Electronic] Volume 2018, issue 05, 24th Jan. Available at: https://300cubits.tech/pdf/Alphaliner_Newsletter_no_05_2018_p1-3.pdf Accessed: 2018-02-20

Arntz, M., Gregory, T., Zierahn, U. (2016) The Risk of Automation for Jobs in OECD Countries, OECD Social, *Employment and Migration Working Papers*. [Electronic] No. 189, OECD Publishing, Paris. Available at: http://www.ifuturo.org/sites/default/files/docs/automation.pdf Accessed: 2018-04-10

Bali, N., Goyal, M., Watkins, C. (2007) *Advanced engineering mathematics: a computer approach.* [Electronic] 7th edn, Infinity Science Press, Hingham, Mass. Available at: http://common.books24x7.com.proxy.lib.chalmers.se/toc.aspx?bookid=15374 Accessed: 2018-04-08

Banko, M. Brill, E. (2001) *Scaling to very very large corpora for natural language disambiguation.* [Electronic] 39th Annual Meeting on Association for Computational Linguistics. Pp 26-33. Available at: https://dl.acm.org/citation.cfm?id=1073012.1073017 Accessed: 2018-03-11

Berglund, P. (2017) Pros and cons of working with freight forwarders, *Xeneta*. [Electronic] 19th Oct. Available at https://www.xeneta.com/blog/working-with-freight-forwarders-carriers Accessed: 2018-02-03

Bhattacharjee, A. (2016) Developing Machine Learning and Deep Learning Algorithms Using Matlab, *Mathworks Academy*. [Electronic] 3rd Nov. Available at: https://se.mathworks.com/videos/developing-machine-learning-and-deep-learning-algorithms-using-m-1481315519196.html Accessed: 2018-05-01

Bishop, C.M. (2006) *Pattern recognition and machine learning.* Springer, New York, NY.

BLACK, H.C. (1968) *Black's Law Dictionary*, [Electronic] 4th edn. West Publishing Company. St Paul: Minnesota. Available at: http://heimatundrecht.de/sites/default/files/dokumente/Black%27sLaw4th.pdf Accessed: 2018-05-11

Box, G.E.P., Leonard, T., Wu, C. (1983) *Scientific inference, data analysis, and robustness.* [Electronic] Academic Press, London, England; New York, New York. Available at: https://www-sciencedirect-com.proxy.lib.chalmers.se/science/book/9780124381506 Accessed: 2018-04-07

Brooks, R. (2017) The seven deadly sins of Ai prediction, *MIT Technology Review*. [Electronic] 6th Oct. Available at: https://www.technologyreview.com/s/609048/the-seven-deadly-sins-of-ai-predictions/ Accessed: 2018-03-20

Bryman, A., Bell, E. (2011) *Business research methods*. 3rd edn. Oxford University Press, Oxford.

Bughin, J., Catlin, T., Hirt, M., Willmott, P. (2018) Why digital strategies fail, *McKinsey&Co.* [Electronic] Jan. Available at https://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/why-digital-strategies-fail Accessed: 2018-04-20

Chapmann, J. (2017) *Machine Learning: fundamental algorithms for supervised and unsupervised learning with real-world applications*. CreateSpace Independent Publishing Platform.

Chaudhuri, A., Mandaviya, K., Badelia, P., Ghosh, S. (2017) *Optical Character Recognition Systems for Different Languages with Soft Computing*. Springer International Publishing, Cham.

Cormen, TH., Leiserson, CE., Rivest, RL. (2014) *Introduction to Algorithms*. [Electronic] Press, Cambridge Available at: https://ebookcentral.proquest.com/lib/chalmers/detail.action?docID=3339142# Accessed: 2018-04-08

Dawson, A., Hirt, M., Scanlan, J. (2016) The Economic Essentials of Digital Strategy, *McKinsey&Co*. [Electronic] March. Available at: https://www.mckinsey.com/business-functions/strategy-and-corporate-finance/our-insights/the-economic-essentials-of-digital-strategy Accessed: 2018-03-20

DHL Trend Research (2018) Blockchain In Logistics, *DHL Trend Research*. [Electronic] Available at: https://www.logistics.dhl/content/dam/dhl/global/core/documents/pdf/glo-core-blockchain-trend-report.pdf Accessed: 2018-04-15

EESC (2017) Impact of digitalisation and the on-demand economy on labour markets and the consequences for employment and industrial relations, *European Economic and Social Committee*. [Electronic] Available at: https://www.eesc.europa.eu/resources/docs/qe-02-17-763-en-n.pdf Accessed: 2018-04-27

Géron, A. (2017) *Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems.* O'Reilly Media, Sebastopol.

