Big Data and its implications for the statistics profession and statistics education

Busayasachee Puang-Ngern

A Thesis submitted in fulfilment of the requirements for the degree of

Master of Research

Department of Statistics, Faculty of Science

Macquarie University

Sydney, NSW, Australia

October 2014

Contents

Abstract

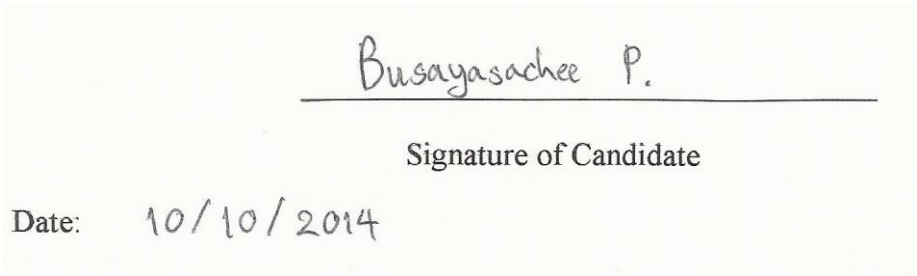
The convergence of computer and communications technology, the processing speed of the computers, and the widespread use of the internet around the world have led us to the age of Big Data. Massive volumes of data have been and are being collected and this far outstrips the capacity to analyse the data and convert it into useful and usable information. The need for people who can analyse the data and obtain useful and usable information from it has led to a new career path: the "data scientist". The data scientist is a kind of hybrid of a statistician and a computer scientist / IT professional. Big Data and its analysis is an area in which industry is leading and academia seems to be playing catch up. Universities are responding to the changing needs of industry and government by introducing new degree programs and units of study. In this thesis, we investigate the perceptions of graduates working in the analysis of Big Data and the perceptions of academics about the types of expertise and the types of software skills required for working in this new field via online surveys. This facilitates comparison between the perceptions of statisticians and of computer scientists about what expertise and software skills are required, and provides information useful for the purpose of developing the curriculum for new degrees in data science and statistics which is urgently needed.

The survey results showed that Statistical Analysis and Statistical Software Skills were the most necessary type of expertise required for working in the Big Data field. SQL and R programming were the most necessary software tools for Big Data analysis as rated by both graduates and academics respectively.

Statement by Candidate

I hereby certify that this thesis has not been submitted for a higher degree to any other university or institution.

The ethical aspects of this study have been approved by the Macquarie University Human Research Ethics Committee. The protocal number is 5201400282.



Acknowledgements

I wish to thank all of the academics and graduates who took part in this study by completing the survey and everyone who helped to distribute the surveys. Their opinions are valuable.

Warmest thanks to my perfect Supervisors, Dr Ayse Bilgin and Dr Timothy Kyng, for all of their kindness and support. I appreciate to all of their tireless and knowledgeable advice. This thesis would not be completed without their attention.

I would like to thank Chulalongkorn University as my scholarship provider to permit me to study here. Thanks to lecturers and officers at Faculty of Commerce and Accountancy, Chulalongkorn University, Thailand, as well as officers at Office of Educational Affairs, Royal Thai Embassy, Australia for their support.

Special thanks to Dr Sirikul Klongkumnuankarn, Dr Adele Thomas, and Miss Sompatu for their generous suggestion on thesis writing.

Finally, I am very grateful to my lovely family; my beloved dad, sister, and especially my mom, Ms Patumarn Puang-Ngern; and friends for all of their patience and support.

1. Introduction

In the past, gathering data to analyse was a major issue. For example, collecting data about people and their spending patterns was difficult to do and expensive to obtain. However, nowadays, in the world of information technology, data generation is pervasive. Data about people and their spending decisions are recorded from different sources such as customer contact centers, customer loyalty cards, websites and social media such as google, facebook, linkedin. There are billions to trillions of records about millions of people. The Computer Age, or Information Age in respect to technology are transforming society to the Big Data Era due to the massive rapid growth in data collection and storage (Boyd & Crawford, 2011; Brown, Chui, & Manyika, 2011; McAfee & Brynjolfsson, 2012; Zikopoulos & Eaton, 2011).

The volume of data being collected is huge, however the capacity to analyse the data is limited. The number of people with the right skills in statistical analysis and computing who could work in this area is low relative to the demand for them.

