

BIG DATA ANALYSIS IN FINANCIAL MARKETS

**A Dissertation Presented in Partial Fulfillment of the
Requirements for the Degree of
Doctor of Computer Science**

By

Tamaro J. Green

Colorado Technical University

November, 2018

ProQuest Number: 10980927

All rights reserved

INFORMATION TO ALL USERS

The quality of this reproduction is dependent upon the quality of the copy submitted.

In the unlikely event that the author did not send a complete manuscript and there are missing pages, these will be noted. Also, if material had to be removed, a note will indicate the deletion.



ProQuest 10980927

Published by ProQuest LLC (2019). Copyright of the Dissertation is held by the Author.

All rights reserved.

This work is protected against unauthorized copying under Title 17, United States Code
Microform Edition © ProQuest LLC.

ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 – 1346

Committee

Dr. Daphne DePorres, EdD, Research Mentor and Chair

Dr. James Webb, PhD, Committee Member

Dr. Kelly Hughes, DS, Committee Member

November 3, 2018

© Tamaro J. Green, 2018

Abstract

This dissertation researched topics in the financial analysis of publicly traded companies. The opportunity for this dissertation was to address the potential of increasing value for public companies that implemented big data technologies. The participants for this research were public companies that included “big data” in their periodic filings with the United States Securities and Exchange Commission. The research methodology includes quantitative methods and features statistical analysis of financial markets. The findings of the research were that although there was not a statistically significant increase of the implementation of big data technologies for public companies, there was a general increase over the performance of other companies as measured by the major indices of the financial markets. The theories that formed the basis of this topic included the technology acceptance model and complexity theory (Davis, 1989; Davis Jr, 1986; Hartmanis & Stearns, 1965).

Dedication

This dissertation is dedicated to my family who has made education a priority and always showed me the importance of education in achieving success. I would like to thank my family for always providing patience, encouragement, support, and guidance throughout my education.

Acknowledgements

I would like to acknowledge the professors at Colorado Technical University for challenging me to improve my writing skills and always being there with professionalism and guidance. Every professor at Colorado Technical University shows extreme dedication in ensuring that students achieve success and understand the material.

Table of Contents

Acknowledgements.....	iv
Table of Contents.....	v
List of Tables.....	viii
List of Figures	x
Chapter One	1
Topic Overview/Background	1
Opportunity Statement	2
Purpose Statement	3
Research Question	3
Hypotheses	3
Assumptions/Biases	7
Significance of the Study	7
Delimitations	8
Limitations	9
Definition of Terms	9
General Overview of the Research Design	10
Summary of Chapter One	11
Chapter Two.....	12
Quantitative Analyses of Stock Markets.....	12

Statistical Analysis.....	16
Big Data Technologies.....	24
Modern Applications of Conceptual Theories	29
Influence of Big Data on Information Technology	32
Conceptual Framework.....	33
Summary of Literature Review	35
Chapter Three.....	37
Research Tradition.....	37
Research Questions and Hypotheses	38
Hypotheses	38
Research Design	38
Population and Sample.....	39
Sampling Procedure.....	40
Instrumentation.....	40
Validity.....	43
Reliability	44
Data Collection	44
Data Analysis.....	46
Ethical Considerations	47
Summary of Chapter Three.....	49

Chapter Four.....	50
Participant Demographics	50
Presentation of the Data	58
Presentation and Discussion of Findings	64
Summary of Chapter	76
Chapter Five	77
Limitations of the Study.....	77
Implications for Practice	77
Implications of Study and Recommendations for Future Research	78
Reflections.....	79
Conclusion.....	80
References	82

List of Tables

Table 1	<i>Analysis Points</i>	41
Table 2	<i>Sectors for Stocks Associated with Big Data</i>	51
Table 3	<i>Sectors and Industries for Stocks Associated with Big Data</i>	51
Table 4	<i>Exchanges for Big Data Companies</i>	52
Table 5	<i>Status 2013-2018</i>	53
Table 6	<i>Sectors for Public Companies with "Big Data" in Description</i>	53
Table 7	<i>Indices 2013-2018</i>	54
Table 8	<i>Summary Data for Returns</i>	54
Table 9	<i>Big Data Companies on Major Exchanges</i>	55
Table 10	<i>Yearly Percent Change of Three Major Indices</i>	55
Table 11	<i>Yearly Close of Three Major Indices</i>	55
Table 12	<i>Yearly Returns for Three Major Indices</i>	56
Table 13	<i>Statistically Significant Differences in Stock Value Changes with DJIA</i>	57
Table 14	<i>Statistically Significant Difference with NASDAQ Index</i>	57
Table 15	<i>Statistically Significant Difference with Standard and Poor 500 Index</i>	58
Table 16	<i>Kernel Density Estimation Default</i>	60
Table 17	<i>Logistic Model for Influence on Big Data Company Price Increase</i>	61
Table 18	<i>Logistic Model for Sector</i>	62
Table 19	<i>Analysis of Variance for Models</i>	62
Table 20	<i>Analysis of Variance of Models</i>	64
Table 21	<i>Bayesian Generalized Linear Model</i>	65
Table 22	<i>Market Capitalization Model</i>	66

Table 23	<i>Raftery Diagnostics of Sector Model</i>	66
Table 24	<i>Raftery Diagnostics of Sector Model</i>	67
Table 25	<i>Quartiles for Sector Markov Chain Monte Carlo Model</i>	67
Table 26	<i>Raftery Diagnostics of Markov Chain Monte Carlo Simulation 3</i>	68
Table 27	<i>Markov Chain Monte Carlo Simulation Empirical Mean</i>	69
Table 28	<i>Markov Chain Monte Carlo Simulation Quartiles</i>	69
Table 29	<i>Percent Change by Sector</i>	69
Table 30	<i>Market Capitalization by Sector (Billions)</i>	70
Table 31	<i>Top Ten Industries by Market Capitalization</i>	71
Table 32	<i>Decision Tree Market</i>	71
Table 33	<i>Decision Tree Market and Sector</i>	72
Table 34	<i>Decision Tree Market and Market Capitalization</i>	73
Table 35	<i>Number of Companies in Each Sector with a Percent Change Greater than 200%</i>	75
Table 36	<i>Number of Companies in Each Sector with a Percent Change Greater than 300%</i>	75
Table 37	<i>Number of Companies in Each Sector with a Percent Change Greater than 400%</i>	75
Table 38	<i>Number of Companies in Each Sector with a Percent Change Greater than 500%</i>	76

List of Figures

<i>Figure 1.</i> The components of financial analysis and statistical models and their influence on the development of big data technologies.	35
<i>Figure 2.</i> Histogram of the percent change of the stock prices between 2013 and 2018.	59
<i>Figure 3.</i> Kernel destiny estimation of the percent change of the stock prices.	60
<i>Figure 4.</i> True Positive vs False Positive Rates.	63
<i>Figure 5.</i> Receiving operating curve plot for market and sector.	64
<i>Figure 6.</i> Random forest of linear models with 100 trees.	74

CHAPTER ONE

The research outlined in this dissertation explored the extent to which there was a statistically significant increase, from 2013 to 2018, of the market value of companies that had “big data” in the company description, which is part of the public filing document registered with the United States Securities and Exchange Commission. This research examined statistically significant changes in the market value of publicly traded companies in the United States’ that utilized big data, specifically, looking for increases in the market value of these companies.

This research examined data about publicly traded companies that were listed on the major US exchanges, the New York Stock Exchange and the NASDAQ exchange and examined how public companies have increased the value of their stock price, for a period of 5 years, exploring whether or not there is a difference between these companies and other publicly traded companies.

Included in this chapter are the topic overview, problem statement, purpose statement, research question and hypotheses, theoretical perspectives, researcher assumptions, and biases. Also included in this chapter is the significance of the study, delimitations, limitations, definitions of terms. This chapter also includes an overview of the research design.

Topic Overview/Background

This research proposed an exploration of the extent to which there is a statistically significant increase in impact on market value of companies that have “big data” in the company description, for the period from 2013-2018 as denoted by revenue for publicly traded technology companies in the United States featuring “data analysis” in the company description. This topic was of interest as many corporations have been exploring opportunities to achieve value from raw data (Wamba et al., 2017). Research on the potential for companies to seize opportunities from the advancement in big data technologies has become a research topic in both academia and

practice (Wamba et al., 2017). The research of the potential of big data technologies has also led to the development of innovative implementations of the technology (CACM Staff, 2017).

As of May of 2018, there were over 6,800 publicly traded companies that were listed on the two major exchanges the New York Stock Exchange and the NASDAQ exchange (Nasdaq.com, 2018). Of those 6,800 publicly traded companies, there were approximately 350 companies that are categorized as technology companies (Nasdaq.com, 2018). Of all the publicly traded companies, companies that had “big data” in their company description were compared to the other companies by examining the major indices for the exchanges. The comparison included the price of the stock and the market capitalization. The analysis also compared the sectors of companies that utilized big data, evaluating how they performed for a 5 year period. The analysis identified if these companies profited in the implementation of big data in their organizations.

Opportunity Statement

The opportunity represented by this study was to explore the extent to which there was a statistically significant increase, from 2013 to 2018, of the market value of companies that had “big data” in the company description, which was part of the periodic filings with the United States Securities and Exchange Commission. The proposed study explored whether public companies that included “big data” in their description for potential investors experienced an increase in market value (Wamba et al., 2017). Du, Deng, and Qian (2018) identified that 63% of researched companies had business initiatives on topics around big data technologies. This knowledge may contribute to what investors knew about the market value of publicly traded companies. Specifically, the research explored data from companies that utilized big data in their analysis processes and have converted raw data into information and knowledge to gain a competitive advantage over companies that have not utilized big data or have not used it

effectively (CACM Staff, 2017). The competitive advantage has been explored as an increase in the market value of the company that utilized big data (Wamba et al., 2017). The competitive advantage can also lead to an identification of the challenges that deter practical implementations of big data technologies for widespread applications (Alexander, Das, Ives, Jagadish, & Monteleoni, 2017; Tabona & Blyth, 2016).

Purpose Statement

The purpose of this research was to explore the extent to which there was a statistically significant increase, from 2013 to 2018, of the market value of companies that had “big data” in the company description. An attempt to find and substantiate a statistically significant increase of the market value of companies that had “big data” in the company description, which revealed the extent to which there was a statistically significant increase, from 2013 to 2018, of the market value of these companies. The strategy of inquiry included time series analysis, linear models, Bayesian analysis, decision trees, random forests, and ensemble methods. The raw data was collected from the United States Securities and Exchange Commission Electronic Data Gathering and Retrieval system, EDGAR, and from the major United States financial markets. The phenomenon of the study was the increase in the financial value of the public companies that included “big data” in their periodic filings.

Research Question

The research question for this study was, to what extent was there a statistically significant increase, from 2013 to 2018, of the market value of companies that have “big data” in the company description, which is part of the public filing document registered with the United States Securities and Exchange Commission?

Hypotheses

The null and alternative hypotheses for this study were:

H1₀: There is no statistically significant increase, from 2013 to 2018, of the market value of companies that have “big data” in the company description, which is part of the public filing document registered with the United States Securities and Exchange Commission.

H1_a: There is a statistically significant increase, from 2013 to 2018, of the market value of companies that have “big data” in the company description, which is part of the public filing document registered with the United States Securities and Exchange Commission.

Theoretical Perspectives

The theoretical frameworks for the study include the technology acceptance model and computational complexity theory. The technology acceptance model is a model that describes the relationship between perceived ease of use and perceived usefulness (Davis, 1989; Davis Jr, 1986). Davis Jr (1986); Venkatesh, Morris, Davis, and Davis (2003) expanded their original technology acceptance model to describe a unified view of technology acceptance.

The technology acceptance model provides a guide on how new technology is accepted by the end user and the factors that influence the acceptance (Davis Jr, 1986). The technology acceptance model describes factors that influence the acceptance of technology (Davis Jr, 1986). Davis (1989) introduced the technology acceptance model which describes the intention of behavior in the use of technology as a product of perceived usefulness and perceived ease of use of the technology. These factors affect the desired behavior of the system and influence the reaction to the system behavior (Davis Jr, 1986). The technology acceptance model applied to these changes in programming design can identify metrics for measuring the reaction to change for the design of new advanced computer systems. The improvements to the capabilities of these computer systems can enhance the actual system use (Ming et al., 2013). These enhancements

are influenced by the algorithmic design of how computers process information (Plekhanov, 2017).

Complexity theory attempts to bring order to the large number of algorithmic problems we face in computer science (Clarke, O'Connor, & Leavy, 2016; Rzevski, 2016; Zhang & Liu, 2016). Computational complexity theory provides mathematical calculations for the requirements of computer problems such as disk storage space and time of execution (Ambainis, Filmus, & Gall, 2015; Clarke et al., 2016). Hartmanis (1975) described the difference between context-sensitive and context-free languages. The complexity theory assists computer scientists in discovering the time and space requirements of algorithms and functions (Blaser & Manthey, 2015; Endriss, Haan, & Szeider, 2015; Hubie, 2017; Landsberg & Ressayre, 2016).

The conceptual constructs for the study included hypothesis testing, models, and simulations in order to explore the relationship between market value and big data analytics, indicative of how value was added for technology companies that included “big data” in their descriptions of periodic filings with the United States Securities and Exchange Commission. The research coordinated theories in computer science with contemporary advancements in big data research.

Tang, Pan, and Yao (2018) proposed a combination of principal component analysis with the k-nearest neighbor algorithm that is capable of financial prediction. Piao and Breslin (2018) tested 9 algorithms for financial sentiment prediction with data from a social media corpus. Brinkman and Wellman (2017) implemented an empirical game-theoretic analysis for a game based simulation of financial markets. Baralis, Cagliero, and Cerquitelli (2016) implemented a news crawler that gets information about the stock market from news sources. Hayashi and

Baranauskas (2013) explained how the conceptual framework should be a dynamic map of the research that studies and identifies relationships.

This research attempted to identify the influence of the increase in the value of a company based on its ability to collect and analyze big data. One of the problems for big data in the technology industry was the need to deliver mainstream technological practical solutions in the with revolutionary new technology innovation (Al-Htaybat & von Alberti-Alhtaybat, 2017). With the technological advancements available with big data technologies, companies were able to improve their products and services (Binbin & Xiaoyan, 2017). This research attempted to identify if there was a correlation between the increase in the value of a company and the decision of companies that adopted big data analysis technologies. If companies in recent years have achieved value from their ability to enhance their products and services through big data analytics, there may have been a potential for understanding how these solutions have attracted investments for these organizations. This research identified whether investors saw the potential and value that big data technologies created for companies that seized the opportunity to gain insight from these technologies.

The quantitative research was an appropriate research methodology for studies exploring mature theories that have been studied in the field (Edmondson & McManus, 2007). Edmondson and McManus (2007) described three stages for the methodological fit of research, nascent, intermediate, and mature. Hypothesis testing and statistical inference qualify this research as mature research, a quantitative study. Edmondson and McManus (2007) described mature research to include focused questions, surveys, interviews, systematically coded observations, formal hypothesis testing, and statistical inference.

Assumptions/Biases

The researcher understood that bias and assumptions may have impacted the research, therefore it was necessary to identify potential bias in the research (Noble & Smith, 2015). The researcher was in the field of information technology for over fifteen years and had experience in the education, science, finance, media, and publishing industries. The researcher understood that experience with and exposure to certain technologies may have influenced the perception of how participants accepted technology. The researcher also worked primarily for large corporations which may have impacted the perception of how participants at smaller organizations accept technology. The bias of preconceived ideas on big data analytics in the research was removed or reduced by taking a systematic approach to the research design and implementation.

There were several assumptions the researcher made related to the proposed research. Assumption one was big data frameworks added value to the data being collected by individuals and organizations (Samadi, Zbakh, & Taddonki, 2016). Assumption two was that the confidentiality and anonymity issues that may arise from research studies research are removed by how the research is designed (Creswell, 2014). Assumption three was that the number of publicly traded companies with “big data” in their descriptions was large enough to represent the population of public companies that may have implemented big data (Creswell, 2014). Assumption four assumed that a quantitative study was appropriate for the study of big data analysis in computer science (Edmondson & McManus, 2007).

Significance of the Study

The significance of the study was that if there was a statistically significant increase, from 2013 to 2018, of the market value of companies that had “big data” in the company description, the outcomes of the study would contribute to what was known about the relevance of implementing big data analytics by organizations to optimize their market value. The research

also attempted to identify areas that may be improved with existing technologies for the analysis of data by identifying the components that influence publicly traded organizations when choosing big data technologies (Glicksman, Marked, & Monteleoni, 2017; Gorisek & Pahor, 2017; Kaal & Vermeulen, 2017). The research may have identified the value that has been added for publicly traded companies with regards to big data and showed the benefits of investing in the new technology as well as providing guidance for computer science and software and information technology specialist in improving future technologies (Bradlow, Gangwar, Kopalle, & Voleti, 2017; Graham, Kim, DeVasto, & Keith, 2015; Halford & Savage, 2017; Pusala, Salehi, Katukuri, Xie, & Raghavan, 2016).

Some research was already underway in defining new programming languages and programming concepts to approach the latest trends in computer systems (Praveen & Chandra, 2017). NoSQL provided new tools for advanced data management systems (Yassien & Desouky, 2016). Rust is a programming language that also looks to improve resource management for communication-driven applications (Taft, Moore, Pinho, & Michell, 2014). Blockchain technologies have also created new avenues for supporting distributed computing applications (Xu et al., 2017; Yasaweerasinghelage, Staples, & Weber, 2017).

Delimitations

Delimitations are factors that may constrain the researcher from learning from the data that is analyzed (Sampson, 2016). Constraints for the research included that the research was only limited to public companies that had “big data” in the descriptions of the periodic filings with the United States Securities and Exchange Commission. The population was further constrained by companies that were listed on the major markets and did not include over the counter markets. The population was also constrained by those companies that had listings for the 5 year period from 2013 to 2018.

Limitations

Limitations are influences the ability of the researcher to align the research study to the research questions (Sampson, 2016). Limitations of the study included that the study compared the performance of these companies to the major indices of the financial markets. This limitation requires that the indices are representative of the publicly traded companies on the financial markets. The validity of the study is limited to the reliability of the market indices, the instruments for collecting and analyzing data, and the reliability of the information gathered from the tools provided by the major markets and the United States Securities and Exchange Commission.

Definition of Terms

Several terms that are key to understanding the proposed study are defined below:

Big data – Describes the increasing volume, velocity, variety, veracity, and value of data for analysis and processing (Kung, 2017; Ravi & Kamaruddin, 2017).

EDGAR – The Electronic Data Gathering And Retrieval system, EDGAR, provides free access to the filings of public companies by the United States Securities and Exchange Commission (U.S. Securities and Exchange Commission).

IPO – An Initial Public Offering describes when a company initial becomes listed on an exchange as a publicly traded company (U.S. Securities and Exchange Commission, 2013).

SEC – The United States Securities and Exchange Commission was formed by the Securities Exchange Act of 1934 and enforces the Securities Act of 1933 (U.S. Securities and Exchange Commission, 2013).

Value – Defines a characteristic of big data as an increase in the value of analyzed data (Kung, 2017; Ravi & Kamaruddin, 2017).

Variety – Defines a characteristic of big data as an increase in the types of data sources of collectible data (Kung, 2017; Ravi & Kamaruddin, 2017).

Velocity– Defines a characteristic of big data as an increase in the speed at which data is transferred in information systems (Kung, 2017; Ravi & Kamaruddin, 2017).

Veracity – Defines a characteristic of big data as the ability to verify the correctness of the data (Kung, 2017; Ravi & Kamaruddin, 2017).

Volume – Defines a characteristic of big data as an increase in the amount of collectible data (Kung, 2017; Ravi & Kamaruddin, 2017).

General Overview of the Research Design

The research design was a quantitative study of publicly available stock market data of the entire population of companies with “big data” in the company’s description filed with the SEC. The participants included publicly traded technology companies in the United States. The United States Securities and Exchange Commission provided the search of periodically reported filings by publicly traded companies (U.S. Securities and Exchange Commission, 2007). The Electronic Data Gathering, Analysis, and Retrieval system, EDGAR, is a free tool provided by the SEC to allow the public to review information about publicly traded companies.

The data collection consisted of a combination of collecting search results from the United States Securities and Exchange Commission EDGAR tool and combining the data with publicly available information collected from the major exchanges. Du et al. (2018) conducted a similar study identifying innovative technologies that were scopes of research for public companies listed on the major financial markets. The data analysis consisted of hypothesis testing and statistical inference. The research methodology was appropriate for a quantitative research study.

Summary of Chapter One

Publicly traded companies may benefit from the analysis of big data by being able to convert raw data into information and knowledge (Lahmidani & Beqqali, 2016). Big data can benefit non-technical personnel from data reporting and data visualization (Lan, 2017). The importance of big data to technology is that it enables more advanced computing for both individuals and organizations (Wamba, Akter, Edwards, Chopin, & Gnanzou, 2015; Zhang & Vorobeychik, 2016). Thus, this research explored the extent to which there was a statistically significant increase, from 2013 to 2018, of the market value of companies that have “big data” in the company description, which is part of the public filing document registered with the United States Securities and Exchange Commission. The next chapter will provide a review of the literature on big data technologies and statistical inference methodologies.

Organization of Dissertation

Chapter one is an introduction to the research, the topics, and the conceptual framework. Chapter two is a literature review of financial theories, statistical models, and big data analytics. Chapter three is a description of the research methodology and quantitative correlational study. Chapter four is a discussion of the findings of the research. Chapter five is a summary and conclusion of the research and opportunities for further research.

CHAPTER TWO

The purpose of this research was to explore the extent to which there was a statistically significant increase, from 2013 to 2018, of the market value of companies that had “big data” in the company description. This chapter describes a review of the literature from academic journals and conference papers for the research of testing the hypothesis of the increase of value to public traded companies that implement big data technologies. The chapter reviews academic literature in quantitative market analysis, statistical analysis, and big data technologies. The chapter then includes theoretical perspectives for the analysis of public companies that implement big data technologies. The chapter concludes with a summary of analytical tools in big data.

Quantitative Analyses of Stock Markets

Many researchers have analyzed various sources of data to examine influences on the performance of the stock market (Alpanda, 2013; Klein, Nikitina, & Curtillet, 2016; Sicherman, Loewenstein, Seppi, & Utkus, 2016). A variety of studies have implemented methods such as model averaging, data cleansing, and data integration to achieve this goal. Klein et al. (2016) applied model averaging for analyzing macroeconomic data. Cerniglia, Fabozzi, and Kolm (2016) explained techniques for identifying good data for analysis. Quercia, Pennington-Cross, and Tian (2016) studied different approaches for the analysis of mortgage default. Sicherman et al. (2016) analyzed relationships between investor behavior and investment returns. Briguglio (2016) developed a weighting scheme to measure economic vulnerability as an index. Bin (2017) explained the concept behind new market failure theory and techniques for predicting economic crisis. Alpanda (2013) identified risks in the stock market based on business cycles. Alpanda (2013) measured the debt financing ratio between financial and non-financial corporations. Alpanda (2013) calculated the root mean squared error of the forecasts. Harry M and Nilufer

(1996) applied the principles of entrance to calculate the likelihood of stock market return distributions. Sharma (2017) explained Bayes Estimation combining a Weibull Distribution.

Research has also consisted of analyzing investor behavior, market behavior, and simulating market activity (Barmish & Primbs, 2016; Lambert, Ostrovsky, & Panov, 2014; Mauri & Tettamanzi, 2012). Lambert et al. (2014) analyzed trading behavior and the relationship between investment behavior and share prices in complex markets. Mauri and Tettamanzi (2012) implemented simulations in evaluating how investment behavior influences financial markets. Goswami, Han, Wang, and Jiang (2015) studied how controlled experiments such as A/B tests have improved the ability of e-commerce companies to market their products. Barmish and Primbs (2016) proposed techniques for testing new trading algorithms through simulation. Sim, Sim, and Hang-Hui (2014) implemented a standard deviation approach to forecasting volatility in commodity markets. Berndt, Boogers, Chakraborty, and McCart (2017) applied agent-based modeling to corporate bond markets. Lee, Kweon, Kim, and Chai (2017) described two measures for determining the economic value of information technology investments, the return on assets and the return on stock price. Lee et al. (2017) applied cumulative abnormal returns to measure the returns on the stock price. Lee et al. (2017) concluded that investors in stock markets positively consider investments in big data analytics. Linton and Mahmoodzadeh (2018) evaluated risks and opportunities in high-frequency trading. Cooke (2017) explained how high-frequency trading has influenced the development of new economic zones.

Tigani and Saadane (2018) analyzed historical currency exchange data to predict the best hours to trade currencies. Ha, Lee, and Moon (2016) constructed a rule-based algorithm for identifying patterns in stock market prices. Yelin, Hui, and Junjie (2017) applied financial theories and gradient boosting decision trees for developing investment strategies. Al-augby,

Majewski, Nermend, and Majewska (2016) proposed a decision support system method for choosing investments based on text mining. Neto and Ferreira (2016) implemented a genetic algorithm and an artificial neural network to forecast stock market prices. Dong, Yang, and Tian (2015) applied autoregressive moving averages and Glosten-Jagannathan-Runkle models for predicting high-frequency stock market data. Boonpeng and Jeatrakul (2016) tested one-against-one and one-against-all multi-binary classification methods for developing investment strategies.