Gesing, B. Peterson, J. S. Michelsen, D. (2018) Artificial intelligence in logistics, *DHL Innovation and Trend Research.* [Electronic] Available at: https://www.logistics.dhl/content/dam/dhl/global/core/documents/pdf/glo-artificial-intelligence-in-logistics-trend-report.pdf?j=129072&sfmc_sub=64474602&l=59_HTML&u=7424877&mid=7275327&jb=28 Accesses on: 2018-04-15

Glave, T., Joerss, M., Saxon, S. (2014) The hidden opportunity in container shipping, *McKinsey&Co*. [Electronic] Nov. Available at https://www.mckinsey.com/business-functions/strategy-and-corporate-finance/our-insights/the-hidden-opportunity-in-container-shipping Accessed: 2018-03-04

Habib, M., Pathik, B.B., Maryam, H. (2014) *Research methodology-contemporary practices: guidelines for academic researchers*. Cambridge Scholars Publishing, Newcastle, England.

Hayslett, H.T., Murphy, P. (1995) *Statistics, Made Simple.* [Electronic] Books, Oxford, England. Available at: https://doi-org.proxy.lib.chalmers.se/10.1016/B978-0-7506-0481-9.50005-X Accessed: 2018-04-08

Holley, R. (2009) How Good Can It Get? Analysing and Improving OCR Accuracy in Large Scale Historic Newspaper Digitisation Programs, *D-Lib Magazine*. [Electronic] Vol 15, number 3 / 4, March/April. Available at: http://www.dlib.org/dlib/march09/holley/03holley.html Accessed: 2018-04-01

Hosmer, D.W., Lemeshow, S., Sturdivant, R.X. (2013) *Applied logistic regression.* [Electronic] 3rd edn, Wiley, Hoboken, New Jersey. Available at: https://ebookcentral.proquest.com/lib/chalmers/detail.action?docID=1138225# Accessed: 2018-03-28

Hristea, F.T. (2013) *The Naïve Bayes model for unsupervised word sense disambiguation: aspects concerning feature selection.* [Electronic] Springer, London;Berlin. Available at: https://doi-org.proxy.lib.chalmers.se/10.1007/978-3-642-33693-5 Accessed 2018-04-23

Kerzner, H. (2015) Project Management 2.0 - Leveraging Tools, Distributed Collaboration, and Metrics for Project Success. [Electronic] Wiley, Hoboken, New Jersey. Available at: https://app.knovel.com/hotlink/pdf/id:kt011B8QE1/project-management-2/pareto-analysis Accessed: 2018-07-01

Kleinbaum, D.G., Klein, M. (2002) *Logistic Regression: A Self-Learning Text. Statistics for Biology and Health*. Springer, New York.

Kolbjørnsrud, V., Amico, R., Thomas, J.R. (2016) How Artificial Intelligence Will Redefine Management, *Harvard Business Review*. [Electronic] 2nd Nov. Available at: https://hbr.org/2016/11/how-artificial-intelligence-will-redefine-management Accessed: 2018-03-15

Koperniak, S. (2017) Artificial data give the same results as real data - without compromising privacy. *MIT News*. [Electronic] 3rd Mar. Available at: http://news.mit.edu/2017/artificial-data-give-same-results-as-real-data-0303 Accessed: 2018-03-26

Krizhevsky, A., Sutskever, I. and Hinton, G.E. (2012) Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems* [Electronic] pp. 1097-1105. Available at: https://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf Accessed: 2018-05-18

Lee, C., Meng, Q. (2015) *Handbook of Ocean Container Transport Logistics: Making Global Supply Chains Effective.* Springer International Publishing, Cham.

