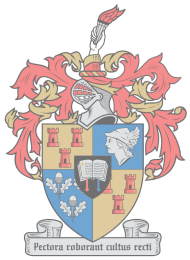
Effective governance through implementation of appropriate algorithms in share trading



UNIVERSITEIT
iYUNIVESITHI
STELLENBOSCH
UNIVERSITY

by

Anna Elizabeth (Nannette) Botha

Thesis presented in partial fulfilment of the requirements for the degree of Masters in Commerce (Computer Auditing) in the Faculty of Economic and Management Sciences at Stellenbosch University

Supervisor: Lize-Marie Sahd
School of Accountancy

December 2018

Declaration

By submitting this thesis electronically, I declare that the entirety of the work contained therein is my own, original work, that I am the authorship owner thereof (unless to the extent explicitly otherwise stated) and that I have not previously in its entirety or in part submitted it for obtaining any qualification.

Date: December 2018

Acknowledgements

A sincere and heartfelt thank you to:

- my Lord and Saviour, for blessing me with a curious mind and an incredible supporting team. And for providing me with an abundance of blessings, even in my sleep (Psalms 127:2).

- my supervisor, Lize-Marie Sahd, for teaching me how to do research with endless patience, and has the ability to make governance sound like poetry.

- the love of my life, Francois, for making me feel like I can do anything, and helping me create time to do it. I will never get over the excitement of finding you.

- Cari, who is teaching me more about love and life than I could ever imagine.

- my mum and dad, for endless and unconditional love, support and cheering. And so many cups of coffee.

- my sister, Sarli, for providing a fresh perspective to this document, and all facets of life (and Albie, I love you too!)

- Liezl and Jana, for so much patience and information, and for so many other colleagues who became friends and blur the lines between learning, working and having fun.

Abstract

Advancement in computer technology enabled an evolution in share trading. This brought such an increase in available data that manual analysis can no longer provide accurate, timeous results. Many share traders have found a solution in the implementation of algorithms.

To effectively govern algorithms and ensure the control objectives of validity, accuracy and completeness are met, the life cycle of an algorithm must be considered: the input data, analysis and results must be governed.

The choice of algorithm is fundamental to effectively govern its analysis and results, since an algorithm is not always appropriate for implementation. The algorithm must be appropriate for the available data, the requirements of the analysis, as well as the required algorithm result in order to meet the control objectives.

To investigate the applicability of algorithms, this research provides an understanding of the evolution in the share trading industry, algorithms and the enabling technologies of big data and machine learning. The study considers both qualitative and quantitative algorithms: statistical characteristics of predictive algorithms are identified, which indicate if the algorithm is appropriate for implementation based on the nature of the data available, the required analysis as well as the results the algorithm can achieve. The research will also investigate how nonpredictive algorithms' outcome determine if it will be useful and appropriate to the data scientist.

Based on the investigation, an applicability model was designed to map the investigated statistical characteristics with the indicators found. This model will provide guidance to data scientists and other users to assess their data and algorithm needs to what the available algorithms can provide, therefore determining which algorithm characteristics will be most appropriate for implementation.

Uittreksel

Die vooruitgang in rekenaartegnologie het ’n evolusie in die verhandeling van aandele moontlik gemaak. Met die toename in beskikbare data, is dit nie meer moontlik om ’n analise per hand te ondersoek en akkurate resultate betyds te kry nie. Baie aandele-makelaars het gevind dat die implementering van algoritmes ’n oplossing hiervoor bied.

Om algoritmes effektief te beheer en te verseker dat die kontroledoelwitte van geldigheid, akkuraatheid en volledigheid behaal word, moet die lewenssiklus van ’n algoritme in ag geneem word: die inset data, analise en resultate moet beheer word.

’n Fundamentele keuse is watter algoritme om te implementeer om die analise en die resultate daarvan te beheer, aangesien algoritmes nie altyd gepas is vir implementering nie. Die algoritme moet gekies word volgens die beskikbare data, die vereistes van die analise, sowel as die resultaat wat van die algoritme vereis word.

Om die toepaslikheid van algoritmes te ondersoek, bied hierdie navorsing ’n begrip van die evolusie in die industrie van aandele-verhandeling, algoritmes en die tegnologieë van ‘big data’ en masjienleer. Hierdie studie neem beide kwalitatiewe en kwantitatiewe algoritmes in ag: dit identifiseer statistiese karaktereienskappe van voorspellende algoritmes, wat gebruik kan word om te bepaal of die algoritme gepas is vir implementering. Dit word bepaal deur die aard van die beskikbare data, die ontleding wat die algoritme moet uitvoer en die resultate wat die algoritme moet verkry. Hierdie studie ondersoek ook die doelwit van algoritmes wat nie waardes voorspel nie, bepaal of dit nuttig en gepas is vir die gebruiker.

Volgens die bevindinge van die ondersoek is ’n model van toepaslikheid ontwerp om die statistiese eienskappe wat ondersoek is, met die aanwysers wat gevind is, te karteer. Hierdie model verskaf riglyne aan die gebruikers om die beskikbare data en behoeftes vir die algoritme te vergelyk met wat die algoritme kan verskaf, en dus te kan bepaal watter algoritme-eienskappe gepas is vir implementering.

Contents

List of tables

List of graphs

List of figures

Chapter 1: Introduction

1.1 Introduction and background information

Success in share trading demands great skill and knowledge of its traders. Not only does it require a thorough knowledge of the industry and a detailed understanding of the different factors impacting companies' share prices, it also requires continuous awareness of the developments and changes in each of these factors. The share market changes constantly and this must be studied, analysed and acted on continuously and accurately and timeously. Failure to do this leads to lost profit opportunities and instant losses (Khan, Alin & Hussain, 2011; Richardson, Gregor & Heany, 2012).

In order to achieve all of this, traders must do extensive research on an ever-changing and ever-expanding data set (Hu, Liu, Zhang, Su, Ngai & Liu, 2015). Using the available data to perform a predictive analysis is done with complex calculations which predicts an expected future value of shares. This expectation is then used as a basis for the required share trades to enhance the share portfolio and its profits. These calculations are based on the underlying companies' own forecasts, combined with the traders' insight in the market and the industry in which the company operates. This is a very time-consuming and data-intensive process due to the great number of shares available for trading, as well as the vast volume of data available for each of these shares (Bloomberg, 2017; Khan *et al.*, 2011).

What further complicates the analysis of shares, is that there is no set of proven rules to maximise profits and avoid losses. While traders have developed trusted methods, it is not a guarantee for profits (Khan *et al.*, 2011). Therefore traders need to be aware of unexpected opportunity or loss indicators, which they might not have used before.

Traders also need to be aware of any movement in the shares' values. When a share price changes it creates a big enough difference between the expected (calculated) share value and actual share price according to the share trader's threshold, there is an opportunity to buy shares at a price lower than its estimated value, or to sell it at a

price higher than the estimated value. Monitoring share prices also avoids the risk of buying shares at an inflated share price, which occurs when a share is sold for a market value higher than its estimated value. These changes can occur over time or at a moments' notice, which necessitates ongoing assessment of changes. It is essential that the share analysis process must be concluded in a continuous and timely manner. Success in share trading is very time sensitive, since any delay in trading affects the price applicable at the exact time at which the trade instruction is posted and actioned. It has been proven that time delays in share trading significantly decreases the quality of decisions made (Bloomberg, 2017; Khan *et al.*, 2011).

This puts pressure on traders to not only complete a comprehensive and accurate analysis, but also conclude and action the required trade in a time-efficient manner. Because this is not humanly possible, another solution had to be found (Bloomberg, 2017). With the advancement in computer and other technologies, more and more companies are trusting algorithms to be that solution. While some companies use the algorithms to only assist in the analyses of shares, others implement algorithms to automate both the analysis and execution: after analysing all the available data, it uses the share trader's preferred trading strategy or its own learning to decide which share trades would be advantageous, and then executes this decision automatically – all without human intervention (Hu *et al.*, 2015).

As with all technologies, the implementation of share trading algorithms must be carefully governed. The King IV report identifies technology as a competitive advantage and emphasises the importance of creating value through technology by managing the associated risks and opportunities. This can only be achieved through effective governance of an entity's information and technology (IODSA, 2016).

Effective control is required in order to appropriately govern the outcomes achieved by algorithms (IODSA, 2016). Effective control requires of algorithms to adhere to the control objectives of validity, accuracy and completeness (Von Wielligh & Prinsloo, 2014). The only way these control objectives can be achieved is if the available data set, its analysis and the results and actions produced all adhere to these control objectives (Deloitte, 2017).

One of the key governance factors in the implementation of algorithms, is choosing an appropriate algorithm. If the implemented algorithm does not suit the available data, the analysis requirements and the required results, it will not deliver valid, accurate and complete results.

This research will investigate and assess characteristics of the algorithms available for share trading, in order to design a model which will assist users in choosing which of these algorithms would be most suited for their needs and circumstances. This is necessary to ensure governance of the technology.

1.2 Research objective

This research will focus on effective governance of algorithms by designing a model to ascertain which algorithm characteristics will be appropriate for addressing the needs or requirements of the user and ensure that the control objectives of completeness, accuracy and validity are achieved.

Governance of a share trading algorithm requires that the result of its implementation – the trades it recommends or automates – must be complete, accurate and valid. Therefore, to effectively govern algorithms it must be appropriate for the quality and quantity of data, then provide an accurate and valid analysis, and finally deliver a valid, accurate and complete action or result.

This leads to the research objective: Designing a model that will ensure effective governance of rule-based (algorithmic) share trading by identifying the appropriate algorithm for implementation.

1.3 Value add research motivation

Governance is an ongoing challenge for users of technologies. This research will design a model of its research findings to provide useful information to users to assess if their chosen algorithm is appropriate. This model will show the nature and applicability of the algorithm characteristics investigated in this research, to assist in assessing if an appropriate algorithm was chosen for the available data, as well as the required analysis and results, to address the governance control objectives.

Because the technology of algorithms is still evolving, this research can assist first time implementers thereof to understand the available algorithms, as well as the limitations of its application. It can also assist these users to assess if implementing this technology will be feasible and useful to address its stakeholders' requirements, especially with regards to governing the technology.

It will also provide guidance to those users who have already implemented algorithms, to assess if their chosen algorithm is still appropriate and effective for any changes in their available data and needs.

1.4 Scope limitation

In order to provide guidance on the governance of share trading technology, there are many components to consider: the requirements of governance, the share trading industry, the statistical nature and purpose of the identified algorithms, as well as the technologies required to enable effective functioning of the algorithm.

However, it is not the aim of the research to be a technical analysis of rule based share trading, and will not assess which share trading strategy should be used for optimal profit. As discussed in chapter 1.1, this is an extremely technical question for which there is no definitive answer for success – it remains a strategy the share trader chooses (Khan *et al.*, 2011). Therefore this research will focus on assisting the share trader in choosing an appropriate algorithm, rather than providing guidance on the trading strategy the algorithm encompasses.

This research will also not be a detailed and technical analysis of the statistical methods for identifying and executing a share trading strategy. While it will investigate the statistical nature and purpose of the algorithm on a high level, it will not provide a technical guide to the statistics thereof. Furthermore, this research will also not provide a technical explanation of the writing (coding) of algorithms, or offer any practical assistance in the programming of algorithms.

While the algorithms' enabling technologies are included in this research for an understanding of its nature and investigation of its risks, this research will not aim to be a complete and detailed study of either big data or machine learning.

Lastly, this research is also not a study of data governance or the implementation of data integrity, data security and other data governance techniques. This research scope will include only the investigation of data characteristics which indicate which algorithm would be appropriate for analysis of that data, and appropriate for user's requirements for the analysis and the expected or required outcome.

1.5 Methodology

In order to achieve the research objective of governing the implementation of share trading algorithms, this research methodology will include the following steps:

1. A literature review will be performed to provide the necessary context to define and understand governance, with specific reference to IT governance. The literature review will also investigate the history and nature of share trading to understand the environment of the research objective. This will highlight the issues share traders face, and identify how the implementation of algorithms can be the solution to these issues.

The evolution in share trading closely coincides with the advancement in technology; this research will also investigate the enabling technologies of big data and machine learning, which often enables the implementation of algorithms. It will also investigate the nature and possibilities of algorithms, and show why it is suitable to address the issues of the share trading industry.

Lastly, there are many inherent risks of implementing algorithms and its enabling technologies. A literature review will be performed to identify these risks to provide context for the implementation of governing algorithms.

2. The research will then investigate identified algorithm characteristics to identify the limitations and risks the share trading industry is exposed to when implementing those algorithms. This will include an understanding of what data qualities are required for the implementation of a specific algorithm; it will also investigate the analysis the algorithm performs, and what it can achieve.

3. Based on the data available and the analysis and results required, not all algorithms will be able to successfully achieve the control objectives of validity, accuracy and completeness for a specific data set. This research will assess what requirements and limitations each algorithm has in order to assess which algorithm would be appropriate for implementation, based on the nature, potential and restrictions of the technology and its components.

4. Lastly, this research will design an applicability model, which will provide a guide to users implementing algorithms to to assess which algorithm characteristic is required or appropriate for the data it has available, and the analysis and results the algorithm must achieve to ensure effective governance thereof by achieving the required control objectives of accuracy, validity and completeness.

1.6 Structure of research chapters

Following the methodology explained in 1.5, and to achieve its research objective, this research will be structured in the following five chapters:

- **Chapter 2: Literature review**
  The literature review will provide the necessary background and context to understand governance and IT governance, as well as the history and nature of share trading, and the evolution thereof. It will also investigate the enabling technologies of big data and machine learning, as well as the inherent risks which the implementation of any share trading algorithm will expose the user to.

- **Chapter 3: Understanding share trading algorithms**
  Considering the context provided in chapter 2, the research will investigate the nature and characteristics of algorithms in chapter 3, in order to design an applicability model in chapter 4.

  In order to govern share trading algorithms by ensuring that it achieves the control objectives required of it, the appropriate algorithm must be implemented. The algorithms most pertinent to share trading are identified, and this chapter provides a high-level explanation of the statistical nature, applicability and limitations of identified algorithms, in order to identify the appropriate algorithm for implementation based on its nature and purpose.

- **Chapter 4: Applicability and limitations of algorithms**
  With the research findings provided by chapter 3, this chapter will design a model to show which algorithm would be appropriate for implementation. It will provide users with a table of data qualities, as well as requirements of the analysis and algorithm results, in order to guide users in assessing the data they have available for assessment by the algorithm, as well as the requirements they have of its results. This will be used to determine which algorithm is appropriate for implementation.

- **Chapter 5: Conclusion**
  In this final chapter, this research and its findings are summarised and concluded.

1.7 Conclusion

This chapter has shown the objective, methodology and layout of this research. In chapter 2, this research will gain an understanding of the governance environment of algorithms and its enabling technologies, by investigating the nature and risks thereof.

Chapter 2: Literature Review

2.1 Introduction

In order to design a model addressing governance of the analysis and results of share trading algorithms by choosing the appropriate algorithm, this research first needs to investigate corporate governance, and more specifically IT governance and how it applies to share trading algorithms. The research will also investigate the evolution of share trading, as well as the technologies of algorithms, big data and machine learning.

2.2 Overview of literature review

It is essential to create synthesis in the research, since there are so many components to the research objective. This requires integration of different components from different study fields, and creating a uniform conclusion from all (Fink, 2005; Cooper, 1998). The following stages proposed by Cooper (1998) were followed for the collection and integration of sources to achieve synthesis in this literature review:

- **Literature search:** Possible sources are investigated which includes the identification of components of the research objective as possibilities to pursue as sources of research articles, after which research is collected.
- **Data evaluation stage:** The quality of collected articles and its sources are assessed to ensure that only relevant, appropriate sources are included in the literature review.
- **Analysis and interpretation stage:** The collected, appropriate sources are collated to form one uniform viewpoint. For this research, the chosen appropriate sources are combined to address the research objective by performing the literature review of chapter 2, and to analyse and interpret the information to form the research findings of chapter 3.
- **Public presentation stage:** A document is prepared to present the research and its findings to the public, which is done with the formalisation of this document (Cooper, 1998).

For the literature search and data evaluation stages, existing research is considered. To assist, Webster & Watson (2002) identified two types of literature reviews:

- A literature review of a **mature topic** which will have a large amount of research available to examine and synthesize. This will lead to a thorough literature review.
- A literature review of an **emerging topic** will have less available literature; however, the field will benefit from this exposure to new theoretical research. The literature reviews of an emerging topic will be shorter and less robust in nature.

The literature review for this research falls in the latter category. However, regardless of the maturity of the research topic, the research sources must be of a high quality to ensure an accurate literature review (Fink, 2005). Furthermore, since the information systems research is often interdisciplinary, a systematic approach is required to obtain all relevant sources (Webster & Watson, 2002).

According to the research of Fink (2005) and Webster and Watson (2002), selecting appropriate databases and articles as research sources is important for an accurate and complete literature review. For this reason, the scope of this research search was identified by using the fundamental components of the research objective: share trading, algorithms, data science and governance. Though these strings were intentionally short in order to expand the search result, databases such as Elsevier and Scopus revealed limited appropriate research articles of which most were of a very technical statistical or computer programming nature. Therefore the scope for research included in this literature review was extended to not only include articles published in accredited journals, but also articles published on the internet by credible sources and authors with the necessary qualifications and experience.

2.3 Corporate governance

One of the most important components of the research objective to consider is the concept of governance. Fundamentally, it dictates effective, ethical leadership and assigns the responsibility of providing direction and an example to its organisation to the governing body (Deloitte, 2018). South African organisations are not left to their own devices to address this comprehensive task; the King IV report is the leading authority on corporate governance in South Africa. It provides practical guidelines and principles to governing bodies to address the increasing challenges of governance, as well as guidelines to report their successes and shortcomings to their stakeholders in an integrated report (PWC, 2017).