1.1 What is Big Data?

The term “Big Data” is believed to have originated with web search companies who had to query very large distributed aggregations of loosely-structured data (Datameer, 2014). There is no single perfect definition of Big Data. In general, Big Data refers to the data sets that contain a huge volume of diverse and complex data. They are difficult to process and to be analysed using traditional statistical techniques and ordinary database tools. Big Data refers not only to the volume of data, but also to the technology (including tools and processes) required to access, store and analyse the data. New technology and methods have been created to deal with these issues. Manyika et al. (2011) proposed techniques to handle Big Data such as Association rule learning, Classification, Cluster analysis, Crowdsourcing, Data fusion and data integration, Data mining, Ensemble learning, Genetic algorithms, Machine learning, Natural language processing (NLP), Neural networks, Network analysis, Optimization, and Visualization. They also suggested analysis and management of Big Data technologies e.g. Business intelligence, Cloud computing, HBase, Hadoop, MapReduce, R programming and SQL. Furthermore, Hurwitz et al. (2013) proposed that Big Data is the capability to manage and analyse these data at the appropriate velocity within a suitable time frame that allows for reasonable reaction to events on a real-time basis. An alternative definition is that Big Data can be anything that will not fit into an Excel spreadsheet (Batty, 2014). Batty claimed that the data may have been collected traditionally by manual labour from many

sources instead of having been streamed in real time such as the information in the Census of the Australian population.

The "3 V's" proposed by Laney (2001) have become ubiquitous in defining Big Data. The three characteristics of Big Data are data *volume*, *velocity*, and *variety*. Volume defines vast size of data records, or transactions in Terabytes, Petabytes, Zettabytes. Velocity is how fast that data is processed e.g. in batch, real time or streaming. Variety specifies the various types of data. Data can be from diverse types of sources e.g. social media, social networks, e-mail, and text streams. The databases can be structured (DB data, TSV data, and CSV data), unstructured (streaming live data, sensor data, spatial data, blog data, web logs, videos, and audios), or semi-structured (XML data, JSON data). Data that can be stored in fields within a relational database is classified as structured data, while the data that does not fit into a pre-defined data model is regarded as unstructured data (Hurwitz et al., 2013; Lunet-Levi, 2013). Big Data is mainly unstructured data which has been estimated to be 90 percent of organizations' data (Professional Advantage, 2014).

The 3 V's have been extended: the 4 V's, the 5 V's, and even up to the 12 V's of Big Data have been suggested. For the 4 V's, the additional V is *veracity* (IBM Big Data & Analytics, 2013). It means the various levels of data uncertainty and reliability. It indicates the level of data accuracy in predicting business value. However, the most significant aspect of Big Data is *value*, the fifth of the V's of Big Data (The Data Alchemists, 2013). It is pointless to put money and effort into Big Data analytics if it cannot generate good value for business. By using Big Data and analyzing the information, companies can gain actionable insight.

Data science is another term for Big Data. Apart from data analysis, data science also includes inference and confidence in the inferences (Dhar, 2013).

However, Lambert (2003) presented Data Science from a data mining perspective as being processing and tabulating data to find associations, rules, and interesting patterns from massive sets of data. It is generally not about making predictions, or modeling or statistical inference. From a database perspective, data are not uncertain or random. Everything that has happened was taken into account in the data (Lambert, 2003).

Knights (2013) indicated that Adrian Simpson, chief innovation officer at technology firm SAP, addressed the differences between Big Data and statistical analysis. First of all, Big Data analysis is used to find trends and patterns, and to predict outcomes based on those trends and patterns. Moreover, Big Data involves a blend of various disciplines and existing technologies, whereas statistics relies more on definitions and exactness. In addition, Big Data analysis makes use of data volumes across a diversity of sources to generate prescriptive or predictive analyses, while the application of traditional statistical approaches is used for descriptive or diagnostic analysis.