Levins, R. (1966) The strategy of model building in population biology. *American Scientist*. [Electronic] *54*(4), pp.421-431. Available at: https://uberty.org/wp-content/uploads/2015/07/Levins-1966-Model_Building.pdf Accessed: 2018-04-05

Libert, B., Beck, M. (2017) The Rise of AI Makes Emotional Intelligence More Important, *Harvard Business Review*. [Electronic] 15th Feb. Available at: https://hbr.org/2017/02/the-rise-of-ai-makes-emotional-intelligence-more-important Accessed: 2018-02-27

Lyfenko, N.D. (2014), Automatic classification of documents in a natural language: A conceptual model, *Automatic Documentation and Mathematical Linguistics.* [Electronic] Vol. 48, no. 3, pp. 158-166. Available at: https://doi.org/10.3103/S0005105514030030 Accessed: 2018-03-15

Manikadan, S. (2011) Frequency distribution, Journal of pharmacology & pharmacotherapeutics. Vol. 2, no. 1, pp. 54. Available at: https://search-proquest-com.proxy.lib.chalmers.se/docview/855954136?pq-origsite=summon Accessed: 2018-03-20

Mason v Lickbarrow, (1787) 2 TR 63, 100 ER 35. [Electronic] Available at: http://maritimelawdigital.com/uploads/HTML/Lickbarrow_Mason_1787.htm Accessed: 2018-05-11

McDonald, M. (2015) What is a digital strategy?, *Accenture*. [Electronic] 3rd Mar. Available at: https://www.accenture.com/us-en/blogs/blogs-digital-what-is-digital-strategy Accessed: 2018-03-20

McKevitt, J. (2018) Agility will collaborate with Maersk-IBM on blockchain initiative, *Supply Chain Drive* [Electronic] 8th Feb. Available at https://www.supplychaindive.com/news/agility-will-collaborate-with-maersk-ibm-on-blockchain-initiative/516581/. Accessed: 2018-03-03

Mitchell, T.M. (1997) *Machine Learning*. McGraw-Hill, New York.

Mohammed, M., Khan, M.B., Bashier, E.B.M. (2017) *Machine learning: algorithms and applications.* CRC Press, Boca Raton.

Nessler, D. (2016) How to apply a design thinking, HCD, UX or any creative process from scratch, *HyperIsland.* [Electronic] 26th May. Available at: https://www.hyperisland.com/community/news/how-to-apply-a-design-thinking-hcd-ux-or-any-creative-process-from-scratch Accessed: 2018-03-03

Ng, A. (2018) *Machine Learning Yearning*. Forthcoming 2018. [Electronic] Available at: http://www.mlyearning.org/ Accessed: 2018-05-18

Olive, D.J. (2017) *Linear Regression.* [Electronic] Springer International Publishing, Cham. Available at: https://rd-springer-com.proxy.lib.chalmers.se/book/10.1007%2F978-3-319-55252-1 Accessed: 2018-04-07

Paluszek, M., Thomas, S. (2017) *MATLAB Machine Learning.* Apress, Berkeley, CA.

Robert, C.P. (2007), *The Bayesian choice: from decision-theoretic foundations to computational implementation.* [Electronic] 2nd edn, Springer, New York. Available at: https://doi-org.proxy.lib.chalmers.se/10.1007/0-387-71599-1 Accessed: 2018-05-15

Rogers, E.M. (1962) *Diffusion of innovations*. Free Press of Glencoe, New York.

Rogers, E.M. (2003) *Diffusion of innovations*. 5th edn, Free Press, New York.

Samuel, A.L. (2000) Some Studies in Machine Learning Using the Game of Checkers, *IBM Journal of Research and Development.* [Electronic] vol. 44, no. 1 / 2, pp.206. Available at: https://search-proquest-com.proxy.lib.chalmers.se/docview/220681210?pq-origsite=summon Accessed: 2018-03-01

Schreck, B., Kanter, M. Veeramachaneni, K. Vohra, S. Prasad, R. (2018) Getting value from Machine Learning Isn't about fancier algorithms - it's about making it easier to use, *Harvard Business Review*. [Electronic] 6th Mar Available at: https://hbr.org/2018/03/getting-value-from-machine-learning-isnt-about-fancier-algorithms-itsabout-making-it-easier-to-use?utm_source=twitter.com&utm_medium=social&utm_campaign=buffer&utm_content=bufferb2b6e https://www.featuretools.com/wp-content/uploads/2018/03/ml20.pdf Accessed: 2018-03-20

Sebastiani, F. (2002) *Machine Learning in automated text categorization*. Consiglio Nazionale delle Ricerche, Italy.