The principles of King IV focus on the following four outcomes of good corporate governance (IODSA, 2016):

- ethical culture,
- good performance,
- effective control and
- legitimacy.

When considering the governance of the share trading algorithms, there are two outcomes of good corporate governance which are prevalent:

- **good performance**: In the share trading environment, good performance translates to trading profits. This must be achieved by identifying trading opportunities for profits and identifying loss indicators, and reacting to it timeously in order to maximise profits.
- **effective control**: Because there are many pitfalls in maximising share trading profits through careless speculating, ensuring effective control is very important in the share trading environment. Therefore the governing body must ensure that appropriate and effective internal controls are in place to ensure an effective control environment. This is particularly important to the research objective, which will provide guidelines to choose an appropriate algorithm for implementation to ensure that control objectives can be met.

Furthermore, one of the main focuses of the best practices provided by King IV, is to assist entities in realising how much potential corporate governance has for the creation of value. Therefore following these guidelines will enable users to not only achieve corporate governance, but also to harness its advantages and opportunities (IODSA, 2016).

2.4 Information Technology (IT) governance

The sentiment of corporate governance to create value is also extremely pertinent to the governance of information technology (IT). Gartner (2018) defines IT governance as *"the processes that ensure the effective and efficient use of IT in enabling an organization to achieve its goals"*.

As with corporate governance, the King IV report also provides guidelines and best practices for achieving governance of information and technology. It also emphasises the importance of IT governance by identifying it as too pivotal to the operations and success of an entity not to pay specific and detailed attention to it. It also proposes that it must be considered as a separate and regular item on any governing body's agenda (IODSA, 2016).

For the governance of IT, the King IV report focuses on the expected results of its implementation as an incentive to the best practices and policies it provides. In the report's foreword, the governance and security of IT is already established as critical components of corporate governance: it is no longer considered only a tool aiding in business practices; it is rather identified as an opportunity for growth, opportunity and value creation. Building on the value created by achieving corporate governance, King IV recognises the opportunity to not only create a competitive advantage through information and technology, but also to avoid disruption and other risks caused by the mismanagement of information and technology (IODSA, 2016).

Before the King IV report, the King III report was issued to provide governance guidelines to South African companies. In the King III report, the following objectives for IT governance were provided: strategic alignment, risk management, value

delivery, resource management and performance management (IODSA, 2009). While the King III report's approach was different, the principles of King IV does not contradict those of King III (PWC, 2017). Butler and Butler (2010) found that by following the guidelines provided in the King III report, South African organisations will also achieve key performance areas of international best practices for IT governance.

While King IV does not provide such a detailed approach as King III did, its focus is still on achieving value creation through the entity's investment in IT, as well as the management of the opportunities and risks it entails. The King IV report includes 17 principles for achieving corporate governance; principle 12 was introduced to provide guidelines for effective IT governance. It assigns the responsibility for IT governance to the governing body to provide leadership and oversight of IT, as well as ensuring that technology and information is governed to assist in achieving the organisation's strategic objectives (IODSA, 2016). This means that the governing body must provide oversight and direction to IT by putting the required strategies and activities in place. While the management of these governance activities can be delegated to the management of the IT department, it remains the responsibility of the governing body (Deloitte, 2018; IODSA, 2016). This creates a great challenge by assigning the responsibility for IT to the governing body, who are most likely not specialists in IT, IT governance or the implementation and management of technology (Boshoff, 2016).

This confirms one of the main challenges of IT governance: alignment. In most cases, the board of directors does the operational and strategic planning for an entity, and it usually consists of specialists in business management. When these business decisions are then communicated to IT management to be actioned and implemented, business terms are used to describe it. Furthermore, the IT needs arising from these business decisions are also communicated to the IT department in business terms. These IT needs is then translated into IT terms and executed according to the IT department's understanding of what those business terms entails. This issue with alignment between what the business needs, requests and expects, and what IT provides as a solution is referred to as the IT gap (Boshoff, 2016; Goosen & Rudman, 2013).

This also applies to share trading companies. The danger in this industry is that IT and business departments often have different ways of measuring business value. What IT could view as valuable, could be disregarded by business. Or IT could misinterpret data value, and not report it to business who could use it to identify opportunities or threats to create value (Boshoff, 2016; Luftman, 2003). To address the IT gap and IT governance requirements, it is fundamental to ensure communication and a mutual understanding of business value, which must be an ongoing strategy (Luftman, 2003).

2.5 Governance of algorithms

Building on the concept of governance as explained above, it is important to consider the practical implications of creating value through share trading algorithms. Considering the corporate governance goals of performance and effective controls of the King IV report (IODSA, 2016) as identified in chapter 2.3, this research will focus on the following components of the governance of algorithms:

- **Performance:** In order to ensure profit, the algorithm must be able to identify share trading opportunities and risks. Choosing the appropriate algorithm is key in achieving this; depending on the nature and amount of data available, not all algorithms will be able to achieve useable, meaningful results. This will also be the case if there are any requirements or constraints for the analyses (such as cost or time), or any expectations of what the result must be (Brownlee, 2014). These requirements will be investigated in chapter 3 of this research.

- **Effective control:** A share trading algorithm has a twofold purpose: not only does it perform the actions of achieving useful information and/or automated trades, it also serves as an internal control. According to Von Wielligh and Prinsloo (2014), in order for it to be considered an effective control, the following control objectives must be considered:

Control objective	Definition
Validity	Transactions are authorised and according to company policy. Transactions occurred during the period, and has supporting documentation.
Completeness	All transactions are recorded, in a timely manner, and none were omitted.
Accuracy	Transactions are recorded at the correct amount, classified correctly and posted correctly.

Table 2-1: Definitions of control objectives for governance

Applying these definitions to share trading algorithms, this research derived the following requirements of algorithms in order to achieve the governance goal of effective control:

Control objective	Application to algorithms
Validity	The algorithm must act and create results according to the entity's strategy and/or its investment management agreement (instruction from client). Results must also be meaningful and add value.
Completeness	All data items must be considered by the algorithm to ensure a complete and meaningful conclusion.
Accuracy	The analysis must be done accurately to ensure an accurate recommendation or conclusion.

Table 2-2: Application of control objectives to algorithms

In order to ensure governance of algorithms, the programmer and algorithm users must ensure that algorithm can achieve these control objectives for the analysis, as well as its results. To understand how this applies to the share trading industry, the evolution of share trading must be considered.

2.6 The evolution of share trading

In order to address the control objectives of share trading, it is important to understand the history and evolution of share trading. The face of the share trading industry has changed significantly from the initial physical negotiations with cries to sell and buy. Its modernisation was enabled by the advance in technology – first by introducing telegraphs in 1856, and again with the telephone in 1876 (Forex Capital Markets, 2018; Flinders 2007). In 1986 the London Stock Exchange introduced a quotation system using computer screens to match sellers and buyers. Today, share traders worldwide silently use a modern, computerised system which matches sellers and buyers. This has inherently changed the nature of share trading, as well as the controls required for it (Stoll, 2006).

Computer systems enabled a more scientific approach to share trading with much higher trade volumes, significantly quicker initiation and completion of trades, and cheaper trading costs. For example, in 1987 monthly trading volumes were equivalent to the annual trading volumes of the pre-1986 era (Flinders, 2007).

These significant changes did not occur without problems. Progress in the operations of share trading with the implementation of technology also introduced significant risks. Not understanding and addressing these risks resulted in the stock market crash on 19 October 1987, which is now known as Black Monday. On this day, share values of the major indexes lost more than 30% of its value. While there were other contributing factors, the fairly new concept of computer trading was cited as one of its main causes. Many analysts blamed computer programming for the market crash, which automatically continued to trade large volumes of shares when its prescribed conditions were met (Flinders, 2007; Itskevich, 2002).

While other experts now say that it was not the programmed trading itself which caused the market crash, they agree that it did enable the speed and severity of the decline in the share prices. It was the immaturity of the implemented technology which caused issues: it was not advanced enough to analyse and consider factors other than share prices (Flinders, 2007).

Following the market crash, technology was quickly updated to avoid a repeat of the disaster. The most important aspect was the introduction of trading curbs, also referred to as circuit breakers (Flinders, 2007; Itskevich, 2002). It is a regulatory tool to avoid speculation and major losses, by monitoring and allowing only a fixed percentage difference between the trade price and the reference price as determined on the day before. The price used as reference is usually the Volume Weighted Average Price (VWAP). This limits all trading if these circumstances are met, especially automated trading which might not take all aspects of trading into account for its conclusions (Johannesburg Stock Exchange, 2016).

This potential limitation creates issues for the required high volume of research which must be completed with speed and accuracy; with the advancement in technology, share trading algorithms became the alternative to manual calculations and analysis.

2.7 Understanding algorithms

Computer algorithms is one of the basic principles of computer programming. An algorithm is a specific set of instructions, which are programmed for the computer to understand and execute the instructions and achieve the required outcome. It instructs the computer on how its required functions should be accomplished and prioritised and how the predetermined input must be handled to achieve a predetermined outcome (Murty, 1997).

While the implementation of computer algorithms is not a new technology, the advancement in computing capacity enabled it to be implemented to assist in or automate the analysis of share data, and to derive conclusions and decisions and even automate the trading of shares (Deloitte, 2017; Hu *et al.*, 2015; Khan *et al.*, 2011).

Though researchers who tested profitability of algorithms in the Australian market twenty years ago found initial algorithms unsuccessful, they concluded that the introduction and progress of machine learning would assist (Pereira, 2002). While not all algorithms require machine learning, it provides additional predicting power. It was found that algorithms combined with machine learning can solve the issue of which

strategy to use between alternative rules in the algorithm, since this technology can add any hidden or unknown data patterns to the programmed algorithms and therefore identify its own optimal trading strategy (Hu *et al.,* 2015; Khan, 2011).

There are many algorithms available and choosing one superior algorithm which will be best under all circumstances is not possible. The reason for this is twofold: Firstly, to obtain a better-quality outcome, the data quality is very important. While algorithms can be updated to better suit the problem, poor quality data will always lead to confusing or irrelevant results (EliteDataScience, 2017; Brownlee, 2017).

Secondly, the inherent nature of algorithms must be understood to ascertain if it is applicable to address not only the problem, but also the data set under analysis (Brownlee, 2017). To achieve the research objective, the characteristics of algorithms prevalent to share trading will be investigated in chapter 3, and further guidance on the applicability of the investigated algorithms will be provided in chapter 4 of this research.

Considering algorithms to address these issues cannot be done in isolation. The implementation of share trading algorithms is an efficient but data intensive process (Hu *et al., 2015)*. Because of this, the advance in predictive algorithms cannot be investigated without considering the advance in big data, as well as machine learning (Mnich, 2018; Vorhies, 2017). The development in the technologies of big data as well as machine learning has now progressed far enough to enable algorithms as a share trading solution, while achieving the control objectives of validity, accuracy and completeness, which are required for effective governance of algorithms.

2.8 Understanding big data

The evolution of computerised share trading not only impacted how shares are traded, but also the data available for the analyses to do so. There has been exponential growth in the quantity of available data, as well as in the technologies available for performing an analysis of such a large volume of data (Hu *et al.*, 2015). Analysing such a big data set cannot be completed manually; even traditional computer software struggled and failed to handle such large volumes of data. A solution to this is provided in big data technologies.

Investigating the research of Mnich (2018), Jain (2017), Marr (2015), IBM (2014) and Mayer-Schonberger and Cukier (2013), the following aspects of understanding the nature and scope of big data, and defining it, was identified:

- **volume** of data: volumes of data too large for normal computers tools to process or analyse it;
- **variety** of data: data sets which contains complicated, unstructured data has the potential to show trends and insights which were previously hidden;
- **velocity** of data: very diverse data, which changes often and quickly; and
- **value** of data: big data must address accuracy of results through the data quality, while performing data collection.

These qualities of big data solutions enables it to provide real-time processing solutions (IBM, 2014). This is essential to the share trading industry, since data quality impacts on the resulting trends and insights obtained, and these form the basis of automated conclusions on which share trades will be actioned. Resources and time to confirm findings is not available, since real-time insight is required to action share trades. However, it must be emphasised that the definition and emphasis of big data is on the data itself, and not the tools used to analyse it (Mayer-Schonberger & Cukier, 2013).

Big data is collected when every online action is captured by compatible technology and filed and stored continuously, which builds a vastness of available data. The potential in this data can only be harnessed if it is analysed accurately and timeously to provide meaningful information (Mnich, 2018; Jain, 2017; Marr, 2015; IBM, 2014). While the tools are now available to enable this analysis, the biggest challenge is still to interpret the findings and obtain insight through the analysis of big data (Deloitte, 2017; Marr, 2015).

Simply gathering the data would not be useful unless it is tidied and put in order, to enable meaningful analysis thereof (Jain, 2017). Share trading requires continuous research of very large data sets to obtain meaningful insight, which makes big data very relevant to this industry. With the share trading evolution described above, the introduction and development of computers allowed share traders to analyse significantly more data than is manually possible (Flinders, 2007).

Managing large sets of data requires careful data governance. As part of this, it is important to continuously assess the performance management of big data, to test if it is still achieving its set goals. The following are proposed as relevant big data objectives (Ryan, 2018):

- quality of insights gained;
- quality of forecasting;
- automation of business processes; and
- providing detailed and timely performance measurement.

As per these objectives, not all gathered data is relevant or useful. It is key to distinguish which data can create value for the business' strategies, and what data should not be investigated further, in order to govern big data (Gantz & Reinsel, 2012). The King IV report also prescribes the management of value creation through information and technology as an important aspect of IT governance (IODSA, 2016). Creating value is particularly important to the share trading industry, where such a great amount of data is available, which is constantly increasing and changing (Mnich, 2017; Hu *et al.*, 2015). Therefore, the challenge is to harness this data while it still has value during its life-cycle (Mnich, 2017).

However, big data cannot operate as a stand-alone technology. For big data technology to add value, algorithms and machine learning must be applied to enable accurate insights and results.

2.9 Understanding machine learning

The developments in big data technology enabled share traders to use computers to gather vast quantities of data, and to organise it into meaningful information which can be used for trading insights.

Not all algorithms require machine learning. However, harnessing value from a vast amount of data can be problematic. In such cases, machine learning provides a solution by automating the investigation process: depending on the level of automation, limited or no human intervention is required. Machine learning can automatically analyse data to obtain meaningful trends, anomalies and other insights (Shalev-Shwartz & Ben-David, 2014).

This characteristic makes machine learning technology very different from other technologies, since user intervention cannot change or redirect its processes. While the user could remain in control of the instructions and prescribed outcomes achieved by the technology of big data, machine learning uses its own experience as a learning opportunity and continuously updates and betters its own processes to obtain further understanding of the data it analyses. (Shalev-Shwartz & Ben-David, 2014). This is in stark contrast to most other technologies which only follows strict prescribed instructions. While initial instructions are set to initiate the machine learning tool, it will use its experiences to develop itself without further instructions or prescriptions from its owner (Shalev-Shwartz & Ben-David, 2014; Langley, 1996).

Machine learning refers to computers and other computing tools which are able to find meaningful trends in data automatically through its own learning abilities. Therefore there are two requirements to machine learning: these tools must be able to learn, and to adapt (Shalev-Shwartz & Ben-David, 2014). While the concept of learning is so broad that it would be limited by trying to fit it in a single definition, it is explained as the "improvement of performance" by obtaining knowledge through experience in a field (Shalev-Shwartz & Ben-David, 2014; Langley, 1996).

Machine learning imitates human learning: training data is provided to introduce the full data set. The technology investigates the training data to draw conclusions on any trends, outliers and other insights. This is then tested on another data set: the evaluation set (also called the validation data set or testing data set), which tests the insights and conclusions of the machine learning with predetermined outcomes. When these results are satisfactory, machine learning insights can be generalised to the full data set, or other data sets obtained (Microsoft, 2018; Brownlee, 2017; Shalev-Shwartz & Ben-David, 2014).

The concept of computer learning is wider than just memorising theory and acquiring knowledge; it requires the skill to apply theory and its findings to more data. Machine learning requires both the acquisition of knowledge, as well as refining those acquired skills for effective results; in order to learn, machine learning must understand the source, type and applicability of the data items, group or summarise it accordingly and finally draw meaningful conclusions from the obtained insights. The quality or depth of learning accomplished cannot be measured, only the achieved results. A deeper level of learning is required for an effective analysis to add value by achieving useful, accurate results (Witten, Franck, Hall & Pal, 2017; Michalski, Carbonell & Mitchell, 1983).

Most of machine learning's conclusions are made by making rules: Machine learning analyses the data, finds trends and insights and use these to build rules for optimal decision-making: clear conclusions of "yes" or "no", rather than considering grey areas. An area of concern are those cases where a judgement call is necessary, such as fields where there are borderline cases or other areas where human intervention might traditionally be preferred (Witten *et al.*, 2017; Essinger, 1990).

However, because machine learning uses its learning insight rather than judgement, it avoids bias and make better quality decisions. If sufficient training data is provided and the machine learning was not too restricted to be able to learn, machine learning will produce better founded decisions based on past results than human judgement can achieve (Witten *et al.*, 2017; Essinger, 1990). Therefore machine learning makes judgements with the advantage of avoiding bias, making quicker decisions and always being available, which leads to better quality estimates and decisions. This elevates the algorithmic instructions from a static system to an evolving and always relevant system (Essinger, 1990).