Algorithms that have been implemented in financial analysis include artificial neural networks, k-nearest neighbor, discrete wavelets transform, and support vector machines (Ahmed, Choudhury, & Uddin, 2017; Golbayani & Bozdog, 2017; Samarakoon & Athukorala, 2017). Samarakoon and Athukorala (2017) described different techniques for anomaly detection in machine learning such as artificial neural networks with Boltzmann machines and hidden Markov models that can identify the root causes of faults in unsupervised machine learning. Samarakoon and Athukorala (2017) implemented three classification algorithms for detecting anomalies in stock markets, C4.5, Naïve Bayes, and ensemble techniques. Yang, Yang, and Li (2017) implemented multifractal detrended fluctuation analysis for financial time series analysis of share price data. Yang et al. (2017) obtained data from Yahoo finance for the stock market share prices. Das, Sadiq, Mirjalili, and Noraziah (2017) applied the grey wolf optimizer classification algorithm and the nonlinear autoregressive exogenous algorithm for predicting stock prices. Golbayani and Bozdog (2017) implemented the zonoid trimming algorithm for high-frequency financial data. Tang et al. (2018) implemented principal component analysis and k-nearest neighbor algorithm for predicting currency exchange rates. Raimundo and Okamoto (2018) implemented the discrete wavelets transform to predict currency exchange rates. Ahmed et al. (2017) explained a number of methods that can be implemented in evaluating stock market

anomalies such as the k-nearest neighbor, local outlier factor, connectivity outlier factor, principal component analysis, and support vector machines.

The effects of news and social media has also been analyzed to surface relationships between what is reported in the media and the performance of the stock market (Cerchiello & Nicola, 2018; Li, Chen, Wang, Chen, & Chen, 2018; Raschid, Xu, & Zotkina, 2017). Li et al. (2018) correlated social media feeds with the value of the DJIA with self-organizing fusion neural networks. Li et al. (2018) applied support vector matrices and k-nearest-neighbor algorithms for studying the impact of news on stock price movements. Cresci, Lillo, Regoli, Tardelli, and Tesconi (2018) studied the effects of microblogs on the prices and traded volumes of stocks in the financial markets. Cerchiello and Nicola (2018) provided a structural topic model linear discriminant analysis of financial news topics. Raschid et al. (2017) implemented topic models for discovering topics of discussion for financial institutions. Wu, Zheng, and Olson (2014) provided a technique for evaluating sentiment of posts about stock markets. Wu et al. (2014) identified a correlation between stock price volatility and stock forum sentiment. Cortez, Oliveira, and Ferreira (2016) measured the correlation of two technology firms and daily sentiment in social media. Baralis et al. (2016) addressed the issues of analyzing stock trading news information that is written in different languages.

Analysis of influences on the stock market have also included the application of algorithms to simulate the creation of patterns that may be predictive of trends that impact stock market prices (Cazzoli, Sharma, Treccani, & Lillo, 2016; Rathod & Garg, 2017; Siddiqui & Patil, 2017). Siddiqui and Patil (2017) proposed a technique for simulating the value of stock prices with Monte Carlo simulations. Sekar, Srinivasan, Ravidiandran, and Sethuraman (2017) explained a number of techniques for predicting stock market prices such as hidden Markov

models, artificial neural networks, fuzzy logic, and support vector machines. Vitt and Xiong (2015) evaluated the relationship between patent activities and stock prices for technology companies. Jin et al. (2017) implemented naïve Bayes models for predicting financial market fluctuations based on social media. More, Rathod, Patil, and Sarode (2018) applied the Naïve Bayes theorem for predicting stock market data with the Hadoop framework. Cazzoli et al. (2016) described a number of studies that apply sentiment analysis of social media data for stock market prediction.

Researchers have applied a number of methods, algorithms, and simulations to examine the behavior of the stock market as impacted by environmental factors (Alpanda, 2013; Klein et al., 2016; Sicherman et al., 2016). Some relationships between market performance and different environmental factors have been found (Costa & Silva, 2018; Lee & Paik, 2017; Sandhu & Mankotia, 2018). Costa and Silva (2018) applied sentiment analysis and supervised machine learning to demonstrate a relationship between positive news and social media feeds about a company and stock performance. Lee and Paik (2017) identified that there was not a strong relationship between real-time social media feeds and stock performance. Sandhu and Mankotia (2018) describe the progressive relationship between the stock market and economic growth and between the Indian stock market and other markets.

Statistical Analysis

Different methods for estimating parameters for statistical error include generalized least squares and maximum likelihood (Moreno, Harwell, Guzey, Phillips, & Moore, 2016). The generalized least square model is based on the statistical technique, generalized least squares (Liu, Okui, & Yoshimura, 2016). In statistics, the generalized least squares model is a technique for estimating unknown values in regression models (Liu et al., 2016). The two-step generalized method of moments method provides a covariance matrix and standard errors that can auto-

correlate the estimation value (Álvarez-Ayuso, Condeço-Melhorado, Gutiérrez, & Zofío, 2016). Erdinç and Gurov (2016) implemented the generalized method of moments estimation to identify regression in financial data. Erdinç and Gurov (2016) evaluated the correlation between ratings and the data. Racicot, Rentz, and Kahl (2017) also implemented both the generalized method of moments method and the ordinary least squares estimates in regression analysis and correlation of financial data.

Liu et al. (2016) proposed a method for averaging generalized least squares estimators for linear regression models by applying a weighted average of generalized least squares estimators. Kalkan and Kelecioğlu (2017) listed additional types of analysis such as principal component analysis, principal axis factoring, unweighted least squares, alpha factoring, and image factoring techniques. Jones, Lomas, Moore, and Rice (2016) listed semiparametric models as extended estimating equations, finite mixture models, and conditional density approximation estimators. Jones et al. (2016) described generalized linear models as a predecessor for extended estimating equations.

Liu, Okui, and Yoshimura (2013) demonstrated that generalized least squares estimators can provide better prediction than ordinary least squares estimators as generalized least squares estimators have smaller variances. Liu et al. (2016) explained that the generalized least squares estimator is also more efficient than the ordinary least squares estimator. Generalizability error can occur when the estimate is based on a sample from a different population than the actual population (Rhoads, 2014). Sampling error can occur when the sample does not accurately represent the population (Rhoads, 2014).

Outliers have a great effect when implementing maximum likelihood inference with generalized linear models and generalized linear mixed models (Chambers, Salvati, & Tzavidis,

2016). Fitting these models with variations of Monte Carlo methods of estimation has been developed for handling outliers (Chambers et al., 2016). Semiparametric and nonparametric regression techniques such as generalized additive models or local linear kernel regression can be implemented to disregard observations that are not close to the cutoff (Kim & Steiner, 2016).

Yongzhong, Danping, Min, and Levy (2017) described the advantage of maximum likelihood estimation to support consistency, asymptotic normality, and efficiency. Some of the recent developments in maximum likelihood parameter estimation include Lagrange multipliers, least squares methods, and Monte Carlo expectation maximization algorithms (Yongzhong et al., 2017). Gibbons, Richards, Valderas, and Campbell (2017) explained how algorithms can be combined to train datasets for machine learning. Algorithms that can be combined can include random classification, regression forests, classification and regression trees, maximum entropy, and generalized linear models (Gibbons et al., 2017).

Schochet (2013) implemented a generalized estimated equation that finds parameter vectors that minimize the weighted differences and allows for within-group correlations. The generalized least squares estimator can be implemented for random effect designs with continuous outcomes (Schochet, 2013). Álvarez and Griffin (2016) described a new process for regression coefficient estimation implementing partial least squares. Álvarez and Griffin (2016) combined both partial least squares and singular value decomposition methods.

Kisi and Ozkan (2017) described multiple linear regressions as having two or more independent variables against one dependent variable. The linear multiple regression model has coefficients where the response is a continuous variable and predictors are discrete or continuous constants (Kisi & Ozkan, 2017). Mohan (2016) described the locally weighted scatterplot smoothing (LOWESS) estimator as a non-parametric estimator directed by a smoothing constant.

Al-Badarenah and Alsakran (2016) explain the k-nearest-neighborhood algorithm in collaborative filtering. LOWESS estimation can have applications in prediction and machine learning. Kisi and Ozkan (2017) listed various types of linear regression analysis such as simple linear regression, multiple linear regression, logistic regression, ordinal regression, multinomial regression, and discriminant analysis. Amigud, Arnedo-Moreno, Daradoumis, and Guerrero-Roldan (2017) provided the steps of machine learning as collecting the raw data, processing data, extracting features, computing feature importance, reducing feature space, splitting data, training classifiers, testing classifiers, and reporting accuracy.

The linear multiple regression model also has many applications in education. da Costa, Pelissari, and Gonzalez (2018) investigated the relationship between different factors in distance learning education. Sinharay (2016) described various regression problems in education such as predictive validity studies, prediction of test scores, classification essay scoring, high-stakes testing prediction, course placement, prediction of dropouts, and prediction of program completion. Bırcan and Sungur (2016) applied multiple linear regression analysis for the prediction of student achievement based on factors of self-efficacy, task-value, and cognitive engagement. Immekus and Atitya (2016) applied hierarchical multiple linear regression in the evaluation of assessment score data.

Tan Boon (2016) explained how the philosophy of Occam's razor can choose the simplest as the correct theory of competing theories when hypothesis testing. Mjöberg (2018) explained parsimony as a scientific method to choose the simplest explanation that fits the evidence. Rossmo (2016) explained that Occam's razor can also be an excellent investigative technique, focusing on the most probable hypothesis in light of evidence. Drton and Plummer (2017) explained that prior distributions have a great influence on learning coefficients causing

the Occam's razor effect when integrating parameters. Briguglio (2016) explained that this principle can be applied to decisions on a weighting scheme.

Sicherman et al. (2016) provided an application of Occam's razor in financial modelling. McNally, Martin, Honig, Bergmann, and Piperopoulos (2016) promoted developing more parsimonious scales in measuring entrepreneurial intentions. Ghosh, Julliard, and Taylor (2017) applied Occam's razor to asset pricing. Khoja, Chipulu, and Jayasekera (2016) applied a parsimonious logit model for analyzing corporate insolvency.

Some models are sensitive to parsimony such as root mean square error of approximation which will have lower values with fewer parameters and more degrees of freedom (Smith, Fulcher, & Sanchez, 2017). Deacy, Gayes, De Lurgio, and Wallace (2016) compared the comparative fit index and parsimonious normed fit index for evaluating models. Ting-Chia and Gwo-Jen (2017) applied various two-group structural models and showed that parsimony normed fit index provided a better fit value for two-groups than the normed fit index, non-normed fit index, comparative fit index, and relative fit index. Brown and Lambert (2016) proposed a smooth parameter count model that creates parsimonious parameterization of coefficients.

Drton and Plummer (2017) explained that hierarchical models include order-restricted parameters and multivariate variance components models. Hierarchical models include a poset structure with active or inactive inequalities (Drton & Plummer, 2017). Leppink and van Merriënboer (2015) applied hierarchical linear modeling to compare the regression of a control group and an experimental group. Miller and Goodchild (2015) explained that data-driven models can be too complex and argue that relying on the data for the explanation may still be a future stage of computers.

Li (2016) explained logistic regression models as a generalized linear model that studies the relationships between a dichotomous outcome variable and predictor variables. Khoja et al. (2016) explained the criteria for logistic regression to include the lack of restrictive distributional assumptions, high confidence in prediction accuracy, and the ability to interpret the significance of the ratios. Coughenour, Paz, de la Fuente-Mella, and Singh (2016) observed logit models generated better goodness of fit than probit models.

Sato, Takano, Miyashiro, and Yoshise (2016) described binary classification methods as classic discriminant analysis, logistic regression, and support vector machine. Vaitheeswaran, Subbiah, Ramakrishnan, and Kannan (2016) compared binary logistic regression and ordinal logistic regression and explain that ordinal logistic regression can evaluate the significance of predictor variables. Ocampo-Vega et al. (2016) discovered that artificial neural networks had more predictive capabilities than decision trees and logistic regression models.

Hsieh, Lee, Li, and Tu (2016) applied a randomized response technique for evaluating a maximum likelihood estimator and logistic regression to evaluate the prevalence of the estimation. Louzada and Granzotto (2016) proposed a transmuted log-logistic regression model based on a quadratic transmutation map. Son, Brennan, and Zhou (2016) described a mixed effects logistic regression model with a logarithm of odds as the logit function. Li (2016) proposed a semiparametric logistic regression model that can test for linearity by issuing a penalty in the likelihood ratio test statistic.

Rembart and Soliman (2017) explained logistic regression as the most common tool used in credit scoring. Bates, Saria, Ohno-Machado, Shah, and Escobar (2014) explained that health care systems often choose decision trees and logistic regression because they can be implemented quickly in health care. Yu (2017) applied logistic regression models for capital

asset pricing. Raja, Çiçek, Türkoğlu, Aydin, and Kawasaki (2017) applied logistic regression for determining landslide susceptibility and describe binary logistic regression as the most common technique for this purpose.

Brady and Irwin (2011) described spatial econometric models and methods that can evaluate data based on spatial weights matrices or isolate causal effects from unobserved spatially correlated variables. Flaster and DesJardins (2014) explained the regression discontinuity model as a tool that can estimate treatment effects by modelling the relationship between control groups and treatment groups to find the probability of unknowns. Chambers (2016) explained the advantages of regression discontinuity design techniques to ANCOVA. A regression discontinuity model can show how the effects of the survey can be estimated among other factors that may influence decisions for technology.

Gashler, Smith, Morris, and Martinez (2016) described an unsupervised backpropagation that allows for imputing missing values in classification. Scantamburlo (2016) described the main goal of classification is to develop a classification rule. Elsayed and Qasem (2014) made the distinction between inductive and deductive clustering in classification. Wang et al. (2016) described how classification can be implemented in feature selection and techniques for measuring the classification ability of each feature. Gibbons et al. (2017) implemented a classification in text mining.

Hadioui, El Faddouli, Touimi, and Bennani (2017) presented a study that shows how classification can analyze large datasets. Hadioui et al. (2017) reviewed studies that that combine a Naïve Bayes classification with predictive analysis. Depren, Aşkın, and Öz (2017) compared various classification algorithms for their accuracy. Depren et al. (2017) compared a decision tree, a Bayesian network, and artificial neural network algorithms. Depren et al. (2017)

compared Kappa statistics, mean absolute error, root-mean-square error, true positive rates, false positive rates, Matthews correlation coefficients, f-measures, receiver operating characteristics, and precision-recall curves.

George, Janaki, and Parameswaran Gomathy (2016) described local polynomial regression to identify non-linear relationships compared to traditional linear regression. Local polynomial regression identified nearest neighbors of predictor points for the noted bandwidth (George et al., 2016). Brabanter et al. (2013) described two methods for nonparametric derivative estimation, which are smoothing splines and local polynomial regression. Brabanter et al. (2013) mentioned challenges in choosing smoothing parameters. Local polynomial regression fits polynomial models to subsets of data points using weighted least squares. George et al. (2016) compared local polynomial regression to multiple linear regression and artificial neural network models and found that local polynomial regression performed better in geological forecasting.

Advantages for local polynomial regression are that it allows the estimation locally at a point without the influence of outliers and can calculate non-linear relationships at each estimation point (George et al., 2016). Local polynomial regression has many applications in analyzing images. Tan, Sun, and Pham (2016) applied local polynomial regression in approximation for image filtering. Kachouie, Gerke, Huybers, and Schwartzman (2015) explained local polynomial regression in analyzing satellite images. Bharath and Rajalakshmi (2016) applied local polynomial regression in analyzing ultrasound images. Rong, Wenjing, Zhang, Nguyen Quoc, and Akcan (2016) also implemented local polynomial regression in determining shipping routes.

Big Data Technologies

As emerging data standards are available on the market, there are more formats for storing data (Demchenko, Turkmen, de Laat, Blanchet, & Loomis, 2016). Blockchain technologies, such as bitcoin, are similar to a distributed system as an accessible record of transactions (Glaser, 2017). Transactions broadcast to the entire network. New transactions check the blockchain first to ensure availability (Hanifatunnisa & Rahardjo, 2017). Bitcoin implements public-key technology to prevent double spending (Chuen, 2015). Blockchain technology has also introduced new studies in cryptology such as cryptocurrency (Hanifatunnisa & Rahardjo, 2017). Cryptocurrency is a digital alternative to currency. Account holders verify transactions through a peer-to-peer system. To prevent fraud, digital transactions sign with a private key. Similar digital currencies to bitcoin are Geist Geld, Tenebrix, and Litecoin (Chuen, 2015).

The increasing amount of data storage capabilities provides new opportunities for data analysis (Demchenko et al., 2016). The amount of data formats has been increasing creating a demand for such tools. An increase in network capabilities provides opportunities for data transactions to occur in real-time or as streams. All these factors combine to form new opportunities for data analysis in big data. The development of relational databases and NoSQL databases as created new opportunities for research design (Xu et al., 2016).

Various statistical methods exist for calculating quantitative studies in science. de Siqueira Santos, Takahashi, Nakata, and Fujita (2014) compared and classify correlation and dependence in biological studies. de Siqueira Santos et al. (2014) classified under the categories monotonic, non-monotonic, functional, non-functional, small sampling, and large sampling. de Siqueira Santos et al. (2014) described methods for correlation such as Pearson, Spearman, and Kendell, and classify them as linear or non-linear monotonic. de Siqueira Santos et al. (2014)

classified distance correlation methods as HHG measure, Hoeffding's D measure, and MI, and classify them to identify non-monotonic and non-functional relationships.

Corcuera (2012) described implementations of Malliavin calculus for the purpose of finding new results to stochastic processes. Malliavin calculus differs from variable calculus and allows derivatives to be calculated from random values. Corcuera (2012) defines Malliavin calculus as an infinitely dimensional differential calculus. Corcuera (2012) also calculated a Mallavin derivative as a Fréchet derivative. A Fréchet derivative is a derivative defined on a Banach space, a vector space of real or complex numbers.

Sagar, Nerurkar, and Sharma (2010) explained the advantages of fuzzy logic over data-driven techniques for estimating the reusability of components. Liu, Yu, Miao, and Kot (2013) described fuzzy logic as a rule-based system. As some of the factors of the system may not be linear, these rules may be nonlinear functions that represent failure probability or component lifespan. Liu et al. (2013) described how a Gaussian fuzzier maps values to a fuzzy set. Flores and Pineda (2016) identified the benefits of fuzzy logic in approximating the behavior of a system.

Keshavarz and Abadeh (2016) described how to implement Bayes classification in subject classification. Keshavarz and Abadeh (2016) tested an implementation of Bayes Network with genetic algorithms. The Bayes Theorem is based on probabilities of prior occurrences and likely occurrences. Overholt and Ezekoye (2015) explained how Bayesian interface models can be used in identifying the probabilities of various scenarios. Priyadharshini and Arivazhagan (2014) described a Naïve-Bayes Nearest Neighbor classifier that uses nearest neighbor distances in trees for image recognition. Vateekul, Kubat, and Sarinnapakorn (2014) also described a similar technique for the Naïve-Bayes Nearest Neighbor classification in hierarchical tree

classification. Sindhu, Riaz, Aslam, and Ahmed (2016) explained mixture distributions and implement the Gumbel model for estimating product life with classical and Bayesian estimation. Sindhu et al. (2016) explained the estimation to be used widely in engineering and business for environment, reliability, and life testing. Sindhu et al. (2016) proposed extending the study to other distributions such as inverse Weibull and inverse Rayleigh distributions.

Bouades, Kaszczyc, and Pellissier (2015) presented how managing one to one relationships in types can increase overall system performance when there are a high number of transactions in a bank. Kim, Hwang, Jae, Jun, and Kwon (2016) discussed some of the techniques that can be implemented to introduce biometrics into financial architectural layers. Kim et al. (2016) described a message queue for managing authentication during transactions. Halstead (2015) tested the performance of Field Programmable Gate Arrays (FPGA) that can be implemented in graphical processing units. Radojković et al. (2016) mentioned thread assignments should be cognizant of the hardware architecture. Three common types of thread assignments include manual thread assignments, performance prediction assignments, and load balanced assignments.

One method for describing systems is with a diagram known as a Petri Net. A petri net has four components, the sets of places, the transitions, flow relations with arcs, and mapping. Hong, Hou, Jing, Wang, and Litvin (2015) described computational complexity, structural complexity, and behaviors as factors to consider when designing a deadlock prevention mechanism. Hong et al. (2015) explained how petri nets can be implemented in deadlock prevention. Sandén (2011) described how simultaneous access to resources creates the potential for deadlock. One example of deadlock is when two resources are waiting infinitely for each

other to free up a resource. Software designers can implement methods to prevent deadlock in their applications.

Algorithms can execute many calculations simultaneously with parallel programming (Fujimoto et al., 2017; Jammula, Chockalingam, & Aluru, 2017; Wang, Zhou, Li, & Wu, 2016). Solving problems with algorithms can solve optimization problems and also create opportunities for training systems. Machine learning algorithms can be effective for enhancing and validating system designs and prototypes (Kang & Lee, 2016; Nawaratne et al., 2017).

Li (2015) included definitions for solving a problem as a constraint region, a feasible region, projection of decision space, rational reaction set, and an inducible region. Li (2015) suggested that the crossover operator can be identified with the pivot algorithm and mutation can be selected by random indices of the parent set. Li (2015) proposed an algorithm that contains six steps, the initial population, fitness evaluation, crossover, mutation, evaluation, and stopping criterion. Li (2015) defined the initial population as matrices and mention that the pivoting algorithm can obtain the pivot algorithm with the simplex method. Hordijk, Hordijk, and Heidergott (2015) provided the six steps for their genetic algorithm, initialization, fitness, selection, crossover, mutation, and repeat. Hordijk et al. (2015) found a “near-optimal deterministic sequence” for the optimization.

Hordijk et al. (2015) also described stochastic models for solving Markov decision processes. Hordijk et al. (2015) described the need for simple random policies with good performance in server optimization. Hordijk et al. (2015) described the genetic algorithm as a stochastic search method modeled after natural evolution. Hordijk et al. (2015) developed a multi-server queuing system framework that can calculate wait times for customers in a queue.

Fowe, Nouiri, Ibrahim, Karambiri, and Paturel (2015) described many implementations of genetic algorithms that can be implemented in various optimization problems. These algorithms include various implementations of genetic algorithm models. Genetic algorithms include both sequential and sorting algorithms (Fowe et al., 2015). Fowe et al. (2015) optimized a number of variables in their model including factors and constraints that can be influential on the outcome of the system.

Candan and Yazgan (2015) described the steps of the genetic algorithm to include determining the job sequence, calculating the fitness value, implementing crossover and mutation methods, checking the threshold, and finding the best solution. Candan and Yazgan (2015) implemented the Taguchi method which may potentially increase the efficiency of implementation and evaluation of experiments. The implementation included five parameters, the derivation method, population size, crossover method, mutation rate, and crossover rate (Candan & Yazgan, 2015). Candan and Yazgan (2015) mentioned that space may not be optimized in the solution, but time and cost may be optimized.

Wei, Tang, and Liu (2015) showed advantages for graphs in genetic algorithm implementations to include ordered distribution of line segments, graphs connected with subgraphs, searching within graph structures, graphs to implement the depth of connections between nodes, and subgraphs consistent with genetic processes. Wei et al. (2015) proposed a novel crossover operation with varying probability in exchanges. Wei et al. (2015) had seven parameters for configuring the algorithm: initial population, initial partition size, crossover probability, mutation-deleting probability, mutation-appending probability, number of elite individuals, and selection method. Wei et al. (2015) implemented the statistical signal-to-noise ratio curve to determine that the results were consistent with the original graphs (Wei et al.,

2015). The experiments successfully constructed image edges and created object contours as vectors.

Tozan and Donmez (2015) described the steps of the genetic algorithm as initialization, evaluation, selection, exploration, exploitation, replacement, and termination. Tozan and Donmez (2015) suggested that the genetic algorithm can be implemented in solving deterministic problem situations. Tozan and Donmez (2015) described the advantage of the genetic algorithm as able to obtain a qualified and quick solution by not solving for the global optimal solution. Tozan and Donmez (2015) implemented the algorithm to find the maximum coverage of population and location.

Modern Applications of Conceptual Theories

There have been many advances in data analysis with improvements to textual analysis, data visualization, and parallel and distributed computing. Sun, Bie, and Zhang (2016) defined five operations for semantic relations to include generalization, addition, difference, complement, and inverse. Associate various relations for studying the semantic relational basis in scientific research for researchers, publications, emails, and publications. Sun et al. (2016) demonstrated that searching relevant information can be accomplished by increasing the connection between resources semantically and including relevant results from queries. Goudarzi and Pedram (2016) proposed an algorithm that improves efficiency for cloud data centers. The algorithm ranks solutions for virtual machines based on containers, racks, and chasses. The algorithm also optimizes virtual machines based on migrations and objectives. Goudarzi and Pedram (2016) also enhanced emergency operations by reducing over-provisioning of resources through monitoring and reactive decision-making capabilities. Goudarzi and Pedram (2016) stressed the cost of energy to be empirical for designing workload distribution and calculate how to optimize costs while minimizing disruption to the service level agreements.

Leijon, Henter, and Dahlquist (2016) explained how mutual information techniques quantify information received between recipients in communication. Values can be measured as the score of correct responses and the probability of correct responses. Leijon et al. (2016) explained methods of analyzing confusion data with Bayesian statistics, applying consonant recognition testing, mutual information simulations, and confusion matrices. Leijon et al. (2016) supported the Bayesian distribution over the null hypothesis to provide estimates of the conditional probability of observed data. Leijon et al. (2016) tested both singular and multiple conditions and approximate samples for data sets. Leijon et al. (2016) implemented the Dirichlet distribution model which is often applied in machine-learning algorithms. Leijon et al. (2016) calculated standard differences from observed matrices and test significance at various levels. Krulewich, Yin, Bundick, and Zeng (2015) tested parallel computing with supercomputing resources. Krulewich et al. (2015) described financial analysis data as high-frequency which can be analyzed with frequency equations. Krulewich et al. (2015) compared MPI Fortran with a multi GPU implementation and experimented with multiple GPU configurations. Krulewich et al. (2015) demonstrated that parallel computations can be implemented to compute stochastic volatility. Soares, Canizes, Vale, and Venayagamoorthy (2016) explained the potential of solving nonlinear programming problems that require a heavy work-load for solving energy resource management problems. These solutions are beneficial for managing power in energy storage systems. Soares et al. (2016) managed the demand response and the load control of power sources with optimization techniques and demonstrate the capabilities of nonlinear programming on data analysis systems.