Shumsky, T. (2017) U.S. Consulting Spending Tops $58 billion in 2016, *The Wallstreet Journal*. [Electronic] 23rd May. Available at: http://on.wsj.com/2qQXiQi Accessed: 2018-03-14

Sinclair, I. (1991) *Computer science: a concise introduction.* [Electronic] Newnes, Oxford, England. Available at: https://doi-org.proxy.lib.chalmers.se/10.1016/B978-0-7506-0252-5.50005-X Accessed: 2018-03-22

Song, D., Panayides, P.M. (2015) *Maritime logistics: a guide to contemporary shipping and port management*. 2nd edn, Kogan Page, London, UK;Philadelphia, PA.

Strannegård, L., Styhre, A. (2013) *Management: an advanced introduction.* Studentlitteratur, Lund.

Tarnopol, R. (2017) How to OCR Documents for Free in Google Drive, *Envatotuts Business Intelligence.* [Electronic] 9th May. Available at: https://business.tutsplus.com/tutorials/how-to-ocr-documents-for-free-in-google-drive--cms-20460 Accessed: 2018-03-03

Thakur, Y. (2016) Data Analytics in the Maritime Industry, *Oceanmanager.* [Electronic] 14th Jun. Available at: https://www.oceanmanager.com/data-analytics-in-the-maritime-industry Accessed: 2018-01-17

The Economist (2018) The global logistics business is going to be transformed by digitisation, *The Economist*. [Electronic] 26th Apr. Available at: https://www.economist.com/news/briefing/21741139-will-be-bad-news-some-global-logistics-business-going-be-transformed?frsc=dg%7Ce Accessed: 2018-05-10

Tidd, J. (2010) *Gaining momentum: managing the diffusion of innovations*. Imperial College Press, Hackensack, NJ;London;Singapore.

Tullverket (2018) Styrkande handlingar, *Tullverket*. [Electronic] Available at: http://www.tullverket.se/sv/foretag/importeravaror/deklareravarorvidimport/styrkandehandlingar.4.6ac2c843157b7beb007406.html Accessed: 2018-03-13

UK Design Council (2007) Eleven lessons: managing design in eleven global brands*, UK Design Council* [Electronic] Available at: https://www.designcouncil.org.uk/sites/default/files/asset/document/ElevenLessons_Design_Council%20(2).pdf Accessed: 2018-03-14

UNCTAD (2017) Review of maritime transport 2016, *United Nations Conference on Trade And Development* [Electronic] Available at: http://unctad.org/en/PublicationsLibrary/rmt2016_en.pdf Accessed: 2018-02-14

Wang, Y., Tseng, M.M. (2015) A Naïve Bayes approach to map customer requirements to product variants, *Journal of Intelligent Manufacturing*. [Electronic] Vol. 26, no. 3, pp. 501-509. Available at: https://rdcu.be/RbJy Accessed: 2018-05-15

Witten, I.H., Frank, E., Hall, M. (2017) *Data mining: practical machine learning tools and techniques*. 4th edn, Elsevier, Amsterdam.

World Economic Forum (2016) The Future of Jobs Executive summary, *World Economic Forum* [Electronic] January. Available at http://www3.weforum.org/docs/WEF_FOJ_Executive_Summary_Jobs.pdf Accessed: 2018-04-02

Wright, D.B. (2003) Making friends with your data: Improving how statistics are conducted and reported, *British Journal of Educational Psychology*. Vol. 73, pp.123-136.

Yan, X., Su, X. (2009) *Linear regression analysis: theory and computing.* [Electronic] World Scientific, Hackensack, NJ;Singapore. Available at: https://ebookcentral.proquest.com/lib/chalmers/detail.action?docID=477274# Accessed: 2018-03-28