Furthermore, there are few grey areas in the share trading industry: While judgement is required when initially choosing or reassessing a share trading strategy, its execution is managed by rules rather than guidelines. Therefore the execution by algorithms and machine learning provides for few areas for judgement, and makes it an ideal field for rule based technologies such as algorithms (Essinger, 1990).

Some algorithms can only function effectively when implemented with a component of machine learning, as will be investigated in chapter 3. This is fundamental for the algorithms implemented in the share trading industry: share traders do not always want to prescribe to the computer what should be achieved through the provided programmed instructions, but also need the meaningful insight from the conclusions of machine learning (Deloitte, 2017).

2.10 Inherent risks of share trading algorithms

The nature and characteristics of algorithms will be investigated in chapter 3 to ascertain its applicability and limitations for share trading. However, there are inherent risks of the share trading industry and its enabling technologies which will apply to all algorithms, not only to a specific type of algorithm as will be investigated in chapter 3.

This research will not attempt to create a complete set of risks, but rather list those risks most relevant to address the research subject of governance of share trading algorithms, to understand the nature of algorithms, and what the implementation thereof entails. After consideration of available research (Deloitte, 2017; Cox, 2016; Boshoff, 2016; Goosen & Rudman, 2013; Richardson *et al.*, 2012; Mackenzie, 2011), the following inherent risks of implementing any share trading algorithm, regardless of which type of algorithm is chosen to be implemented, were identified:

- **Alignment of operational and programming strategies:** When writing and programming an algorithm, there is a risk that the instructions from the share trader and the execution of the programmer does not align. Not only must the algorithm be carefully aligned to the share trader's strategies, it must be written to execute exactly what is expected of it. Inappropriate or inaccurate algorithms will lead to misleading and inaccurate conclusions. This refers to the issue of alignment, as described in chapter 2.4 (Boshoff, 2016; Goosen & Rudman, 2013).

- **Programming skill:** Insufficient or inaccurate programming will lead to invalid and inaccurate conclusions. Writing an algorithm is not a simple process; it must be carefully planned, executed and monitored. It is imperative that the entire process must be completed by a knowledgeable, skilled programmer, who understands the implications of what is included and excluded from the algorithm (Deloitte, 2017).

- **Trading costs of small orders:** On small trading orders, the cost of each trade can exceed the value thereof. Ensuring the timeous and automated completion of a trade must be balanced with the cost of trading, since the cost of continuous small trades can exceed the gains made if not managed carefully (Cox, 2016).

- **Trading cost of large orders:** With large trade orders, the risk of an inflated trading price is created through a sudden spike in demand. To trade shares, a purchase order must be matched with an order to sell. This creates an issue called "slippage": when a large order to purchase is noticed, it causes an increase in the asking selling price. This can be avoided by programming the algorithm to purchase the order in parts by using execution algorithms (Mackenzie, 2011).

- **Share trading regulations:** Share trading is a very closely regulated industry and the rules will not be excused because it is an algorithm planning and executing the trades rather than a human trader. This means that these rules must be carefully programmed to be a priority, regardless of machine learning insights. Failure to do so could cause significant reputational and financial damage (Deloitte, 2017).

- **Value and accuracy of conclusions:** Because of the time-sensitivity of share trading decisions, conclusions cannot be manually checked or recalculated. Therefore the risk with algorithms is that conclusions could be inaccurate or irrelevant, and that this is not detected by the programmer or trader (Deloitte 2017; Richardson *et al.*, 2012).

- **Decision quality:** One of the unavoidable risks of manual share trading is poor decision quality under time pressure. A computerised decision support tool can assist in providing timely information to ease the burden of decision-making, and result in better quality trading decisions (Richardson *et al.*, 2012). However, because of the time-sensitivity of pricing, it is important to ensure not only the timely execution of algorithms, but also integration with others systems to automate trading. If there are any communication issues in the execution of the trades, it will cancel any success the system had in identifying the trading opportunity or risk (Deloitte 2017; Cox, 2016).

2.11 Additional risks introduced by using big data technology

Algorithms cannot function optimally in circumstances with large, complex data sets without the implementation of big data. While big data is important to enable the technology of algorithms, it also exposes the implementer thereof to additional risks. In the investigation of the research of SAS (2018), Deloitte (2017), Boshoff (2016), Tene and Polonetsky (2013), Tallon (2013), and Bantleman (2012), the following risks which are inherent to big data and apply to the share trading industry, were identified:

- **Inaccurate results due to data quality:** Poor data quality leads to inaccurate results; well governed, high quality, reliable data is required to obtain well-balanced, quality insights. Therefore it is important to understand the data item types and data contributors to ensure compatibility, eliminate data noise (irrelevant items) and other issues which can cause the analysis and its conclusions to be inaccurate (SAS, 2018; Deloitte, 2017).

- **Data privacy:** If data privacy is compromised, it causes legal and reputational damage. One of the main concerns with gathering vast amounts of data, is the privacy thereof. Collecting or using private data from third parties without their knowledge or consent is problematic, and illegal in most cases. Therefore users must be careful of their data collection methods, and if their intended use of the data is allowed (Tene & Polonetsky, 2013).

- **Data security:** Even if consent was obtained to use data, security of that data remains a risk. With international privacy laws continuously increasing, there is great technical, reputational and economic risk if its requirements are not met. Therefore an important part of data governance is the safekeeping of data, and protecting it from intruders and other unauthorised users (Tallon, 2013).

- **Hardware cost:** It is very expensive to implement big data – to obtain the required processing power, expensive hardware is required. Therefore the cost and benefit of big data must be carefully managed, to ensure that value created is higher than the investment and running costs of big data technology (Tallon, 2013; Bantleman, 2012).

- **Assessing data value:** The risk in assessing the value of data, is that it might not be apparent initially and useful data can be disregarded, or useless data pursued. The usefulness of data is often determined by how well it can predict future values; however, this is only determined by exploratory analysis (Smeda, 2015). Therefore it is difficult to ascertain beforehand if there is value in data and information which can be obtained from it, or if it should be discarded (Boshoff, 2016; Tallon, 2013).

- **Value creation and alignment:** If the IT department and share traders do not have the same strategy for which data will be valuable, it will lead to an issue of alignment; business requirements of what big data should achieve must reconcile with IT's capabilities to analyse and store data to create value for the business (Boshoff, 2016; Goosen & Rudman, 2013).

- **Cost of data retention:** If the value of data can be ascertained, the risk is that its value might not exceed the costs of its retention. It is important to consider the life cycle of information: As it loses value towards the end of its useful life, it must be reassessed to see if it is still useful and relevant for decision-making and should still be retained. If it is no longer useful, it must be archived or even deleted in order to save the retention costs. Furthermore, all gathered data cannot be retained for the chance that it could be useful later; retained data must be governed to ensure that it remains relevant and useful to enable optimal functioning of this technology (Tallon, 2013).

2.12 Additional risks introduced by implementing machine learning

Machine learning is valuable to the implementation of many algorithms. The following additional risks, pertinent to the share trading industry, which relate to the introduction of machine learning when implementing an algorithm were obtained from the research of Zhou (2018), Stobart (2018), Serialmetrics (2018), the AI Congress (2017), Brownlee (2017), Coglianese and Lehr (2017) and Deloitte (2017), and relates to the data, learning and the outcomes achieved by machine learning:

- **Sufficiency of training data:** For effective machine learning, the designing of detailed algorithms can be replaced by machine learning if sufficient training data is provided. The risk is that the training data could be insufficient in volume, or in nature; if it does not possess all the characteristics of the population, it will lead to inappropriate and irrelevant insights and solutions by machine learning (Zhou, 2018). This will also result in an inherent discrimination in decisions made if the training data contains these inherent biases (AI Congress, 2017; Brownlee, 2017).

- **Biases in training data:** Any biases in training data will be extrapolated to the entire population, resulting in irrelevant or inaccurate conclusions for the population. This will especially be the case if the training data does not include final conclusions, but rather preliminary information which might not be accurate according to the final information (Coglianese & Lehr, 2017; Deloitte, 2017).

- **Data storage:** Storing machine learning data is problematic since memory networks need so much working memory to store data. This is an ongoing challenge with machine learning, with advancements in data storage required for optimal governance of machine learning (Stobart, 2018).

- **Cost of machine learning:** Machine learning algorithms require a large amount of time and resources to train. Although it can be very effective, it remains a time-consuming and expensive solution (Zhou, 2018; Stobart, 2018; Deloitte, 2017).

- **Human intervention:** One of the biggest issues of machine learning lies in human intervention: depending on the depth of machine learning, the data scientist still provides a human understanding of the problem, the available data and in choosing the most appropriate function or model for its analysis (Serialmetrics, 2018; Deloitte, 2017). If the data scientist made any mistakes or omissions in the assessments, it will lead to invalid, incomplete and inaccurate results.

- **Interpretation of results:** Furthermore, even if the entire process is completed using machine learning, the users of the final solutions will be responsible to understand and interpret the machine learning conclusions. Therefore the risk remains that human intervention and biases could lead to inappropriate interpretations and conclusions of the machine learning results (Deloitte, 2017).

- **Assessment of machine learning quality:** Monitoring of machine learning algorithms is especially complicated: these algorithms are so opaque that users cannot see why deductions are made from training and actual data, and what its conclusions are founded on. Therefore it is very difficult to ascertain if the conclusions are relevant and accurate – especially because deep learning is an ongoing process, which must be monitored continuously (Deloitte, 2017).

The risks discussed above are inherent in the nature of computerised share trading and the required technologies to enable the implementation of algorithms. While it is unavoidable, it is manageable. The value created through the implementation of share trading algorithms will outweigh the cost of managing its risks, if its users remain aware of the risks and govern it by implementing the necessary controls to address it.

2.13 Conclusion

The research objective of designing a model for effective governance of share trading algorithms by identifying which algorithm is appropriate for implementation can only be achieved by understanding the components of governance, share trading, algorithms, big data and machine learning, as well as the inherent risks each of these components entails. This chapter has provided the necessary background information and explanation of the nature of these different components of the research objective. This can now be used in chapter 3 to investigate the nature and purpose of predictive and nonpredictive algorithms, which will assist in achieving governance by choosing to implement an algorithm which is appropriate for the available data, and the required outcome of the algorithm.

Chapter 3: Understanding Share Trading Algorithms

3.1 Introduction

Chapter 2 has provided the background information and an understanding of the components of the research objective: governance, IT governance, the share trading evolution, algorithms, big data and machine learning. This has provided the necessary context to address the research objective in providing guidance on governing the use of algorithms by identification of an appropriate algorithm to provide valid results in an industry where there are many types of algorithms available for implementation. To choose which of these algorithms would be appropriate, it is necessary to understand the limitations and implementation risks of these algorithms. This will determine which algorithm will be applicable and appropriate for the nature of the available data, as well as the intended result the algorithm must achieve. This requires a basic understanding of the statistical nature and purpose of these algorithms.

Share traders choose a strategy to follow in order to obtain profits and avoid losses. According to the chosen strategy, share trading can be categorised according to the intention of the share trader (Kirilenko, Kyle, Samadi & Tuzun, 2017; Van Winkle, 2011; Chan, 2015; Hu *et al.*, 2015):

- **Quantitative trading** is focused on short term transactions, where predictive research is applied. The share data is analysed to provide predictive share prices by analysing the data distribution and trends in each share's price to identify a momentum curve (Van Winkle, 2011). This occurs when the share price is increasing and is expected to keep growing, creating an opportunity for profit in the long position. Alternatively, potential losses can be indicated if the price is decreasing, and expected to continue losing value. It uses historic share pricing and share price trends as a basis (Marx, Mpofu, De Beer, Nortjé, 2013).

- **Qualitative trading** has a long-term focus, for which fundamental research is used. For fundamental research, traders use a company's financial information such as its balance sheet, income statement and forecasted statements, to analyse the credibility of the current market value of the shares. This is not an audit of accuracy, but rather uses the traders' knowledge of the industry and the companies' circumstances to assess share prices to see if all factors which contribute to the current and future profitability of these companies are taken into account and result in a reasonable share price. If any areas for adjustments are identified by the qualitative research, the resulting theoretical share price calculated by the share trader, is compared to the actual market price thereof. A difference between these values creates an opportunity for a trade profit, or a risk for future losses when the required adjustments are incorporated in its market value, assuming a rational market where all information will eventually be incorporated in the share price (Marx *et al.*, 2013).

Using these categories, this research will group the algorithms available for implementation in share trading into two statistical groups, based on the intention of its implementation:

- quantitative, predictive trading algorithms
- qualitative, non-predictive trading algorithms

Because of its predictive nature, there are many statistical tools and methods available for use in quantitative trading. This means that the algorithms used for quantitative, predictive trading can be further classified according to its statistical characteristics. Understanding the statistical characteristics of the algorithm, and assessing if it is appropriate for the intended use of the algorithm, will assist in ensuring effective governance of these algorithms by identifying an appropriate algorithm for a specified data set to achieve valid, accurate and complete results and appropriate conclusions and decisions in share trading. This research will investigate the statistical characteristics of the algorithms in this chapter. These characteristics will then be used to design the applicability model in chapter 4, by using the research findings of this chapter to show which circumstances are appropriate for each algorithm characteristic.

Choosing which algorithm to implement is a very important decision, not only impacting how expensive and time-consuming the algorithm analysis is, but also the efficiency thereof to achieve its intended purpose (EliteDataScience, 2017; Brownlee, 2017). Therefore, the consideration is not with the classification and attributes of the algorithms itself, but rather to first consider the intended purpose of its implementation. For this reason, this research will not only aim to understand the nature of algorithms, but also the applicability of the algorithm according to what the user's intended outcome with the implementation of the algorithm is (EliteDataScience, 2017).

Furthermore, if an algorithm uses machine learning to predict values, it requires training data. Different algorithms act differently, by using different assumptions and biases when analysing training data – therefore achieving different results, even with an identical training set. Based on this, it is imperative to also understand the nature and extent of the available data in order to choose the most appropriate algorithm for the available data, in order to achieve governance of the algorithm technology and ensure valid, accurate and complete results. Based on the outcomes achieved, the most appropriate algorithm can be identified and implemented to live data (Brownlee, 2017; Shalev-Shwartz & Ben-David, 2014).

The importance of understanding of the available data and intended outcome is confirmed by consideration of the life cycle of an algorithm (Deloitte, 2017; Hu *et al.*, 2015):

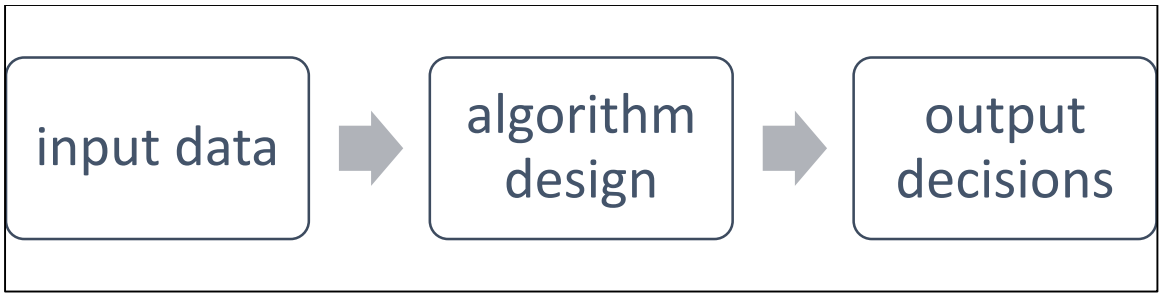


Figure 3-1: Life cycle of an algorithm

Based on this life cycle, and the investigation performed in chapter 2, the following stages of achieving governance of algorithms were derived:

- governance of input data to ensure the data set remains valid, accurate and complete
- governance of the analysis, in order to achieve meaningful insight
- governance of results, which must be valid, accurate and complete

The governance of input data is specifically excluded from the scope of this research, and therefore in order to achieve the research objective this research must investigate how these control objectives can be achieved for the analysis and the results by implementing the appropriate algorithm. This can be ensured by assessing if the chosen algorithm's characteristics are appropriate for the available data, analysis and results, to be able to achieve the control objectives.

3.2 Quantitative / Predictive algorithms:

Performing a statistical analysis of predictive algorithms provides guidance on how appropriate these algorithms are, based on their statistical characteristics and the required outcome of its implementation.

When considering the characteristics applicable to the nature of available data, the following fundamental data characteristics considered in this chapter is: linear or nonlinear, and parametric or nonparametric (EliteDataScience, 2017; Brownlee, 2017; Coffin & Saltzman, 1999).

For both linear and nonlinear, as well as parametric or nonparametric algorithms, machine learning is not required but can be implemented to support the algorithm. However, if the data analysis does include a component of machine learning, the impact of its implementation on applicability of the algorithm must be considered. Therefore, the characteristics of supervised or unsupervised machine learning algorithms are investigated in this chapter (Kolanovic & Krishnamachari, 2017; Microsoft Azure, 2017; Brownlee, 2016d).

Lastly, the required outcome also determines how successful the implementation of an algorithm will be. The characteristics of error of bias and error of variance, as well as overfit and underfit must be investigated for inclusion in the applicability model (EliteDataScience, 2017; Brownlee, 2017; Coffin & Saltzman, 1999).