Chen, Mao, and Liu (2014) explained hybrid data warehouses implementing various types of analysis including structured, network, mobile, text, and web data analysis. Chen et al.

(2014) also described traditional data analysis techniques that can be implemented including cluster analysis, factor analysis, correlation analysis, regression analysis, A/B Testing, Statistical Analysis, Data Mining Algorithms, Bloom Filter, and hashing. Chen et al. (2014) listed some of the challenges of big data to include data representation, redundancy reduction, data confidentiality, data life-cycle management, and analytics. Include sources of data to include log files, sensors, and network data. Chen et al. (2014) also mentioned some of the parallel processing models to include Message Passing Interface (MPI), MapReduce, and Dryad. Tian, ÖZcan, Tao, Goncalves, and Pirahesh (2016) explained many join algorithms that can be implemented in data warehousing. Many distributed file systems fetch tables and conduct joins in the database. Some distributed file systems conduct the joins in the file system, also known as a broadcast join. To increase selectivity in certain instances, additional joins include repartition joins and zigzag joins (Tian et al., 2016). A repartition join transfers a database table to a distributed file system and conducts the join. A zigzag join algorithm alternates between the database tables and the distributed file system. Liu, Golab, Golab, Ilyas, and Jin (2016) analyzed and evaluate five big data frameworks including Matlab, PostgreSQL, System C, Spark, and Hive. Liu et al. (2016) discovered that C performs very well but requires a higher programmer skillset. Liu et al. (2016) also identified that PostgreSQL with MADlib and Matlab are user-friendly processing tools. Liu et al. (2016) found Hive more user-friendly than Spark.

Parallel and distributed computing play an important role in modern data warehouses (Krulwich et al., 2015). Join algorithms are a common task in data warehouses and can be improved with distributed file systems (Tian et al., 2016). Algorithms can improve the efficiency of data management in cloud centers (Goudarzi & Pedram, 2016). Business intelligence

platforms can be implemented in text mining, knowledge management, and semantic discoveries (Sun et al., 2016).

Influence of Big Data on Information Technology

Software-defined networking, cloud computing, infrastructure-as-a-service, platform-as-a-service and other advancements in information technology have improved service management and the ability to provide customer relationship management and business relationship management (Benzekki, El Fergougui, & Elbelrhiti Elalaoui, 2016; Chang, Kuo, & Ramachandran, 2016; Dayal, Maity, Srivastava, & Khondoker, 2016). Service management has become an integral part of business intelligence and enterprise resource planning solutions (Chin Lay & Balakrishnan, 2017). In a variety of industries, enterprise resource planning solutions have added organizational efficiency to existing processes (Chin Lay & Balakrishnan, 2017; Dayal, Castellanos, Simitsis, & Wilkinson, 2009).

This research demonstrated how algorithms and statistics can promote the acceptance of big data technology in information systems and data analytics (Davis Jr, 1986; Hartmanis & Stearns, 1965; Venkatesh et al., 2003). Big data analysis implementations can be improved with a greater technological adoption of big data (Abbasi, Sarker, & Chiang, 2016; Wamba et al., 2015). Kandogan, Balakrishnan, Haber, and Pierce (2014) described the improvements to create data-driven enterprises that can bridge the divide between people and data. Kandogan et al. (2014) mentioned new research in scientific fields such as human-computer interaction, information visualization, information management, computer storage, and service sciences. Hancu (2012) presented a model for sharing intelligence among businesses. The model relies on building trust between businesses and providing reliable and reputable services. Hancu (2012) suggested future collaborative systems may include pricing information, stocks, and human resource information.

Conceptual Framework

The conceptual framework for the research described how statistical analysis and financial analysis identified the value of big data to publicly traded companies. Figure 1 displays a general overview of how the statistical analysis and the financial analysis can identify the value of big data to publicly traded companies. The combined framework can be a basis for the adoption of big data technologies by demonstrating how organizations have been able to profit from implementing big data technologies.

Big data is often classified by the V's that represent an assortment of characteristics of big data such as variety, value, volume, veracity, and velocity (Flood, Jagadish, & Raschid, 2016). Variety refers to the different types of data sources. Conventional data usually consisted of relational data stores which were primarily SQL data. Big data can represent a variety of data sources including SQL, NoSQL, JSON, XML, structured, and unstructured data sets (Gadepally, Kepner, & Reuther, 2016).

The value represents the value of data to the organization. Big data has increased the monetary value of data and the need for protection of the data (Bucur, 2015). Data breaches cost organizations more and more as the amount of personal data that is being stored is increasing (Bumblauskas, Nold, Bumblauskas, & Igou, 2017). It is necessary for companies to assess the value that their data has to the organization and understand the value lost when that data is breached (Chen et al., 2014).

The volume of data represents the amount of data being stored (Shneiderman & Plaisant, 2015). Conventional data represented data that was primarily measured in GB and TB (Dede, Govindaraju, Gunter, Canon, & Ramakrishnan, 2013; Miyauchi, Takahashi, & Kinomura, 2015). Big data has introduced higher measurements of data including petabytes, exabytes, and zettabytes (Abdulkadri, Evans, & Ash, 2016; Graham et al., 2015). Solid state storage and shared

storage on commodity devices have increased the capability of storage systems and allowed for the storage and analysis of more data.

The veracity refers to the correctness of the data being collected and analyzed (Jain, Gyanchandani, & Khare, 2017). Chandra, Ray, and Goswami (2017) described the importance of trustworthiness in data sources and stressed the reliability of the data source. Chandra et al. (2017) also explain trustworthiness of the data becomes a challenge as the variety of the data increases.

The velocity of data represents the speed at which data is stored and analyzed (Venkatraman, Fahd, Kaspi, & Venkatraman, 2016). Conventional data often stored data in near to real-time transactions with some lag between the request for storage and the write to the disk (Miyauchi et al., 2015). This lag has often been an additional layer of data management to improve the reliability of transactions. With improvements to processing and storage computing, real-time transactions become more and more feasible and have lessened the need for conventional transactional management (Cai, Gallina, Nyström, & Seceleanu, 2016). The architecture of these systems has also changed with the advent of parallel and distributed computing architectures (Huang, Mozafari, Schoenebeck, & Wenisch, 2017).

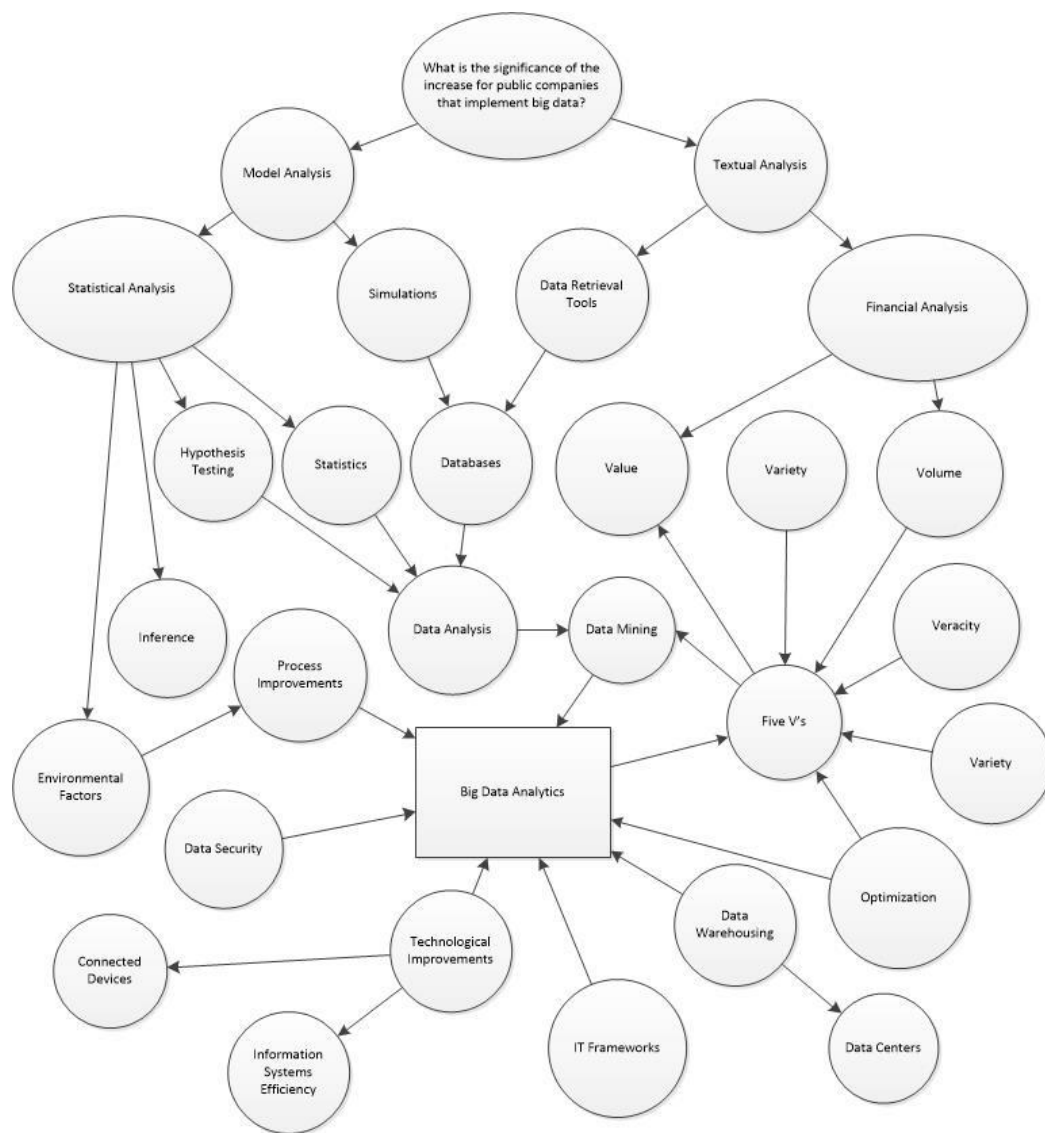


Figure 1. The components of financial analysis and statistical models and their influence on the development of big data technologies.

Summary of Literature Review

Research in the area of financial analysis of stock and currency markets, statistical simulation, and sentiment analysis of stock market data are examples of how research in this area has provided opportunities for evaluating big data in financial markets (Olaniyan, Stamate, Ouarbya, & Logofatu, 2015; Singh & Thakral, 2017). Uncertainties about measuring the increase

in value to the publicly traded companies that implement big data technologies highlight a potential research void that demonstrated the need for the proposed research. Chapter three will describe the methodology of the research of how the advance of big data technologies can potentially be measured by the increase in the stock market value of companies that include “big data “ in their periodic filings with the United States Securities and Exchange Commission.

CHAPTER THREE

The purpose of this research was to explore the extent to which there was a statistically significant increase, from 2013 to 2018, of the market value of companies that had “big data” in the company description. This chapter discusses the research methodology for the research of the value of big data analysis to public companies. The research tradition discusses the quantitative research tradition. The research question highlights the central research question of the research. The research design details the population, sampling, instrumentation, data collection, data analysis, and ethical considerations for the research. The chapter is completed with a summary of the chapter.

Research Tradition

The research tradition for the research was the quantitative tradition. Edmondson and McManus (2007) suggest a continuum for research that traverses from nascent research to intermediate research to mature research. Testable hypotheses and standard statistical procedures categorized the research as a quantitative study for the research continuum (Edmondson & McManus, 2007). The population for the study was publicly traded companies in the United States with “big data” in the description of the periodic filings with the United States Securities and Exchange Commission. The topic was the value of big data analysis to publicly traded companies (Wamba et al., 2017).

The research methodology and design of hypothesis testing and statistical inference were chosen for this research were empirical methods for measuring the impact that big data technologies have had on publicly traded companies in the United States (Wamba et al., 2017). Other methodology options were not chosen because, with the variety of technology tools available, there was no other methodology or design that would provide a far reaching assessment of the publicly traded companies (Durcikova, Lee, & Brown, 2018). The

methodology design provided a rigorous statistical approach to the research opportunity and purpose of identifying the value of big data technologies to publicly traded companies (Durcikova et al., 2018).

Research Questions and Hypotheses

The central research question for this study was:

RQ1: To what extent is there a statistically significant increase, from 2013 to 2018, of the market value of companies that had “big data” in the company description, which is part of the public filing document registered with the United States Securities and Exchange Commission?

Hypotheses

H1₀: There is no statistically significant increase, from 2013 to 2018, of the market value of companies that had “big data” in the company description, which is part of the public filing document registered with the United States Securities and Exchange Commission.

H1_a: There is a statistically significant increase, from 2013 to 2018, of the market value of companies that had “big data” in the company description, which is part of the public filing document registered with the United States Securities and Exchange Commission.

Research Design

The research design included a quantitative analysis that identified any statistically significant differences between the stock price of publicly traded companies that had “big data” in their periodic filings with the United States Securities and Exchange Commission. The intent of the study was to identify if big data analysis has increased the value of the publicly traded companies that implement big data technologies (Gepp, Linnenluecke, Terrence, & Smith, 2018). The intent was aligned with the problem by collecting information on how public companies make decisions about technology (Gepp et al., 2018). The research design was a quantitative correlational analysis consisting of hypothesis testing and statistical inference

(Franke et al., 2016). The dependent variables in the study were the percent change of the stock price for the five year period of 2013 and 2018. The independent variables were the sectors, industries, market capitalization, and markets for public companies. Nichols et al. (2017) described techniques for optimizing a correlational model by including all covariates in a single model. Nelson (2017) highlighted that correlation is not necessarily causation. The proposed design accomplished the research goals by identifying patterns and relationships in the observed data (Nelson, 2017). Ha et al. (2016) provided a similar study of identifying correlational patterns in a stock market analysis.

This proposed research fit with other research in computer science that identified how technology is adopted and how to improve the technology of data analysis (Pusala et al., 2016; Wamba et al., 2015). Potential issues for this research were that some of the technologies were relatively new and continued to change and develop (Weinberg, 2016). New technologies were introduced and replaced outdated technologies. There were also varying perspectives on how these new technologies may have been implemented (Weinberg, 2016). Controversies in the field included security issues of data analysis, privacy issues related to data, and ethical concerns relating to data analysis (Fox & Vaidyanathan, 2016; Glicksman et al., 2017).

Population and Sample

The population of the study consisted of public companies in the United States that included “big data” in their periodic filings with the United States Securities and Exchange Commission. The data was collected from the Electronic Data Gathering and Retrieval system of the United States Securities and Exchange Commission (U.S. Securities and Exchange Commission, 2018). Data was also collected from the major markets (Nasdaq.com, 2018). There were 204 publicly listed companies that fell under this category. The sample for this study was the population as the data is available publicly.

Sampling Procedure

Access to the entire population was available through information that was available from the markets of the NASDAQ and the New York Stock Exchange (Nasdaq.com, 2018). The data was accessed with spreadsheets and raw data available from the market web sites (Nasdaq.com, 2018). The public companies with “big data” in the description of their periodic filings with the SEC that were listed on the major exchanges were selected for the population. There were approximately 6,800 publicly traded companies on the three major exchanges (Nasdaq.com, 2018). Of the 6,800 publicly traded companies, 377 unique companies had “big data” in their periodic filings with the SEC during the time period of 2013 and 2018 (Nasdaq.com, 2018; U.S. Securities and Exchange Commission, 2018). The listings by the major exchanges also provided the sector and industry for the companies (Nasdaq.com, 2018).

Instrumentation

Lodewyk and Mandigo (2017) explained how an instrument can be validated based on its application. This instrument was designed for the specific purpose of evaluating the value for public companies with “big data” in the description of their periodic filings during the period of 2013 and 2018. The instrument was an analysis of the public information for the publicly traded companies in the United States that had “big data” in the description of the periodic filing of the United States Securities and Exchange Commission. Raw data was collected from publicly traded companies that were listed on the two major exchanges, the New York Stock Exchange and the NASDAQ Exchange (Nasdaq.com, 2018). Of the approximate 6,800 publicly traded companies, a textual analysis analyzed descriptions for the public companies with the Electronic Data Gathering And Retrieval system of the United States Securities and Exchange Commission (U.S. Securities and Exchange Commission, 2018). Of the public companies, companies with “big data” in their company description were analyzed for their stock price for the period from

2013 to 2018. Some companies were delisted in the past five years but that represented a small percentage of companies. The comparison included the value of their stock price, the industry, the sector, and market capitalization. The analysis also compared the sub-sectors with other sub-sectors to evaluate the significance of any outperformance for a period of five years. The analysis identified components that have influenced the advance of these companies and how big data technologies may have increased the value of these organizations. The analysis points for the raw data are listed in Table 1.

Table 1
Analysis Points

Data Point	Responses
Company Name	
Company Listing	
Exchange	[NASDAQ] [NYSE]
IPO Year	
Stock Price	
Industry	
Sector	
Market Capitalization	[Small-Cap] [Medium-Cap] [Large-Cap]

All of the data was collected and analyzed on a Lenovo PC with a dual-core processor, 8 GB of RAM, and a 750GB hard drive. Preliminary tests found that the PC provides adequate processor and storage capabilities for the research. The data was analyzed with RStudio 1.1.383. RStudio is an open-source data analysis tool that includes many statistical libraries (Fotache, 2016; Kamakshaiah, 2013). The RStudio application was capable of a variety of statistical functions such as correlation, ANOVA, and regression (Fotache, 2016; Kamakshaiah, 2013). The researcher was the only person with physical and logical access to the machine.

Two programming languages that were useful for data analysis are R and Python. R is a programming language that was specifically designed for data analysis (Balan & Perez-Carballo, 2016). The R programming language was capable of storing data in structures such as data frames or tables (Gwynllyw, Weir, & Henderson, 2016). These tables could be analyzed with statistical functions native to R or with a large number of libraries that have been developed for the R programming languages (Gwynllyw et al., 2016; Muller & Kidd, 2014). Relational database developers have also created packages for R to integrate with relational databases. R also included a library for connecting to JDBC connections to connect to almost any data source (Lightstone et al., 2017).

Amigud et al. (2017) described the SciKit toolkit available for Python. Luukko, Helske, and Räsänen (2016) explained the statistical tools available in both R and Python. Both R and Python had machine learning libraries which could be used for training data sets for analysis (Arcila-Calderón, Barbosa-Caro, & Cabezuelo-Lorenzo, 2016). R and Python also had text mining libraries for text mining (Arcila-Calderón et al., 2016; Rose & Lennerholt, 2017). An advantage of R was that it was designed for data analysis and easy to learn. Another advantage of R was that the data frames were a useful tool for the storage of data sets (Gwynllyw et al., 2016). A disadvantage for R was that the syntax could be difficult to learn when accustomed to other programming languages. Another disadvantage was that it may be tough to implement in a production system.

Programming languages have developed many features with the availability of open source developers to provide continuous improvements to the language (DeLuca, 2016; Eckert, Meyer, & Stuermer, 2017). R had developed as an open source alternative to commercial statistical packages such as SPSS and MatLab (Park, 2009; Razali, Shamsudin, Maarof, Hadi, &

Ismail, 2012). The R programming language had also developed as an alternative for the structured query language for business analytics (Fotache, 2016). A common format for datasets was the NetCDF format (Poorthuis, Goergen, Sharples, & Kollet, 2016; Zhuang, North, & Stevens, 2017). The data collection could have also implemented a NetCDF format for storing data.

Validity

Internal validity demonstrates that the particular practice has a causal relationship between variables (Schalock, Gomez, Verdugo, & Claes, 2017). The research demonstrated internal validity by identifying variables that have a relationship such as the market price and the market capitalization, or the sector and the industry. The research also demonstrated internal validity by providing a consistent source for the raw data.

External validity demonstrates that the practice can be generalized to other populations (Schalock et al., 2017). The research study demonstrated external validity by identifying a common strategy that was implemented by all public companies in the United States for measurement. The periodic filings and market values were a requirement for all public companies and are therefore an adequate variable for measurement.

Coutinho, Perrone-McGovern, and Gonçalves (2017) recommended methods for ensuring internal and external validity in clinical research such as beginning with a well-defined research question, starting with internal validity, controlling factors that can influence results, and reporting procedures for quality control. The implementation of the major indices, analyzing the data from the SEC and markets, and reporting the procedures applied this approach for the research (Coutinho et al., 2017).

To ensure validity, the data collection should be aligned with the research questions (Tavakol & Sandars, 2014). Stock market value was a valid measure for this research as the

market value represented the financial value of the company. The stock market value of the traded companies was representative of the value that investors have placed in these companies. Validity ensures that the instrument measures what it is supposed to measure (Tavakol & Sandars, 2014). In this research, the measurement of the stock market value represented a measurement of the value of the company. Levy, Livingood, and Maranga (2016) presented how to record demographic information, identify themes and conduct data analysis.

Reliability

Koehler and Meixner Jr (2016) proposed methods for establishing guidelines of reliability in forensic sciences. Maier, Hampe, and Jahn (2016) identified assessments for reliability such as test-retest, split-half, and parallel-test designs. To ensure reliability the research design should be systematic and consistent throughout the research process (Waidi, 2016). The same measurements were applied to all the participants in the research. Reio (2016) highlighted the importance of reliability in secondary data. The raw data was collected from the United States Securities and Exchange Commission and major markets to ensure the reliability of the data sources (Nasdaq.com, 2018; U.S. Securities and Exchange Commission, 2018).

There was a logical consistency for the research design, data collection, and data analysis. For reliability in the research analysis, a textual analysis of the company descriptions was implemented for ensuring that all companies were selected through the same criteria. The data collected for each company was for the same time period. The companies all received the same data points for collection and analysis.

Data Collection

The publicly available information for each company included historical data for the year from the time the company went public, the initial listing price of the company, the historical market capitalization, a description of the company, sectors for each company, industries for the

companies, the historical open, close, high, and low share price for the company stock, the 52-week highs and lows for the company, quarterly and annual revenue for each company.

The first step of the data collection procedure was to collect the data from the EDGAR tool of the SEC. The next step was to collect data from the major financial markets on publicly traded companies. The next step in the data collection process was to combine the data from the SEC with the data from the major markets. The final step in the data collection process was to pre-process the data for analysis with the R programming tool.

Pre-processing the data included converting all the data to formats for analysis. The ticker symbols were identified for each of the companies with the data from the markets. The textual content of the market capitalization was converted into numerical format. Variations of spelling and case were identified for the markets. The data was stored on a computer that did not have outside access by the Internet. The data was also backed up regularly.

The data was collected on a secure computer and all information about the companies was kept confidential. The data was collected with RStudio and PSPP. RStudio is an open source application that allows for the running of R programs. R is a programming language that has developed for the purpose of statistical analysis (Wu, Li, & Chang, 2015). PSPP is an open source program for statistical analysis (Arcila-Calderón et al., 2016; Wu et al., 2015). PSPP allows for data to be viewed in a table format. PSPP also has many statistical functions for the analysis of data. Another program that was available for statistical analysis is Libre Office Calc. Libre Office Calc is a spreadsheet program that can handle multiple file formats (Favario & Masala, 2017; Gurbuz & Aydin, 2017; Sankari & Bose, 2016). Once the data was analyzed, the resulting tables and charts were saved on a USB drive for transfer to another computer for

documenting and presenting the information. One year after final publication, all raw data will be destroyed.

Data Analysis

The research methodology for measuring the performance of the sector implemented Bayesian estimation (Haq & Al-Omari, 2016). Bayes Theorem predicts variables based on prior distributions and probabilities (Seal & Hossain, 2015). The information available on publicly traded companies was analyzed and combined with probabilities obtained from the various indices and evaluated with sectors, industries, and market capitalization (Huang, 2018).

Data analysis choices were justified by reporting analysis in detail (Köhler, Landis, & Cortina, 2017). The goodness of fit details were also provided for data analysis (Rosal et al., 2012). In data analysis, the classification of variables of measurement such as nominal, ordinal, interval, and ratio influence how the data can be analyzed (Tavakol & Sandars, 2014).

Siriwardena and Gillam (2014) suggested a logic model of baseline surveys, observation, recording, data analysis, post-intervention survey and outcomes analysis. Siriwardena and Gillam (2014) described an analysis approach of interrupted time series, pattern matching, cross-case synthesis, explanation building, and an explanatory logic model. Two types of analysis that were conducted were a classification for classifying the technology companies and a regression for testing the performance of the technology company data. One of the most accurate classifiers is the Bayesian belief network, a probabilistic network, which includes a directed acyclic graph and a conditional probability table. Eleye-Datubo and Wall (2008) described the fuzzy logic and Bayesian networks as some of the most effective methods for solving problems of uncertainty. Coussement, Benoit, and Antioco (2015) described a Bayesian approach to decision making for big data mining. Zhou, Tian, and Lim (2015) compared the naïve Bayesian, tree augmented naïve Bayesian, and fuzzy naïve Bayesian algorithms. Nikoo, Kerachian, Malakpour-Estalaki,

Bashi-Azghadi, and Azimi-Ghadikolaee (2011) combined fuzzy inference systems, Bayesian networks, and probabilistic neural networks.

The supervised learning model can provide guidance on statistical analysis by training the model with statistical tests. Landset, Khoshgoftaar, Richter, and Hasanin (2015) described three phases of supervised machine learning models which are, build the model, evaluate the model, and execute the model. Gibbons et al. (2017) described the importance of testing the efficiency of supervised training models. Pedro and Estevam R. Hruschka (2012) implemented a machine learning implementation that validates results with a rule learning algorithm.