1.2 Application(s) of Big Data

The application(s) of Big Data have been widely used in business. The flood of data now available and stored provides opportunities to gain unprecedented insight to transform business, government, and individual life. Marr (2013) presented the utilization of Big Data in ten ways. Firstly, one of the largest and most widespread uses of Big Data is to better understand and target customers. Businesses can collect information about customer behaviour and preferences to generate predictive models to target customers. Secondly, it is widely used to understand and optimize business processes. For example, by using predictive models created from social media data, web search trends and weather forecasts, retailers are able to optimize their inventory. Big Data analytics is also applied in delivery route optimization (supply chain management). Cargo can be tracked by geographic positioning and radio frequency identification sensors. Integrating live traffic data can be used to optimize delivery routes. Thirdly, Big Data is used for personal quantification and performance optimization. An example of this is the armband that keeps data on personal activity levels, calorie consumption, and sleep patterns. Individual users can gain advantages from finding new insights by analyzing the large amount of collective data. Another aspect of the potential benefits of Big Data is in improvements in Healthcare and Public Health. Big Data can be used to better understand and predict disease patterns, and find new treatments. For instance, the algorithms developed by recording and analyzing every heart beat and breathing pattern(s) of every baby monitored in a specialist premature and sick baby unit can be used to predict infections 24 hours before any physical symptoms appear. Sports performance improvement is another example of application of Big Data. Big Data analytics have been adopted in elite sports such as the performance tracking of football or baseball players using video analytics, and sensor technology in sports equipment such as basket balls or golf clubs. In addition, Big Data study leads to science and research improvements. Moreover, Big Data analysis is applied to optimize machine and device performance to become smarter and more autonomous such as the driverless car. Furthermore, it has been used in security and law enforcement where it improves detection and prevention of cyber-attacks, or detection of fraudulent transactions by credit card companies. Another application is in improving and optimizing cities and countries. Smart Cities are applying Big Data analytics in joining up the transport infrastructure and utility processes. Financial trading is also being affected by Big Data. Big Data algorithms are being applied to make trading decisions.

1.3 Literature Review

1.3.1 Big Data in Industry

Many industry studies have reported significant growth in data storage, as well as in jobs related to data analytics, Big Data analysis. Ninety percent of global data has been generated over the last two years (Bradshaw, 2013; SINTEF, 2013). Moreover, the world's data is predicted to experience forty percent growth per year (Manyika et al., 2011). These data are produced from many sources in our interconnected world today, such as mobile phone interactions, social media sites, online searches, and sensor technologies.

An expansion of Big Data leads to increasing job opportunities. Gartner (2012) forecasts that the growth in Big Data will produce 4.4 million IT jobs globally. Bradshaw (2013) also says for each new IT job created there will be three new non-IT positions. Deep analytical skills are needed to analyse and extract useful and useable information from this massive volume of data. These developments have created a new career role, the “data scientist”. The McKinsey Global Institute (Manyika et al., 2011) reported that by 2018, for the United States alone, there will be a rising shortage in Big Data experts. They are predicting a 50 – 60 percent gap between demand for and supply of the required personnel. The McKinsey Global Institute also forecasted a shortage of 1.5 million data-savvy managers with proficiency in effective decision making using Big Data analysis.

1.3.1.1 Skills demand

The e-skills UK (2013) report on Big Data Analytics: Adoption and Employment Trends, 2012–2017 states that the technical software skills most often demanded by Big Data recruiters in the third quarter of 2012 was NoSQL, followed by Oracle, Java, SQL, Linux, Hadoop, MySQL, JavaScript, UNIX, and Python. In addition, the common functional knowledge or skills demanded by Big Data recruiters in that sector were Business Intelligence, Big Data, Data Warehouse, Web Services, Analytics, E-commerce, Internet, Social Media, Cloud Computing, CRM, and SaaS, respectively.

None of these technologies can handle all 3 of the V’s of Big Data at once. There is no one-size-fits-all solution (Molinari, 2012). Multiple Big Data technologies are applied to solve data analytics problems. The solutions involved in addressing Big Data and data analytics problems are still being investigated and developed.

1.3.1.2 Big Data and IT Businesses

IT Businesses have been vigilant in responding to the growth in required storage of data. However, overall, Universities in Australia have been slow in responding to these new developments and in providing education to their students to enable them to be work ready in Big Data related jobs (i.e. as a data scientist). Many corporations have been organizing their own training courses for their employees in this area. Some corporations such as Intel, IBM, Oracle, Quantium, Cloudera, Hortonworks, Datameer and Gartner also provide a Big Data knowledge base on their websites. These materials are provided not only as a service for the community but also to provide information to potential clients about their products and services in relation to Big Data. Please see Appendix B for the list of business community web base providing knowledge on Big Data.