Appendices

Appendix A – Bill of Ladings

1. Black's Law dictionary

"*Bill of lading. In common law. The written evidence of a contract for the carriage and delivery of goods sent by sea for a certain freight. Mason v. Lickbarrow, 1 H.B1. 359. A written memorandum, given by the person in command of a merchant vessel, acknowledging the receipt on board the ship of certain specified goods, in good order or "apparent good order," which he undertakes, in consideration of the payment of freight, to deliver in like good order (dangers of the sea excepted) at a designated place to the consignee therein named or to his assigns. Devato v. Barrels, D.C.N.Y., 20 Fed. 510. The term is often applied to a similar receipt and undertaking given by a carrier of goods by land. A bill of lading is an instrument in writing, signed by a carrier or his agent, describing the freight so as to identify it, stating the name of the consignor, the terms of the contract for carriage, and agreeing or directing that the freight be delivered to the order or assigns of a specified person at a specified place. See Civil Code Cal. § 2126a; Aman v. Dover & Southbound R. Co., 179 N. C. 310, 102 S.E. 392, 393; Rudin v. King-Richardson Co., 143 N.E. 198, 201, 311. Ill. 513. It is receipt for goods, contract for their carriage, and is documentary evidence of title to goods. Schwalb v. Erie R. Co., 293 N.Y.S. 842, 846, 161 Misc. 743*"

2. Mason v Lickbarrow, (1787) 2 TR 63, 100 ER 35

a. Second trial at the Court of King's Bench (1794) 5 TR 693, 101 ER 380, KB
"*If a bill of lading be a negotiable instrument, and convey an indefeasible* [i.e. not able to be lost, annulled, or overturned] *property in the goods, it must be so by the custom of merchants: but such custom is not to be found in any of the books treating upon the subject. There are cases which establish a contrary doctrine, in which the Courts have held that the rights of the assignees* [i.e. person to whom the right or liability is legally transferred] *are the same as the rights of the original consignee. It cannot indeed be disputed but that, as between the assignee and the indorsee* [i.e. the person to whom the negotiable document is endorsed], *the indorsement of a bill of lading is a complete transfer of the property which the consignee has in it* [.]"

b. Ashhurst J
   i. *"*[Between] *the vendor and third persons, the delivery of a bill of lading is a delivery of the goods themselves: if not, it would enable the consignee to make the bill of lading an instrument of fraud. The assignee of a bill of lading trusts to the indorsement; the instrument is in its nature transferable; in this respect therefore this is similar to the case of a bill of exchange. If the consignor had intended to restrain the negotiability of it, he should have confined the delivery of the goods to the vendee only: but he has made it an indorsable instrument."*
   ii. *"The rule is founded purely on principles of law, and not on the custom of merchants. The custom of merchants only establishes that such an instrument may be indorsed; but the effect of that indorsement is a question of law, which is, that as between the original parties the consideration may be inquired into; though when third persons*

are concerned it cannot. This is also the case with respect to a bill of lading. Though the bill of lading in this case was at first indorsed in blank, it is precisely the same as if it had been originally indorsed to this person; for when it was filled up with his name, it was the same as if made to him only."

c. Grose J
   i. *After this case has been so elaborately spoken to by my brethren, it is not necessary for me to enter fully into the question, as I am of the same opinion with them. But I think that the importance of the subject requires me to state the general grounds of my opinion. I conceive this to be a mere question of law, whether, as between the vendor and the assignee of the vendee, the bill of lading transfers the property. I think that it does.*
   ii. *A bill of lading carries credit with it; the consignor by his indorsement gives credit to the bill of lading, and on the faith of that money is advanced. The first case that I find, where an attempt was made to introduce the same law between the consignor and the indorsee of the consignee, is that of Snee v Prescot; but as my brother Buller has already made so many observations on that case, it would be but repetition in me to go over them again, as I entirely agree with him in them all, as well as in those which he made on the other cases.*

d. Buller J
   "An argument was used with respect to the difficulty of determining at what time a bill of lading shall be said to transfer the property, especially in a case where the goods were never sent out of the merchant's warehouse at all: the answer is, that under those circumstances a bill of lading could not possibly exist, if the transaction were a fair one; for a bill of lading is an acknowledgment by the captain, of having received the goods on board his ship: therefore it would be a fraud in the captain to sign such a bill of lading, if he had not received goods on board; and the consignee would be entitled to his action against the captain for the fraud."