A clear purpose in supervised learning is necessary for the data not to malign the original intention of the data analysis. Random sampling ensures that the data being sampled is unbiased. Radojković et al. (2016) implemented random sampling with the analysis of multiple core processors. With big data, it is possible to evaluate a large amount of data, but it is determined that all the possible population is available on the dataset. A statistical sampling of the entire dataset may not necessarily be significant if the dataset does not represent the entire population.

Haug (2016) discussed the importance of data quality and issues with analyzing big data such as integrating data from various sources and varying file formats. Haug (2016) also described how data may not necessarily be modeled in the format for it to be analyzed. Defining a clear purpose in data collection can ensure data quality for obtaining a random sampling for quality analysis of big data.

Ethical Considerations

Barsdorf and Millum (2017) explained how research should be socially valuable to be valuable. The Belmont Report underlines principles for respect for persons, beneficence, and justice (Goldstein, 2017). Resnik (2016) outlined the history of the Belmont Report with the passage of the National Research Act which led to the development of the National Commission

for the Protection of Human Subjects of Biomedical and Behavioral Research. Schnipper (2017) described how the Belmont Report developed after multiple high profile cases of abuses in human study research. The Belmont Report outlines the process that institutional review boards review studies involving human participants. The institutional review board is responsible for determining risks to human participants in research and that these risks are minimized (Schnipper, 2017).

Steps for considering ethical considerations for human participation in the research can include Belmont considerations of autonomy, justice, and beneficence (Ohayon, Cousins, Brown, Morello-Frosch, & Brody, 2017). This research does not include human participants so, therefore, does not include vulnerable populations (Ohayon et al., 2017). The research is aligned with ethical considerations such as choosing an ethical design, ensuring validity and reliability, and choosing a reliable test instrument (Creswell, 2014). The research design will also be submitted for IRB approval (Stark, 2011). The participants of this research, the public companies, will be kept anonymous and all data will be destroyed one year after the final publication of the research.

Moon (2017) described a fundamental principle of research ethics to be justice in the distribution of risks and benefits. Boronow et al. (2017) described the key components of research ethics, autonomy, justice, and beneficence, and the importance of potential harm and collecting data of uncertain consequence to participants. Goldstein (2017) discussed recent developments in changes to the code of federal regulations for the protection of human subjects. Annas (2018) proposed changes to federal codes on informed consent to allow for more genuine discussion between physicians and research subjects.

Schmidt and Schwartz (2016) described strategies and procedures for which ethics committees can arrive at decisions in research review. Bender, Cyr, Arbuckle, and Ferris (2017) described challenges for researchers in meeting the requirements of ethics research boards with varying regulations across disciplines.

Summary of Chapter Three

The research attempted to identify the value that big data analytics has had for publicly traded companies. The steps to quantitative data analysis include preprocessing the data, cleaning the data, finalize the data, presenting the data, and publishing the findings. The steps for research in computer science can include synthesizing a theory, specifying the process mechanisms, developing a simulation, conducting the research, and advancing the paradigm (Niazi, Hussain, & Kolberg, 2017; Zhang & Vorobeychik, 2016). The next chapter will present the findings of the research.

CHAPTER FOUR

This chapter provides the findings on the research on big data and publicly traded companies in the United States. The chapter starts with the results of the analysis that was conducted with publicly traded companies. The chapter then describes the findings of the analysis of market data. The chapter then explains the results of the simulation and modeling analysis. The chapter concludes with a summary of the research findings.

Participant Demographics

The analysis of the publicly traded companies in the United States was provided with the EDGAR tool of the Securities and Exchange Commission (U.S. Securities and Exchange Commission, 2018). The tool allowed the search of company filings for the past four years. A search for the term “big data” provided 1,070 results. Of the 1,070 results, 377 unique companies were identified. 246 companies appeared on the major exchanges of the New York Stock Exchange or the NASDAQ exchange. The NASDAQ exchange provides a listing of each company, its ticker symbol, its IPO date, and its last price.

The ticker symbol was obtained from the NYSE, NASDAQ exchange, and Yahoo! Finance, which maintain data on all the equities for the major markets and over the counter markets. The quantmod library in RStudio allows for historical retrieval of market data from a variety of sources including FRED, Yahoo! Finance, and Google Finance. Market data was obtained for a time span of January 2, 2013, to January 2, 2018, as a measure of how the market value of the company has increased over five years. After an analysis of the companies on the major markets that were available in 2013 and 2018 and removing outliers, the list of companies was processed to 165 companies.

Table 2 displays the sectors for the stocks that were obtained from EDGAR that are associated with big data. The sector data was collected from the markets. The majority of the

companies are in the technology sector with 81. There are also 23 companies in the consumer services sector.

Table 2
Sectors for Stocks Associated with Big Data

Sector	Count
Basic Industries	1
Capital Goods	13
Consumer Durables	2
Consumer Non-Durables	1
Consumer Services	23
Energy	4
Finance	14
Health Care	7
Miscellaneous	14
Public Utilities	8
Technology	81
Other	36

Table 3 displays the sectors and the industries for the public companies that are associated with big data. The largest number of companies is in the technology electronic data processing industry with 26 companies. The next largest number of companies is in the computer prepackaged software industry with 19 companies.

Table 3
Sectors and Industries for Stocks Associated with Big Data

Sector	Industry	Count
Technology	Advertising	1
Technology	Computer Communications Equipment	4
Technology	Computer Manufacturing	5
Technology	Computer Software: Prepackaged Software	19
Technology	Computer Software: Programming, Data Processing	4
Technology	Diversified Commercial Services	1
Technology	EDP Services	26
Technology	Electronic Components	4
Technology	Industrial Machinery/Components	3
Technology	Radio And Television Broadcasting And Communications Equipment	2
Technology	Retail: Computer Software & Peripheral Equipment	1
Technology	Semiconductors	11

U.S. Securities and Exchange Commission (2018) provides a full-text search of company filings for documents filed in the past four years. U.S. Securities and Exchange Commission (2018) provided 1,070 results for the full-text search of “big data” in 10-K filings, a filing required for public companies, within the past four years. From the 1,070 companies, 377 unique companies were found. Ticker symbols for publicly traded companies that list on exchanges were identified for these companies (Nasdaq, 2018; Yahoo, 2018). There were 207 companies that provided historical stock price values for January 2, 2013. There were 279 companies that provided historical price values for January 2, 2018. There were 138 companies listed in the NASDAQ exchange, and 108 companies listed in the NYSE. The other companies were listed on over the counter exchanges, pink listing exchanges, foreign exchanges, or exchange data could not be found. Table 4 displays the exchanges for which the companies associated with big data are listed.

Table 4
Exchanges for Big Data Companies

Markets	Count
NASDAQ	138
NYSE	108
OTC	42
PINX	19
No exchange data	69
STUTTGART	1
Grand Total	377

For the 279 companies, 207 had exchange data for both January 2, 2013, and January 2, 2018. There were 72 companies that had stock price data on January 2, 2018, and did not have data on January 2, 2013. There were four companies that had stock price data on January 2, 2013, and did not have data on January 2, 2018. There were 94 companies for which stock price data could not be found. Table 5 shows the status of the companies for the period 2013 to 2018.

Table 5
Status 2013-2018

Status	Count
Existing in both 2013 and 2018	207
Newly added between 2013 and 2018	72
No exchange data found in the periods from 2013 to 2018	94
Exchange data found in 2013 but not in 2018 (blank)	4
Grand Total	377

For the public companies that had listings from 2013 and 2018, there were 204 companies that fall into this category from the population of companies that had “big data” in their description in the filings of the United States Securities and Exchange Commission. 81 companies were in the technology sector. 36 companies did not have a sector designated in market information. 23 companies were in the consumer services sector. 14 companies were in both the finance and miscellaneous sectors. Eight companies were in the public utility sector. Seven companies were in the health-care sector. Four companies were in the energy sector. Two companies were in the consumer durables sector. There was one company each from the basic industries and consumer non-durables sectors.

Table 6
Sectors for Public Companies with "Big Data" in Description

Sector	Count
Basic Industries	1
Capital Goods	13
Consumer Durables	2
Consumer Non-Durables	1
Consumer Services	23
Energy	4
Finance	14
Health Care	7
Miscellaneous	14
Public Utilities	8
Technology	81
<NA>	36
Total	204

For three major stock market indices, the NASDAQ Composite Index performed at 125% increase, the Dow Jones Industrial Average at an 85% increase, and the Standard and Poor 500 at an 84% increase over the period of January 2, 2013, to January 2, 2018. Table 7 shows the performance of the indices over the five year period.

Table 7
Indices 2013-2018

Index	2013	2018	Change
NASDAQ	3112.26	7006.9	1.251386
DJIA	13412.6	24824.01	0.850805
SP500	1462.42	2695.81	0.84339

There was one company that was added to the NASDAQ market in 2017 that may have been using the ticker symbol of another company in 2013. This company was removed from the existing companies. The listing caused an outlier for the listings. Table 8 shows the summary information for the data on the percent change in stock market prices for the period of January 2, 2013, to January 2, 2018. The first row shows the changes with the outlier. The second row shows the changes without the outlier. The third row shows the percent change in stocks that were listed on the two major exchanges, NASDAQ and NYSE. In the second row, the mean percent change of .77 percent is less than the performance of the three major indices. In the third row, the mean percent change of .89 percent is greater than the performance of the Standard and Poor 500 and the Dow Jones Industrial Average and less than the performance of the NASDAQ Composite Index.

Table 8
Summary Data for Returns

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-1	-0.264	0.4826	24.175	1.3311	2526.289
-1	-0.264	0.4767	0.7668	1.2594	14.6722
-0.9653	-0.0202	0.60176	0.88953	1.34884	14.67217

Table 9 shows a summary of the performance over the five year period for the stock listings with big data on the two major indices. The big data companies performed at an average

of 1.11% on the New York Stock Exchange. The big data companies performed at an average of .78% on the Nasdaq Exchange.

Table 9
Big Data Companies on Major Exchanges

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
NYSE	-0.8444	0.2684	0.895	1.1123	1.6459	4.3524
NASDAQ	-0.9653	-0.1582	0.4781	0.7809	1.0773	14.6722

Table 10 displays the yearly percent change of the three major market indices. The NASDAQ had the highest growth during the five year period from 2013 to 2018, followed by the Dow Jones Industrial Average, and the Standard and Poor 500 Index. The NASDAQ exchange also had consistent growth for all five years from 2013 to 2018.

Table 10
Yearly Percent Change of Three Major Indices

	DJIA	NASDAQ	S&P 500
Percent Change	85.08047	125.1386	84.33897
2013-2014	23.59067	34.19797	26.3905
2014-2015	7.519067	13.39514	11.39063
2015-2016	-2.233291	5.729677	-0.7265997
2016-2017	13.41502	7.503079	9.535023
2017-2018	25.61105	30.16429	20.41155
Average	13.5805032	18.1980312	13.40022066

Table 11 provides the yearly close of the three major indices, the Dow Jones Industrial Average, the NASDAQ, and the Standard and Poor 500 Index. The close value is the end of the year close for each year. The Dow Jones Industrial Average closed at 21750.2 at the end of 2017. The NASDAQ closed at 6235.30. The Standard & Poor 500 Index closed at 2449.08.

Table 11
Yearly Close of Three Major Indices

DJIA	Close
2013-12-31	15009.52
2014-12-31	16777.69
2015-12-31	17587.09

2016-12-30	17927.11
2017-12-29	21750.2

NASDAQ	Close
2013-12-31	3541.292
2014-12-31	4375.104
2015-12-31	4945.554
2016-12-30	4987.793
2017-12-29	6235.299

S&P 500	Close
2013-12-31	1643.799
2014-12-31	1931.376
2015-12-31	2061.068
2016-12-30	2094.651
2017-12-29	2449.076

Table 12 shows the yearly returns for the three major indices. The returns are displayed as a decimal percentage increase. The yearly returns are calculated for the years, 2013 through 2017.

Table 12
Yearly Returns for Three Major Indices

DJI	Returns
2013-12-31	0.216783890
2014-12-31	0.078413898
2015-12-31	-0.010690749
2016-12-30	0.133844513
2017-12-29	0.226066858
NASDAQ	Returns
2013-12-31	0.30137595
2014-12-31	0.13572538
2015-12-31	0.06989207
2016-12-30	0.08500271
2017-12-29	0.25344123
S&P 500	Returns
2013-12-31	0.240127385
2014-12-31	0.114344729
2015-12-31	0.004678027
2016-12-30	0.099646149
2017-12-29	0.179757391

Table 13 shows the eight stocks listed on the major exchanges that had a statistically significant difference for a 95% confidence interval with the Dow Jones Industrial Average. The difference was calculated with a Welch's t-test comparing the daily change of the Dow Jones Industrial Average and the stock prices. Four of the stocks declined in price over the 5 year period. Four of the stocks increased in price over the 5 year period. The stocks that had a statistically significant difference with the Dow Jones Industrial Average did not necessarily have the greatest or lowest percent change. These stocks had performance patterns that deviated from the Dow Jones Industrial Average.

Table 13
Statistically Significant Differences in Stock Value Changes with DJIA

Symbol	T Statistic	P Value	Percent Change
stock1	-4.33530700	0.00001572	-99.99670
stock2	-3.05988300	0.00226634	87.50000
stock3	-2.37203400	0.01783990	11.11111
stock4	-1.97280100	0.04866642	417.49880
stock5	-2.01304500	0.04432188	-63.63636
stock6	2.494734	0.01269502	-63.71078
stock7	-2.01285000	0.04425173	347.76290
stock8	-2.47281600	0.01353650	-85.00000

Table 14 shows 7 stocks with a statistically significant difference between the stocks with “big data” in the description of their periodic filings with the United States Securities and Exchange Commission. Six of the stocks declined over the five year period from 2013 to 2018. One of the stocks advanced in the five year period from 2013 to 2018.

Table 14
Statistically Significant Difference with NASDAQ Index

Symbol	T Statistic	P Value	Percent Change
stock1	-4.335359	0.00001571	-99.9967
stock2	-2.372179	0.01783293	-26.45018
stock3	-1.962734	0.04982346	-81.875
stock4	-2.013221	0.04430328	-96.41667
stock5	2.476919	0.01334270	-15.82003

stock6	-1.996013	0.04605308	197.9037
stock7	-2.472944	0.01353167	-94.90826

Table 15 displays the six stocks that had “big data” in their periodic filings with the United States Securities and Exchange Commission and had a statistically significant difference with the Standard and Poor 500 Index. All six of the stocks that deviated statistically from the index declined from the period from 2013 to 2018.

Table 15
Statistically Significant Difference with Standard and Poor 500 Index

Symbol	T Statistic	P Value	Percent Change
stock1	-4.327631	0.00001627	-99.9967
stock2	-2.344506	0.01920668	-26.45018
stock3	-1.977522	0.04819954	-96.41667
stock4	2.730237	0.00638454	-15.82003
stock5	-2.450491	0.01440128	-94.90826
stock6	1.980877	0.04778720	-56.52174

Although the majority of the stocks, more than 95%, did not have a statistically significant difference in the performance of the stock markets, further analysis showed that there was a general increase in the value of the stocks for the public companies that had “big data” in their periodic filings with the United States Securities and Exchange Commission over the major indices.

Presentation of the Data

Raw data was read into R and stored into CSV with the readr library (Wickham, Hester, & Francois, 2017). The data was filtered by the dplyr library (Wickham, Francois, Henry, & Müller, 2017). The publicly traded companies associated with big data are compared with the performance of the major indices. Time series data for the stock prices were generated with the quantmod library in R (Ryan & Ulrich, 2017).

Figure 2 displays a histogram of the percent change of the stocks for the period from 2013 to 2018. The histogram displays that generally more stocks had an increase than a decrease. The histogram also shows that the increases between zero and 200 percent were more than the increases over 200 percent.

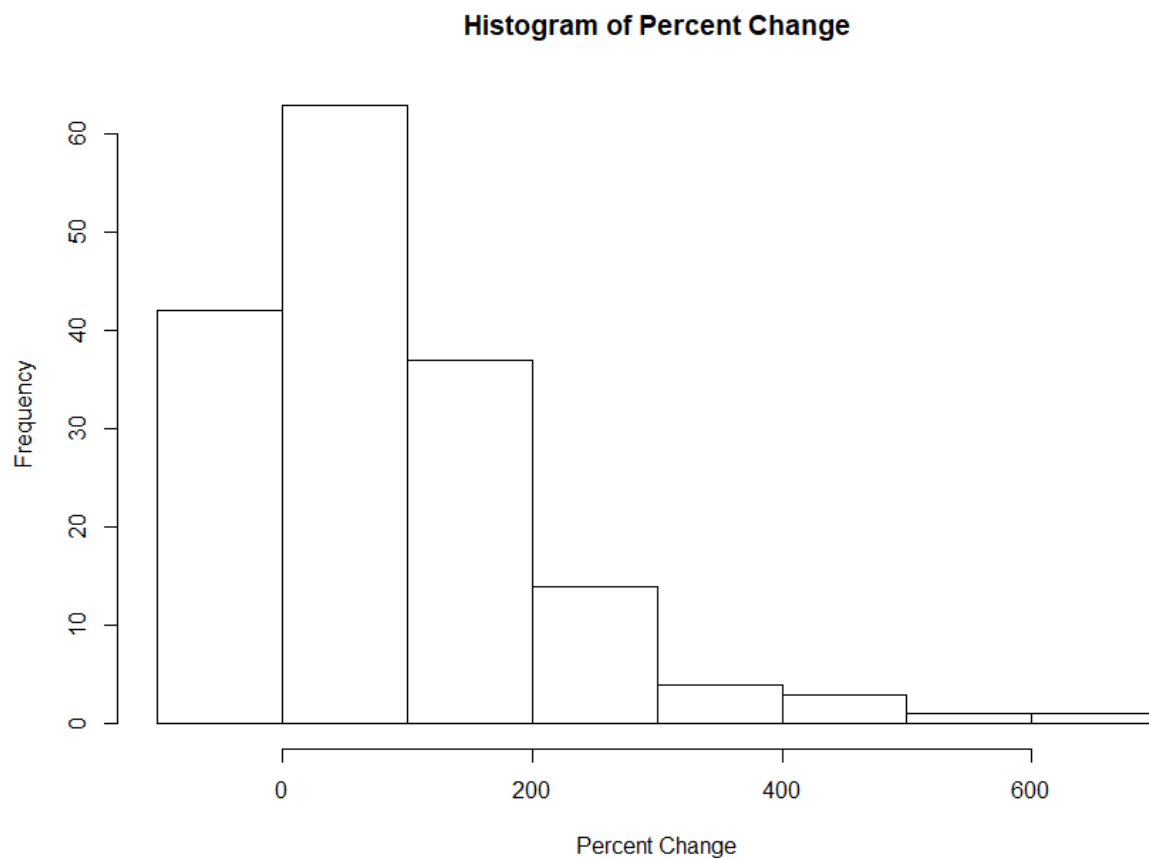


Figure 2. Histogram of the percent change of the stock prices between 2013 and 2018.

Figure 3 displays a kernel density estimation for predicting the percent change of the stocks during the five year period. The thin grey line shows a probability density of 25%. The thick grey shows a probability density of 50%. The black line shows a probability density of 100%.

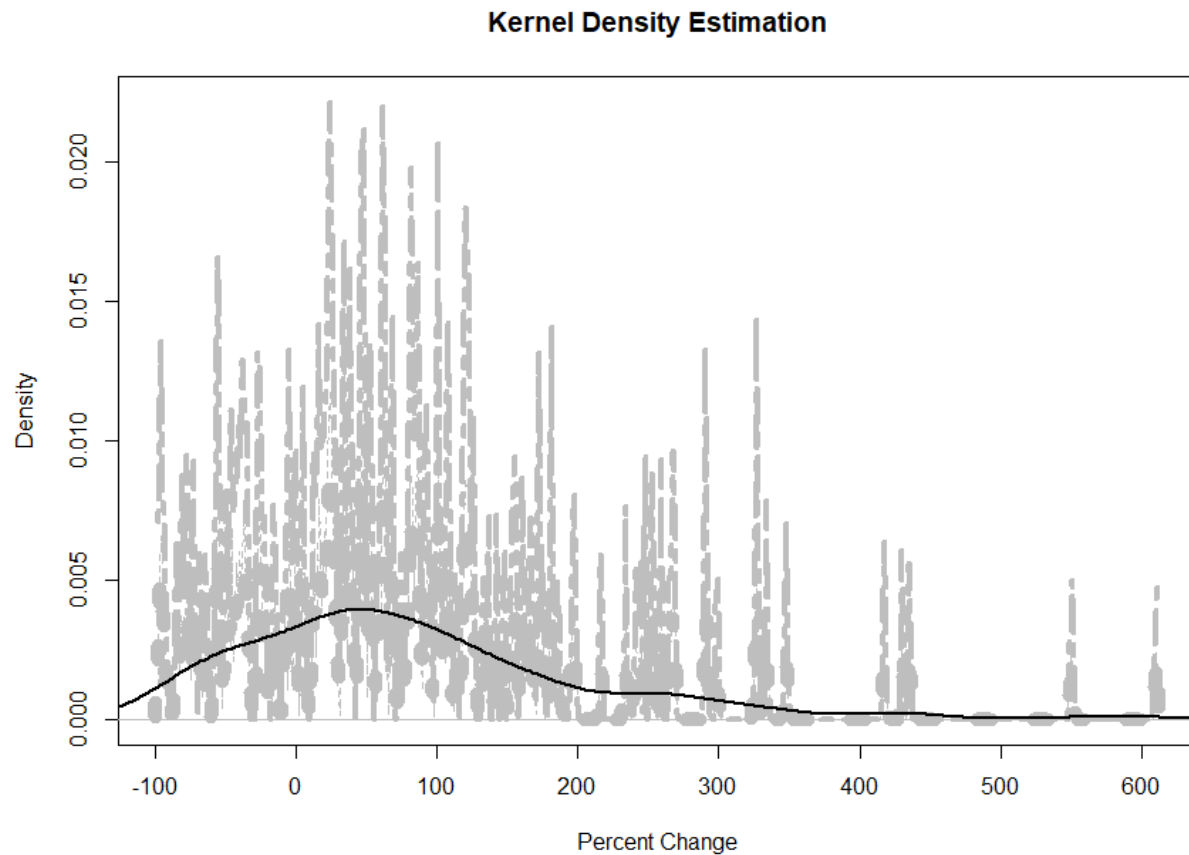


Figure 3. Kernel destiny estimation of the percent change of the stock prices.

Table 16 displays the kernel density estimation of the probability distribution of the companies with “big data” in their periodic filing with the United States Securities and Exchange Commission. The kernel density estimation showed a minimum percent change of -205% and a maximum percent change of 715%. The median and mean percent change was 255%.

Table 16
Kernel Density Estimation Default

X		y	
Min. :	-204.62	Min. :	6.33E-07
1st Qu.:	25.27	1st Qu.:	8.80E-05
Median :	255.17	Median :	4.49E-04
Mean :	255.17	Mean :	1.09E-03
3rd Qu.:	485.06	3rd Qu.:	2.10E-03
Max. :	714.96	Max. :	3.53E-03

Table 17 displays the summary of a logistic model for identifying the factors that influence whether the price of the stock increases or decreases. The logistic model included the sector, industry, and market. Not all of the factors are listed in the table.

Table 17
Logistic Model for Influence on Big Data Company Price Increase

	Estimate	Std. Error
(Intercept)	-4.50E+15	6.71E+07
Market: NYSE	9.40E+14	1.64E+07
Sector: Capital Goods	-3.93E+13	1.09E+08
Sector: Consumer Durables	1.25E+16	1.33E+08
Sector: Consumer Non-Durables	-9.40E+14	1.24E+08
Sector: Consumer Services	2.98E+15	1.16E+08
Sector: Energy	-9.05E+15	1.45E+08
Sector: Finance	-9.40E+14	9.63E+07
Sector: Health Care	3.56E+15	9.63E+07
Sector: Miscellaneous	7.44E+15	1.39E+08
Sector: n/a	-4.70E+14	8.26E+07
Sector: Public Utilities	8.10E+15	1.29E+08
Sector: Technology	-9.40E+14	9.63E+07
Industry: Auto Manufacturing	8.58E+15	9.76E+07
Industry: Auto Parts:O.E.M.	8.11E+15	1.08E+08
Industry: Biotechnology: Commercial Physical & Biological Research	5.44E+15	9.63E+07
Industry: Biotechnology: Laboratory Analytical Instruments	8.11E+15	1.08E+08
Industry: Broadcasting	1.06E+15	1.05E+08
Industry: Business Services	-7.94E+14	1.23E+08
Industry: Catalog/Specialty Distribution	1.53E+15	1.16E+08
Industry: Computer Communications Equipment	3.94E+15	7.92E+07
Industry: Computer Manufacturing	4.97E+15	7.55E+07
Industry: Computer Software: Prepackaged Software	6.49E+15	7.07E+07
Industry: Computer Software: Programming, Data Processing	6.10E+15	7.55E+07

Table 18 displays a logistic model for the sectors for influencing whether the price of the stock increased or decreased in the five year period from 2013 to 2018. The sectors with the highest estimates are consumer durables and consumer non-durable goods. The sectors with the lowest estimates are capital goods and energy.