1.3.2 Big Data in University Education

Over the past two decades, the significance of Big Data analytics is expanding not only in business, but also in academia (Chen, Chiang, & Storey, 2012). However, the business community might not find new graduates well equipped with skill sets for working in the Big Data arena as this is a relatively new multidisciplinary area and academic institutions are only recently responding to the demand for graduates with skills in data science. The shortage of and demand for data scientists and other graduates with the required skills has induced some universities to modify their existing degree programs and create new data science courses (Degree Prospects, LLC., 2013) as well as partnering with industry to initiate courses that focus on the specific skills (Glance, 2013).

IBM and SAS Institute (Nerney, 2013) are examples of university – industry partnering trend. They partner with many universities around the world in creating the Big Data analytics degree programs for undergraduate and graduate students in various disciplines including mathematics, business, marketing, and health services. Both assist educational institutions to design curricula which emphasize both technical and problem solving skills. Furthermore, they provide support such as analytics software, case study projects, and their data scientists as guest lecturers to academic partners. Moreover, they also provide online tutorials and certifications for the data analytics workforce. Examples of such academic partnerships are SAS Institute with the Data Science course at the University of South Australia, Australia and the Master of Applied Data Analytics at Bournemouth University, United Kingdom. IBM Institute (2013) has collaborations about Big Data analytics with over 1,000 universities such as Master of Science degree in Business Analytics at the

George Washington University, United States and Master of Science in Business Analytics (MSBA) degree at the National University of Singapore (NUS), Singapore.

1.3.2.1 Big Data in Australian Universities

Most Australian Universities offer degree programs in statistics, computer science, mathematics, information technology, information systems, and engineering. Although graduates with these university qualifications have been the essential source of entry level employees in the Big Data area in Australia, Glance (2013) pointed out that a combination of computational, statistical and mathematical skills are required in Big Data analysis and data visualization. So expertise in one discipline is not enough. Multidisciplinary degree(s) are required to prepare the students for the graduate entry level positions in Big Data analysis. Recent graduate employees often have to learn technical skills by self-study in online courses, learn new skills on the job, or enrol to study in another degree program (such as Master of Applied Statistics graduate studying Master of Information Technology).

Data scientists would benefit from obtaining skills in computer science, software, statistical theory and analysis and mathematics, since these disciplines' approaches to data analysis are different. For example, the disciplines of Computer Science and Statistics approach data analytics in quite different ways. For instance, the computer science approach involves creating software for pattern recognition and development of models for prediction. This includes machine learning, statistical learning, and genetic algorithms. It tends to produce complex models, sometimes lacking a sensible or intuitive or theoretical interpretation. On the other hand, statistics aims to understand data and the relationships between variables and outcome(s). Statisticians often work with other researchers who have collected the data, have hypotheses and theories they want to test and have knowledge about the area / subject. So it is perhaps more inclined to obtain models which have an intelligible interpretation or which fit into some theoretical conceptual framework.

In Australia, there are some universities currently offering and planning to launch the Data Science degree programs. For the list of Big Data related degrees in Australia, please see Appendix C. The University of South Australia is the first Australian University to offer the Master of Data Science degree (including Graduate Diploma of Data Science and Graduate Certificate in Data Science). Their degree programs have evolved and been designed with the collaboration of industry. The SAS Institute has assisted the University of South Australia to develop the Data Science program (University of South Australia, 2014); therefore industry certification by SAS will also be awarded to graduates of the program.

Other Australian universities with related degree programs include Macquarie University that will launch a new degree named "*Master of Data Science*" (also related graduate diploma and graduate certificate) in 2016 (Macquarie University, 2014) and the University of Technology Sydney that will offer the Masters of Data Science and Innovation (including Graduate Diploma of Data Science and Innovation, and Graduate Certificate in Data Science and Innovation) in 2015 (University of Technology Sydney, 2014).

Some universities have developed the new specialisation from the existing degree program such as the Master of Information and Communications Technology specialisation in Distributed Computing proposed by the University of Western Sydney (University of Western Sydney, n.d.).