Appendix B – Random variables

In the following example, the variables can take any of the values $\{x_i\}$, were $i = 1, \ldots, M$, being denoted by c_i, where the letter c stands for column, as well as the values $\{y_j\}$ respectively, were $j = 1, \ldots, L$, being denoted by r_j, where the letter r stands for rows, for N total amount of trials, which is the number of instances of these variables. This is denoted by n_{ij} which is the number of points in the corresponding cell of the matrix, whereupon the joint probability $p(X = x_i, Y = y_j)$, i.e. that X and Y will take the values x_i and y_j, is the probability of both $X = x_i$ and $Y = y_j$, and is given by $\frac{n_{ij}}{N}$(Bishop, 2006). If the joint distribution of the two variables X and Y factorises into the product of the marginals, that is $p(X, Y) = p(X)p(Y)$, then it is said that the variables are *independent* of each other. When only calculating the probability that X takes the value x_i, that is $p(X = x_i)$, also called the marginal probability as it is obtained by summing out Y, i.e. marginalising Y, this would be given by $\frac{c_i}{n_{ij}}$. Because the following applies,$c_i = \Sigma_j n_{ij}$, Bishop (2006) explains that the sum rule of joint and marginal probability is the following:

$$p(X = x_i) = \sum_{j=1}^{L} p(X = x_i, Y = y_j)$$

When only considering this marginal probability, then the conditional probability of $Y = y_j$given $X = x_i$ is written $p(Y = y_j|X = x_i)$and is obtained by $\frac{n_{ij}}{c_i}$. This is the fraction of those points in c_i that fall in the coordinate (i, j) from which it is possible to derive the product rule of probability, being the relationship of the joint probability equalling the product of the conditional probability and the marginal probability (Bishop, 2006). To simplify the formulations using more compact notations, the two rules presented above can be written in the following way according to Bishop (2006)

Sum rule: $p(X) = \sum_Y p(X, Y)$

Product rule: $p(X, Y) = p(Y|X)p(X)$

Appendix C – Correlation

When an observation is being made, usually denoted by an observation pair (x_n, y_n), these can be considered as being coordinates of a point in a bivariate distribution, which can be plotted into what can be called a scatter diagram (Hayslett and Murphy, 1995). This is shown in Figure X_SCATTER, which represent an example of a scatter plot made in MATLAB, using the formula "scatter(x,y)" of a generated set of random sample data into one figure. Even though an association or relationship between the variables might be ocularly detected, it is not enough to draw any conclusions on the data. In order to determine to which degree there exist a relationship, in other words, if the variables are correlated, a measurable quantity is used for this purpose, i.e. sample correlation coefficient, usually denoted by lower case r (not indicating rows) (Hayslett and Murphy, 1995).

The main definition of correlation is if the change in one variable affects a change in the other (Bali, Goyal, and Watkins, 2007). This quantity r is a value between -1 and $+1$, with both extreme values indicating a linear relationship, and is said to be perfect when the deviation is proportional in both variables (Bali, Goyal, and Watkins, 2007). If the variables deviate in the same direction, the relationship is said to be directly correlated, but if the deviation is opposite, the relationship is then said to be inversely correlated (Bali, Goyal, and Watkins, 2007). The coefficient r is an estimate of a theoretical quantity called population correlation coefficient, usually denoted by the Greek lowercase letter ρ (Hayslett and Murphy, 1995). Larger values of y are associated with larger values of x, and the same goes for smaller values of both variables respectively, i.e. the variables are in a linear relationship, where if the value is near zero it would have signified that there is no such relationship (Hayslett and Murphy, 1995).

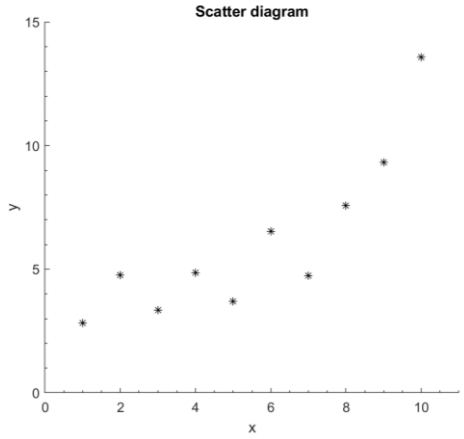

Figure: X_SCATTER

As can be seen by the plot, the diagram is concentrated to a certain extent around a curve, called the curve of regression, meaning that the relationship is expressed by means of curvilinear regression (Bali, Goyal, and Watkins, 2007). If the curve would rather take the form of a straight line, it is then a line of regression instead of a curve, meaning that the regression is linear (Bali, Goyal, and Watkins, 2007). This would give the best fit in the least square sense to the given frequency (Bali, Goyal and Watkins, 2007).