Table 18
Logistic Model for Sector

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-			
	1.66E+01	2.40E+03	-0.007	0.994
Sector: Capital Goods	1.90E+01	2.40E+03	0.008	0.994
Sector: Consumer Durables	3.31E+01	2.94E+03	0.011	0.991
Sector: Consumer Non-Durables	3.31E+01	3.39E+03	0.01	0.992
Sector: Consumer Services	1.77E+01	2.40E+03	0.007	0.994
Sector: Energy	1.66E+01	2.40E+03	0.007	0.994
Sector: Finance	1.81E+01	2.40E+03	0.008	0.994
Sector: Health Care	1.73E+01	2.40E+03	0.007	0.994
Sector: Miscellaneous	1.84E+01	2.40E+03	0.008	0.994
Sector: n/a	2.93E-08	2.94E+03	0	1
Sector: Public Utilities	1.84E+01	2.40E+03	0.008	0.994
Sector: Technology	1.75E+01	2.40E+03	0.007	0.994

Table 19 displays the analysis of variance for the logistic models. The analysis of variance table shows the degrees of freedom and deviance for the two models. The table also displays the chi-squared probability.

Table 19
Analysis of Variance for Models

	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)	
NULL			147	164.22		
Market	1	4.2	146	160.02	0.04044	*
Sector	11	13.91	135	146.11	0.23807	
Industry	42	0	93	1946.36	1	
Percent Change > 100%	1	504.61	92	1441.75	< 2e-16	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Figure 4 displays the area under the curve for the two models. The area under the curve for the market model is 0.8653846. The area under the curve for the sector model is 0.6923077.

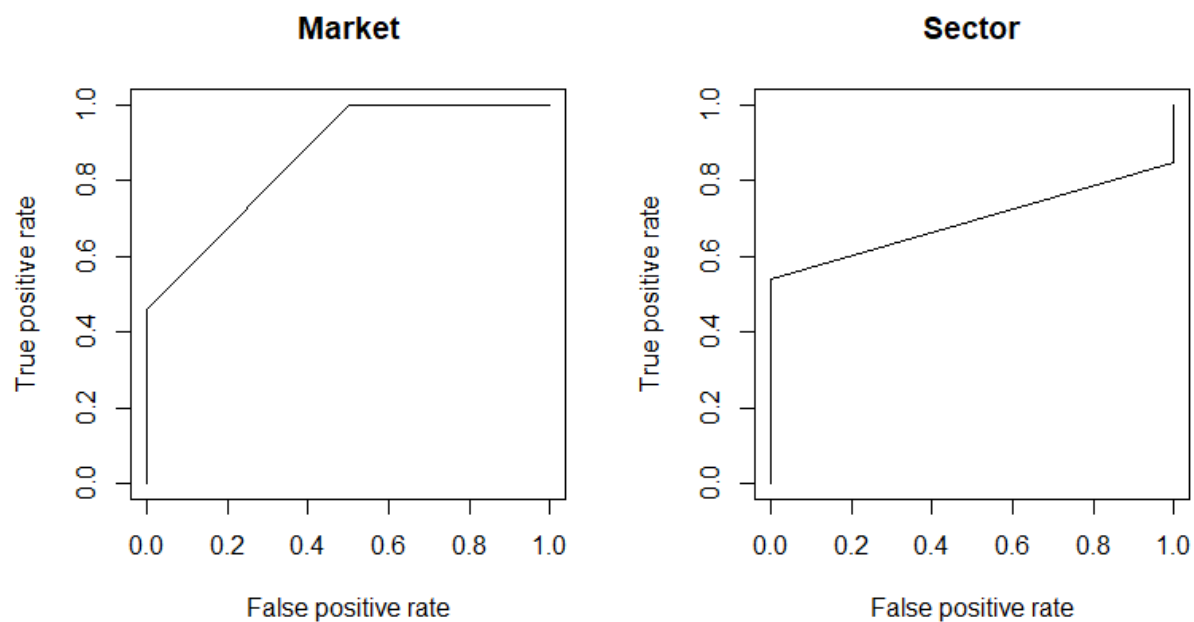


Figure 4. True Positive vs False Positive Rates.

Figure 5 similarly displays the receiving operating curve for the influence of market and sector to influence the percent change for the price of the stock of the company with “big data” in the periodic filings with the United States Securities and Exchange Commission. The receiving operating curve plot was created with generalized linear models of market and sector. The models show that market had more influence on the percent change of the stock price than the sector.

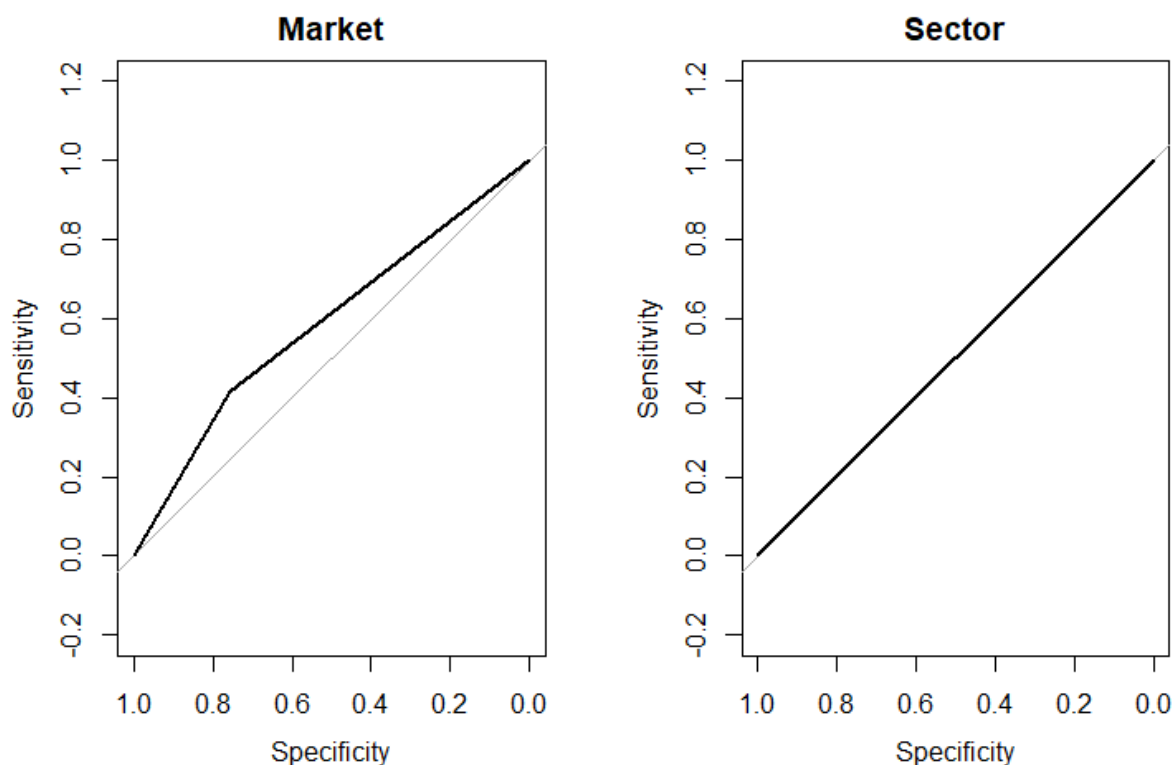


Figure 5. Receiving operating curve plot for market and sector.

Presentation and Discussion of Findings

Two linear models were created to compare the factors that influence the change in percent from 2013 to 2018. The first model evaluated the market, market capitalization, and sector as factors. The second model evaluated the market, market capitalization, sector, and industry as factors. The analysis of variance in the two models is provided in Table 20.

Table 20
Analysis of Variance of Models

Model 1	Model 2
Percent_Change ~ Market + MarketCap + Sector	Percent_Change ~ Market + MarketCap + Sector + industry
Res.Df RSS	Res.Df RSS
151 2209646	107 1432744

Table 21 displays the results of a Bayesian generalized linear model for the factors of sector, market, and market capitalization. The Bayesian generalized linear model shows that market capitalization and market were greater influences on the percent change of the companies that had “big data” in their periodic filing of the United States Securities and Exchange Commission. Whether the company was in the capital goods sector was also an influence of the percent change.

Table 21
Bayesian Generalized Linear Model

	coef. est	coef. Se
(Intercept)	0.03	0.96
Percent Change	0.04	0.01
Market Cap	0	0.01
Market: NYSE	0.29	0.8
Sector: Capital Goods	0.21	1.51
Sector: Consumer Durables	0.84	1.95
Sector: Consumer Non-Durables	0.83	1.95
Sector: Consumer Services	-0.18	1.24
Sector: Energy	-1.73	1.69
Sector: Finance	-0.47	1.31
Sector: Health Care	0.26	1.6
Sector: Miscellaneous	0.08	1.25
Sector: n/a	-0.85	1.92
Sector: Public Utilities	1.19	1.6
Sector: Technology	0.11	0.99

Three additional Markov Chain Monte Carlo models were created to identify which of the sectors and industries influence the change in percent from 2013 to 2018. A Raftery diagnosis was conducted of the models to measure quantiles of 0.025 with an accuracy of +/- 0.005 and a probability of 95%. Table 22 shows a model showing the influence of market

capitalization on the percent change of the companies that have “big data” in the description of the periodic filings with the United States Securities and Exchange Commission.

Table 22
Market Capitalization Model

	Burn-in (M)	Total (N)	Lower bound (Nmin)	Dependence factor (I)	
(Intercept)		2	3741	3746	0.999
Market Cap		2	3680	3746	0.982
sigma2		2	3865	3746	1.03

Table 23 displays Raftery Diagnostics of a Markov Chain Monte Carlo model designed to measure the influence of sectors on the percent change of the public companies that had “big data” in the description of their periodic filings with the United States Securities and Exchange Commission. The Raftery Diagnostics were conducted with quantiles of .025, an accuracy of +/- 0.005 and a probability of 95%.

Table 23
Raftery Diagnostics of Sector Model

	Burn-in (M)	Total (N)	Lower bound (Nmin)	Dependence factor (I)	
(Intercept)		2	3680	3746	0.982
Sector: Capital Goods		2	3802	3746	1.01
Sector: Consumer Durables		2	3802	3746	1.01
Sector: Consumer Non-Durables		2	3834	3746	1.02
Sector: Consumer Services		2	3710	3746	0.99
Sector: Energy		2	3802	3746	1.01
Sector: Finance		2	3771	3746	1.01
Sector: Health Care		2	3710	3746	0.99
Sector: Miscellaneous		2	3741	3746	0.999
Sector: n/a		2	3710	3746	0.99
Sector: Public Utilities		2	3680	3746	0.982
Sector: Technology		2	3741	3746	0.999
sigma2		2	3802	3746	1.01

The iterations for the Monte Carlo Markov Chain model were between 1001 and 11000. The thinning interval was one and the number of chains was one. The sample size per chain was

10,000. Table 24 displays the empirical mean, standard deviation, and standard error of the mean for the sectors of the companies that had “big data” in their description of their periodic filings with the United States Securities and Exchange Commission.

Table 24
Raftery Diagnostics of Sector Model

	Mean	SD	Naive SE	Time-series SE
(Intercept)	0.17278	0.46828	0.0046828	0.004683
Sector: Capital Goods	0.74667	0.49201	0.0049201	0.004990
Sector: Consumer Durables	0.81331	0.60146	0.0060146	0.006015
Sector: Consumer Non-Durables	0.79897	0.70643	0.0070643	0.007064
Sector: Consumer Services	0.58877	0.48462	0.0048462	0.004846
Sector: Energy	0.07283	0.53798	0.0053798	0.005380
Sector: Finance	0.68381	0.48889	0.0048889	0.004993
Sector: Health Care	0.54095	0.51046	0.0051046	0.005105
Sector: Miscellaneous	0.68066	0.49016	0.0049016	0.004986
Sector: n/a -	0.17715	0.60292	0.0060292	0.006029
Sector: Public Utilities	0.69863	0.50382	0.0050382	0.005038
Sector: Technology	0.53281	0.47161	0.0047161	0.004716
sigma2	0.3013	0.03435	0.0003435	0.000370

Table 25 displays the quartiles for the influence of the sectors on the Markov Chain Monte Carlo simulation models of the public companies that had “big data” in the description of the periodic filings of the United States Securities and Exchange Commission. The consumer non-durables sector had the highest value at the 75% quartile with the simulated value of 1.27. The energy sector had the lowest value at .43 for the 75% quartile.

Table 25
Quartiles for Sector Markov Chain Monte Carlo Model

	2.5%	25%	50%	75%	97.5%
(Intercept)	-0.7365	-0.1447	0.17259	0.4822	1.1035
Sector: Capital Goods	-0.2304	0.4249	0.74992	1.0735	1.7022
Sector: Consumer Durables	-0.37	0.4174	0.80969	1.2129	2.0166
Sector: Consumer Non-Durables	-0.5894	0.3239	0.80347	1.2755	2.1919

Sector: Consumer Services	-0.3779	0.2676	0.59010	0.9175	1.5302
Sector: Energy	-0.9777	-0.2851	0.07818	0.4392	1.1167
Sector: Finance	-0.2875	0.3614	0.68613	1.0108	1.6428
Sector: Health Care	-0.4742	0.2049	0.53884	0.8797	1.5439
Sector: Miscellaneous	-0.3151	0.3579	0.68475	1.0046	1.6266
Sector: n/a	-1.3529	-0.5816	-0.17587	0.2226	1.0041
Sector: Public Utilities	-0.3173	0.3631	0.69657	1.0352	1.6921
Sector: Technology	-0.4059	0.2178	0.53159	0.8506	1.4458
sigma2	0.2406	0.2767	0.29871	0.3230	0.3748

Table 26 displays Raftery Diagnostics for the Markov Chain Monte Carlo simulation of the model that shows the influence of the market for the percent change of the stock of the public companies that had “big data” in the description for periodic filings with the Securities and Exchange Commission. The quantiles for the simulation was .025, with an accuracy of +/-0.005 and a 95% probability.

Table 26
Raftery Diagnostics of Markov Chain Monte Carlo Simulation 3

	Burn-in (M)	Total (N)	Lower bound (Nmin)	Dependence factor (I)	
(Intercept)		2	3710	3746	0.99
Market: NASDAQ		2	3710	3746	0.99
Market: NYSE		2	3650	3746	0.974
sigma2		2	3929	3746	1.05

Table 27 displays the result of a Markov Chain Monte Carlo simulation with 1001 to 11,000 iterations, a thinning interval of one, one chain, and 10,000 samples per chain, to measure the influence of the market on price change of public companies that had “big data” in the periodic filings with the United States Securities and Exchange Commission. Table 27 displays the empirical mean, standard deviation, and standard error of the mean, for the variables of the market.

Table 27
Markov Chain Monte Carlo Simulation Empirical Mean

	Mean	SD	Naive SE	Time-series SE
(Intercept)	0.2566	0.2691	0.002691	0.0026756
Market: NASDAQ	0.432	0.2743	0.002743	0.0027451
Market: NYSE	0.5764	0.2767	0.002767	0.0027673
sigma2	0.2996	0.0328	0.000328	0.0003248

Table 28 displays the quantiles for the market variable for the public companies that have “big data” in their periodic filings with the United States Securities and Exchange Commission. For the 97.5% quartile the NASDAQ market had a value of .96. For the 97.5 quartile, the NYSE had a value of 1.11.

Table 28
Markov Chain Monte Carlo Simulation Quantiles

	2.5%	25%	50%	75%	97.5%
(Intercept)	-0.26427	0.07418	0.2569	0.4376	0.7797
Market: NASDAQ	-0.09915	0.25007	0.4337	0.6187	0.9597
Market: NYSE	0.02376	0.39282	0.5784	0.7624	1.1116
sigma2	0.24190	0.27653	0.2975	0.3194	0.3713

Table 29 displays the mean, maximum, and minimum percent change by sector for the public companies that had “big data” in the description of the periodic filing with the United States Securities and Exchange Commission. Basic industries had the lowest mean with -77.4%. Consumer durables the highest mean with 176%.

Table 29
Percent Change by Sector

Sector	Mean	Max	Min
Basic Industries	-77.4	-77.4	-77.4
Capital Goods	163	417	-39.4
Consumer Durables	176	327	25
Consumer Non-Durables	25.4	25.4	25.4
Consumer Services	88.4	435	-81.4

Energy	-14.4	15.6	-41.7
Finance	95.9	198	-25.3
Health Care	167	550	-96.5
Miscellaneous	100	265	-35
No Sector	-50.5	-37.2	-63.7
Public Utilities	62.3	327	-94.9
Technology	66.6	610	-95.9

Table 30 shows the market capitalization by sector. The energy sector had the highest average market capitalization for companies with “big data” in the periodic filing of the United States Securities and Exchange Commission. Basic Industries had the lowest average market capitalization. The market capitalization is in billions of dollars.

Table 30
Market Capitalization by Sector (Billions)

Sector	Mean	Max	Min
Basic Industries	0.01499	0.01499	0.01499
Capital Goods	21.40246	85.48	0.02208
Consumer Durables	0.14549	0.16701	0.12397
Consumer Non-Durables	5.08	5.08	5.08
Consumer Services	8.979006	59.35	0.03137
Energy	73.48	239.06	2.18
Finance	36.78037	191.87	0.26523
Health Care	16.95373	62.45	0.01215
Miscellaneous	11.78724	107.43	0.04108
No Sector	0.1707	0.1912	0.1502
Public Utilities	28.36494	205.6	0.01955
Technology	15.32001	244.65	0.00317

Table 31 displays the top ten industries by market capitalization of the public companies that had “big data” in the periodic filings of the United States Securities and Exchange Commission. The leading industries in market capitalization by average were integrated oil companies, savings institutions, and financial consumer services. The market capitalization is displayed in billions of dollars.

Table 31
Top Ten Industries by Market Capitalization

Industry	Mean	Max	Min
Integrated oil Companies	239.06	239.06	239.06
Savings Institutions	191.87	191.87	191.87
Finance: Consumer Services	83.65	83.65	83.65
Television Services	59.35	59.35	59.35
Computer Communications Equipment	51.05185	203.16	0.0189
Military/Government/Technical	43.295	85.48	1.11
Marine Transportation	37.08	56.25	9.48
Medical Specialties	35.54667	62.45	1.84
Computer Manufacturing	34.23	129.69	1.04
Real Estate Investment Trusts	33.28	33.28	33.28

Table 32 displays a decision tree that shows the probability that the stock price would increase over the five year period for the companies that had “big data” in the description of the periodic filings with the United States Securities and Exchange Commission. The decision tree provides the deviance and probability for the influence of the market. The NASDAQ market had a higher deviance and lower probability than the New York Stock Exchange.

Table 32
Decision Tree Market

Decision Tree Market		
Node	Deviance	Y Value
Root	31.3091	0.74545
Market=NASDAQ	21.2929	0.68687
Market=NYSE	9.16667	0.83333

Table 33 displays a decision tree that shows the probability that the stock price would increase over the five year period for the companies that “big data” in the description of the periodic filings with the United States Securities and Exchange Commission. The decision tree provides the deviance and probability for the influence of market and sector. The sectors of capital goods, consumer durables, consumer non-durables, consumer services, finance, and health care had a higher deviance than other sectors.

Table 33
Decision Tree Market and Sector

Decision Tree Market and Sector		
Node	Deviance	Y Value
Root	31.3091	0.74545
Sector=Basic Industries, Energy, No Sector	0.85714	0.14286
Sector=Capital Goods, Consumer Durables, Consumer Non-Durables, Consumer Services, Finance, Health Care, Miscellaneous, Public Utilities, Technology	27.7975	0.77215
Market=NASDAQ	20.3299	0.70103
Sector=Capital Goods, Health Care, Miscellaneous, Public Utilities, Technology	17.443	0.67089
Sector=Consumer Durables, Consumer Non-Durables, Consumer Services, Finance	2.5	0.83333
Market=NYSE	6.19672	0.88525
Sector=Consumer Services, Finance, Technology	5.55882	0.79412
Sector=Capital Goods, Health Care, Miscellaneous, Public Utilities	0	1

Table 34 displays a decision tree that shows the probability that the stock price would increase over the five year period for the companies that “big data” in the description of the periodic filings with the United States Securities and Exchange Commission. The decision tree provides the deviance and probability for the influence of market and market capitalization. The market capitalization is displayed in billions of dollars.

Table 34
Decision Tree Market and Market Capitalization

Decision Tree Market and Market Capitalization		
Node	Deviance	Y Value
Root	31.3091	0.74545
Market Capitalization< 0.281445	7.05882	0.29412
Market Capitalization< 0.032175	0	0
Market Capitalization>=0.032175	5.2381	0.47619
Market Capitalization>=0.281445	15.5267	0.8626
Market Capitalization< 2.76	9	0.75
Market Capitalization< 0.758965	2.54546	0.63636
Market Capitalization>=0.758965	6.27027	0.78378
Market Capitalization>=1.195	5.7931	0.72414
Market Capitalization< 1.195	0	1
Market Capitalization>=2.76	5.56627	0.92771
Market Capitalization>=118.56	1.42857	0.71429
Market Capitalization< 118.56	3.78947	0.94737

Figure 6 shows a plot of random forest simulation of 100 trees for three linear models to influence the increase of the stock of public companies with “big data” in the description of their periodic filings with the United States Securities and Exchange Commission. The solid line displays a linear model of the market to influence the percent change in increase. The dashed line displays a model of the market, market capitalization, and sector to influence the percent change in increase. The dotted line displays the model of the market, market capitalization, sector and industry to influence the percent change in increase. The random forest displays the probable error for the linear model as the number of trees increase. A new random forest was created for each run.

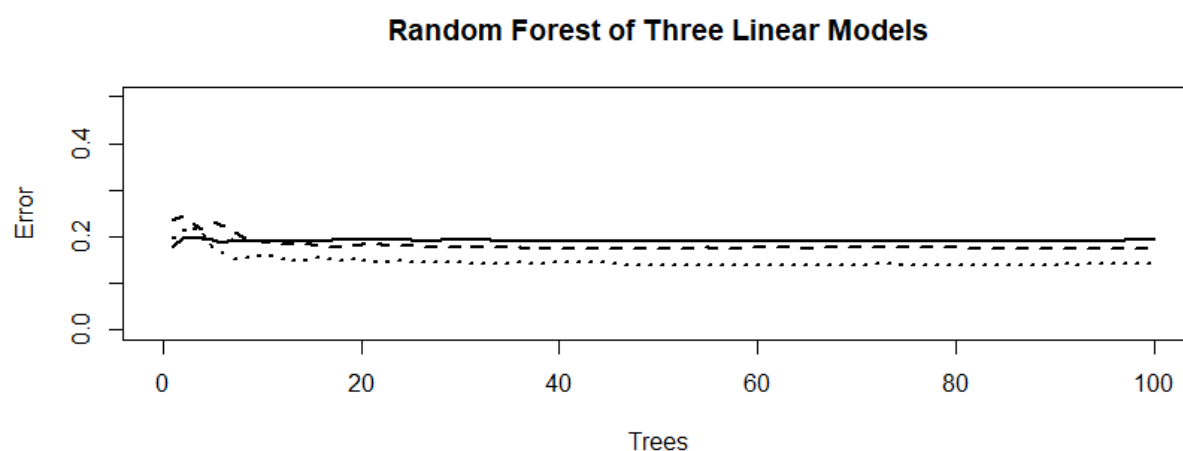


Figure 6. Random forest of linear models with 100 trees.

Table 35 displays a frequency of the companies that had an increase of over 200% of the companies that had “big data” in the periodic filings of the United States Securities and Exchange Commission. The sectors of companies with an increase of over 200% included technology, capital goods, consumer services, health care, and miscellaneous. Consumer durables and public utilities also had companies that increased by over 200% for the period of January 1, 2013, to January 1, 2018.

Table 35
Number of Companies in Each Sector with a Percent Change Greater than 200%

Sector Greater than 200%	
Capital Goods	4
Consumer Durables	1
Consumer Services	3
Health Care	3
Miscellaneous	3
Public Utilities	1
Technology	8

Table 36 displays a frequency of the companies that had an increase of over 300% of the companies that had “big data” in the periodic filings of the United States Securities and Exchange Commission. The largest group of companies was from the technology sector. The next highest group of companies was from the capital goods sector.

Table 36
Number of Companies in Each Sector with a Percent Change Greater than 300%

Sector Greater than 300%	
Capital Goods	2
Consumer Durables	1
Consumer Services	1
Health Care	1
Public Utilities	1
Technology	3

Table 37 shows the frequency of the companies that had an increase of over 400% of the companies that had “big data” in the periodic filings of the United States Securities and Exchange Commission. The largest group of companies was in the technology sector. The other sectors with companies that had an increase of over 400% included the capital goods, consumer services, and health-care sectors.

Table 37
Number of Companies in Each Sector with a Percent Change Greater than 400%

Sector Greater than 400%	
--------------------------	--

Capital Goods	1
Consumer Services	1
Health Care	1
Technology	2

Table 38 displays the frequency of the companies that had an increase of over 500% of the companies that had “big data” in the periodic filings of the United States Securities and Exchange Commission. The two sectors represented by companies that had an increase of over 500% include the health-care and technology sectors.

Table 38
Number of Companies in Each Sector with a Percent Change Greater than 500%

Sector Greater than 500%	
Health Care	1
Technology	1

Summary of Chapter

The chapter displayed the findings of the research of public companies with “big data” in their description of their periodic filings with the United States Securities and Exchange Commission. The chapter explored a statistically significant difference between the performance of these stocks with the major indices. The chapter discovered that there was not a statistically significant difference in the performance of these companies and the major indices, although the stocks generally performed slightly better than the major indices. The next chapter will describe how the research and findings can be applied in practice, future research, and conclusions of the research.

CHAPTER FIVE

This chapter provides the findings and conclusions of the research, the limitations of the study, implications for practice, and recommendations for research. This chapter also provides reflections of the research. The chapter concludes with a summary of the research.