Combining these characteristics, the following structure will be used for predictive algorithms in chapter 3.2:

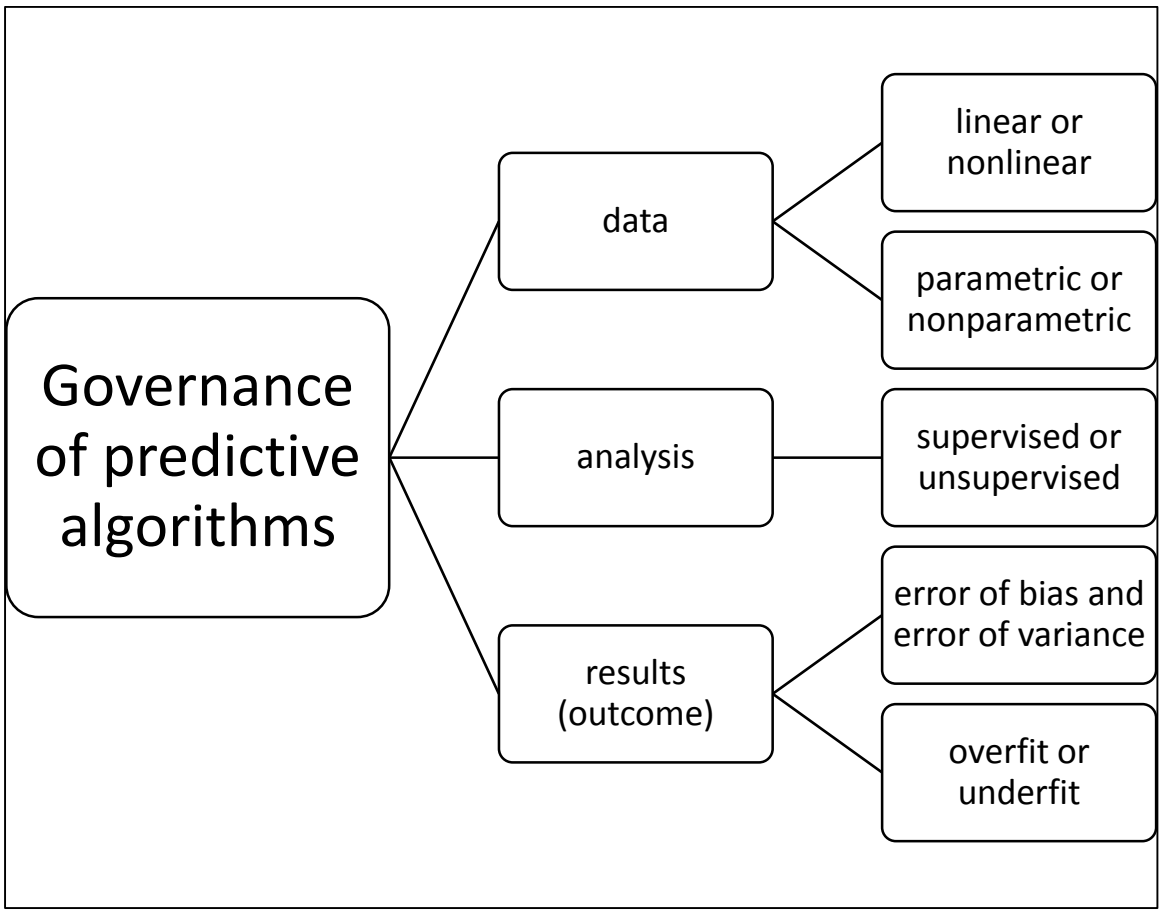


Figure 3-2: Layout of this research: Governance of predictive algorithms

Quantitative trading recalculates an expected or predicted share value which requires the implementation of a predictive algorithm for its estimations. This algorithm is chosen through its applicability to the scenario it is used for, and can be reused for the same situation at a later occasion, which is particularly useful for share traders who use repetitive analyses to monitor share values for changes (MathWorks, 2018a).

Furthermore, when considering applicability to a machine learning scenario, the theorem of “No Free Lunch” is fundamental: algorithms are appropriate for a specific issue, data set or scenario; it cannot be used freely to address other or even similar issues. This is especially true for predictive modelling, and even for the share trading industry where repetitive algorithms will be performed. Any changes or additions in the data will mean that the applicability of the algorithm implemented must be reassessed to ensure that the initial conclusions are still reasonable and true (EliteDataScience, 2017; Brownlee, 2017; Coffin & Saltzman, 1999).

It must be noted that the characteristics of predictive algorithms as discussed in this chapter, are attributes of the method of analysis (the algorithm); it does not describe the data analysed (Altman and Bland, 2009). However, based on the characteristics of the algorithm, it will be either appropriate to use it for the data intended for analysis, or not. This confirms how important it is to understand the available data set and to determine which statistical characteristics of the algorithm would be appropriate for this data set, in order to choose which algorithm would be appropriate for implementation.

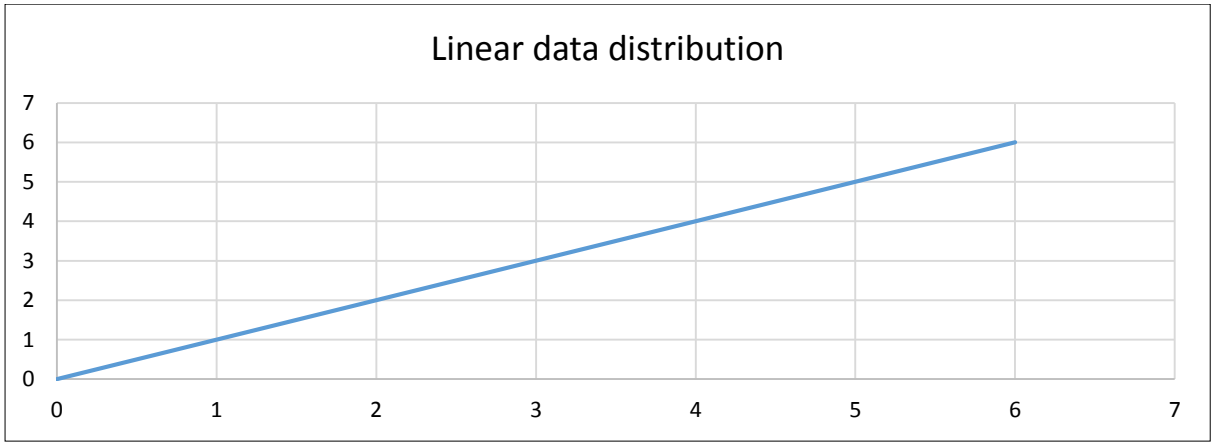
3.2.1 Linear and non-linear algorithms

Understanding the available data is important for choosing an appropriate algorithm and governing its use. This is particularly true when assessing whether a linear or nonlinear algorithm should be implemented; it is not a matter of choosing an algorithm, but rather of understanding the data under scrutiny (Witten *et al*., 2017; Hastie, Tibshirani & Friedman, 2017; Shalev-Shwartz & Ben-David, 2014).

A linear equation has the following function: $y = f(x) = a + bx$. Therefore each term in the model is a constant or the product of a dependant variable (y) and an independent variable (x). The result is linear, since all the terms of the equation are linear (Witten *et al*., 2017; Hastie *et al.*, 2017; Shalev-Shwartz & Ben-David, 2014; Murty, 1997). An example of a graph of a linear function is provided below in graph 3-1 (Colombia, 2018).

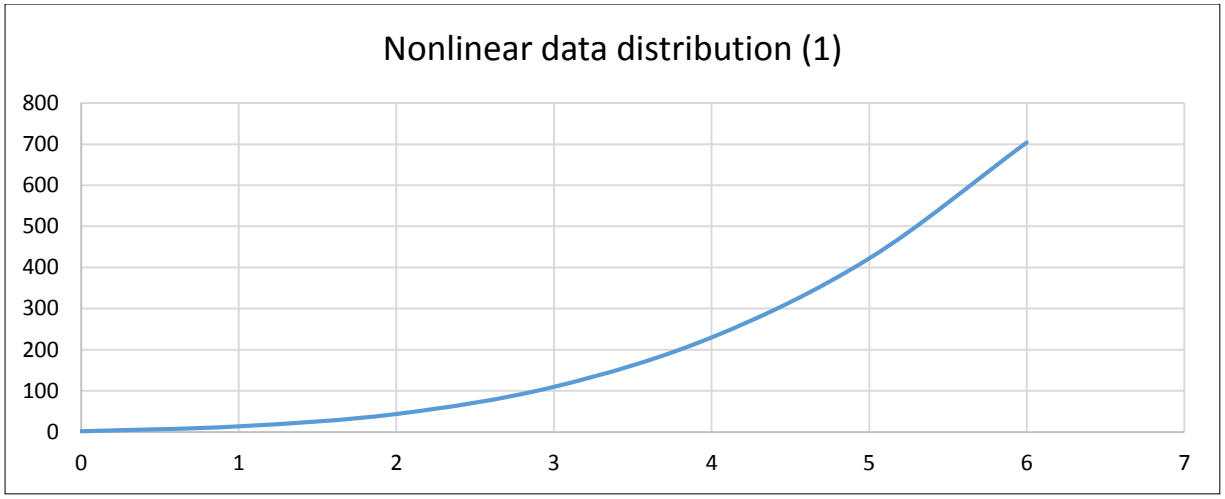
Since all the function terms are linear, a linear relationship is assumed between the input variables and a single output variable. The strength of this relationship is statistically tested using linear regression, which estimates the required coefficients required in order for the model to make predictions. A linear algorithm then uses the linear relationship found in the training data to predict output values for new data (EliteDataScience, 2017; Hastie *et al.*, 2017).

If the algorithm is combined with machine learning technology, this insight is obtained by using a training data set. The linear relationship found in the training data can then also be used for the evaluation data set and actual data sets, to find the final predicted results (EliteDataScience, 2017; Brownlee, 2016c). Assessing this relationship requires calculation of the mean (statistical average), the variance (which is a measure of the distribution of the data) and the covariance (which tests the degree of change between variables) to ascertain the relationship between variables (Oxford University Press, 2018; Brownlee, 2016c).

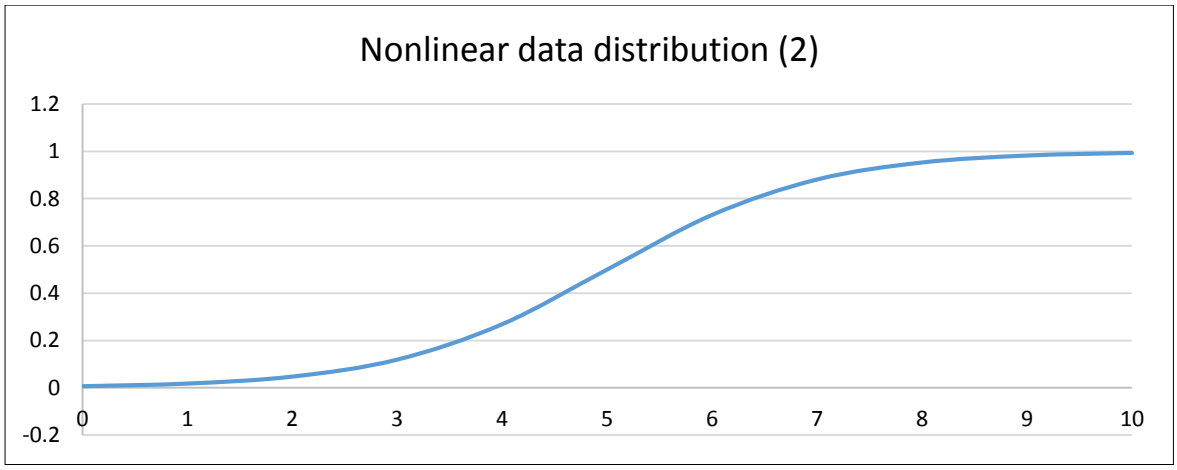


Graph 3-1: Example of linear function

A non-linear algorithm is an equation which is not linear. If the data set does not meet the requirements for a linear algorithm by including terms which are not constant or a single predictor variable and parameter, it is considered non-linear. It can take many forms, as shown in the examples of graphs of nonlinear functions which are provided in graph 3-2 and graph 3-3 below (Monterey Institute for Technology and Education, 2018):



Graph 3-2: Example of nonlinear function (1) (Monterey Institute for Technology and Education, 2018)



Graph 3-3: Example of nonlinear function (2) (Monterey Institute for Technology and Education, 2018)

If the data set is nonlinear, more powerful machine learning will be required to analyse its distribution and identify any trends or outliers (Brownlee, 2017).

Choosing between a linear or non-linear algorithm depends on the context of the problem. The choice does not depend on the attributes of the algorithm, but of the data set which is analysed. If the data is linear, or has a strong linear input variable, a linear algorithm would be more appropriate. If this is not the case, a nonlinear algorithm should be used (Witten *et al*., 2017; EliteDataScience, 2017; Shalev-Shwartz & Ben-David, 2014).

If there are nonlinear variable relationships, it will lead to poor prediction power of a linear algorithm. If non-linear variable relationships exist, transformation of the data to a linear format is required before a linear algorithm can be implement (Witten *et al*., 2017; Shalev-Shwartz and Ben-David, 2014; Russel and Norvig, 2010). The implementation of linear regression is limited to linear problems, where the relationship between variables is linear. A linear regression can be used for nonlinear problems, by technical statistical normalisation methods. This will avoid overfitting of the data (EliteDataScience, 2017). Overfit and underfit of a model is investigated in chapter 3.2.5 in this research.

A great advantage of using linear regression is that it is easier to understand and implement, making it a less time-consuming and less expensive algorithm (Russel & Norvig, 2010). However, because it is a simpler function, the disadvantage of using a linear algorithm rather than a nonlinear algorithm, is the strict limits it places on the machine learning's solution – therefore increasing the error due to bias by decreasing the error due to variance (Hastie *et al*., 2017; Fortmann-Roe, 2012). An investigation of these errors will be conducted in chapter 3.2.4.

Because of its machine learning capabilities, nonlinear algorithms are able to find more complex relationships between the input and output variables than a linear algorithm. It is also more adaptable, and even nonparametric. This means that the algorithm is able to find the number of parameters required for the data set and identify these parameters, and therefore the model (function) of the algorithm. However, this requires sufficient training data, which can be problematic to obtain. If more appropriate data is available for training the nonlinear algorithm, the algorithm's result will be more accurate (Kolanovic & Krishnamachari, 2017; Brownlee, 2016c; Brownlee, 2014).

Another consideration is that non-linear algorithms are able to handle data with high variance. This occurs when different data sets are available which are not identical. This will result in a difference in the prediction for each group (Brownlee, 2017; Fortmann-Roe, 2012). This prediction error, and the impact thereof is investigated in chapter 3.2.4.

Choosing between linear and nonlinear algorithms does not depend on the number of assumptions used (therefore the nature of the algorithm), but rather the nature of the data. Where there are data assumptions required, the decision will be between parametric or nonparametric algorithms.

3.2.2 Parametric and nonparametric algorithms

The number of assumptions used by an algorithm can simplify it significantly. However, if it is implemented with a component of machine learning, this simplification will limit the algorithm's ability to learn, and limit its obtained insights (Brownlee, 2016c; Russel & Norvig, 2010).

A parametric algorithm refers to an algorithm with a function (learning objective) which can be simplified to a known form. This means that the data used for the analysis can be summarised by a fixed number of predetermined parameters which are characteristics of the data population, rather than only of the training data. Therefore, even if data volumes increase, the algorithm will not require additional parameters (Brownlee, 2016c; Russel & Norvig, 2010).

In contrast, a nonparametric algorithm does not have predetermined parameters. This is particularly powerful when combining these algorithms with machine learning, since it is not constrained by predetermined parameter decisions, but rather allows more learning to take place and allows machine learning to identify its own parameters. It can also develop more insights as it learns, and add more parameters when required. The ability to add additional parameters also enables machine learning to change the functional form (the curve) of the algorithm in order to provide the best fit for the data it is analysing – which leads to better quality learning results and therefore better quality decisions (Brownlee, 2016c).

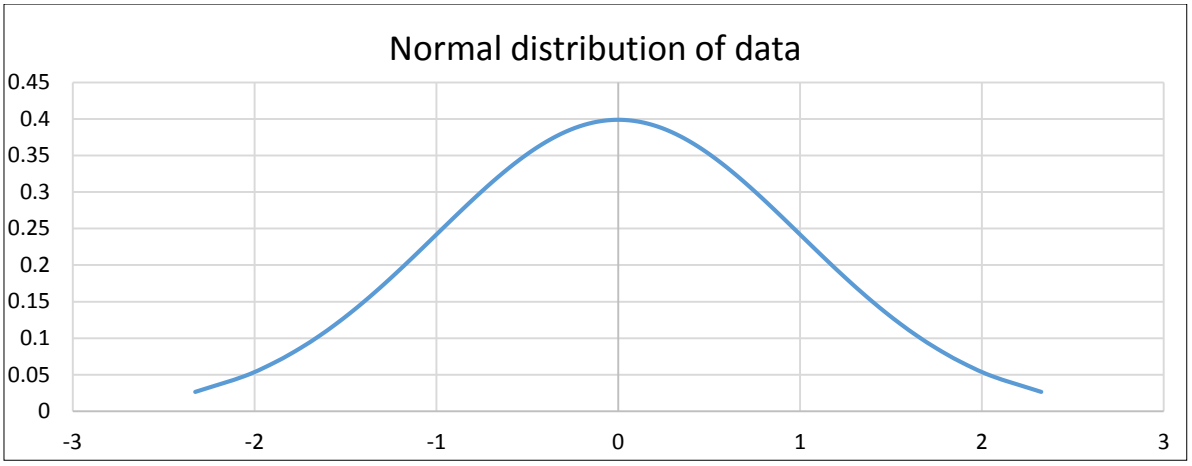
Furthermore, if different groups of data are compared, a parametric algorithm will have to assume that the different groups' data has a similar distribution. Comparisons are very important for share trading, since many strategies involve comparing different versions of companies' share data. Therefore, the implementation of algorithms for many of the share trading analyses will be problematic if parametric algorithms are used, unless the data structure is known to be similar enough (Altman & Bland, 2009). This can also be achieved by transforming both data sets to a normal distribution, to obtain a similar data structure for all (different) data sets.