$$r = \frac{\Sigma(x-\overline{x})(y-\overline{y})}{[\Sigma(x-\overline{x})^2][\Sigma(y-\overline{y})^2]} = \frac{\Sigma xy - \frac{(\Sigma x)(\Sigma y)}{n}}{\sqrt{\left[\Sigma x^2 - \frac{(\Sigma x)^2}{n}\right] * \left[\Sigma y^2 - \frac{(\Sigma y)^2}{n}\right]}}$$

The above formula for the sample correlation coefficient for the Figure X_SCATTER shows that the following is needed: total sums of both variables x and y each; both variables separately squared; as well as the sum of the product of the variables. The computations are shown in Figure TBL_SCATTER, and by substituting the values into the formula, then the value r will be the following:

x	y	x^2	y^2	xy
6	6,53766714	36	42,74109	39,226
1	2,833885015	1	8,030904	2,833885
7	4,741153139	49	22,47853	33,18807
4	4,86217332	16	23,64073	19,44869
9	9,31876524	81	86,83939	83,86889
5	3,692311704	25	13,63317	18,46156
8	7,566407978	64	57,25053	60,53126
3	3,342624467	9	11,17314	10,02787
10	13,57839694	100	184,3729	135,784
2	4,76943703	4	22,74753	9,538874
55	61,24282197	385	472,9079	412,9091

Figure: TBL_SCATTER

$$r = \frac{412{,}9091 - 336{,}8355}{\sqrt{(385 - 302{,}5) * (472{,}9079 - 97{,}8395)}} = \frac{76{,}0736}{89{,}8430} = 0{,}8467$$

To avoid making these computations manually, it is possible to use the formula "corrcoef(x,y)" in MATLAB, which in this case would give us the same value as the one computed above. From the result it is now possible to infer a positive linear relationship, i.e. direct correlation, between the random sample variables x and y:

r =

1.000000000000000 0.846738959762191

0.846738959762191 1.000000000000000

If this direct correlation would be graphically represented on the X_SCATTER diagram, it would look like in the Figure X_rSCATTER by using the “lslines” formula in MATLAB, which would represent the least squares line of the figure:

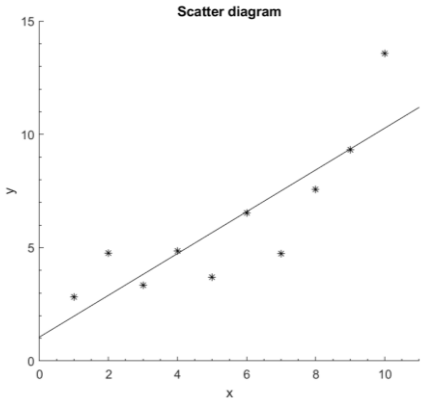

Figure: X_rSCATTER

Appendix D – Bayes' theorem

From the sum and product rule (Appendix B) it is possible to obtain the relationship between conditional probabilities, being the following:

$$p(Y|X) = \frac{p(X|Y)p(Y)}{p(X)}$$

This is commonly known as the Bayes' theorem, being the required normalisation constant to make certain that the conditional probability $p(Y|X)$ sums up to one. The theorem, being a logical argument based on the sum and product rule, is about Y and X being two distinct events, and $P(Y)$ and $P(X)$ being the class-, respectively predictor-, prior probabilities of Y and X without regard to each other (Bishop, 2006). It is used to convert this prior probability into $P(Y|X)$, also called posterior probability, which is the probability of observing event Y given that X is true, and $P(X|Y)$, being the likelihood, which is the probability of observing event X given that Y is true (Bishop, 2006).

Appendix E – List of invited speakers

The first conference was arranged by Chalmers University of Technology on the 15th of March 2018, called 'Digitalisation Conference', in which the following speakers made contributions:

- Dr Elena Fersman, PhD in Computer Science from the University of Uppsala, Adjunct Professor in Cyber-Physical Systems specialised in Automation at KTH, currently Research Leader in Machine Intelligence at 'Ericsson Research';
- Dr Angeliki Lazaridou, PhD in Cognitive Science, NLP, from the University of Trento, currently Research Scientist at Google DeepMind;
- Dr Richard Johansson, Postdoctoral and Associate Professor in Data Science at the University of Gothenburg;
- Dr Fredrik Kahl, Professor in Computer Vision from Chalmers University of Technology.