Interesting that the new degree programs are not strictly to only statistics and computing disciplines but they are also spread to business and marketing disciplines. The example is Master of Business Analytics at Deakin University with the collaboration of IBM Institute, Microsoft, SAS, Accenture, Altis Consulting, Deloitte, Ernst & Young, PBT Group and PwC (Bender, 2013; Deakin University, 2014b). This course is offered both on campus and online study. Moreover, Deakin University also has planned to launch the new Bachelor of Computer Science with major sequence in Data Science in 2015 (Deakin University, 2014a).

The other degree programs in Big Data analysis are the Bachelor of Computer Science with new major in Big Data (RMIT University, 2014a), Master of Information Technology and Master of Computer Science with specialisation in Big Data Management (RMIT University, 2014b) at RMIT University.

1.3.2.2 Big Data in Overseas Universities

In New Zealand, there are two universities offering the data science course (see Appendix C). The University of Otago has offered the Masters of Business Data Science (University of Otago, 2013). This degree is providing knowledge from multidisciplines of Statistics, Big Data management, and Business knowledge. Another university providing data science course is the University of Auckland. The Master of Professional Studies in Data Science (University of Auckland, 2013) is a collaboration of departments of computer science, statistics, and Information Systems and Operations Management.

In United States, many universities have enthusiastically embraced the concept of Big Data. They have launched a wide range of university degree programs covering data science with a variety of degree program names in various disciplines. For example, the University of Washington offers a new PhD in Big Data with collaboration between the Computer Science & Engineering and Statistics departments (Billhowe, 2013). A Master of Science in Information Science (MSIS) degree

program that provides a specialisation in Big Data Analytics is offered by the University Of Pittsburgh (University Of Pittsburgh, 2013). Master of Business Analytics (Bentley University, n.d.a) and Master of Science in Marketing Analytics (Bentley University, n.d.b) are the data analytics course in business discipline offered by Bentley University (for the list of Big Data related degrees in US, please see Appendix C).

In United Kingdom, there are many universities offering Big Data Analytics courses. Examples include the Master of Applied Data Analytics at Bournemouth University involving collaboration with the SAS Institute (Bournemouth University, 2014), Master of Big Data Analytics at Sheffield Hallam University (Sheffield Hallam University, 2014), and the Master in Data Science at City University London (City University London, 2014).

1.3.3 Big Data in MOOCs

There are many massive open online courses (MOOCs), both paid and free online courses, offering courses about Big Data. These online courses are another option for people to learn Big Data techniques. Udacity, a MOOC provider, offers a variety of courses starting from the beginner level (Introduction to Computer Science, and Introduction to Statistics) to intermediate level (Introduction to Data Science, Introduction to Hadoop and MapReduce, Data Wrangling with MongoDB, Exploratory Data Analysis, and Machine Learning). They also collaborate with Cloudera, a Big Data database provider, to provide some courses on Big Data (Glance, 2013). Another MOOC provider, Coursera, has a data science program (nine courses) with instructions provided by Johns Hopkins University's professors in Biostatistics. There are also other specific courses in data science, for instance, the Big Data University which offers courses mainly focused on software programs such as Hadoop, Hive, IBM Data Studio, Java, R, JaQL and SQL.

Recent literature on Big Data and data analytics covers issues such as the growth of Big Data (Boyd & Crawford, 2011; Bradshaw, 2013; Brown, Chui, & Manyika, 2011; Manyika et al., 2011; McAfee & Brynjolfsson, 2012; SINTEF, 2013; Zikopoulos & Eaton, 2011), the importance and usage of data analytics (Gilpin, 2014; Jeske, Grüner, & Weiß, 2013; Marr, 2013; Mayer-Schönberger & Cukier, 2013; Shaw, 2014), attempts to improve the efficiency of Big Data techniques (Manyika et al., 2011; Kepner et al., 2014) and collaboration between business and

academia in designing education and training courses in the area (Glance, 2013; IBM Institute, 2013; Nerney, 2013).

This thesis examines the existence of a mismatch between the needs of industry and government for appropriately trained people and the supply of such people, particularly in Australia. This poses a challenge for universities in how to respond to this and design new degree and diploma programs. An important issue not yet adequately researched is how to design such programs and the role and relative importance of mathematical and statistical education versus computer science education in these degrees.