Limitations of the Study

The limitations of the study included only stocks listed in the period of January 2, 2013, to January 3, 2018, were included in the population. Also, the population was reduced to only the companies that were listed on the major exchanges of the New York Stock Exchange and NASDAQ exchange. An additional limitation of the study was that the study only addressed public companies so does not evaluate all companies. The periodic filings were limited to quarterly and annual filings which are required per the regulations of the United States Securities and Exchange Commission for public companies.

Implications for Practice

The topic for this study was an evaluation of the stock market performance of public companies that have “big data” in their description for periodic filings with the United States Securities and Exchange Commission. The population for this study included companies with “big data” in periodic filings that were listed on the major exchanges of the NYSE and the NASDAQ during the period from 2013 to 2018 (Nasdaq, 2018; U.S. Securities and Exchange Commission, 2018). The methodology for the study was a quantitative analysis.

The research identified that although there was not a statistically significant difference between the increases in stock market value for the companies that had “big data” in the description of the filings of the periodic filings of the public companies, there was a general increase in the performance of these companies when compared to the major stock market

indices. The research identified that many companies implement big data technologies from various sectors and a large number of companies were from the technology sector.

The study of financial data can be improved with recent advancements in big data technologies (Yang, Huang, Li, Liu, & Hu, 2017). The ability to integrate advances made in machine learning with the improvements made to the components of big data such as integrating data from a variety of sources, analyzing streaming data, and providing real-time analytics has been made possible with the development of various big data analysis tools (Hu et al., 2018; Yang et al., 2017; Yassien & Desouky, 2016). Streaming processor capabilities have allowed for an enhanced system architecture, performance optimization, and scalable systems (Yang et al., 2017). Advancements can be achieved by integrating the public company information with data from the markets for asynchronous analysis of data from a variety of sources (Yang et al., 2017). Data integration can also be improved in distributed environments (Ianni, Masciari, Mazzeo, & Zaniolo, 2018).

Implications of Study and Recommendations for Future Research

This research can be advanced in a number of areas such as content mining, time series analysis, forecasting, and simulation (Bruno, 2016; Cortez et al., 2016; Du et al., 2018). The textual analysis of the company descriptions can be enhanced to reveal more information about how public companies implement big data (Du et al., 2018). The Electronic Data Gathering and Retrieval system, EDGAR, of the United States Securities and Exchange Commission provides free access to the public to search information about the filings of public companies (U.S. Securities and Exchange Commission, 2018). The EDGAR system also provides access to the raw data of the filings (U.S. Securities and Exchange Commission, 2018). One potential for future research is a textual analysis of the raw data (Du et al., 2018).

Another area for future research would be to implement big data technologies to store more data about the public companies and drill down on the analysis of the stock market performance (Du et al., 2018; Elagamy, Stanier, & Sharp, 2018; Khedr & Yaseen, 2017). Increasing the storage capabilities of the analysis environment could allow for a longer range of years to analyze and more in-depth study of the daily opening, high, low, and close prices of the companies with “big data” in the description (Sethia, Sheoran, & Saran, 2017). Drilling down may provide more information about the time periods when companies that utilize big data improve in market performance (Ferrández et al., 2016). The third area for future research would be to combine the analysis of the stock market value with an analysis of the sentiment from other sources such as news feeds (Ren, Wu, & Liu, 2018; Zhang et al., 2018). A sentiment analysis may indicate public influences to the performance of the stock market price of the public companies that have “big data” in the description of their periodic filings (Ren et al., 2018).

Reflections

The research provided guidance on how big data technologies have become influential in many industries. The research also provided the opportunity to demonstrate how to implement some of the techniques of big data analysis, such as combining data from a variety of sources, handling time series data, and gathering information from raw data. The research also indicates that big data technologies have allowed some companies to achieve profitability and sustainability through the analysis of big data.

Identifying the value of big data analysis may possibly lead to the development of more efficient and user-friendly software applications for analyzing data (Quboa & Mehandjiev, 2017; Sowmya & Suneetha, 2017). Further analysis of the descriptions may have led to a different approach for determining the population of companies and produced different results.

Examining the raw data as a streaming process and combining with real-time data may also reveal more analysis possibilities. The research may contribute to determine if analysis of data can improve the value for companies that implement big data.

Conclusion

Although there was not a statistically significant difference between the performance of the public companies with “big data” in their description of their periodic filings, there was an observable increase in the performance of these companies when compared to the indices of the major markets. Investing in big data has been profitable for some companies for the researched time period and may continue to be profitable in the near future (Du et al., 2018).

The research of publicly traded companies with big data in their periodic filings with the United States Securities and Exchange Commission provided an opportunity to explore some of the emerging technologies with big data such as quantitative analysis tools, text mining, Monte Carlo simulation, and time series forecasting (Bruno, 2016; Du et al., 2018; Singh & Thakral, 2017). The support of data from the Electronic Data Gathering and Retrieval system and the markets provided possibilities for implementing application programming interfaces for accessing raw data (U.S. Securities and Exchange Commission, 2018).

Analyzing the data can be improved with developing workflows for steps of data analysis (Padmavathi & Krishnamurthy, 2017). The possibility of developing a graphical user interface with drag and drop capabilities for manipulating the data and operations to supplement the R programming interface may be a future project (Huang, Lange, & Auer, 2015; Liang, Yeh, & Wu, 2018).

The R libraries that were implemented in the research were quite capable throughout the stages of the data analysis (Iancu & Georgescu, 2018). The computational and plotting features provided a number of analytical tools that allowed for detailed assessments of the data (Zaman &

Lung, 2018). Some of the tools were CPU resource heavy and may have been improved if they were performed as background operations (Rodríguez-Narciso, Christen, Rodríguez-Narciso, & Christen, 2016). Running some of the operations on a separate server allowed for more efficient operations as multiple instances of the R programming language could be created through terminal connections. However, these operations could be improved if the connections had access to the same real-time data (Phillips, Neth, Woike, & Gaissmaier, 2017). One opportunity for sharing the real-time data could be through creating an environment which could be loaded through each session and store common variables.

Transferring plot results from the server to the local machine can also be integrated into the platform (Laborie, Manzat, & Sedes, 2018). Remote plotting is possible through an API integration (Hu et al., 2016; Wozniak et al., 2015). Multimedia API integrations may allow for remote plotting to which plots can be transferred from one computer to another through an interface (Laborie et al., 2018). Additionally, saving the analysis output can also be further streamlined if there were more convenient methods for formatting text output (Arulraj, Pavlo, & Menon, 2016; Dogan, 2017). Generally, the R programming language provided a very sophisticated tool for the tasks of the data analysis.

References

- Abbasi, A., Sarker, S., & Chiang, R. (2016). Big data research in information systems: Toward an inclusive research agenda. *Journal of the Association for Information Systems*, 17(2), 3.
- Abdulkadri, A., Evans, A., & Ash, T. (2016). An assessment of big data for official statistics in the Caribbean: Challenges and opportunities.
- Ahmed, M., Choudhury, N., & Uddin, S. (2017). *Anomaly detection on big data in financial markets*. Paper presented at the Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017, Sydney, Australia.
- Al-augby, S., Majewski, S., Nermend, K., & Majewska, A. (2016, 9-10 May 2016). *Proposed investment decision support system for stock exchange using text mining method*. Paper presented at the 2016 Al-Sadeq International Conference on Multidisciplinary in IT and Communication Science and Applications (AIC-MITCSA).
- Al-Badarenah, A., & Alsakran, J. (2016). An automated recommender system for course selection. *International Journal of Advanced Computer Science and Applications*, 7(3), 166-175.
- Al-Htaybat, K., & von Alberti-Alhtaybat, L. (2017). Big data and corporate reporting: Impacts and paradoxes. *Accounting, Auditing & Accountability Journal*, 30(4), 850-873.
- Alexander, L., Das, S. R., Ives, Z., Jagadish, H., & Monteleoni, C. (2017). Research challenges in financial data modeling and analysis. *Big Data*, 5(3), 177-188.
- Alpanda, S. (2013). Identifying the role of risk shocks in the business cycle using stock price data. *Economic Inquiry*, 51(1), 304-335. doi:10.1111/j.1465-7295.2011.00445.x
- Álvarez-Ayuso, I. C., Condeço-Melhorado, A. M., Gutiérrez, J., & Zofío, J. L. (2016). Integrating network analysis with the production function approach to study the spillover effects of transport infrastructure. *Regional Studies*, 50(6), 996-1015. doi:10.1080/00343404.2014.953472
- Álvarez, W., & Griffin, V. J. (2016). Estimation procedure for reduced rank regression, PLSSVD. *Statistics, Optimization & Information Computing*, 4(2), 107-117. doi:http://dx.doi.org/10.19139/soic.v4i2.146
- Ambainis, A., Filmus, Y., & Gall, F. L. (2015). *Fast matrix multiplication: Limitations of the Coppersmith-Winograd method*. Paper presented at the Proceedings of the Forty-Seventh Annual ACM Symposium on Theory of Computing, Portland, Oregon, USA.
- Amigud, A., Arnedo-Moreno, J., Daradoumis, T., & Guerrero-Roldan, A.-E. (2017). Using learning analytics for preserving academic integrity. *International Review of Research in Open & Distance Learning*, 18(5), 192-210.

- Annas, G. J. (2018). Beyond Nazi war crimes experiments: The voluntary consent requirement of the Nuremberg Code at 70. *American Journal of Public Health*, 108(1), 42-46. doi:10.2105/AJPH.2017.304103
- Arcila-Calderón, C., Barbosa-Caro, E., & Cabezuelo-Lorenzo, F. (2016). Técnicas big data: Análisis de textos a gran escala para la investigación científica y periodística. *Big data techniques: Large-scale text analysis for scientific and journalistic research.*, 25(4), 623-631. doi:10.3145/epi.2016.jul.12
- Arulraj, J., Pavlo, A., & Menon, P. (2016). *Bridging the archipelago between row-stores and column-stores for hybrid workloads*. Paper presented at the Proceedings of the 2016 International Conference on Management of Data, San Francisco, California, USA.
- Balan, S., & Perez-Carballo, J. (2016). Investigating data analytic tools used in american higher education. *Global Education Journal*, 2016(2), 64-74.
- Baralis, E., Cagliero, L., & Cerquitelli, T. (2016). *Supporting stock trading in multiple foreign markets: A multilingual news summarization approach*. Paper presented at the Proceedings of the Second International Workshop on Data Science for Macro-Modeling, San Francisco, CA, USA.
- Barmish, B. R., & Primbs, J. A. (2016). On a new paradigm for stock trading via a model-free feedback controller. *IEEE Transactions on Automatic Control*, 61(3), 662-676. doi:10.1109/TAC.2015.2444078
- Barsdorf, N., & Millum, J. (2017). The social value of health research and the worst off. *Bioethics*, 31(2), 105-115. doi:10.1111/bioe.12320
- Bates, D. W., Saria, S., Ohno-Machado, L., Shah, A., & Escobar, G. (2014). Big data in health care: Using analytics to identify and manage high-risk and high-cost patients. *Health Affairs*, 33(7), 1123-1131.
- Bender, J. L., Cyr, A. B., Arbuckle, L., & Ferris, L. E. (2017). Ethics and privacy implications of using the internet and social media to recruit participants for health research: A privacy-by-design framework for online recruitment. *Journal of Medical Internet Research*, 19(4), 1-1. doi:10.2196/jmir.7029
- Benzekki, K., El Fergougui, A., & Elbelrhiti Elalaoui, A. (2016). Software-defined networking (SDN): a survey. *Security & Communication Networks*, 9(18), 5803-5833. doi:10.1002/sec.1737
- Berndt, D. J., Boogers, D., Chakraborty, S., & McCart, J. A. (2017). *Using agent-based modeling to assess liquidity mismatch in open-end bond funds*. Paper presented at the Proceedings of the Summer Simulation Multi-Conference, Bellevue, Washington.
- Bharath, R., & Rajalakshmi, P. (2016, 14-16 Sept. 2016). *Non-local means kernel regression based despeckling of B-mode ultrasound images*. Paper presented at the 2016 IEEE 18th

International Conference on e-Health Networking, Applications and Services (Healthcom).

- Bin, Y. (2017). Stock market crash, new market failure theory and extremely great economic and financial crises. *World Review of Political Economy*, 8(1), 4-34.
- Binbin, W., & Xiaoyan, L. (2017). Big data, platform economy and market competition: A preliminary construction of plan-oriented market economy system in the information era. *World Review of Political Economy*, 8(2), 138-161.
- Bİrcan, H., & Sungur, S. (2016). The role of motivation and cognitive engagement in science achievement. *Science Education International*, 27(4), 509-529.
- Blaser, M., & Manthey, B. (2015). Smoothed complexity theory. *ACM Trans. Comput. Theory*, 7(2), 1-21. doi:10.1145/2656210
- Boonpeng, S., & Jeatrakul, P. (2016, 14-16 Feb. 2016). *Decision support system for investing in stock market by using OAA-Neural Network*. Paper presented at the 2016 Eighth International Conference on Advanced Computational Intelligence (ICACI).
- Boronow, K. E., Susmann, H. P., Gajos, K. Z., Rudel, R. A., Arnold, K. C., Brown, P., . . . Brody, J. G. (2017). DERBI: A digital method to help researchers offer "right-to-know" personal exposure results. *Environmental health perspectives*, 125(2), A27-A33. doi:10.1289/EHP702
- Boudes, P., Kaszczyc, A., & Pellissier, L. (2015). *Monetary economics simulation: Stock-flow consistent invariance, monadic style*. Paper presented at the 11th Artificial Economics Conference.
- Brabanter, K. D., Brabanter, J. D., Moor, B. D., Ir, #232, & Gijbels, n. (2013). Derivative estimation with local polynomial fitting. *J. Mach. Learn. Res.*, 14(1), 281-301.
- Bradlow, E. T., Gangwar, M., Kopalle, P., & Voleti, S. (2017). The role of big data and predictive analytics in retailing. *Journal of Retailing*, 93(1), 79-95. doi:http://dx.doi.org/10.1016/j.jretai.2016.12.004
- Brady, M., & Irwin, E. (2011). Accounting for spatial effects in economic models of land use: Recent developments and challenges ahead. *Environmental and Resource Economics*, 48(3), 487-509. doi:http://dx.doi.org/10.1007/s10640-010-9446-6
- Briguglio, L. P. (2016). Exposure to external shocks and economic resilience of countries: evidence from global indicators. *Journal of Economic Studies*, 43(6), 1057-1078.
- Brinkman, E., & Wellman, M. P. (2017). *Empirical mechanism design for optimizing clearing interval in frequent call markets*. Paper presented at the Proceedings of the 2017 ACM Conference on Economics and Computation, Cambridge, Massachusetts, USA.

- Brown, J. P., & Lambert, D. M. (2016). Extending a smooth parameter model to firm location analyses: The case of natural gas establishments in the United States. *Journal of Regional Science*, 56(5), 848-867. doi:10.1111/jors.12280
- Bruno, G. (2016, 5-8 Dec. 2016). *Text mining and sentiment extraction in central bank documents*. Paper presented at the 2016 IEEE International Conference on Big Data (Big Data).
- Bucur, C. (2015). Using big data for intelligent businesses. *Scientific Research & Education in the Air Force - AFASES*, 2, 605-612.
- Bumblauskas, D., Nold, H., Bumblauskas, P., & Igou, A. (2017). Big data analytics: Transforming data to action. *Business Process Management Journal*, 23(3), 703-720.
- CACM Staff. (2017). Big data. *Commun. ACM*, 60(6), 24-25. doi:10.1145/3079064
- Cai, S., Gallina, B., Nyström, D., & Seceleanu, C. (2016). *A formal approach for flexible modeling and analysis of transaction timeliness and isolation*. Paper presented at the Proceedings of the 24th International Conference on Real-Time Networks and Systems, Brest, France.
- Candan, G., & Yazgan, H. R. (2015). Genetic algorithm parameter optimisation using Taguchi method for a flexible manufacturing system scheduling problem. *International Journal of Production Research*, 53(3), 897-915. doi:10.1080/00207543.2014.939244
- Cazzoli, L., Sharma, R., Treccani, M., & Lillo, F. (2016, 5-7 Sept. 2016). *A large scale study to understand the relation between Twitter and financial market*. Paper presented at the 2016 Third European Network Intelligence Conference (ENIC).
- Cerchiello, P., & Nicola, G. (2018). Assessing news contagion in finance. *Econometrics*, 6(1), 5.
- Cerniglia, J. A., Fabozzi, F. J., & Kolm, P. N. (2016). Best practices in research for quantitative equity strategies. *Journal of Portfolio Management*, 42(5), 135-143. doi:http://dx.doi.org/10.3905/jpm.2016.42.5.135
- Chambers, R., Salvati, N., & Tzavidis, N. (2016). Semiparametric small area estimation for binary outcomes with application to unemployment estimation for local authorities in the UK. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 179(2), 453-479. doi:10.1111/rssa.12123
- Chambers, S. (2016). Regression discontinuity design: A guide for strengthening causal inference in HRD. *European Journal of Training and Development*, 40(8/9), 615-637.
- Chandra, S., Ray, S., & Goswami, R. T. (2017, 5-7 Jan. 2017). *Big data security: Survey on frameworks and algorithms*. Paper presented at the 2017 IEEE 7th International Advance Computing Conference (IACC).

- Chang, V., Kuo, Y.-H., & Ramachandran, M. (2016). Cloud computing adoption framework: A security framework for business clouds. *Future Generation Computer Systems*, 57, 24-41.
- Chen, M., Mao, S., & Liu, Y. (2014). Big data: A survey. *Mobile Networks and Applications*, 19(2), 171-209. doi:http://dx.doi.org/10.1007/s11036-013-0489-0
- Chin Lay, G., & Balakrishnan, V. (2017). Predicting acceptance of mobile technology for aiding student-lecturer interactions: An empirical study. *Australasian Journal of Educational Technology*, 33(2), 143-158.
- Chuen, D. L. K. (2015). *Handbook of digital currency: Bitcoin, innovation, financial instruments, and big data*: Academic Press.
- Clarke, P., O'Connor, R. V., & Leavy, B. (2016, 14-15 May 2016). *A complexity theory viewpoint on the software development process and situational context*. Paper presented at the 2016 IEEE/ACM International Conference on Software and System Processes (ICSSP).
- Cooke, P. (2017). 'Digital tech' and the public sector: What new role after public funding? *European Planning Studies*, 25(5), 739-754.
- Corcuera, J. (2012). New central limit theorems for functionals of gaussian processes and their applications. *Methodology & Computing in Applied Probability*, 14(3), 477-500. doi:10.1007/s11009-011-9236-9
- Cortez, P., Oliveira, N., & Ferreira, J. P. (2016). *Measuring user influence in financial microblogs: Experiments using StockTwits data*. Paper presented at the Proceedings of the 6th International Conference on Web Intelligence, Mining and Semantics, Nîmes, France.
- Costa, D. d. F., & Silva, N. F. F. d. (2018). *INF-UFG at FiQA 2018 Task 1: Predicting sentiments and aspects on financial tweets and news headlines*. Paper presented at the Companion Proceedings of the The Web Conference 2018, Lyon, France.
- Coughenour, C., Paz, A., de la Fuente-Mella, H., & Singh, A. (2016). Multinomial logistic regression to estimate and predict perceptions of bicycle and transportation infrastructure in a sprawling metropolitan area. *Journal of Public Health*, 38(4), e401-e408. doi:10.1093/pubmed/fdv179
- Coussement, K., Benoit, D., & Antioco, M. (2015). A Bayesian approach for incorporating expert opinions into decision support systems: A case study of online consumer-satisfaction detection. *Decision Support Systems*, 79, 24-32.
- Coutinho, J. F., Perrone-McGovern, K. M., & Gonçalves, Ó. F. (2017). The use of neuroimaging methodology in counselling psychology research: Promises, pitfalls, and recommendations. *Canadian Journal of Counselling and Psychotherapy*, 51(4), 327-348.

- Cresci, S., Lillo, F., Regoli, D., Tardelli, S., & Tesconi, M. (2018). Cashtag piggybacking: Uncovering spam and bot activity in stock microblogs on Twitter. *arXiv preprint arXiv:1804.04406*.
- Creswell, J. W. (2014). *Research design: Qualitative, quantitative, and mixed methods approaches*. Los Angeles: Sage publications.
- da Costa, F. R., Pelissari, A. S., & Gonzalez, I. V. D. P. (2018). Corporate image of public higher education institutions: Relevant factors to distance learning students. *Turkish Online Journal of Distance Education (TOJDE)*, 19(1), 117-135.
- Das, D., Sadiq, A. S., Mirjalili, S., & Noraziah, A. (2017). *Hybrid Clustering-GWO-NARX neural network technique in predicting stock price*. Paper presented at the Journal of Physics: Conference Series.
- Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, 13(3), 319-340.
- Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, 319-340.
- Davis Jr, F. D. (1986). *A technology acceptance model for empirically testing new end-user information systems: Theory and results*. Massachusetts Institute of Technology,
- Dayal, N., Maity, P., Srivastava, S., & Khondoker, R. (2016). Research trends in security and DDoS in SDN. *Security & Communication Networks*, 9(18), 6386-6411. doi:10.1002/sec.1759
- Dayal, U., Castellanos, M., Simitsis, A., & Wilkinson, K. (2009). *Data integration flows for business intelligence*. Paper presented at the Proceedings of the 12th International Conference on Extending Database Technology: Advances in Database Technology, Saint Petersburg, Russia.
- de Siqueira Santos, S., Takahashi, D. Y., Nakata, A., & Fujita, A. (2014). A comparative study of statistical methods used to identify dependencies between gene expression signals. *Briefings in Bioinformatics*, 15(6), 906-918.
- Deacy, A. D., Gayes, L. A., De Lurgio, S., & Wallace, D. P. (2016). Adaptation of the state-trait inventory for cognitive and somatic anxiety for use in children: A preliminary analysis. *Journal of Pediatric Psychology*, 41(9), 1033-1043. doi:10.1093/jpepsy/jsw030
- Dede, E., Govindaraju, M., Gunter, D., Canon, R. S., & Ramakrishnan, L. (2013). *Performance evaluation of a MongoDB and Hadoop platform for scientific data analysis*. Paper presented at the Proceedings of the 4th ACM workshop on Scientific cloud computing, New York, New York, USA.
- DeLuca, L. (2016). Statistics for library and information services: A primer for using open source r software for accessibility and visualization. *Library Quarterly*, 86(4), 468-469.