The second conference was arranged by the Meetup group 'Machine Learning and Data Science (GBG)' on the 10th of April 2018, called Gothenburg Artificial Intelligence Alliance (GAIA), in which the following speakers made contributions:

- Mr Mikael Kågebäck, PhD student in Data Science at Chalmers University of Technology, currently active with Machine Learning & Natural Language Processing at CSE;
- Dr Kristy Simmons, PhD in Neuroscience from the University of Pennsylvania, currently Senior Data Scientist at Recorded Future;
- Dr Daniel Gillblad, M.Sc. in electrical engineering and a PhD in Computer science from KTH, as well as Director of the Decisions, Networks and Analytics Laboratory at RISE SICS;

Appendix F – Citations from interviews

[C.1] “Sjöfartsbranschen är inte intresserade av nya system och det är väldigt osexigt när man kommer från ett IT företag och försöker sälja in tjänster, de vet knappt själva vad de behöver för de har inte rätt kompetens.”

[C.2] “De möjligheter vi ser idag till när det kommer till tekniska lösningar fanns ju inte när de här människorna satt på operations, det gör det väldigt svårt för dem att se varför det skulle behövas”

[C.3] “AI inom shipping är lite som ‘sex in high school’, alla snackar om det men ingen vet vad det är”

[C.4] “Många problem i branschen är idag alls problem, där allas problem är ingens problem. Så länge alla har samma problem är alla på samma nivå”

[C.5] “Det finns många andra branscher där man kommit betydligt längre med smarta system och automatisering, det är bra för då kan vi bara kolla vad de har gjort och göra samma sak”

[C.6] “Vissa kunder kommer och säger att de vill ha AI när de inte ens vet vad det är. Det är någon som har läst en artiklen någon gång och tyckt att det här låter bra.”

[C.7] “Det är ungefär så här det såg ut inom andra branscher för tio femton år sedan. Vi har tagit fram liknande lösningar till andra branschen i åratal”

[C.8] “Vi är ju redan digitala, när jag började här hade vi inte ens email, då fixade vi och skickade dokument med cykelbud. Nu kan vi ju göra allt själva i datorn”

[C.9] “Kolla bara hur det gick när X började med automatisk fakturering, det var ju inte en jävla faktura som var korrekt på flera månader”

[C.10] “Kolla på Y när vi (branschen) hade cyberattacken förra året, alla andra rullade ju på som vanligt och de kunde inte göra någonting för att de förlitar sig för mycket på tekniken”

[C.11] “Det är inte alltid våra lokala mål stämmer överens med företagets globala mål, det dom ser från huvudkvarteret är ju ofta något helt annat än vad vi ser här nere på golvet, om vi säger att vi behöver en sak kommer de säga att vi behöver något helt annat”

[C.12] “Tur för dem (HQ) att vi fixar stålarna så att andra kan leka med utveckling”

Appendix G – Questionnaire for Tech industry

How would you rate/describe the shipping sector in terms of usage of IT solutions?

How would you rate/describe the shipping sector in terms of their receptivity of IT in general?

How would you rate/describe your customers interest in AI solutions?

How would you rate/describe your customers knowledge and understanding of AI solutions?

Which are the main problems when providing software solutions?

Appendix H – Questionnaire for Shipping industry

How would you rate/describe your organisation at IT solutions overall?

How would you rate/describe the external IT solutions towards the customer?

How would you rate/describe the internal IT solutions, back-office, e.g.?

How would you rate/describe your organisation compared to other companies in the same industry?

How would you rate/describe the transport industry compared to other industries, e.g. banking/telecom?

How would you rate/describe the IT within ocean transport compared to other transport modes, air/truck/rail?

Do you think your IT solutions could be improved, if yes, how?

Do you think your current processes allow for more IT solutions?

How hard do you think it would be to implement new IT solutions at your office / dept.?

Where do you see improved IT solutions having the most impact on your org / on your daily work?

What is your understanding of how autonomous learning solutions could benefit you as a manager?

How likely do you see yourself as a manager acquiring skills within IT solutions?

Which potential benefits have your organisation identified in digitalisation?

Which potential risks have your organisation identified in digitalisation?