Contrary to expectation, parametric algorithms can perform well for data which does not have a normal distribution if the data is sufficient and the test data volume is sufficient to allow it. However, if there is insufficient data available to meet the volume requirement of training data, a nonparametric algorithm will be better suited. Because it needs fewer assumptions, a nonparametric algorithm is more robust in nature and a valid solution to a wider array of data sets (Brownlee, 2016c; Altman & Bland, 2009).

There is another important factor about the analysed data which must be considered for the decision between a parametric or nonparametric algorithm: whether the mean or the median is the best central tendency of the data (Memorial University, 2018; Dzikiti & Girdler-Brown, 2017; Brownlee, 2016c). For this, there are three data points to consider concerning the data's central tendency:

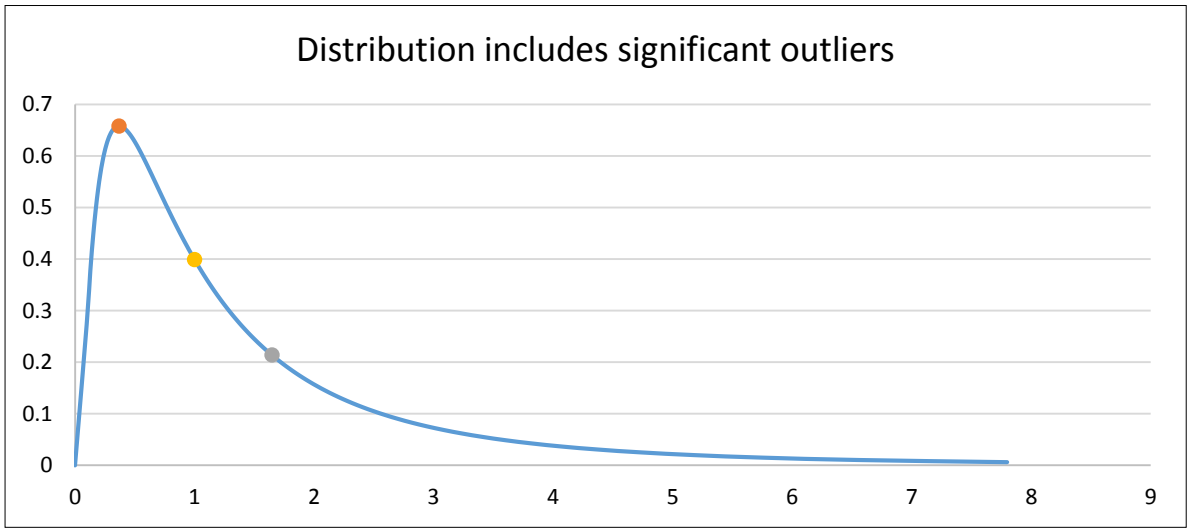
- the mode: the data item which occurs most;
- the mean: the average of the data items; and
- the median: 50% of data items are to this data point's left, and 50% of data items are to its right on the function's curve. Therfore it is the best central point of the data curve (Memorial University, 2018).

When the data is normally distributed as in graph 3-4, the mean, mode and median would all be the same (Memorial University, 2018). In graph 3-4, the mean, mode and median are all 0. Therefore, the mean is acceptable as the central tendency of the curve. In such a case, a parametric algorithm would be more suitable for analyses (Memorial University, 2018; Dzikiti & Girdler-Brown, 2017).



Graph 3-4: Central tendency for normal data distribution is its mean

In contrast, in a data set which includes significant outliers, these outliers will affect the mathematical mean (average) of the data. The result is a distribution such as shown in graph 3-5 below, which indicates from left to right the mode, median and mean. Without changing the average result for most of the data points, the median would be a better representation of the central tendency. In this case, a nonparametric algorithm will be advised for data with significant outliers which cannot be removed from the data set (Memorial University, 2018; Dzikiti & Girdler-Brown, 2017).



Graph 3-5: Central tendency for distribution which includes significant outliers is its median

When performing statistical analyses, most of the analyses will be expected to be parametric, rather than non-parametric. This is because non-parametric tests do not require data to have a specific distribution, contrary to the usual assumption of data following a particular (and often normal) distribution which would then use a parametric test. Nonparametric algorithms do not assume that the user does not know anything about the data set, but rather that the user knows that it does not follow a particular distribution (Dzikiti & Girdler-Brown, 2017; Brownlee, 2016c).

If the data items are values (such as a calculation or measurement), there will most likely be a wide spread of data values included in the data. However, if a scoring system was used, and the number of data options are limited, it will affect how complicated the analyses of the data will be. While this is not a prerequisite for analyses, Altman and Bland (2009) found that values are often analysed using a parametric algorithm, while limited values (such as scores), are rather analysed using nonparametric values. However, parametric algorithms allow for calculations such as estimates, confidence intervals and other more complex analyses, which might be preferable for some data analyses.

The trade-off of using more complicated nonparametric algorithms is that it requires more training data. Also, because more data must be analysed, nonparametric algorithms' process of learning is slower than that of parametric algorithms. Therefore, given the same computing power, it will take longer to derive meaningful conclusions, which is not ideal for share trading since this industry requires quick and timely conclusions and actions when market changes occur. For this reason, parametric algorithms appear to be favoured if it is appropriate for the data set it must analyse (Brownlee, 2016c).

Another reason why parametric algorithms are favoured when possible, is because the many assumptions simplifying the data makes it more probable that a parametric algorithm will detect a significant trend for data with a normal distribution; a non-parametric algorithm will be better suited and have better prediction power than a parametric algorithm for data with a different distribution (Kasparis, Andreou & Phillips, 2015; Brownlee, 2016c).

Since the choice between a parametric and nonparametric algorithm depends on the data distribution and attributes, the importance of understanding the available data and its distribution is reconfirmed. The element of machine learning must also be considered, since the choice between a parametric and nonparametric algorithm impacts the algorithm's ability to learn. Another decision when incorporating machine learning technology with an algorithm, is the level of user prescription. To assist with this decision, supervised and unsupervised algorithms will be investigated.

3.2.3 Supervised and unsupervised algorithms

Considering the nature of both linear or nonlinear algorithms and parametric or nonparametric algorithms, as investigated in 3.2.1 and 3.2.2, the nature of the data under analysis determines which algorithm would be most appropriate. However, when the implementation of an algorithm is done with a component of machine learning, the choice between supervised and unsupervised algorithms must also be considered. Therefore, when machine learning is also implemented, the level of machine learning as well as the expected outcome of the algorithm must be considered (Jain, 2017; Brownlee, 2016d; Shalev-Shwartz & Ben-David, 2014; Manning, Raghavan & Schutze, 2009).

For supervised algorithms, both the input variables and the output variable(s) are predetermined. Therefore, machine learning is used to map predetermined input variables to predict the predetermined output variable(s). The name, "supervised algorithm", is very descriptive as it indicates that the user provides guidance and instructions for the algorithm to follow, rather than using its own learning power and experience to draw conclusions (Jain, 2017; Manning, Raghavan & Schutze, 2009). Because of its lack of prescription, unsupervised algorithms are sometimes referred to as "learning from examples" rather than from prescribed instructions (Langley, 1996). Therefore, the user guides the algorithm to the extent and scope of learning it must achieve. Once the algorithm has achieved the predetermined, required objective, learning will cease since no further insight and objectives are required of it (Brownlee, 2016d).

For unsupervised algorithms, only input variables are determined by the user thereof and output variables are found through machine learning. This is particularly useful where the data distribution and output variables are unclear and can be identified by implementing an unsupervised algorithm to find the distribution and variables (Brownlee, 2016d).

Therefore, unsupervised algorithms more closely resemble true artificial intelligence: without human interference it allows for machine learning to find its own insight to address issues humans would usually disregard because it would not be cost-efficient or time-efficient to try to obtain a solution. It ignores human bias or preconceived ideas, and objectively analyses the data. However, the challenge is that, without a predetermined expected outcome, it is difficult to test the success of an unsupervised algorithm (Witten *et al.*, 2017; Brownlee, 2016d; Shalev-Shwartz & Ben-David, 2014; Essinger, 1990).

To further clarify what supervised algorithms can achieve, it will be further classified according to the characteristics of its output. Likewise, unsupervised algorithms can also be classified by what the user wants to achieve by the implementation of machine learning (Brownlee, 2016d):

Supervised algorithms		**Unsupervised algorithms**	
Classification	**Regression**	**Clustering**	**Association**
The expected output variable is a group or category.	The expected output variable is a calculated value.	The algorithm and machine learning aim to group the data to obtain a common inherent trait.	The algorithm and machine learning aim to obtain a data characteristic which describes a large portion of the data.

Table 3-1: Guidance on classification of supervised and unsupervised algorithms

In practice, most of the scenarios where algorithms can be implemented as a solution requires attributes of both supervised and unsupervised algorithms. This solution is considered as a semi-supervised algorithm. This is particularly important if there is a large volume of input data to analyse, but not all of the data items are labelled (Brownlee, 2016d; Manning *et al*., 2009). Labelled data refers to metadata encoded on digital documents, in a machine-readable format. This metadata provides additional fields of information about the data item, such as the creation date or format. Therefore, the possible variations can be considered as finite (Manning *et al.*, 2009).

Labelled data makes supervised learning easier, since it would assist the machine learning in following the user's instructions. Unsupervised learning will investigate the data and assign its own attributes to data items (Brownlee, 2016d).

Labelling data can be expensive and could require access by domain specialists. In contrast, unlabelled data is cheap and easily obtainable. If the data under scrutiny is large and not all its data is labelled, it will be better to implement a semi-supervised algorithm which will be able to find the inherent structure of the data rather than a supervised algorithm for which it must be prescribed. In such a case, unlabelled data can then be used as training data for an unsupervised algorithm, which can then obtain the insight required to implement a supervised algorithm, to obtain the output variables required (Brownlee, 2016d; Manning *et al*., 2009).

Data scientists who want to obtain well-rounded insight and solutions, often implement a combination of supervised and unsupervised algorithms. This will also be appropriate for the share trading industry, where alternative insight provides more opportunity of identifying opportunities and threats. However, the nature of the investigated data structure and its volume will determine if this is possible (Witten *et al*., 2017; Brownlee, 2016d; Russel & Norvig, 2010; Manning *et al*., 2009).

3.2.4 Bias and variance in algorithms

Since most models require assumptions, predictions remain estimates and can differ from true results. This is especially true for the share trading industry, where many non-quantifiable factors influence share prices. Because of this, users of predictive algorithms must consider the following differences between true results and the algorithm's predictions (Brownlee, 2016a; Gutierrez, 2014):

- error due to bias,
- error due to variance, and
- irreducible error.

These errors are not "mistakes"; it is inherent in the nature of building statistical models. For example, simplifying assumptions in order to simplify an algorithm has many advantages, such as easier learning, but will introduce an error of bias (Brownlee, 2016a). However, there are measures to reduce these errors, and the aim when building a model using algorithms is to obtain a low error of bias, a low error of variance and also an accurate prediction (Brownlee, 2016a; Gutierrez, 2014).

When building an algorithm, it is important to understand and consider those statistical errors which can be avoided and must be minimized, and also to understand that there are inherent errors which cannot be avoided, regardless of which algorithm is implemented (Brownlee, 2016a; Gutierrez, 2014). The irreducible error is caused by how the problem is framed or unknown variables which impacted the mapping of input to output variables. This is inherent to the process of building a model for problem solving, and cannot be avoided (Brownlee, 2016a).

However, both the error of bias and the error of variance can be managed. Error due to bias is considered the difference between the expected, predicted value according to the implemented model, and the actual value of the item the user aims to find. This might appear problematic when only one data set is available and used for the prediction and cannot be tested on another data set (Fortmann-Roe, 2012).

Considering the same conditions as above: if new data sets are available, and different models can be used to predict a specific value through these modes, the error due to variance is the variation between the predicted values provided by different data sets and models for the same prediction (Fortmann-Roe, 2012). It is the difference found in the prediction, if a different training set was used (Brownlee, 2016a).

Therefore, bias shows how far the model's prediction is from being accurate to its true value; variance shows how much the predictions vary between different models. This can be explained further by the following visual representation (Fortmann-Roe, 2012):

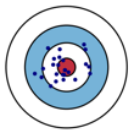
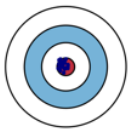
Visual representation of errors due to bias and variance:	
	Bias is high; variation is low
	Bias is high; variation is high
	Bias is low; variation is high
	Bias is low; variation is low

Table 3-2: Graphs showing errors of bias and variance (Fortmann-Roe, 2012)

Considering the table above, there is an important link between the error of bias, and choosing between a parametric algorithm which has predetermined parameters, and a nonparametric algorithm, where its parameters are unknown and must be determined by machine learning (Brownlee, 2016a):

Level of bias:	Low error of bias	High error of bias
Type of algorithm	Nonparametric	Parametric
Nature	Suggests less assumptions about the form of the target function - slower learning - more accurate predictions	Suggests more assumptions about the distribution of data - faster learning - less accurate predictions on complex problems

Table 3-3: Relationship between parametric algorithms, nonparametric algorithms and the error of bias

Furthermore, nonparametric algorithms are often more flexible in nature. Because of this, the error due to variance is most likely higher when using a nonparametric algorithm, rather than a parametric algorithm which has more prescribed parameters (Brownlee, 2016c).

When building a model using algorithms, it is important to consider both the error of bias and error of variance, since these errors are linked, and a trade-off between it is often required. For the error of bias to be low, the algorithm model needs to be flexible enough to fit all data types, and complex enough to include all parameters required to solve the problem. However, these attributes need to be balanced to manage the reducible errors (Gutierrez, 2014).

To avoid the error of bias, more training data sets can be introduced to obtain a more accurate answer (closer to the true value). However, using more training data sets will create more prediction models, which in turn increase the variance in the models' results by obtaining more predicted values through these additional models. This shows that whenever a solution is introduced to address one of these reducible errors, it causes an increase in the other error. Therefore, a trade-off between the errors of bias and variance is required, and the expense of decreasing one error to the increased in the other error must be considered (Fortmann-Roe, 2012).

Likewise, the complexity of the algorithm must be considered: if the implemented algorithm is too rigid (inflexible) because of using too few parameters it will cause inaccuracy in its prediction; therefore increasing the error due to bias. If the model is too flexible, it will allow for very different predicted answers, causing a high variance. Therefore a trade-off in flexibility is required to obtain as accurate a prediction as possible (Brownlee, 2016a).

There is also an important implication for the bias-variance trade-off when using a supervised algorithm: the error trade-off can be managed by introducing elements to the algorithm which can be adjusted to lower the error due to bias, or lower the error due to variance, depending on what is required to balance these errors and obtain a desired outcome. This adjustment can also be automated to require less input from the data scientist (Gutierrez, 2014).

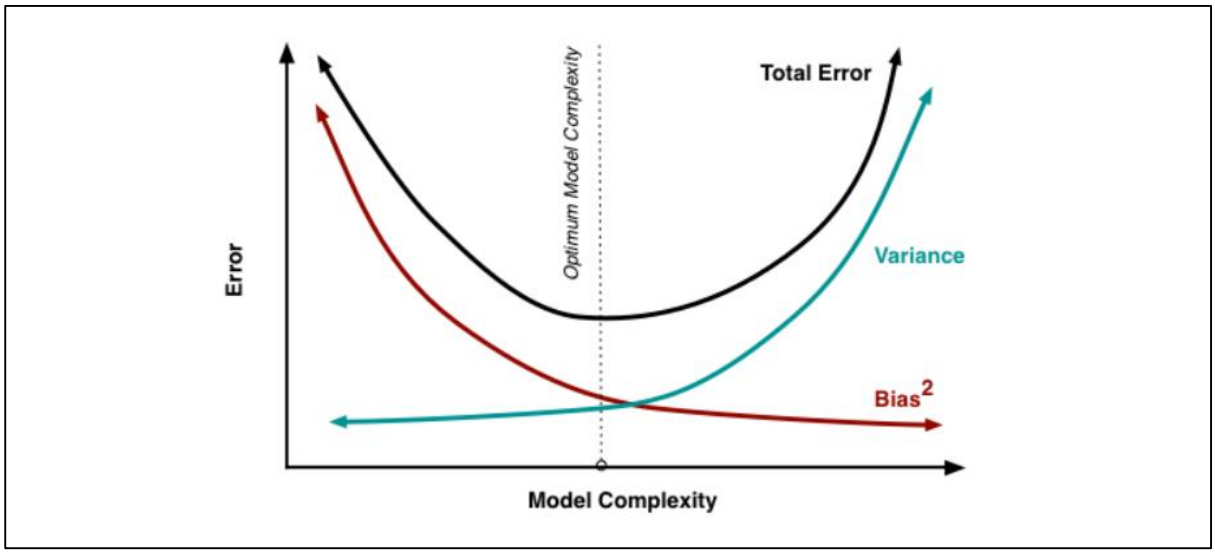
There is a method available to try to manage the trade-off risk. It is referred to as “bagging”: essentially, it entails building multiple derivative data sets from the single data set which is analysed, but repeatedly resampled to replace data items. This forces multiple data sets which can be used for building new models to use for the analysis of the data, which can be combined to reconcile the results, and obtaining an “average” result to use as a predicted value (Fortmann-Roe, 2012).