- Demchenko, Y., Turkmen, F., de Laat, C., Blanchet, C., & Loomis, C. (2016). *Cloud based big data infrastructure: Architectural components and automated provisioning*. Paper presented at the High Performance Computing & Simulation (HPCS), 2016 International Conference on.
- Depren, S. K., Aşkın, Ö. E., & Öz, E. (2017). Identifying the classification performances of educational data mining methods: A case study for TIMSS. *Educational Sciences: Theory & Practice*, 17(5), 1605-1623. doi:10.12738/estp.2017.5.0634
- Dogan, C. D. (2017). Applying bootstrap resampling to compute confidence intervals for various statistics with R. *Bootstrap Örnekleme Yönteminin Çeşitli İstatistikler için Güven Aralığının Hesaplanmasında R Yazılımı ile Kullanımı*. (68), 1-17. doi:10.14689/ejer.2017.68.1
- Dong, T., Yang, B., & Tian, T. (2015, 19-21 Dec. 2015). *Volatility analysis of Chinese stock market using high-frequency financial big data*. Paper presented at the 2015 IEEE International Conference on Smart City/SocialCom/SustainCom (SmartCity).
- Drton, M., & Plummer, M. (2017). A Bayesian information criterion for singular models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(2), 323-380. doi:10.1111/rssb.12187
- Du, X., Deng, L., & Qian, K. (2018). Current market top business scopes trend: A concurrent text and time series active learning study of NASDAQ and NYSE Stocks from 2012 to 2017. *Applied Sciences* (2076-3417), 8(5).
- Durcikova, A., Lee, A. S., & Brown, S. A. (2018). Making rigorous research relevant: Innovating statistical action research. *MIS Quarterly*, 42(1), 241-A213.
- Eckert, R., Meyer, S. K., & Stuermer, M. (2017). *How are open source practices possible within a medical diagnostics company?: Developing and testing a maturity model of inner source implementation*. Paper presented at the Proceedings of the 13th International Symposium on Open Collaboration, Galway, Ireland.
- Edmondson, A. C., & McManus, S. E. (2007). Methodological fit in management field research. *Academy of management review*, 32(4), 1246-1264.
- Elagamy, M. N., Stanier, C., & Sharp, B. (2018, 25-26 April 2018). *Stock market random forest-text mining system mining critical indicators of stock market movements*. Paper presented at the 2018 2nd International Conference on Natural Language and Speech Processing (ICNLSP).
- Eleye-Datubo, A. G., & Wall, A. (2008). Marine and offshore safety assessment by incorporative risk modeling in a fuzzy-Bayesian network of an induced mass assignment paradigm. *Risk Analysis: An Official Publication Of The Society For Risk Analysis*, 28(1), 95-112. doi:10.1111/j.1539-6924.2008.01004.x

- Elsayed, M. N., & Qasem, M. M. (2014). Analysis of some algorithms for clustering data objects. *International Journal of Machine Learning and Computing*, 4(1), 99-105. doi:http://dx.doi.org/10.7763/IJMLC.2014.V4.394
- Endriss, U., Haan, R. d., & Szeider, S. (2015). *Parameterized complexity results for agenda safety in judgment aggregation*. Paper presented at the Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems, Istanbul, Turkey.
- Erdoğan, D., & Gurov, A. (2016). The effect of regulatory and risk management advancement on non-performing loans in European banking, 2000-2011. *International Advances in Economic Research*, 22(3), 249. doi:10.1007/s11294-016-9591-y
- Favario, L., & Masala, E. (2017). A new architecture for cross-repository creation and sharing of educational resources. *International Journal of Emerging Technologies in Learning*, 12(2), 185-209. doi:10.3991/ijet.v12i02.6058
- Ferrández, A., Maté, A., Peral, J., Trujillo, J., De Gregorio, E., & Aufaure, M.-a. (2016). A framework for enriching data warehouse analysis with question answering systems. *Journal of Intelligent Information Systems*, 46(1), 61-82. doi:http://dx.doi.org/10.1007/s10844-014-0351-2
- Flaster, A., & DesJardins, S. L. (2014). Applying regression discontinuity design in institutional research. *New Directions for Institutional Research*, 2014(161), 3-20. doi:10.1002/ir.20064
- Flood, M. D., Jagadish, H., & Raschid, L. (2016). Big data challenges and opportunities in financial stability monitoring. *Banque de France, Financial Stability Review*, 20.
- Flores, W. C., & Pineda, G. M. (2016, 2-4 Nov. 2016). *A type-2 fuzzy logic system approach to train Honduran coffee cuppers*. Paper presented at the 2016 IEEE Latin American Conference on Computational Intelligence (LA-CCI).
- Fotache, M. (2016). Data processing languages for business intelligence. SQL vs. R. *Informatica Economica*, 20(1), 48-61. doi:http://dx.doi.org/10.12948/issn14531305/20.1.2016.05
- Fowe, T., Nouiri, I., Ibrahim, B., Karambiri, H., & Paturel, J.-E. (2015). OPTIWAM: an intelligent tool for optimizing irrigation water management in coupled reservoir-groundwater systems. *Water Resources Management*, 29(10), 3841-3861.
- Fox, M., & Vaidyanathan, G. (2016). Impacts of healthcare big data: A framework with legal and ethical insights. *Issues in Information Systems*, 17(3).
- Franke, B., Plante, J. F., Roscher, R., Lee, E. s. A., Smyth, C., Hatefi, A., . . . Reid, N. (2016). Statistical inference, learning and models in big data. *International Statistical Review*, 84(3), 371-389. doi:10.1111/insr.12176

- Fujimoto, R. M., Bagrodia, R., Bryant, R. E., Chandy, K. M., Jefferson, D., Misra, J., . . . Unger, B. (2017). *Parallel discrete event simulation: The making of a field*. Paper presented at the Simulation Conference (WSC), 2017 Winter.
- Gadepally, V., Kepner, J., & Reuther, A. (2016). Storage and database management for big data. *Big Data: Storage, Sharing, and Security*, 15.
- Gashler, M. S., Smith, M. R., Morris, R., & Martinez, T. (2016). Missing value imputation with unsupervised backpropagation. *Computational Intelligence*, 32(2), 196-215. doi:10.1111/coin.12048
- George, J., Janaki, L., & Parameswaran Gomathy, J. (2016). Statistical downscaling using local polynomial regression for rainfall predictions - a case study. *Water Resources Management*, 30(1), 183-193. doi:http://dx.doi.org/10.1007/s11269-015-1154-0
- Gepp, A., Linnenluecke, M. K., Terrence, J. O. N., & Smith, T. (2018). Big data techniques in auditing research and practice: Current trends and future opportunities. *Journal of Accounting Literature*, 40, 102-115. doi:http://dx.doi.org/10.1016/j.acclit.2017.05.003
- Ghosh, A., Julliard, C., & Taylor, A. P. (2017). What Is the Consumption-CAPM Missing? An Information-Theoretic Framework for the Analysis of Asset Pricing Models. *Review of Financial Studies*, 30(2), 442-504. doi:10.1093/rfs/hhw075
- Gibbons, C., Richards, S., Valderas, J. M., & Campbell, J. (2017). Supervised machine learning algorithms can classify open-text feedback of doctor performance with human-level accuracy. *Journal of Medical Internet Research*, 19(3), 1-1. doi:10.2196/jmir.6533
- Glaser, F. (2017). Pervasive decentralisation of digital infrastructures: A framework for blockchain enabled system and use case analysis.
- Glicksman, R. L., Marked, D. L., & Monteleoni, C. (2017). Technological innovation, data analytics, and environmental enforcement. *Ecology Law Quarterly*, 44(1), 41-88. doi:10.15779/Z38GX44T46
- Golbayani, P., & Bozdog, D. (2017, Nov. 27 2017-Dec. 1 2017). *Detection of rare events in multidimensional financial datasets with zonoid depth functions*. Paper presented at the 2017 IEEE Symposium Series on Computational Intelligence (SSCI).
- Goldstein, M. M. (2017). Revising the common rule: Ethics, scientific advancement, and public policy in conflict. *Journal of Law, Medicine & Ethics*, 45(3), 452-459. doi:10.1177/1073110517737546
- Gorisek, A., & Pahor, M. (2017). Missing value imputation using contemporary computer capabilities: An application to financial statements data in large panels. [Vstavljanje manjkajocih vrednosti v podatke z uporabo sodobnih racunalniških zmožnosti: Uporaba na velikih panelih financnih izkazov podjetij]. *Economic and Business Review for Central and South - Eastern Europe*, 19(1), 97-119,125. doi:http://dx.doi.org/10.15458/85451.38

- Goswami, A., Han, W., Wang, Z., & Jiang, A. (2015, Oct. 29 2015-Nov. 1 2015). *Controlled experiments for decision-making in e-Commerce search*. Paper presented at the 2015 IEEE International Conference on Big Data (Big Data).
- Goudarzi, H., & Pedram, M. (2016). Hierarchical SLA-driven resource management for peak power-aware and energy-efficient operation of a cloud datacenter. *IEEE Transactions on Cloud Computing*, 4(2), 222-236. doi:10.1109/TCC.2015.2474369
- Graham, S. S., Kim, S.-Y., DeVasto, D. M., & Keith, W. (2015). Statistical genre analysis: Toward big data methodologies in technical communication. *Technical Communication Quarterly*, 24(1), 70-104. doi:10.1080/10572252.2015.975955
- Gurbuz, S., & Aydin, G. (2017). *Classification of scientific papers with big data technologies*. Paper presented at the Computer Science and Engineering (UBMK), 2017 International Conference on.
- Gwynllyw, D. R., Weir, I. S., & Henderson, K. L. (2016). Using DEWIS and R for multi-staged statistics e-Assessments. *Teaching Mathematics & its Applications*, 35(1), 14-26. doi:10.1093/teamat/hrv018
- Ha, M. H., Lee, S., & Moon, B.-R. (2016). *A genetic algorithm for rule-based chart pattern search in stock market prices*. Paper presented at the Proceedings of the Genetic and Evolutionary Computation Conference 2016, Denver, Colorado, USA.
- Hadioui, A., El Faddouli, N.-e., Touimi, Y. B., & Bennani, S. (2017). Machine learning based on big data extraction of massive educational knowledge. *International Journal of Emerging Technologies in Learning*, 12(11), 151-167. doi:10.3991/ijet.v12.i11.7460
- Halford, S., & Savage, M. (2017). Speaking sociologically with big data: Symphonic social science and the future for big data research. *Sociology*, 51(6), 1132-1148.
- Halstead, R. J. (2015). Using multithreaded techniques to mask memory latency on FPGA accelerators.
- Hancu, L. (2012). *Acquiring business ecosystem's intelligence through Large-Scale Collaborative Systems*. Paper presented at the Information Technology Interfaces (ITI), Proceedings of the ITI 2012 34th International Conference on.
- Hanifatunnisa, R., & Rahardjo, B. (2017). *Blockchain based e-voting recording system design*. Paper presented at the Telecommunication Systems Services and Applications (TSSA), 2017 11th International Conference on.
- Haq, A., & Al-Omari, A. I. (2016). Bayes estimation and prediction of a three component mixture of Rayleigh distribution under type-i censoring. *Investigación Operacional*, 37(1), 22-37.
- Harry M, M., & Nilufer, U. (1996). The likelihood of various stock market return distributions, part 1: Principles of inference. *Journal of Risk & Uncertainty*, 13(3), 207-219.

- Hartmanis, J. (1975). Formal languages (Arto Salomaa). *SIAM Review*, 17(1), 179-172. doi:<http://dx.doi.org/10.1137/1017022>
- Hartmanis, J., & Stearns, R. E. (1965). On the computational complexity of algorithms. *Transactions of the American Mathematical Society*, 117, 285-306.
- Haug, F. S. (2016). *Bad big data science*. Paper presented at the Big Data (Big Data), 2016 IEEE International Conference on.
- Hayashi, E. C. S., & Baranauskas, M. C. C. (2013). Affectibility in educational technologies: A socio-technical perspective for design. *Journal of Educational Technology & Society*, 16(1), 57-n/a.
- Hong, L., Hou, Y., Jing, J., Wang, A., & Litvin, D. A. (2015). Deadlock prevention policy with behavioral optimality or suboptimality achieved by the redundancy identification of constraints and the rearrangement of monitors. *Discrete Dynamics in Nature and Society*, 2015.
- Hordijk, W., Hordijk, A., & Heidergott, B. (2015). A genetic algorithm for finding good balanced sequences in a customer assignment problem with no state information. *Asia-Pacific Journal of Operational Research*, 32(3), -1. doi:10.1142/S0217595915500153
- Hsieh, S.-H., Lee, S.-M., Li, C.-S., & Tu, S.-H. (2016). An alternative to unrelated randomized response techniques with logistic regression analysis. *Statistical Methods & Applications*, 25(4), 601-621. doi:10.1007/s10260-016-0351-1
- Hu, F., Yang, C., Schnase, J. L., Duffy, D. Q., Xu, M., Bowen, M. K., . . . Song, W. (2018). ClimateSpark: An in-memory distributed computing framework for big climate data analytics. *Computers & Geosciences*, 115, 154-166.
- Hu, Y., Sulek, D., Carella, A., Cox, J., Frame, A., Cipriano, K., & Wang, H.-x. (2016). Efficient Distributed Vulnerability Assessment by Utilizing Miniaturized Computers.
- Huang, J.-Y., Lange, C., & Auer, S. (2015). *Streaming transformation of XML to RDF using XPath-based mappings*. Paper presented at the Proceedings of the 11th International Conference on Semantic Systems, Vienna, Austria.
- Huang, J., Mozafari, B., Schoenebeck, G., & Wenisch, T. F. (2017). *A top-down approach to achieving performance predictability in database systems*. Paper presented at the Proceedings of the 2017 ACM International Conference on Management of Data.
- Huang, X. (2018). Macroeconomic news announcements, systemic risk, financial market volatility, and jumps. *Journal of Futures Markets*, 38(5), 513-534.
- Hubie, C. (2017). The tractability frontier of graph-like first-order query sets. *Journal of the ACM*, 64(4), 1-29. doi:10.1145/3073409

- Iancu, B., & Georgescu, T. M. (2018). Saving large semantic data in cloud: A survey of the main DBaaS solutions. *Informatica Economica*, 22(1), 5-16.
doi:10.12948/issn14531305/22.1.2018.01
- Ianni, M., Masciari, E., Mazzeo, G. M., & Zaniolo, C. (2018). *Efficient big data clustering*. Paper presented at the Proceedings of the 22nd International Database Engineering & Applications Symposium, Villa San Giovanni, Italy.
- Immekus, J. C., & Atitya, B. (2016). The predictive validity of interim assessment scores based on the full-information bifactor model for the prediction of end-of-grade test performance. *Educational Assessment*, 21(3), 176-195.
doi:10.1080/10627197.2016.1202108
- Jain, P., Gyanchandani, M., & Khare, N. (2017). Privacy and security concerns in healthcare big data: An innovative prescriptive. *Journal of Information Assurance & Security*, 12(1), 18-30.
- Jammula, N., Chockalingam, S. P., & Aluru, S. (2017). *Distributed memory partitioning of high-throughput sequencing datasets for enabling parallel genomics analyses*. Paper presented at the Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics, Boston, Massachusetts, USA.
- Jin, F., Wang, W., Chakraborty, P., Self, N., Chen, F., & Ramakrishnan, N. (2017). *Tracking multiple social media for stock market event prediction*. Paper presented at the Industrial Conference on Data Mining.
- Jones, A. M., Lomas, J., Moore, P. T., & Rice, N. (2016). A quasi-Monte-Carlo comparison of parametric and semiparametric regression methods for heavy-tailed and non-normal data: an application to healthcare costs. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 179(4), 951-974. doi:10.1111/rssa.12141
- Kaal, W. A., & Vermeulen, E. P. M. (2017). How to regulate disruptive innovation -- from facts to data. *Jurimetrics: The Journal of Law, Science & Technology*, 57(2), 170-190.
- Kachouie, N. N., Gerke, T., Huybers, P., & Schwartzman, A. (2015). Nonparametric regression for estimation of spatiotemporal mountain glacier retreat from satellite images. *IEEE Transactions on Geoscience and Remote Sensing*, 53(3), 1135-1149.
doi:10.1109/TGRS.2014.2334643
- Kalkan, Ö. K., & Kelecioğlu, H. (2017). The effect of sample size on parametric and nonparametric factor analytical methods. *Educational Sciences: Theory & Practice*, 17(1), 153-171. doi:10.12738/estp.2016.1.0220
- Kamakshaiah, M. (2013). Data mining and analysis by R language for business research: A case study on stress and its influence on health. *International Journal of Business Analytics and Intelligence*, 1(2), 14-21.

- Kandogan, E., Balakrishnan, A., Haber, E. M., & Pierce, J. S. (2014). From data to insight: work practices of analysts in the enterprise. *IEEE computer graphics and applications*, 34(5), 42-50.
- Kang, M., & Lee, J.-G. (2016). *A comparative analysis of iterative MapReduce systems*. Paper presented at the Proceedings of the Sixth International Conference on Emerging Databases: Technologies, Applications, and Theory, Jeju, Republic of Korea.
- Keshavarz, H., & Abadeh, M. S. (2016, 9-11 March 2016). *SubLex: Generating subjectivity lexicons using genetic algorithm for subjectivity classification of big social data*. Paper presented at the 2016 1st Conference on Swarm Intelligence and Evolutionary Computation (CSIEC).
- Khedr, A. E., & Yaseen, N. (2017). Predicting stock market behavior using data mining technique and news sentiment analysis. *International Journal of Intelligent Systems and Applications*, 9(7), 22.
- Khoja, L., Chipulu, M., & Jayasekera, R. (2016). Analysing corporate insolvency in the Gulf Cooperation Council using logistic regression and multidimensional scaling. *Review of Quantitative Finance & Accounting*, 46(3), 483-518. doi:10.1007/s11156-014-0476-y
- Kim, K.-H., Hwang, M.-S., Jae, E.-Y., Jun, S.-H., & Kwon, M.-C. (2016). A study on message queue safe proper time for open API Fast IDentity online Fintech architecture. *International Journal of Software Engineering and Its Applications*, 10(5), 33-44.
- Kim, Y., & Steiner, P. (2016). Quasi-experimental designs for causal inference. *Educational Psychologist*, 51(3/4), 395-405. doi:10.1080/00461520.2016.1207177
- Kisi, O., & Ozkan, C. (2017). A new approach for modeling sediment-discharge relationship: Local weighted linear regression. *Water Resources Management*, 31(1), 1-23. doi:http://dx.doi.org/10.1007/s11269-016-1481-9
- Klein, D., Nikitina, E., & Curtillet, J.-C. (2016). Forecasting swap spreads: A Bayesian approach. *Journal of Fixed Income*, 26(2), 40-53.
- Koehler, J. J., & Meixner Jr, J. B. (2016). An empirical research agenda for the forensic sciences. *Journal of Criminal Law & Criminology*, 106(1), 1-33.
- Köhler, T., Landis, R. S., & Cortina, J. M. (2017). From the editors: Establishing methodological rigor in quantitative management learning and education research: The role of design, statistical methods, and reporting standards. *Academy of Management Learning & Education*, 16(2), 173-192. doi:10.5465/amle.2017.0079
- Krulewich, D., Yin, J., Bundick, B. H., & Zeng, Y. (2015). *Performance assessment of real-time estimation of continuous-time stochastic volatility of financial data on GPUs*. Paper presented at the Proceedings of the 2015 XSEDE Conference: Scientific Advancements Enabled by Enhanced Cyberinfrastructure, St. Louis, Missouri.

- Kung, S.-y. (2017). Discriminant component analysis for privacy protection and visualization of big data. *Multimedia Tools and Applications*, 76(3), 3999-4034.
doi:<http://dx.doi.org/10.1007/s11042-015-2959-9>
- Laborie, S., Manzat, A., & Sedes, F. (2018). Managing and querying efficiently distributed semantic multimedia metadata collections. *IEEE MultiMedia*, 1-1.
doi:10.1109/MMUL.2009.68
- Lahmidani, H., & Beqqali, O. E. (2016). *Mining public administration big data for a financial benefit Word*. Paper presented at the Proceedings of the International Conference on Big Data and Advanced Wireless Technologies, Blagoevgrad, Bulgaria.
- Lambert, N. S., Ostrovsky, M., & Panov, M. (2014). *Strategic trading in informationally complex environments*. Paper presented at the Proceedings of the fifteenth ACM conference on Economics and computation, Palo Alto, California, USA.
- Lan, L. (2017). Data visualization and retrieval based convenience store location model of fresh products. *Journal of Digital Information Management*, 15(3), 125-134.
- Landsberg, J. M., & Ressayre, N. (2016). *Permanent v. Determinant: An exponential lower bound assuming symmetry*. Paper presented at the Proceedings of the 2016 ACM Conference on Innovations in Theoretical Computer Science, Cambridge, Massachusetts, USA.
- Landset, S., Khoshgoftaar, T. M., Richter, A. N., & Hasanin, T. (2015). A survey of open source tools for machine learning with big data in the Hadoop ecosystem. *Journal of Big Data*, 2(1), 1.
- Lee, C., & Paik, I. (2017, 8-10 Nov. 2017). *Stock market analysis from Twitter and news based on streaming big data infrastructure*. Paper presented at the 2017 IEEE 8th International Conference on Awareness Science and Technology (iCAST).
- Lee, H., Kweon, E., Kim, M., & Chai, S. (2017). Does implementation of big data analytics improve firms' market value? Investors' reaction in stock market. *Sustainability*, 9(6), 978.
- Leijon, A., Henter, G. E., & Dahlquist, M. (2016). Bayesian analysis of phoneme confusion matrices. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 24(3), 469-482.
doi:10.1109/taslp.2015.2512039
- Leppink, J., & van Merriënboer, J. J. G. (2015). The beast of aggregating cognitive load measures in technology-based learning. *Journal of Educational Technology & Society*, 18(4), 230-245.
- Levy, D., Livingood, R., & Maranga, K. M. (2016). *Perspectives of qualitative research methods*. Schaumburg, IL: Words of Wisdom.

- Li, C.-S. (2016). A test for the linearity of the nonparametric part of a semiparametric logistic regression model. *Journal of Applied Statistics*, 43(3), 461-475. doi:10.1080/02664763.2015.1070803
- Li, H. (2015). A genetic algorithm using a finite search space for solving nonlinear/linear fractional bilevel programming problems. *Annals of Operations Research*, 235(1), 543-558. doi:10.1007/s10479-015-1878-5
- Li, Q., Chen, Y., Wang, J., Chen, Y., & Chen, H. (2018). Web media and stock markets : A survey and future directions from a big data perspective. *IEEE Transactions on Knowledge and Data Engineering*, 30(2), 381-399. doi:10.1109/TKDE.2017.2763144
- Liang, T.-Y., Yeh, L.-W., & Wu, C.-H. (2018). *A visual MapReduce program development environment for heterogeneous computing on clouds*. Paper presented at the Proceedings of the 2018 International Conference on Computing and Data Engineering, Shanghai, China.
- Lightstone, S., Ohanian, R., Haide, M., Cho, J., Springgay, M., & Steinbach, T. (2017). *Making big data simple with dashDB local*. Paper presented at the Data Engineering (ICDE), 2017 IEEE 33rd International Conference on.
- Linton, O., & Mahmoodzadeh, S. (2018). *Implications of high-frequency trading for security markets*. Retrieved from
- Liu, Q., Okui, R., & Yoshimura, A. (2013). Generalized least squares model averaging. In. St. Louis: Federal Reserve Bank of St Louis.
- Liu, Q., Okui, R., & Yoshimura, A. (2016). Generalized least squares model averaging. *Econometric Reviews*, 35(8-10), 1692-1752. doi:10.1080/07474938.2015.1092817
- Liu, S., Yu, H., Miao, C., & Kot, A. C. (2013). *A fuzzy logic based reputation model against unfair ratings*. Paper presented at the Proceedings of the 2013 international conference on Autonomous agents and multi-agent systems, St. Paul, MN, USA.
- Liu, X., Golab, L., Golab, W., Ilyas, I. F., & Jin, S. (2016). Smart meter data analytics: Systems, algorithms, and benchmarking. *ACM Trans. Database Syst.*, 42(1), 1-39. doi:10.1145/3004295
- Lodewyk, K. R., & Mandigo, J. L. (2017). Early validation evidence of a canadian practitioner-based assessment of physical literacy in physical education: Passport for life. *Physical Educator*, 74(3), 441-475. doi:10.18666/TPE-2017-V74-I3-7459
- Louzada, F., & Granzotto, D. (2016). The transmuted log-logistic regression model: a new model for time up to first calving of cows. *Statistical Papers*, 57(3), 623-640. doi:10.1007/s00362-015-0671-5

- Luukko, P. J., Helske, J., & Räsänen, E. (2016). Introducing libeemd: A program package for performing the ensemble empirical mode decomposition. *Computational Statistics*, 31(2), 545-557. doi:http://dx.doi.org/10.1007/s00180-015-0603-9
- Maier, J., Hampe, J. F., & Jahn, N. (2016). Breaking out of the lab. *Public Opinion Quarterly*, 80(2), 542-553. doi:10.1093/poq/nfw010
- Mauri, A., & Tettamanzi, A. G. B. (2012). *Investigating the role of social behavior in financial markets through agent-based simulation*. Paper presented at the Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems - Volume 3, Valencia, Spain.
- McNally, J. J., Martin, B. C., Honig, B., Bergmann, H., & Piperopoulos, P. (2016). Toward rigor and parsimony: a primary validation of Kolvereid's (1996) entrepreneurial attitudes scales. *Entrepreneurship & Regional Development*, 28(5/6), 358-379. doi:10.1080/08985626.2016.1154985
- Miller, H. J., & Goodchild, M. F. (2015). Data-driven geography. *GeoJournal*, 80(4), 449-461. doi:http://dx.doi.org/10.1007/s10708-014-9602-6
- Ming, Z., Luo, C., Gao, W., Han, R., Yang, Q., Wang, L., & Zhan, J. (2013). *BDGS: A scalable big data generator suite in big data benchmarking*. Paper presented at the Workshop on Big Data Benchmarks.
- Miyauchi, Y., Takahashi, R., & Kinomura, M. (2015). Vertically integrated database system combining advanced hardware with database technologies. *FUJITSU Sci. Tech. J*, 51(2), 55-61.
- Mjöberg, B. (2018). Particle disease, late loosening and Occam's razor. *Acta Orthopaedica*, 89(1), 136-136.
- Mohan, P. (2016). Caribbean diversification and development. *World Economy*, 39(9), 1434-1453. doi:10.1111/twec.12387
- Moon, M. R. (2017). Can central IRBs replace local review? *Journal of Law, Medicine & Ethics*, 45(3), 348-351. doi:10.1177/1073110517737533
- More, A. M., Rathod, P. U., Patil, R. H., & Sarode, D. R. (2018). Stock market prediction system using Hadoop. *International Journal of Engineering Science*, 16138.
- Moreno, M., Harwell, M., Guzey, S. S., Phillips, A., & Moore, T. J. (2016). Complex applications of HLM in studies of science and mathematics achievement: Cross-classified random effects models. *School Science & Mathematics*, 116(6), 338-351. doi:10.1111/ssm.12191
- Muller, C. L., & Kidd, C. (2014). Debugging geographers: Teaching programming to non-computer scientists. *Journal of Geography in Higher Education*, 38(2), 175-192. doi:10.1080/03098265.2014.908275

- Nasdaq. (2018). Nasdaq. Retrieved from <https://www.nasdaq.com/>
- Nasdaq.com. (2018). Company List (NASDAQ, NYSE, & AMEX). Retrieved from <https://www.nasdaq.com/screening/company-list.aspx>
- Nawaratne, R., Bandaragoda, T., Adikari, A., Alahakoon, D., Silva, D. D., & Yu, X. (2017, Oct. 29 2017-Nov. 1 2017). *Incremental knowledge acquisition and self-learning for autonomous video surveillance*. Paper presented at the IECON 2017 - 43rd Annual Conference of the IEEE Industrial Electronics Society.
- Nelson, R. (2017). Five-minute science lesson: Bayesian and frequentist statistics - what's the deal? *Polygraph*, 50(4), 89-95.
- Neto, P. S. G. d. M., & Ferreira, T. A. E. (2016, 24-29 July 2016). *Applying a general hybrid intelligent system for ultra-high-frequency stock market forecasting*. Paper presented at the 2016 International Joint Conference on Neural Networks (IJCNN).
- Niazi, M. A., Hussain, A., & Kolberg, M. (2017). Verification & validation of agent based simulations using the VOMAS (virtual overlay multi-agent system) approach. *arXiv preprint arXiv:1708.02361*.
- Nichols, T. E., Das, S., Eickhoff, S. B., Evans, A. C., Glatard, T., Hanke, M., . . . Poline, J.-B. (2017). Best practices in data analysis and sharing in neuroimaging using MRI. *Nature neuroscience*, 20(3), 299.
- Nikoo, M. R., Kerachian, R., Malakpour-Estalaki, S., Bashi-Azghadi, S. N., & Azimi-Ghadikolaee, M. M. (2011). A probabilistic water quality index for river water quality assessment: a case study. *Environmental Monitoring And Assessment*, 181(1-4), 465-478. doi:10.1007/s10661-010-1842-4
- Noble, H., & Smith, J. (2015). Issues of validity and reliability in qualitative research. *Evidence Based Nursing*, 18(2), 34-35.
- Ocampo-Vega, R., Sanchez-Ante, G., de Luna, M. A., Vega, R., Falcón-Morales, L. E., & Sossa, H. (2016). Improving pattern classification of DNA microarray data by using PCA and logistic regression. *Intelligent Data Analysis*, 20, S53-S67. doi:10.3233/IDA-160845
- Ohayon, J. L., Cousins, E., Brown, P., Morello-Frosch, R., & Brody, J. G. (2017). Researcher and institutional review board perspectives on the benefits and challenges of reporting back biomonitoring and environmental exposure results. *Environmental Research*, 153, 140-149.
- Olaniyan, R., Stamate, D., Ouarbya, L., & Logofatu, D. (2015, 19-21 Oct. 2015). *Sentiment and stock market volatility predictive modelling: A hybrid approach*. Paper presented at the 2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA).
- Overholt, K., & Ezekoye, O. (2015). Quantitative testing of fire scenario hypotheses: A Bayesian Inference approach. *Fire Technology*, 51(2), 335-367. doi:10.1007/s10694-013-0384-z