In choosing which algorithm to use, it is important to be able to measure the accuracy of the models available. When building a model, it is not possible to calculate the actual errors of bias and variance, since the true value of the prediction is usually unknown. However, it is important to understand and measure it to pursue optimal prediction power for the model (Brownlee, 2016a; Gutierrez, 2014). Two popular ways to calculate this, is training error and test error, which tests data both on a training data level and once the algorithm is implemented, and includes technical validation tests, additional completeness validation checks and accuracy tests (Gutierrez, 2014). This will determine if the model fit is appropriate for the underlying data and the intended outcome.

3.2.5 Overfit and underfit in algorithms

When considering the errors of bias and variance, the concepts of overfitting and underfitting are directly related. As discussed above, model complexity dictates the errors of bias and variance, with the number of parameters directly affecting complexity and the error of variance, and therefore also inversely affecting the error of bias (Brownlee, 2016b; Fortmann-Roe, 2012).

However, only considering the errors of bias and variance is not sufficient: the total error must also be taken into account. While chapter 3.2.4 has shown that there is a trade-off between the error of bias and the error of variance, there is an optimal point in the complexity of the model where the total error will be lowest. The following graph shows how the model complexity impacts the errors of bias and variance, and how these errors correlate with the total error (Fortmann-Roe, 2012):



Graph 3-6: The errors of bias and variance, and the total error (Fortmann-Roe, 2012)

One of the fundamental aspects of creating an accurate model, is the model fit. It refers to the ability of the model to generalise the insight learnt from the training data to the data population tested, since the training data is only a sample or example of the population. Therefore it has a direct impact on the prediction performance of a model. The training data is not necessarily complete and can include irrelevant data items or outliers (Brownlee, 2016b; Gutierrez, 2014; Fortmann-Roe, 2012).

One of the most important considerations when assessing the model's accuracy (and therefore also its total error) is to determine whether the model is overfitting or underfitting its training data, by calculating a prediction error between the training data and the evaluation data (Hastie *et al.*, 2017; Fortmann-Roe, 2012).

Underfitting occurs when the model struggles to find the relationship between the input variables, and the required output variables in the training data. Because it cannot learn the required insight on the training data, it is also not able to successfully generalise its insight to the evaluation data set (Brownlee, 2016b). This could be because the model is not flexible enough, and more input variables/parameters are required (Amazon Web Services, 2018; Gutierrez, 2014; Fortmann-Roe, 2012). Because underfitting means that the model could not even address the problem for the training data, it is usually considered irrelevant and is often disregarded because a different model would be more appropriate to use (Brownlee, 2016b).

On the other hand, overfitting happens when the model manages to link the input and output variables when assessing the training data, but cannot repeat its success when assessing evaluation data. This occurs when machine learning was not successful, and is trying to replicate data links, rather than generalising and applying its learning to another data set. In lay man's terms: machine learning took into account all irrelevant data items or random fluctuations. However, these items are not present in the evaluation data, which affected its ability to generalise, and to predict accurately (Brownlee, 2016b). This issue can be addressed by reducing parameters to make the model less flexible, or by introducing more training data (Amazon Web Services, 2018; Hastie *et al.*, 2017).

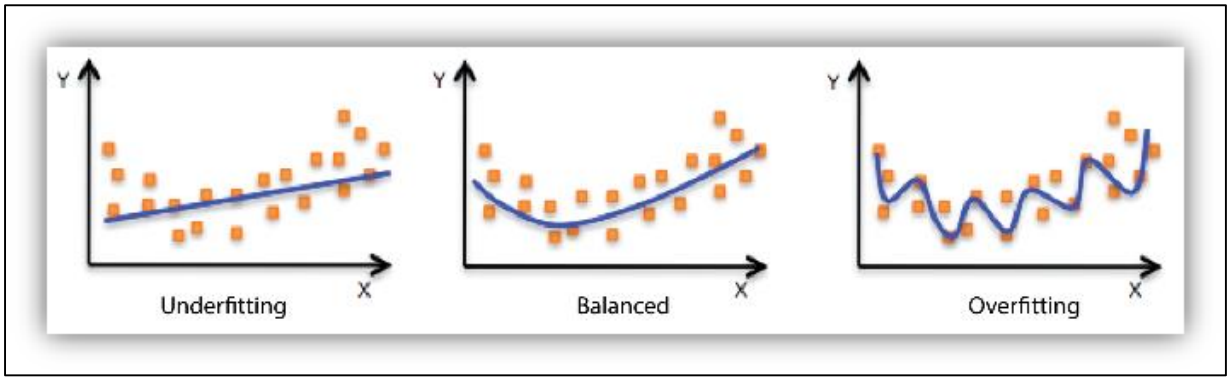
Because nonparametric and nonlinear models are more flexible, overfitting is more likely for these algorithms. For this reason, parameters are included to limit the detail many nonparametric algorithms are able to learn (Hastie *et al.*, 2017; Brownlee, 2016c; Gutierrez, 2014).

Because the errors of bias and variance, and the model fit are affected by the complexity of the model through the number of parameters (input variables) included in the model, it is possible to map these issues. If there is a small variance but high bias, the model was unable to obtain insight from the training data, and therefore also cannot use this to make predictions for another data set. This is an underfit model. In contrast, the model has small bias but high variance, it has customised the model to such an extent that it makes very accurate predictions for the training data, but cannot replicate this success for the evaluation data, or other data sets. This results in an overfit model (Hastie *et al.*, 2017; Gutierrez, 2014):

	Small variance	**High variance**
Small bias	*n/a*	overfit model
High bias	underfit model	*n/a*

Table 3-4: Linking the errors of bias and variance, with model fit

The concepts of overfitting and underfitting are illustrated on the following graphs, which shows the ability of the function to follow the data distribution. In an underfit model, the machine learning insight is insufficient and results in an overly simple model which cannot predict accurately. For an overfit model, it is so customised to the training data, that it is overly specific for that data set, and therefore unable to make accurate predictions for a different data set (Amazon Web Services, 2018):



Graph 3-7: Graphs showing model fit (Amazon Web Services, 2018)

Ensuring a balanced model fit is especially important for a supervised algorithm, since the mapping of input variables to output variables is controlled by the data scientist rather than letting machine learning determine the most appropriate fit. A further contributor to obtaining a balanced algorithm, is to allow only enough time for the algorithm to train itself by using the training data. If too little time is allowed, it will not learn sufficiently to predict values, or to generalise its insight to new data. If too much time is allowed, it will complicate its model with too many parameters, which will lead to an overfit model which also cannot generalise insight to other data sets (Brownlee, 2016b).

There are two techniques available to limit overfitting and obtain a balanced algorithm, (Brownlee, 2016b; Fortmann-Roe, 2012):

- A resampling technique can be used, using subsets of the data set. Therefore the full training data is divided in different subsets, to obtain different training data sets which are not identical.
- A validation dataset should be held back. Therefore, not all training data is introduced at once, which provides the data scientist with time to evaluate the algorithm's ability to perform on unseen (training) data.

Therefore consideration of the model fit by avoiding overfitting or underfitting, as well as the errors of bias and variance, will assist in the aim of governance of the algorithm technology by achieving validity, accuracy and completeness in the analysis of the data, as well as the result achieved by the technology.

3.3 Qualitative / Nonpredictive algorithms

Effectively governing qualitative algorithms has the same objective as quantitative algorithms: ensuring a valid, accurate and complete result (conclusion). The difference is that the user does not require an estimation or prediction from nonpredictive algorithms, but rather achieving another outcome conclusion. This is an important component of share trading research and complements quantitative research by harnessing alternative sources of profit or loss indicators (Marx *et al.*, 2013).

Because the required outcomes of these algorithms are different, the inherent nature and workings thereof will also differ from the predictive algorithms researched in chapter 3.2. This research will focus on two of the most prevalent examples of nonpredictive algorithms to investigate its nature, workings and purpose in order to include it in the applicability model: (Ain *et al.*, 2017; Hu *et al.*, 2015; Medhat, Hassan and Korashy, 2014; Verma & Verma; Yu *et al.,* 2013*;* McCall, 2005; Coffin & Saltzman, 1999).

- **optimisation algorithms**
  These algorithms use a technique which is able to select the best element from the items under scrutiny (MathWorks, 2018b; Verma & Verma, 2012; McCall, 2005).

- **search heuristics**
  These algorithms are able to discover or learn something for themselves (Oxford Dictionary, 2018).

Therefore, the layout of this investigation of nonpredictive algorithms are as follows:

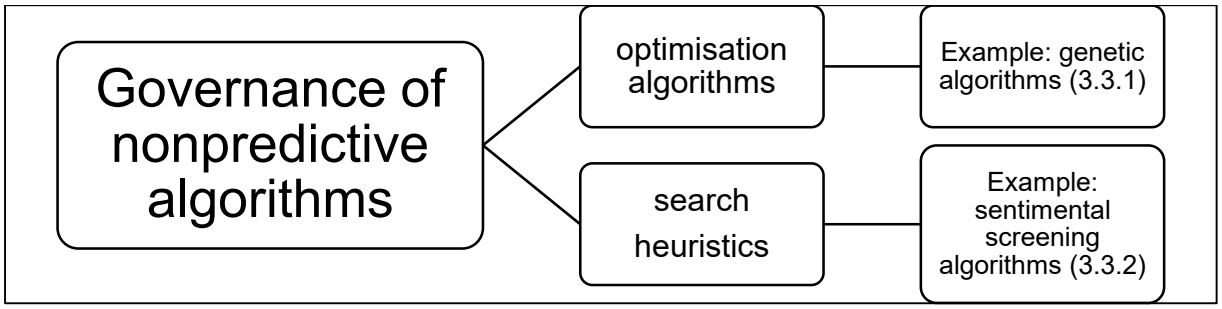


Figure 3-3: Layout of this research: Governance of nonpredictive algorithms

3.3.1 Genetic algorithms

Genetic algorithms are an optimisation technique, which tries to find the values of input to create the best output or results (Verma & Verma, 2012; McCall, 2005). Rather than only analysing the data to find one solution, it tests many solutions and finds the optimal solution for the data set, according to the prescribed characteristics of what an optimal solution entails (MathWorks, 2018b; Verma & Verma, 2012).

The genetic algorithm's process resembles the process of natural selection, which is a biological process based on Darwin's theory of evolution. In this process, only those organisms who are best suited to their environment survives and transfers their superior genes to their offspring. This is often referred to as survival of the fittest (Oxford Dictionary, 2018).

A genetic algorithm also follows this process to solve problems. In this process, the population of data items is continuously updated and improved, by identifying superior data items according to its requirements, and pairing those superior items at random to produce the next generation of data items. This is reperformed until this evolution shows the best remaining solution (MathWorks, 2018b; McCall, 2005).

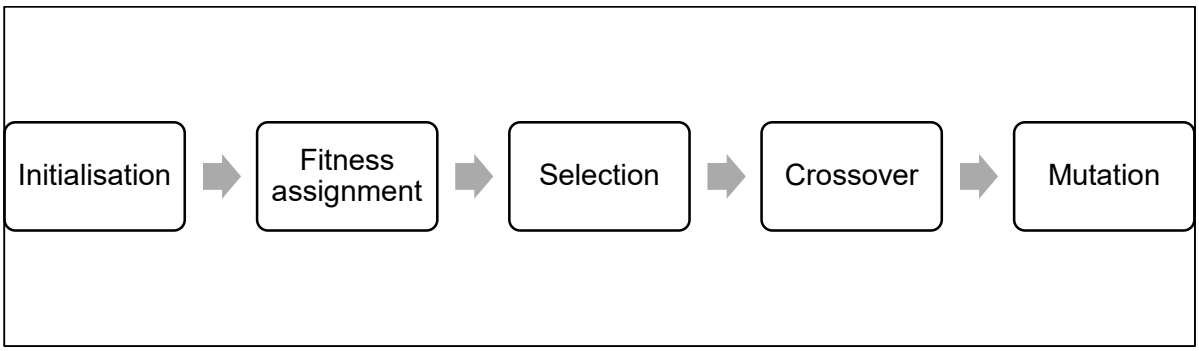


Figure 3-4: Phases of a genetic algorithm

Referring to figure 3-4 above (Mallawaarachi, 2017; Verma & Verma, 2012; McCall, 2005), the process and phases of a genetic algorithm can be explained as follows:

1. **Initialisation**:
   First of all, the initial population is defined. At the start of the process the population includes all possible solutions available for the data under scrutiny. Each of these possible solutions has certain variables which serve as parameters, which can be joined to form an optimal solution (Mallawaarachchi, 2017).

2. **Fitness assignment**:

   Next, “good fitness” is defined. This is a characteristic of the solution which is in line with the predetermined outcome of what the user wants to achieve. This definition of fitness is what the parameters identified in step 1 will be tested against to assess its fitness according to the algorithm (Mallawaarachchi, 2017; McCall, 2005).

3. **Selection**:

   Based on the prescribed fitness characteristic(s) defined in step 2, data items with good fitness are selected from the population to create the next generation of solutions, acting as “parents” of the evolution. The selection rules are applied to the parents when they are chosen: these are the data items which will be combined to form the next generation of solutions. The better the fitness of the parents, the better fitness the next generation will have, therefore strengthening the population towards better fitness (Mallawaarachchi, 2017; McCall, 2005).

4. **Crossover**:

   The crossover rules is a critical point in a genetic algorithm, and occurs when the parents are matched (McCall, 2005). The crossover rules are applied to choose the characteristic from the parents to form a recombination – the next generation. In contrast to classic algorithms which are based on instruction, this crossover in a genetic algorithm occurs at a random point (MathWorks, 2018).

5. **Mutation**:

   Finally, the new generation is now part of the population, replacing items with poor fitness. Therefore, the number of items in the population remain constant as the new population is generated. This is because the least fit individuals will die when the new generation is created. When the population converged and the new generation no longer differ significantly from the previous generation, the process will cease and produce a solution to the investigated problem (Mallawaarachchi, 2017).

This process shows how genetic algorithms are parallel in nature – rather than only considering data in serial order like most other algorithms, genetic algorithms are able to assess different directions of the solution space at the same time (through its multiple next generations). If a direction does not provide an optimal or useful solution, it is disregarded and the next direction is assessed. This makes genetic algorithms more powerful, by making its chances of identifying the optimal solution much greater (Verma & Verma, 2012).

A genetic algorithm follows this process to solve problems which contain both constrained and unconstrained variables, the latter of which have restrictions on its range of possible values (MathWorks, 2018b; McCall, 2005). It is also fairly easy to implement, since it is a very modular algorithm (McCall, 2005).

Application scenarios where genetic algorithms have been proven as efficient and accurate, are: data mining, classification through inductive learning, and optimisation of data preparation and data transformation. These are all applications which are very useful for the share trading industry. For example, it can be used as an optimisation technique to maximise business profits through data mining; genetic algorithms are able to ascertain what the optimal values of variables are and find the optimal association rules between those variables (MathWorks, 2018b; Verma & Verma, 2012).

There are also applications to use genetic algorithms in creating optimal technical trading rules by identifying parameters in the data and testing all parameter combinations to find which will be optimal for profit. While this is very useful to the share trading industry, it requires a large amount of time and memory to run the algorithm (Verma & Verma, 2012).

3.3.2 Sentimental analysis algorithms

There are more factors than share value that provides data to share traders. Another contributing factor is the impact of human sentiment. Consumer attitude, emotions and opinions have an impact on the selling power of the shares, which in turn impacts the share value. Monitoring news articles can also encourage or discourage investors, and affect share values. Monitoring development of these emotions can indicate opportunities for profits, or potential losses. The sheer volume of publications and social media makes it impossible for a human trader to scan all the available sources. However, it is possible through a sentimental analysis algorithm (Ain, Ali, Riaz, Noureen, Kamran, Hayat & Rehman, 2017; Yu, Wu, Change & Chu, 2013).

The analysis of sentiment harnesses the analysing power of many techniques to perform this opinion mining. It uses language processing, statistical analyses and the text screening on social media to categorise the sentiment into a positive, negative or neutral category. This is usually coded as a binary conclusion to indicate opportunities or threats (Gage, 2018). Therefore sentimental analysis algorithms can be defined as opinion mining, which finds and extracts language, analyse it to find if it contains any opinions, and then classifies it accordingly (Ain *et al.*, 2017; Medhat, Hassan & Korashy, 2014; Yu *et al.,* 2013).

Sentimental analysis algorithms apply different techniques in order to screen and identify the required attributes. According to the research by Gage (2018) and Ain *et al.*, (2017), this includes the following techniques:

- **social sentiment algorithm:** This analyses news articles, social media updates and other online publications for indication of a reaction or other sentiment of the public towards a company. It can also categorise the emotion as a potential indicator of a trading opportunity or disregard it, according to its conclusion of the emotion.

- **benchmarking sentiment algorithms:** This is a language processor used to monitor social media platforms and other texts, to gain insight about how the brand is viewed. It analyses all text posts, assigns emotion to each and use it to build an overall view of the sentiment towards a company or brand.
- **deep learning for sentiment analysis:** This type of algorithm is implemented with machine learning to classify the emotion of the text, and can also be able to predict what the sentiment or emotion will be. This is extremely useful to pre-empt a share trader of opportunities in expected changes in share values.

Even though the algorithms can be combined with machine learning to predict the expected sentiment, it is at its core a search algorithm. While classifying emotion for an article or social update which contains positive words or phrases is simple, these algorithms are also able to ascertain the intensity of the emotion (Medhat *et al.*, 2014; Yu *et al.*, 2013). This is very useful to traders, to base their review of results on.

One of the biggest challenges of sentiment analysis, is that people do not always write what they mean. While more formal sources such as newspapers and research papers write in a more straightforward manner, social media often includes sarcastic or ironic remarks which could lead to an incorrect categorisation by the algorithm (Gates, 2018).

Combining sentimental analysis with machine learning, allows deep learning to find the optimal solution.

3.4 Conclusion

Achieving the research objective of governance of share trading algorithms requires an understanding of the data, its analysis and the type of result the user requires from its analysis. This chapter has shown how the nature and purpose of the algorithm dictates which algorithm is appropriate for implementation, based on those characteristics. These characteristics will now be used to design an applicability model, mapping the circumstances with the characteristics of the algorithms, to show which algorithm would be most appropriate for implementation.

Chapter 4: A Model of the Applicability of Share Trading Algorithms

4.1 Introduction

In chapter 3 the nature and statistical characteristics of algorithms were investigated to understand what type of data it is appropriate for, and what outcomes can be achieved by implementing it. This was done based on five statistical characteristics of predictive algorithms, as well as two examples of nonpredictive algorithms.

In this chapter, the findings of chapter 3 will be applied to design a model of applicability. This model will use the nature and scope of the available data, as well as the required analysis and the intention with the algorithm results, as indicators and map this to the algorithm characteristics to identify which algorithm is appropriate for implementation in order to govern the algorithm technology.

The users of this model include:

- Users of the algorithm, including those writing it, can determine which algorithm is most appropriate for implementation, and ensure effective governance by doing so.

- Users in share trading who have already implemented algorithms can evaluate whether the algorithm in use is the most appropriate, or if a different algorithm should be written to be more appropriate for their data sets and intended result.

The model indicators are grouped according to the life cycle of an algorithm as described in chapter 3.1:

- nature of the available input data
- requirements for algorithm analysis
- reguirements for algorithm results

As per the scope limitation, the governance of data is not included in the scope of this research. Therefore this model does not address data governance; it shows attributes of data which indicates which of the algorithm characteristics are applicable for the intended analysis and outcome of the algorithm.

What is evident from the model and the findings of chapter 3, is that the following algorithm characteristics are determined by the nature of the investigated data: linear or unlinear algorithms, parametric or nonparametric algorithms and error of bias and error of variance.

The level of machine learning in the data analysis will determine of the algorithm is supervised or unsupervised (or semi-supervised, using a combination of attributes from supervised and unsupervised algorithms). The characteristics of the data and level of machine learning will then impact the error of bias, error of variance and possibly the overfit or underfit of the algorithm function.

The models in chapter 4.2, chapter 4.3 and chapter 4.4 shows the mapping of each of the indicators of applicability, to the appropriate characteristic of predictive algorithms, in order to address the research objective of designing a model for effective governance of rule-based (algorithmic) share trading by identifying the appropriate algorithm.

Where the data or outcome indicator applies to an algorithm characteristic, it is indicated by an “x”. To indicate the impact on the error of bias and the error of variance (which is inversely correlated, and usually managed by a trade-off), it will be indicated by a “↑” where the data indicator causes an increase of the error, or “↓” where it causes a decrease in the error.

4.2 Applicability model of predictive algorithms according to available data

	Linear	**Nonlinear**	**Parametric**	**Nonparametric**	**Supervised**	**Unsupervised**	**Error of bias**	**Error of variance**	**Overfit**	**Underfit**
Reference to investigation of nature and purpose of characteristic in chapter 3	*3.2.1*		*3.2.2*		*3.2.3*		*3.2.4*		*3.2.5*	
Input and output variables predetermined	x				x					
Only input values pre-determined						x				
Machine learning determines output variables		x				x				

	Linear	Nonlinear	Parametric	Nonparametric	Supervised	Unsupervised	Error of bias	Error of variance	Overfit	Underfit
Each term is a constant / data is linear.	x									
Linear relationship between input and output variables	x									
Single output variable	x									
Data with high variance		x					↓	↑	x	
Complex relationship between input and output variables		x		x		x				
Data parameters unknown or determined by machine learning		x		x		x				

	Linear	Nonlinear	Parametric	Nonparametric	Supervised	Unsupervised	Error of bias	Error of variance	Overfit	Underfit
Limited training data available	x		x				↑	↓		x
Sufficient training data available		x					↓	↑	x	
Fixed number of parameters			x							
Known distribution of data, usually normal			x		x					
Data distribution unclear/unknown or determined by machine learning		x		x		x				
Data set can be segregated into groups with non-identical distribution				x						

	Linear	Nonlinear	Parametric	Nonparametric	Supervised	Unsupervised	Error of bias	Error of variance	Overfit	Underfit
Normalisation of data possible	x		x						x	
Mean is central tendency of data curve			x							
Median is central tendency of data curve / data has significant outliers				x						
Unlimited spread of data items / data items are values			x							
Spread of data items limited (for whole numbers or scores)				x						
All data items are labelled					x					

	Linear	Non-linear	Parametric	Nonparametric	Supervised	Unsupervised	Error of bias	Error of variance	Overfit	Underfit
Machine learning assigns data labels						x				
Data assumptions simplified for algorithm			x				↑	↓		x
Fewer assumptions made about data				x			↓	↑	x	
More assumptions made about data			x				↑	↓		x
Data contains irrelevant or random data items										x

4.3 Applicability model of predictive algorithms according to analysis requirements:

	Linear	Nonlinear	Parametric	Nonparametric	Supervised	Unsupervised	Error of bias	Error of variance	Overfit	Underfit
Easy to understand and implement	x									
Less time-consuming; less expensive	x		x							
Machine learning not required or limited	x						↑	↓		x
Machine learning capabilities required		x		x			↓	↑	x	
More adaptable		x		x			↓	↑	x	
Sufficient training time		x		x						
Linear regression required	x									
Complex calculations and analyses			x							
Avoid human bias		x		x		x				

4.4 Applicability model of predictive algorithms according to required results:

	Linear	Nonlinear	Parametric	Nonparametric	Supervised	Unsupervised	Error of bias	Error of variance	Overfit	Underfit
Accurate predictions required							↓			
Less accurate predictions for complex problems							↑	↓		x

4.5 Applicability model of nonpredictive algorithms:

	Genetic algorithms	**Sentimental screening algorithms**
Reference to investigation of nature and purpose of characteristic in chapter 3	*3.3.1*	*3.3.2*
Easy to implement	x	
Parameters identified by machine learning	x	
Search result required	x	x
Optimised result	x	
Search of words, tone and emotion		x
Able to make predictions through machine learning	x	x
Accurate predictions required	x	

4.6 Conclusion

This research has added research value in this chapter by using its investigation of chapter 3 to design a model which shows which algorithm will be appropriate for implementation to ensure effective governance. This is based on the users' circumstances: the nature and scope of the data available, the requirements or limitations of the analysis as well as what the algorithms must achieve.

Chapter 5: Conclusion

With the advances in technology the share trading industry evolutionised. There has been a significant increase in the amount of data available for share trading research, and a manual share trading system can no longer provide the valid, accurate and complete analysis and conclusion required for successful share trading (Bloomberg, 2017). A solution for this is the implementation of share trading algorithms which can assist or automate the analysis of data, and derive trading conclusions (Deloitte, 2017).

Not all algorithms are appropriate for implementation and achieving effective governance. The data, analysis and required result determines which algorithm would be appropriate to obtain a valid, accurate and complete analysis and algorithm result (EliteDataScience, 2017; Brownlee, 2017). This lead to the research objective to design a model for effective governance of algorithmic share trading by identifying the appropriate algorithm for the available data, necessary analysis and required results.

To provide context for the research objective, the literature review investigated the nature and scope of governance, specifically referring to IT governance. It also investigated the evolution of share trading in order to understand the impact of the advance in technology on this field, and highlighting the inherent issues and challenges of share trading. The research then investigated the nature and workings of algorithm technology, and how it addresses the challenges in the share trading industry. The technologies of big data and machine learning were also included, as it enables the implementation of algorithms in the share trading industry.

There are two governance outcomes prescribed by the King IV report which are pertinent to the implementation of share trading algorithms: performance and effective control (IODSA, 2016). Performance requires meaningful, value-adding decisions based on the algorithm results. Effective control requires of the technology to achieve valid, accurate and complete results (Von Wielligh and Prinsloo, 2014), which can only be the case if the algorithm can achieve these control objectives for the analysis and create conclusions and outcomes which also adhere to these control objectives.

Therefore the chosen implemented algorithm must be appropriate for the data qualities, the intended analysis as well as the required outcome of the algorithm.

In order to design a model for the governance of algorithms through its applicability, a thorough understanding of the data requirements and the intended outcome of the algorithm is required. Two categories of share trading were identified accordingly: quantitative, predictive algorithms, and qualitative, nonpredictive algorithms.

The following five characteristics of predictive algorithms were identified and investigated to determine its applicability based on the nature, requirements and intention of the algorithm (Coffin & Saltzman, 1999; EliteDataScience, 2017; Kolanovic & Krishnamachari, 2017; Microsoft Azure, 2017):

- linear or nonlinear,
- parametric or nonparametric,
- supervised or unsupervised,
- error of bias and error of variance, and
- overfit and underfit.

The investigation of these characteristics identified indicators which assist in identifying if the algorithm would be appropriate for implementation. Based on these indicators, an algorithm would either be appropriate for implementation, or lead to inaccurate or irrelevant results (EliteDataScience, 2017; Brownlee, 2017).

Using the investigated algorithm characteristics, as well as the indicators which shows which algorithm would be appropriate for implementation, a model of algorithm applicability was designed. It provides guidance in assessing which algorithm is appropriate for implementation, based on the nature of the available data as well as the intended analysis and outcome of the algorithm.

Furthermore, two examples of nonpredictive algorithm types were investigated to show how the nature and intention of these algorithms determine its applicability:

- optimisation algorithm: genetic algorithms
- search heuristic genetic algorithm; sentimental screening

The model developed in this research will assist users of algorithms to identify which algorithm to implement based on the data they have available, the constraints of the required analysis as well as the algorithm's intended results. The model is intended at share traders, operational managers and other users who are concerned with governance, but do not have the technical knowledge and experience to understand algorithms. Therefore it can also assist in achieving aligning between the business and IT departments (Boshoff, 2016; Goosen & Rudman, 2015).

The literature review and research has shown that there are many opportunities for future research in the study of algorithms:

- There is little research on the applications of algorithm technology in business. Most available research focus only on a technical statistical or programming analysis of such an application, but does not provide any business guidance.
- There is very little research available on the governance of algorithms.

References:

1. Ain, Q.T., Ali, M., Riaz, A., Noureen, A., Kamran M., Hayat, B. & Rehman, A. 2017. Sentiment Analysis Using Deep Learning Techniques: A Review, *International Journal of Advanced Computer Science and Applications (IJACSA)*, Vol. 8, No. 6, 2017. [Online] Available: http://thesai.org/Downloads/Volume8No6/Paper_57-Sentiment_Analysis_using_Deep_Learning.pdf [2018, August 21]

2. AI Congress. 2017. The Usefulness – and Possible Dangers – of Machine Learning, Transcript of Workshop held on 3 October 2017. [Online] Available: https://theaicongress.com/news/2017/10/3/the-usefulnessand-possible-dangersof-machine-learning [2018, May 31]

3. Altman, D.G., & Bland, J.M. 2009. Parametric *v* non-parametric methods for data analysis. *BMJ* 2009; 338:a3167. [Online] Available: https://www.bmj.com/content/338/bmj.a3167 [2018, May 18]

4. Amazon Web Services, Inc. 2018. Model fit: Underfitting vs Overfitting. [Online] Available: https://docs.aws.amazon.com/machine-learning/latest/dg/batch-predictions.html [2018, May 22]

5. Bantleman, J. 2012. The Big Cost of Big Data. CIO Network. [Online] Available: https://www.forbes.com/sites/ciocentral/2012/04/16/the-big-cost-of-big-data/#33c87f05a3b7 [2018, May 25]

6. Bloomberg. 2017. 3 ways big data is changing financial trading, Data and Tech Operations. [Online] Available: https://www.bloomberg.com/professional/blog/3-ways-big-data-changing-financial-trading/ [2018, August 1]

7. Boshoff, W. 2016. Masters in accounting (Computer Auditing). Unpublished lecture slides. Stellenbosch: University of Stellenbosch

8. Brownlee, J. 2017. How Much Training Data is Required for Machine Learning?, Machine Learning Mastery. [Online]
Available: https://machinelearningmastery.com/much-training-data-required-machine-learning/ [2018, May 25]

9. Brownlee, J. 2016a. Gentle Introduction to the Bias-Variance Trade-off in Machine Learning, Machine Learning Mastery. [Online]
Available: https://machinelearningmastery.com/gentle-introduction-to-the-bias-variance-trade-off-in-machine-learning/ [2018, May 22]

10. Brownlee, J. 2016b. Overfitting and Underfitting with Machine Learning Algorithms, Machine Learning Mastery. [Online]
Available: https://machinelearningmastery.com/overfitting-and-underfitting-with-machine-learning-algorithms/ [2018, May 23]

11. Brownlee, J. 2016c. Parametric and Nonparametric Machine Learning Algorithms, Machine Learning Mastery. [Online]
Available: https://machinelearningmastery.com/parametric-and-nonparametric-machine-learning-algorithms/ [2018, May 18]

12. Brownlee, J. 2016d. Supervised and Unsupervised Machine Learning Algorithms, Machine Learning Mastery. [Online]
Available: https://machinelearningmastery.com/supervised-and-unsupervised-machine-learning-algorithms/ [2018, May 21]

13. Brownlee, J. 2014. The Best Learning Algorithm, Machine Learning Mastery. [Online] Available: https://machinelearningmastery.com/much-training-data-required-machine-learning/ [2018, May 25]

14. Butler, R. & Butler, M.J. 2010. Beyond King III: Assigning Accountability for IT Governance in South African Enterprises, *South African Journal of Business Management*, 2010, 41 (3). [Online]
Available: http://hdl.handle.net/10019.1/16711 [2018, March 3]

15. Chan, Y.E. 2015. IT Value: The Great Divide Between Qualitative and Quantitative and Individual and Organisational Measures, Journal of Management Information Systems, Vol 16, 2000, Issue 4, pp. 225-261. [Online] Available: https://www.tandfonline.com/doi/abs/10.1080/07421222.2000.11518272?casa_token=l4g8l8y2KsUAAAAA:oGrkEE967hFnQgfrMqaLm1JDnXGAruemnzCc2wtOyaLD9DxzufhaKSuPcLt3s2oMp8adNT70rs9V [2018, August 22]

16. Coffin, M. & Saltzman, M.J. 1999. Statistical Analysis of Computational Tests of Algorithms and Heuristics, *INFORMS Journal of Computing*, Vol. 12, No. 1, Winter 2000. [Online] Available: https://pdfs.semanticscholar.org/ef87/199c014a18b8b58e05822c5c05ae743447f4.pdf [2018, August 10]

17. Coglianese, C. & Lehr, D. 2017. Regulating by Robot: Administrative Decision Making in the Machine-Learning Era, Penn Law Faculty Scholarship, 6-2017. [Online] Available: https://scholarship.law.upenn.edu/faculty_scholarship/1734/ [2018, July 4]

18. Colombia University. 2018. Linear Functions. [Online] Available: http://www.columbia.edu/itc/sipa/math/linear.html [2018, August 9]

19. Cooper, H. 1998. Synthesising Research, A Guide for Literature reviews, Third Edition. Sage Publications. [Online] Available: https://books.google.co.za/books?hl=en&lr=&id=ZWvAmbjtE9sC&oi=fnd&pg=PP11&dq=components+of+literature+reviews&ots=pE8iMy1Oz1&sig=k6N8hPZEN2MhIfuSOMpDd9heb2s#v=onepage&q=components%20of%20literature%20reviews&f=false [2018, July 31]

20. Cox, H. 2016. How can you can build your own share-trading algorithm? Times, December 2016. [Online] Available: go.galegroup.com/ps/i.do?p=AONE&sw=w&u=27uos&v=2.1&id=GALE%7CA472489984&it=r&asid=786924f2e6e412e17466f477ae00292b [2017, March 9]

21. Deloitte. 2017. Managing Algorithmic Risks: Safeguarding the use of complex algorithms and machine learning. [Online] Available: https://www2.deloitte.com/us/en/pages/risk/articles/algorithmic-machine-learning-risk-management.html [2018, May 31]

22. Deloitte. 2018. King IV: Bolder Than Ever. [Online] Available: https://www2.deloitte.com/za/en/pages/africa-centre-for-corporate-governance/articles/kingiv-report-on-corporate-governance.html [2018, May 8]

23. Dzikiti, L.N. & Girdler-Brown, B.V. 2017. Parametric hypothesis tests for the difference between two population means, *Strengthening Health Systems Journal*, November 2017. [Online] Available: www.shsjournal.org/index.php/shsj/article/download/50/26 [2018, August 1]

24. EliteDataScience. 2017. Modern Machine Learning Algorithms: Strengths and Weaknesses. [Online] Available: https://elitedatascience.com/machine-learning-algorithms [2018, May 24]

25. Essinger, J. 1990. Ruling markets by machine. Computer Weekly, March 1990. [Online] Available: http://go.galegroup.com.ez.sun.ac.za/ps/i.do?p=AONE&sw=w&u=27uos&v=2.1&it=r&id=GALE%7CA8307028&asid=3601ac7c8bd088d94b52b3e156d928b9 [2017, March 9].