- Padmavathi, T., & Krishnamurthy, M. (2017). Semantic web tools and techniques for knowledge organization: An overview. *Knowledge Organization*, 44(4), 273-290.
- Park, H. M. (2009). Comparing group means: t-tests and one-way ANOVA using Stata, SAS, R, and SPSS.
- Pedro, S. D. S., & Estevam R. Hruschka, J. (2012). *Collective intelligence as a source for machine learning self-supervision*. Paper presented at the Proceedings of the 4th International Workshop on Web Intelligence & Communities, Lyon, France.
- Phillips, N. D., Neth, H., Woike, J. K., & Gaissmaier, W. (2017). FFTrees: A toolbox to create, visualize, and evaluate fast-and-frugal decision trees. *Judgment and Decision Making*, 12(4), 344-368.
- Piao, G., & Breslin, J. G. (2018). *Financial aspect and sentiment predictions with deep neural networks: An ensemble approach*. Paper presented at the Companion Proceedings of the The Web Conference 2018, Lyon, France.
- Plekhanov, D. (2017). *Official statistics embrace big data: a review of current and developing international practice*. Paper presented at the Proceedings of the International Conference on Electronic Governance and Open Society: Challenges in Eurasia, St. Petersburg, Russia.
- Poorthuis, L., Goergen, K., Sharples, W., & Kollet, S. (2016). *Implementation of parallel NetCDF in the ParFlow hydrological model: A code modernisation effort as part of a big data handling strategy*. Paper presented at the NIC Symposium 2016.
- Praveen, S., & Chandra, U. (2017). NoSQL products: IT giants perspectives. *International Journal of Computational Intelligence Research*, 13(8), 2125-2133.
- Priyadharshini, R. A., & Arivazhagan, S. (2014). A quaternionic wavelet transform-based approach for object recognition. *Defence Science Journal*, 64(4), 350-357. doi:10.14429/dsj.64.4503
- Pusala, M. K., Salehi, M. A., Katukuri, J. R., Xie, Y., & Raghavan, V. (2016). Massive data analysis: Tasks, tools, applications, and challenges. In *Big Data Analytics* (pp. 11-40): Springer.
- Quboa, Q., & Mehandjiev, N. (2017). *Creating intelligent business systems by utilising big data and semantics*. Paper presented at the Business Informatics (CBI), 2017 IEEE 19th Conference on.
- Quercia, R., Pennington-Cross, A., & Tian, C. (2016). Differential impacts of structural and cyclical unemployment on mortgage default and prepayment. *Journal of Real Estate Finance & Economics*, 53(3), 346-367. doi:10.1007/s11146-014-9461-4

- Racicot, F.-É., Rentz, W., & Kahl, A. (2017). Rolling regression analysis of the Pástor-Stambaugh model: Evidence from robust instrumental variables. *International Advances in Economic Research*, 23(1), 75. doi:10.1007/s11294-016-9620-x
- Radojković, P., Carpenter, P. M., Moreto, M., Čakarević, V., Verdu, J., Pajuelo, A., . . . Valero, M. (2016). Thread assignment in multicore/multithreaded processors: A statistical approach. *IEEE Transactions on Computers*, 65(1), 256-269.
- Raimundo, M. S., & Okamoto, J. (2018, 23-25 March 2018). *SVR-wavelet adaptive model for forecasting financial time series*. Paper presented at the 2018 International Conference on Information and Computer Technologies (ICICT).
- Raja, N., Çiçek, I., Türkoğlu, N., Aydin, O., & Kawasaki, A. (2017). Landslide susceptibility mapping of the Sera River Basin using logistic regression model. *Natural Hazards*, 85(3), 1323-1346. doi:10.1007/s11069-016-2591-7
- Raschid, L., Xu, Z., & Zotkina, E. (2017). *Exploring financial relationships using probabilistic topic models (demonstration paper)*. Paper presented at the Proceedings of the 3rd International Workshop on Data Science for Macro--Modeling with Financial and Economic Datasets, Chicago, IL, USA.
- Rathod, R. R., & Garg, R. D. (2017). Design of electricity tariff plans using gap statistic for K-means clustering based on consumers monthly electricity consumption data. *International Journal of Energy Sector Management*, 11(2), 295-310.
- Ravi, V., & Kamaruddin, S. (2017). *Big data analytics enabled smart financial services: Opportunities and challenges*. Paper presented at the International Conference on Big Data Analytics.
- Razali, N. M., Shamsudin, N. R., Maarof, N. N. N. A., Hadi, A. A., & Ismail, A. (2012, 10-12 Sept. 2012). *A comparison of normality tests using SPSS, SAS and MINITAB: An application to Health Related Quality of Life data*. Paper presented at the 2012 International Conference on Statistics in Science, Business and Engineering (ICSSBE).
- Reio, T. G. (2016). Nonexperimental research: Strengths, weaknesses and issues of precision. *European Journal of Training and Development*, 40(8/9), 676-690.
- Rembart, F., & Soliman, E. (2017). Loan demand in Jordanian microfinance market: interest rate elasticity and loan-acceptance prediction via logistic regression. *Enterprise Development & Microfinance*, 28(4), 305-322. doi:10.3362/1755-1986.16-00009
- Ren, R., Wu, D. D., & Liu, T. (2018). Forecasting stock market movement direction using sentiment analysis and support vector machine. *IEEE Systems Journal*, 1-11. doi:10.1109/JSYST.2018.2794462
- Resnik, D. B. (2016). Bioethics and climate change: A response to Macpherson and Valles. *Bioethics*, 30(8), 649-652. doi:10.1111/bioe.12261

- Rhoads, C. (2014). Under what circumstances does external knowledge about the correlation structure improve power in cluster randomized designs? *Journal of Research on Educational Effectiveness*, 7(2), 205-224. doi:10.1080/19345747.2013.848962
- Rodríguez-Narciso, S., Christen, J., Rodríguez-Narciso, S., & Christen, J. A. (2016). Optimal sequential Bayesian analysis for degradation tests. *Lifetime Data Analysis*, 22(3), 405-428. doi:10.1007/s10985-015-9339-7
- Rong, W., Wenjing, Y., Zhang, A. N., Nguyen Quoc, C., & Akcan, O. (2016, 9-12 Oct. 2016). *Spatio-temporal route mining and visualization for busy waterways*. Paper presented at the 2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC).
- Rosal, M. C., Heyden, R., Mejilla, R., DePaoli, M. R., Veerappa, C., & Wiecha, J. M. (2012). Design and methods for a comparative effectiveness pilot study: Virtual world vs. Face-to-face diabetes self-management. *Journal of Medical Internet Research*, 14(6), 23-23. doi:10.2196/resprot.2415
- Rose, J., & Lennerholt, C. (2017). Low cost text mining as a strategy for qualitative researchers. *Electronic Journal of Business Research Methods*, 15(1), 2-16.
- Rossmo, D. K. (2016). Case rethinking: a protocol for reviewing criminal investigations. *Police Practice & Research*, 17(3), 212-228. doi:10.1080/15614263.2014.978320
- Ryan, J. A., & Ulrich, J. M. (2017). quantmod: Quantitative Financial Modelling Framework. R package version 0.4-12. Retrieved from <https://CRAN.R-project.org/package=quantmod>
- Rzevski, G. (2016, 12-16 June 2016). *Managing complexity: Theory and practice*. Paper presented at the 2016 11th System of Systems Engineering Conference (SoSE).
- Sagar, S., Nerurkar, N. W., & Sharma, A. (2010). A soft computing based approach to estimate reusability of software components. *SIGSOFT Softw. Eng. Notes*, 35(4), 1-4. doi:10.1145/1811226.1811235
- Samadi, Y., Zbakh, M., & Tadonki, C. (2016, 24-26 May 2016). *Comparative study between Hadoop and Spark based on Hibench benchmarks*. Paper presented at the 2016 2nd International Conference on Cloud Computing Technologies and Applications (CloudTech).
- Samarakoon, P. A., & Athukorala, D. A. S. (2017, 14-15 Sept. 2017). *System abnormality detection in stock market complex trading systems using machine learning techniques*. Paper presented at the 2017 National Information Technology Conference (NITC).
- Sampson, J. P. (2016). Guide to quantitative and qualitative dissertation research.
- Sandén, B. I. (2011). *Design of multithreaded software: The entity-life modeling approach*. Hoboken, NJ: John Wiley & Sons.

- Sandhu, N., & Mankotia, N. (2018). Perceptual impediments to stock market investment: Factor analysis approach. *SCMS Journal of Indian Management*, 15(1), 50-58.
- Sankari, S., & Bose, S. (2016, 8-9 April 2016). *Elaborative survey on storage technologies for XML big data: A real-time approach*. Paper presented at the 2016 International Conference on Recent Trends in Information Technology (ICRTIT).
- Sato, T., Takano, Y., Miyashiro, R., & Yoshise, A. (2016). Feature subset selection for logistic regression via mixed integer optimization. *Computational Optimization & Applications*, 64(3), 865-880. doi:10.1007/s10589-016-9832-2
- Scantamburlo, T. (2016). Machine learning in decisional process: a philosophical perspective. *SIGCAS Comput. Soc.*, 45(3), 218-224. doi:10.1145/2874239.2874270
- Schalock, R. L., Gomez, L. E., Verdugo, M. A., & Claes, C. (2017). Evidence and evidence-based practices: Are we there yet? *Intellectual and Developmental Disabilities*, 55(2), 112-119. doi:10.1352/1934-9556-55.2.112
- Schmidt, H., & Schwartz, J. L. (2016). The missions of national commissions: Mapping the forms and functions of bioethics advisory bodies. *Kennedy Institute of Ethics Journal*, 26(4), 431-VI.
- Schnipper, L. E. (2017). Central IRB review is an essential requirement for cancer clinical trials. *Journal of Law, Medicine & Ethics*, 45(3), 341-347. doi:10.1177/1073110517737532
- Schochet, P. Z. (2013). Statistical power for school-based RCTs with binary outcomes. *Journal of Research on Educational Effectiveness*, 6(3), 263-294. doi:10.1080/19345747.2012.725803
- Seal, B., & Hossain, S. J. (2015). Empirical Bayes estimation of parameters in Markov transition probability matrix with computational methods. *Journal of Applied Statistics*, 42(3), 508-519. doi:10.1080/02664763.2014.963525
- Sekar, K. R., Srinivasan, M., Ravidiandran, K. S., & Sethuraman, J. (2017, 20-22 July 2017). *Gold price estimation using a multi variable model*. Paper presented at the 2017 International Conference on Networks & Advances in Computational Technologies (NetACT).
- Sethia, D., Sheoran, S., & Saran, H. (2017). *Optimized MapFile based storage of small files in Hadoop*. Paper presented at the Proceedings of the 17th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing, Madrid, Spain.
- Sharma, V. K. (2017). Estimation and prediction for type-II Hybrid censored data follow flexible Weibull distribution. *Statistica*, 77(4), 385-414. doi:http://dx.doi.org/10.6092/issn.1973-2201/6779
- Shneiderman, B., & Plaisant, C. (2015). Sharpening analytic focus to cope with big data volume and variety. *IEEE computer graphics and applications*, 35(3), 10-14.

- Sicherman, N., Loewenstein, G., Seppi, D. J., & Utkus, S. P. (2016). Financial attention. *Review of Financial Studies*, 29(4), 863-897. doi:10.1093/rfs/hhv073
- Siddiqui, S. S., & Patil, V. A. (2017, 15-16 June 2017). *Proposed system for estimating intrinsic value of stock using Monte Carlo simulation*. Paper presented at the 2017 International Conference on Intelligent Computing and Control Systems (ICICCS).
- Sim, K. H., Sim, K. Y., & Hang-Hui, P. T. (2014, 3-5 Nov. 2014). *Forecasting price volatility cluster of commodity futures index by using standard deviation with dynamic data sampling based on significant interval mined from historical data*. Paper presented at the 2014 International Conference on Control, Decision and Information Technologies (CoDIT).
- Sindhu, T. N., Riaz, M., Aslam, M., & Ahmed, Z. (2016). Bayes estimation of Gumbel mixture models with industrial applications. *Transactions of the Institute of Measurement & Control*, 38(2), 201-214. doi:10.1177/0142331215578690
- Singh, P., & Thakral, A. (2017, 17-19 Aug. 2017). *Stock market: Statistical analysis of its indexes and its constituents*. Paper presented at the 2017 International Conference On Smart Technologies For Smart Nation (SmartTechCon).
- Sinharay, S. (2016). An NCME instructional module on data mining methods for classification and regression. *Educational Measurement: Issues & Practice*, 35(3), 38-54.
- Siriwardena, A. N., & Gillam, S. (2014). Evaluating improvement. *Quality in Primary Care*, 22(2), 63-70.
- Smith, K., Fulcher, K., & Sanchez, E. (2017). Ethical reasoning in action: Validity evidence for the ethical reasoning identification test (ERIT). *Journal of Business Ethics*, 144(2), 417-436. doi:10.1007/s10551-015-2841-8
- Soares, J., Canizes, B., Vale, Z., & Venayagamoorthy, G. K. (2016, 8-11 March 2016). *Benders' decomposition applied to Energy Resource Management in smart distribution networks*. Paper presented at the 2016 Clemson University Power Systems Conference (PSC).
- Son, J., Brennan, P. F., & Zhou, S. (2016). Rescue inhaler usage prediction in smart asthma management systems using joint mixed effects logistic regression model. *IIE Transactions*, 48(4), 333-346. doi:10.1080/0740817X.2015.1078014
- Sowmya, R., & Suneetha, K. R. (2017, 5-6 Jan. 2017). *Data mining with big data*. Paper presented at the 2017 11th International Conference on Intelligent Systems and Control (ISCO).
- Stark, L. (2011). *Behind closed doors: IRBs and the making of ethical research*: University of Chicago Press.

- Sun, Y., Bie, R., & Zhang, J. (2016). Measuring semantic-based structural similarity in multi-relational networks. *International Journal of Data Warehousing and Mining (IJDWM)*, 12(1), 20-33.
- Tabona, O., & Blyth, A. (2016). *A forensic cloud environment to address the big data challenge in digital forensics*. Paper presented at the SAI Computing Conference (SAI), 2016.
- Taft, S. T., Moore, B., Pinho, L. M., & Michell, S. (2014). Safe parallel programming in ada with language extensions. *ACM SIGAda Ada Letters*, 34(3), 87-96.
- Tan Boon, S. (2016). Accounting research for the management accounting profession. *Journal of Applied Management Accounting Research*, 14(1), 69-77.
- Tan, X., Sun, C., & Pham, T. D. (2016). Edge-aware filtering with local polynomial approximation and rectangle-based weighting. *IEEE Transactions on Cybernetics*, 46(12), 2693-2705. doi:10.1109/TCYB.2015.2485203
- Tang, L., Pan, H., & Yao, Y. (2018). *K-nearest neighbor regression with principal component analysis for financial time series prediction*. Paper presented at the Proceedings of the 2018 International Conference on Computing and Artificial Intelligence, Chengdu, China.
- Tavakol, M., & Sandars, J. (2014). Quantitative and qualitative methods in medical education research: AMEE Guide No 90: Part II. *Medical Teacher*, 36(10), 838-848. doi:10.3109/0142159X.2014.915297
- Tian, Y., Özcan, F., Tao, Z. O. U., Goncalves, R., & Pirahesh, H. (2016). Building a hybrid warehouse: Efficient joins between data stored in HDFS and enterprise warehouse. *ACM Transactions on Database Systems*, 41(4), 1-38. doi:10.1145/2972950
- Tigani, S., & Saadane, R. (2018). *Multivariate statistical model based currency market profitability binary classifier*. Paper presented at the Proceedings of the 2nd Mediterranean Conference on Pattern Recognition and Artificial Intelligence, Rabat, Morocco.
- Ting-Chia, H., & Gwo-Jen, H. (2017). Effects of a structured resource-based web issue-quest approach on students' learning performances in computer programming courses. *Journal of Educational Technology & Society*, 20(3), 82-94.
- Tozan, H., & Donmez, S. (2015). A genetic algorithm based approach to provide solutions for emergency aid stations location problem and a case study for Pendik/İstanbul. *Journal of Homeland Security & Emergency Management*, 12(4), 915-940. doi:10.1515/jhsem-2015-0025
- U.S. Securities and Exchange Commission. Using EDGAR: Researching public companies.

- U.S. Securities and Exchange Commission. (2007). Researching public companies through EDGAR: A guide for investors. Retrieved from <https://www.sec.gov/oiea/Article/edgarguide.html>
- U.S. Securities and Exchange Commission. (2013). Initial Public Offerings (IPO). Retrieved from <https://www.sec.gov/fast-answers/answersipohtm.html>
- U.S. Securities and Exchange Commission. (2013). Rule 144: Selling restricted and control securities. Retrieved from <https://www.sec.gov/reportspubs/investor-publications/investorpubsrule144htm.html>
- U.S. Securities and Exchange Commission. (2018). Full-text search. Retrieved from https://searchwww.sec.gov/EDGARFSCClient/jsp/EDGAR_MainAccess.jsp
- Vaitheeswaran, K., Subbiah, M., Ramakrishnan, R., & Kannan, T. (2016). A comparison of ordinal logistic regression models using Classical and Bayesian approaches in an analysis of factors associated with diabetic retinopathy. *Journal of Applied Statistics*, 43(12), 2254-2260. doi:10.1080/02664763.2016.1140725
- Vateekul, P., Kubat, M., & Sarinnapakorn, K. (2014). Hierarchical multi-label classification with SVMs: A case study in gene function prediction. *Intelligent Data Analysis*, 18(4), 717-738. doi:10.3233/IDA-140665
- Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User acceptance of information technology: Toward a unified view. *MIS Quarterly*, 425-478.
- Venkatraman, S., Fahd, K., Kaspi, S., & Venkatraman, R. (2016). SQL versus NoSQL movement with big data analytics.
- Vitt, C. A., & Xiong, H. (2015, 14-17 Nov. 2015). *The impact of patent activities on stock dynamics in the high-tech sector*. Paper presented at the 2015 IEEE International Conference on Data Mining.
- Waidi, A. A. (2016). Employment of questionnaire as tool for effective business research outcome: Problems and challenges. *Global Economic Observer*, 4(1), 136.
- Wamba, S. F., Akter, S., Edwards, A., Chopin, G., & Gnanzou, D. (2015). How 'big data' can make big impact: Findings from a systematic review and a longitudinal case study. *International Journal of Production Economics*, 165, 234-246.
- Wamba, S. F., Gunasekaran, A., Akter, S., Ren, S. J.-f., Dubey, R., & Childe, S. J. (2017). Big data analytics and firm performance: Effects of dynamic capabilities. *Journal of Business Research*, 70, 356-365.
- Wang, S., Zhou, L., Li, Y., & Wu, J. (2016, 23-26 Aug. 2016). *Parallel Monte Carlo method with MapReduce for Option pricing*. Paper presented at the 2016 IEEE Trustcom/BigDataSE/ISPA.

- Wang, Y., Li, R., Zhou, Y., Ling, Z., Guo, X., Xie, L., & Liu, L. (2016). Motif-based text mining of microbial metagenome redundancy profiling data for disease classification. *BioMed Research International*, 2016, 1-11. doi:10.1155/2016/6598307
- Wei, H., Tang, X.-s., & Liu, H. (2015). A genetic algorithm(GA)-based method for the combinatorial optimization in contour formation. *Applied Intelligence*, 43(1), 112-131. doi:http://dx.doi.org/10.1007/s10489-014-0633-y
- Weinberg, S. (2016). Selling the fourth revolution: The correlation between C-Suite architecture and a big data mindset as portrayed in the letter to the shareholders.
- Wickham, H., Francois, R., Henry, L., & Müller, K. (2017). dplyr: A grammar of data manipulation. Retrieved from <https://CRAN.R-project.org/package=dplyr>
- Wickham, H., Hester, J., & Francois, R. (2017). readr: Read rectangular text data. Retrieved from <https://CRAN.R-project.org/package=readr>
- Wozniak, J. M., Chard, K., Blaiszik, B., Osborn, R., Wilde, M., & Foster, I. (2015). *Big data remote access interfaces for light source science*. Paper presented at the 2015 IEEE/ACM 2nd International Symposium on Big Data Computing (BDC).
- Wu, C.-h., Li, J.-b., & Chang, T.-y. (2015). SLinRA^{sup} 2[^]S: actively supporting regression analysis with R. *Information Systems and eBusiness Management*, 13(2), 309-328. doi:http://dx.doi.org/10.1007/s10257-014-0262-3
- Wu, D. D., Zheng, L., & Olson, D. L. (2014). A decision support approach for online stock forum sentiment analysis. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 44(8), 1077-1087. doi:10.1109/TSMC.2013.2295353
- Xu, J., Shi, M., Chen, C., Zhang, Z., Fu, J., & Liu, C. H. (2016). *ZQL: A unified middleware bridging both relational and NoSQL databases*. Paper presented at the Dependable, Autonomic and Secure Computing, 14th Intl Conf on Pervasive Intelligence and Computing, 2nd Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress (DASC/PiCom/DataCom/CyberSciTech), 2016 IEEE 14th Intl C.
- Xu, X., Weber, I., Staples, M., Zhu, L., Bosch, J., Bass, L., . . . Rimba, P. (2017, 3-7 April 2017). *A taxonomy of blockchain-based systems for architecture design*. Paper presented at the 2017 IEEE International Conference on Software Architecture (ICSA).
- Yahoo. (2018). Yahoo! finance. Retrieved from <https://finance.yahoo.com/lookup>
- Yang, C., Huang, Q., Li, Z., Liu, K., & Hu, F. (2017). Big Data and cloud computing: Innovation opportunities and challenges. *International Journal of Digital Earth*, 10(1), 13-53.
- Yang, Y., Yang, Y., & Li, J. (2017, 15-17 Dec. 2017). *Role of mean in the multifractal analysis of financial time series*. Paper presented at the 2017 14th International Computer

Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP).

- Yasaweerasinghelage, R., Staples, M., & Weber, I. (2017, 3-7 April 2017). *Predicting latency of blockchain-based systems using architectural modelling and simulation*. Paper presented at the 2017 IEEE International Conference on Software Architecture (ICSA).
- Yassien, A. W., & Desouky, A. F. (2016). *RDBMS, NoSQL, Hadoop: A performance-based empirical analysis*. Paper presented at the Proceedings of the 2nd Africa and Middle East Conference on Software Engineering, Cairo, Egypt.
- Yelin, L., Hui, B., & Junjie, W. (2017, 16-18 June 2017). *A two-stage multi-view prediction method for investment strategy*. Paper presented at the 2017 International Conference on Service Systems and Service Management.
- Yongzhong, L., Danping, Y., Min, Z., & Levy, D. (2017). Maximum likelihood parameter estimation of dynamic systems by heuristic swarm search. *Intelligent Data Analysis*, 21(1), 97-116. doi:10.3233/IDA-150340
- Yu, H. X. (2017). Capital asset pricing model – investigation and testing. *Journal of Applied Finance and Banking*, 7(6), 85-97.
- Zaman, M., & Lung, C. H. (2018, 23-27 April 2018). *Evaluation of machine learning techniques for network intrusion detection*. Paper presented at the NOMS 2018 - 2018 IEEE/IFIP Network Operations and Management Symposium.
- Zhang, C., Wang, Y., Chen, C., Du, C., Yin, H., & Wang, H. (2018). *StockAssIstant: A stock AI assistant for reliability modeling of stock comments*. Paper presented at the Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, London, United Kingdom.
- Zhang, H., & Vorobeychik, Y. (2016). Empirically grounded agent-based models of innovation diffusion: A critical review. *arXiv preprint arXiv:1608.08517*.
- Zhang, Z., & Liu, Q. (2016). *Covering number: Analyses for approximate continuous-state pomdp planning (Extended Abstract)*. Paper presented at the Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems, Singapore, Singapore.
- Zhou, X. Y., Tian, X. W., & Lim, J. S. (2015). Fuzzy Naive Bayesian for constructing regulated network with weights. *Bio-Medical Materials And Engineering*, 26 Suppl 1, S1757-S1762. doi:10.3233/BME-151476
- Zhuang, K., North, G. R., & Stevens, M. J. (2017). A NetCDF version of the two-dimensional energy balance model based on the full multigrid algorithm. *SoftwareX*, 6, 198-202.