26. Fink, A. 2005. Conducting Research Literature Reviews, From Internet to Paper. Second edition, Sage Publications. [Online] Available: https://books.google.co.za/books?hl=en&lr=&id=VyROaw-hLJMC&oi=fnd&pg=PA1&dq=components+of+literature+reviews&ots=doq6AWQliy&sig=bmVBJVOOyh61jjgenDHXHY5vayI#v=onepage&q&f=false [2018, August 2]

27. Flinders, K. 2007. The evolution of stock market technology. [Online] Available: https://www.computerweekly.com/news/2240083742/The-evolution-of-stock-market-technology [2018, April 4]

28. Forex Capital Markets. 2018. Evolution of the marketplace: From Open Outcry to Electronic Trading. [Online] Available: https://www.fxcm.com/insights/evolution-of-the-marketplace-from-open-outcry-to-electronic-trading/ [2018, April 12]

29. Fortmann-Roe, S. 2012. Understanding the Bias-Variance Trade-off. [Online] Available: http://scott.fortmann-roe.com/docs/BiasVariance.html [2018, May 22]

30. Gage, J. 2018. Introduction to sentimental screening. [Online] Available: https://blog.algorithmia.com/introduction-sentiment-analysis [2018, May 28]

31. Gantz, J. & Reinsel, D. 2012. The Digital Universe in 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East, *IDC View.* [Online] Available: https://www.emc.com/collateral/analyst-reports/idc-the-digital-universe-in-2020.pdf [2018, May 25]

32. Gartner, 2018. IT Glossary. [Online] Available: https://www.gartner.com/it-glossary/it-governance/ [2018, August 25]

33. Goosen, R. & Rudman, R. 2013. An Integrated Framework To Implement IT Governance Principles at a Strategic and Operational Level for Medium- To Large-Sized South African Businesses, *International Business and Economics Research Journal*, July 2013, Vol. 12, Number 7. [Online] Available: https://www.researchgate.net/publication/297754852_An_Integrated_Framework_To_Implement_It_Governance_Principles_At_A_Strategic_And_Operational_Level_For_Medium-To_Large-Sized_South_African_Businesses [2018, August 15]

34. Gutierrez, D. 2014. Bias vs Variance trade-off. [Online] Available: https://insidebigdata.com/2014/10/22/ask-data-scientist-bias-vs-variance-tradeoff/ [2018, May 22]

35. Hastie, T., Tibshirani, R. & Friedman, J. 2017. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Second Edition. Springer Publishers. [Online] https://web.stanford.edu/~hastie/ElemStatLearn/printings/ESLII_print12.pdf [2018, September 5]

36. Hu, Y., Liu, K., Zhang, K., Su, L., Ngai, E. & Liu, M. 2015. Application of evolutionary Computation for rule discovery in stock algorithmic trading: A Literature review, *Applied Soft Computing*, Vol 36, November 2015 Pages 534 – 551. [Online] Available: https://www.sciencedirect.com/science/article/pii/S156849461500438X [2018, May 25]

37. IBM. 2014. Performance and Capacity Implications for Big Data, First Edition. [Online] Available: https://www.redbooks.ibm.com/redpapers/pdfs/redp5070.pdf [2018, May 24]

38. Informatica. 2017. Holistic Data Governance: A Framework for Competitive Advantage. [Online] Available: https://www.informatica.com/lp/holistic-data-governance-framework_2297.html#fbid=Y3HVTx1upSv [2018, April 18]

39. Institute of Directors in Southern Africa (IODSA). 2009. King Report on Governance for South Africa. [Online] Available: https://cdn.ymaws.com/www.iodsa.co.za/resource/resmgr/king_iii/King_Report_on_Governance_fo.pdf [2017, March 9]

40. Institute of Directors in Southern Africa (IODSA). 2016. King IV report: Corporate Governance. [Online] Available: https://www.pwc.co.za/en/publications/king4.html [2017, March 9]

41. Itskevich, J. 2002. What caused the stock market crash of 1987? [Online] Available: https://historynewsnetwork.org/article/895 [2018, April 4]

42. Jain, V.K. 2017. Big Data & Hadoop. [Online] Available: https://books.google.co.za/books?id=i6NODQAAQBAJ&printsec=frontcover&dq=what+is+big+data&hl=en&sa=X&ved=0ahUKEwjiguLmgoLZAhXHuBQKHd_mCxMQ6AEILzAB#v=onepage&q=what%20is%20big%20data&f=false [2018, April 1]

43. Johannesburg Stock Exchange. 2016. Johannesburg Stock Exchange New Equity Market Trading and Information Solution. [Online] Available: https://www.jse.co.za/content/JSETechnologyDocumentItems/Volume%2000%20-%20Trading%20and%20Information%20Overview%20v3.01. pdf [2018, April 4]

44. Kasparis, I., Andreou, E. & Phillips, P.C.B. 2015. Nonparametric Predictive Regression, *Journal of Econometrics*. 185, (2), 468-494. Research Collection School of Economics. [Online] Available: http://ink.library.smu.edu.sg/soe_research/1836 [2018, August 20]

45. Khan, Z.H., Alin, T.S. & Hussain, M.A. 2011. Price Prediction of Share Market using Artificial Neural Network (ANN), *International Journal of Computer Applications* (0975 – 8887) Volume 22– No.2, May 2011. [Online] Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.206.4394&rep=rep1&type=pdf [2017, April 2]

46. Kirilenko, A., Kyle, A.S., Samadi, M. & Tuzun, T. 2017. The Flash Crash: High-Frequency Trading in an Electronic Market, *The Journal of Finance.* Vol 72:3, June 2017, Pages 967-998. [Online] Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/jofi.12498 [2018, August 22]

47. Ko, Y. & Seo, J. 2000. Automatic Text Categorization by Unsupervised Learning, 18th International Conference on Computational Linguistics. [Online]
Available: http://www.aclweb.org/anthology/C00-1066 [2018, May 29]

48. Kolanovic, M. & Krishnamachari, R.T. 2017. Big Data and AI Strategies: Machine Learning and Alternative Data Approach to Investing, JP Morgan. [Online] Available: http://valuesimplex.com/articles/JPM.pdf [2018, August 10]

49. Langley, P. 1996. Elements of Machine Learning. Morgan Kaufmann Publishers, Inc. [Online] Available: https://books.google.co.za/books?hl=en&lr=&id=TNg5qVoqRtUC&oi=fnd&pg=PR9&dq=what+is+machine+learning&ots=Q3ltZrr0Lr&sig=9U666Bre8voM8GLhNj66cw-HQ5g#v=onepage&q=what%20is%20machine%20learning&f=false [2018, April 13]

50. Luftman, J. 2003. Assessing IT/Business Alignment, *Information Systems Management.* 20:4, 9-15. [Online]. Available: http://www.tandfonline.com/doi/pdf/10.1201/1078/43647.20.4.20030901/77287.2 [2018, April 12]

51. MacKenzie, D. 2011. 'How to Make Money in Microseconds'. *London Review of Books*, Vol 33: 10, pp. 16-18. [Online] Available: http://www.research.ed.ac.uk/portal/files/13410827/How_to_Make_Money_in_Microseconds.pdf [2017, April 2]

52. Mallawaarachchi, V. 2017. Introduction to Genetic Algorithms - Including Example Code, Towards Data Science. [Online] Available: https://towardsdatascience.com/introduction-to-genetic-algorithms-including-example-code-e396e98d8bf3 [2018, May 28]

53. Manning, CD., Raghavan, P. & Schutze, H. 2009. An Introduction to Information Retrieval, p. 110. Cambridge University Press. [Online] Available: https://nlp.stanford.edu/IR-book/pdf/irbookonlinereading.pdf [2018, May 24]

54. Marr, B. 2015. Big data: Using SMART Big Data, Analytics and Metrics to Make Better Decisions and Improve Performance, pp. 9 – 10. [Online] Available: https://books.google.co.za/books?id=p_glBgAAQBAJ&printsec=frontcover&dq=what+is+big+data&hl=en&sa=X&ved=0ahUKEwjiguLmgoLZAhXHuBQKHd_mCxMQ6AEINDAC#v=onepage&q=what%20is%20big%20data&f=false [2018, April 12]

55. Marx, J., Mpofu, R.T., De Beer, J.S., Mynhardt, R.H. & Nortjé A. 2013. Investment Management, Fourth Edition, Van Schaik Publishers

56. MathWorks. 2018a. Predictive Analytics: 3 Things you need to know. [Online] Available: https://www.mathworks.com/discovery/predictive-analytics.html [2018, May 17]

57. MathWorks. 2018b. What is the Genetic Algorithm? [Online] Available: https://www.mathworks.com/help/gads/what-is-the-genetic-algorithm.html [2018, May 25]

58. Mayer-Schonberger, V. & Cukier, K. 2013. 'Big Data: A Revolution That Will Transform How We Live, Work and Think". [Online] Available: https://books.google.co.za/books?id=HpHcGAkFEjkC&printsec=frontcover&dq=what+is+big+data&hl=en&sa=X&ved=0ahUKEwjiguLmgoLZAhXHuBQKHd_mCxMQ6AEIKDAA#v=onepage&q=what%20is%20big%20data&f=false [2018, April 12]

59. McCall, J. 2005. Genetic Algorithms for Modelling and Optimisation, *Journal of Computational and Applied Mathematics*, Vol. 184, Issue 1, December 2005, pp. 205 – 222. [Online] Available: https://www.sciencedirect.com/science/article/pii/S0377042705000774 [2018, May 28]

60. Medhat, W., Hassan, A. & Korashy, H. 2014. Sentiment analysis algorithms and applications: A survey, *Ain Shams Engineering Journal* (2014) 5, 1093–1113. [Online] Available: http://kt.ijs.si/markodebeljak/Lectures/Seminar_MPS/2012_on/Seminars_2015_16/Simon%20Brmez/Bibliography/%5B5%5D%20Sentiment%20analysis%20algorithms%20and%20applications%20A%20survey.pdf [2018, August 21]

61. Memorial University. 2018. Mean, Median, Mode. [Online] Available: http://www.mun.ca/educ/ed4361/virtual_academy/campus_a/losinskim/mean,median,mode.html [2018, August 9]

62. Microsoft. 2018. Training and Testing Data Sets. [Online] Available: https://docs.microsoft.com/en-us/sql/analysis-services/data-mining/training-and-testing-data-sets?view=sql-analysis-services-2017 [2018, May 25]

63. Microsoft Azure. 2017. How to choose algorithms for Microsoft Azure Machine Learning. [Online] Available: https://docs.microsoft.com/en-us/azure/machine-learning/studio/algorithm-choice [2018, August 10]

64. Michalski, R.S., Carbonell, J.G. & Mitchell, T.M. 1983. Machine Learning: An Artificial Intelligence Approach, *Springer-Verlag*. [Online] Available: https://books.google.co.za/books?hl=en&lr=&id=-eqpCAAAQBAJ&oi=fnd&pg=PA2&dq=understanding+machine+learning&ots=WI1SMy7Im7&sig=j3FvtcWtxwqLPG6G3blhtJFc66o#v=onepage&q=understanding%20machine%20learning&f=false [2018, May 25]

65. Mnich, M. 2018. Big data algorithms beyond machine learning, Künstliche Intelligenz (2018) 32: 9. [Online] Available: https://doi.org/10.1007/s13218-017-0517-5 [2018, August, 16]

66. Monterey Institute for Technology and Education. 2018. Nonlinear functions. [Online] Available: https://www.montereyinstitute.org/courses/Algebra1/COURSE_TEXT_RESOURCE/U03_L2_T5_text_final.html [2018, August 9]

67. Murty, K. 1997. Linear Complementarity, Linear and Nonlinear Programming, Internet Edition. [Online] Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.303.3931&rep=rep1&type=pdf [2018, May 25]

68. Oxford University Press. 2018. Oxford Dictionary. [Online] Available: https://en.oxforddictionaries.com/definition/ [2018, April 12]

69. Pereira, R. 2002. Forecasting Ability but No Profitability: An Empirical Evaluation of Genetic Algorithm-Optimised Technical Trading Rules. *Evolutionary Computation in Economics and Finance.* [Online] Available: https://link.springer.com/chapter/10.1007/978-3-7908-1784-3_16#page-2 [2017, April 2]

70. PWC. 2017. Governing structures and delegation – A comparison between King IV and King III. [Online] Available: https://www.pwc.co.za/en/assets/pdf/king-iv-comparison.pdf [2018, August 01]

71. Richardson, A., Gregor, S. & Heany, R. 2012. Using decision support to manage the influence of cognitive abilities on share trading performance, *Australian Journal of Management 37(3).* [Online] Available: http://journals.sagepub.com.ez.sun.ac.za/doi/abs/10.1177/0312896211432942 [2017, April 2]

72. Russel, S. & Norvig, P. 2010. Artificial Intelligence, A Modern Approach. Third Edition, Pearson Education Inc. [Online] Available: http://aima.cs.berkeley.edu/ [2018, August 20]

73. Ryan, N. 2018. Big Data and Performance Management. *ACCA.* [Online] Available: http://www.accaglobal.com/africa/en/student/exam-support-resources/professional-exams-study-resources/p5/technical-articles/big-data-pm.html [2018, May 25]

74. Serialmetrics. 2018. Common Machine Learning Challenges. [Online] Available: http://serialmetrics.com/blog/common-machine-learning-challenges/ [2018, May 31]

75. Shalev-Shwartz, S. & Ben-David, S. 2014. Understanding Machine Learning: From Theory to Algorithms. [Online] Available: https://books.google.co.za/books?id=ttJkAwAAQBAJ&printsec=frontcover&dq=what+is+machine+learning&hl=en&sa=X&ved=0ahUKEwinu7n9kf3YAhUG6RQKHZMAAuIQ6AEIQjAE#v=onepage&q=what%20is%20machine%20learning&f=false [2018, April 12]

76. Smeda, J. 2015. Benefits, business considerations and risks of big data. [Online] Available: http://scholar.sun.ac.za/bitstream/handle/10019.1/96684/smeda_benefits_2015.pdf?sequence=3&isAllowed=y [2018, May 25]

77. SAS. 2018. Big Data Analytics: What It is and Why It Matters. [Online] Available: https://www.sas.com/en_us/insights/analytics/big-data-analytics.html [2018, May 25]

78. Stobart, N. 2018. AI and Machine Learning – What are the Most important Data Storage Requirements? [Online] Available: https://www.cbronline.com/emerging-technology/ai-machine-learning-important-data-storage-requirements/ [2018, August 16]

79. Stoll, HR. 2006. Electronic Trading in Stock Markets, *Journal of Economic Perspectives.* Vol. 20:1, Winter 2006, pp. 153–174. [Online] Available: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=905614 [2017, April 2]

80. Tallon, P. 2013. Corporate Governance of Big Data: Perspectives on Value, Risk and Cost, IEEE, June 2013. [Online] Available: https://ieeexplore.ieee.org/abstract/document/6519236/ [2018, May 25]

81. Tene, O. & Polonetsky, J. 2013. Big Data for All: Privacy and User Control in the Age of Analytics, *Northwestern Journal of Technology and Intellectual Property*, Vol. 11, Issue 5 (2013). [Online] Available: https://scholarlycommons.law.northwestern.edu/njtip/vol11/iss5/1/ [2018, May 25]

82. Van Winkle, E.M. 2011. The Incremental Value of Qualitative Fundamental Analysis to Quantitative Fundamental Analysis: A Field Study. [Online] Available: https://deepblue.lib.umich.edu/bitstream/handle/2027.42/84568/mvanwink_1.pdf;sequence=1 [2018, August 22]

83. Verma, G. & Verma V. 2012. Role and Applications of Genetic Algorithm in Data Mining, International Journal of Computer Applications (0975 – 888), Volume 48 – No. 17, June 2012. [Online] Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.258.9723&rep=rep1&type=pdf [2018, May 28]

84. Von Wielligh, P. & Prinsloo, F. 2014. Auditing Fundamentals in a South African Context, Oxford University Press South Africa

85. Vorhies, W. 2017. Data or Algorithms – Which is More Important? Data Science Central. [Online] Available: https://www.datasciencecentral.com/profiles/blogs/data-or-algorithms-which-is-more-important [2018, August 16]

86. Walaa, M., Hassan, A. & Korashy, H. 2014. Sentiment analysis algorithms and applications: A survey. Ain Shams Engineering Journal, Volume 5, Issue 4, pp. 1093 – 1113. [Online] Available: https://www.sciencedirect.com/science/article/pii/S2090447914000550#b0020 [2018, May 28]

87. Webster, J. & Watson, R. 2002. Analyzing the Past to Prepare for the Future: Writing a Literature Review, MIS Quarterly, Vol. 26, No. 2 (Jun., 2002), pp. xiii-xxiii. [Online] Available: https://www.jstor.org/stable/4132319?seq=1#page_scan_tab_contents [2018, August 2]

88. Witten, I.H., Frank, E., Hall, M.A. & Pal, C.J. 2017. Data Mining: Practical Machine Learning Tools and Techniques, Fourth Edition, Elsevier. [Online] Available: https://books.google.co.za/books?hl=en&lr=&id=1SylCgAAQBAJ&oi=fnd&pg=PP1&dq=components+of+machine+learning&ots=8IBPsenDCc&sig=ZbP_C1Z0qKuBuimV4G75yVddj2M&redir_esc=y#v=onepage&q=components%20of%20machine%20learning&f=false [2018, May 25]

89. Yu, L., Wu, J., Chang, P. & Chu, H. 2013. Using a contextual entropy model to expand emotion words and their intensity for the sentiment classification of stock market news, Elsevier, Knowledge-Based Systems 41, 2013, pp. 89 – 97. [Online] Available: https://ac.els-cdn.com/S095070511300004X/1-s2.0-S095070511300004X-main.pdf?_tid=11606146-55c6-4058-ae32-f32aa20058b4&acdnat=1527512367_1bbac681feb4ac0f89a69109b2f2237b [2018, May 30]

90. Zhou, Z. 2018. Machine Learning Challenges and Impact: An Interview with Thomas Dietterich, National Science Review, Vol. 5, Issue 1, 1 January 2018, pp. 54 – 58. [Online] Available: https://academic.oup.com/nsr/article/5/1/54/3789514 [2018, May 31]