Data Science is a very new area and new degree program(s) which has been established very recently. Such a new area, there is no existing literature about the gap between industry and academia for Australia. We would expect that there would be feasibility studies conducted by universities about establishing new degrees in this area, but these were not in the public domain at the time of writing this thesis.

The gap between the needs of industry and government for graduates equipped with the right knowledge and skills and the supply of graduates from universities in Australia leads to the following research questions:

a) What are the differences between what Australian universities supply and what industry and government needs in terms of graduates trained in the relevant skills?
b) What specific technical skills and software tools are required for Big Data professionals?
c) What are the opinions of academics in relevant disciplines (such as statistics and computing) about Big Data and the extent to which students need to be educated in disciplines other than their own? For example statistics students may need education in computing and marketing.

We have investigated these questions through both academic and graduate surveys as well as analysis of job descriptions in recent job advertisements for data scientist(s). This is exploratory research and a descriptive study. This study will contribute to the understanding of what is needed in design of new degree programs in Big Data, Data Analytics, Statistics, Computing and Mathematics in response to current workplace demands.

2. Methodology and Data Collection

Our main aim is identification of the gap between what industry demands and what universities are supplying, with respect to graduates equipped with Big Data knowledge, skills, and techniques.

We surveyed two groups of participants. The first group was the academics in targeted departments. The second group was the graduates in the data analytics work force. In addition, we analysed skill requirements for data scientists from recent job advertisements. To answer our research questions, online surveys were undertaken in Australia and New Zealand. This approach was appropriate because we could reach all of the targeted academics in Australia and New Zealand via online surveys. Moreover, snowball sampling of respondents for the graduate survey was the best way to reach the graduates we wanted to target. There is no easy way to find the graduates, unlike the academics where can easily identify them via information available on the website. Snowball sampling allowed us to access a wider range of graduates. Using paper based surveys would not have been practical.

2.1 Surveys

The surveys were conducted through online questionnaires via the Qualtrics Surveys website (https://mqedu.qualtrics.com/ControlPanel/). The questionnaires for the academics and graduates were individually designed for these two participant groups. However, the topics addressed in these two questionnaires were quite similar to facilitate comparison between academics' and graduates' (who are working in industry) perspectives.

The pilot surveys for both of these questionnaires were tested by three professors from the disciplines of Statistics, Actuarial Science, and Computer Science. The results from these pilot surveys led to rephrasing some of the questions to clarify them, as well as to add an additional software tool (Python) to questions 17 & 18 of the academic survey, and questions 20 & 21 of the graduate survey.

2.1.1 The academic questionnaire

The academic questionnaire (see Appendix D.1) was conducted online and covers the following topics:

- General information (questions 1 – 4): These questions inquired about gender, age, nationality, and language background. These answers will provide information about participants' backgrounds.
- Professional information (questions 5 - 7, 9 – 10, and 13 - 14): We collected data related to the participants' workplace i.e. their University, faculty, and department, as well as their working experience in the Big Data area. The purpose of these questions was to enable us to classify the survey data by universities or disciplines, and analyse or identify the differences between them. Moreover, these data will provide information about their expertise and experiences relevant to Big Data analytics. In addition, we can identify the importance given to Big Data from different disciplines within academia.
- Degree program information (questions 8, 11 – 12, and 19 - 20): We investigated the degree programs and subjects offered in participants' departments that are relevant to Big Data / Data Analytics/ Data Science, such as Machine Learning, Statistical Learning, Data Mining, and Statistical Analysis. Furthermore, we also asked the academics about the number of students graduating from their departments and the number of students enrolled in those relevant subjects each year. The intention was to examine the existing University degree programs that relate to Big Data / Data Analytics which educates graduates who may be suited to work in this field. In addition, their opinions about what improvement(s) may be needed in their degree programs and subjects taught relating to Big Data / Data Analytics were also investigated.
- Area of expertise (questions 15 – 16): We asked the academics' opinions about the expertise requirement(s) for graduate students to gain employment in the Big Data / Data Analytics field, and whether or not their departments impart these required skills to their students. This will enable us to explore the academics' viewpoint(s) regarding what graduates need to know for a career in this area and their University's response to these needs.
- Software tools (questions 17 – 18): We obtained the academics' opinions about the Big Data analysis' software tools that graduates should be able to use, as well as whether or not it has been used in teaching in their degree programs. This will help us to identify what software skills the academics think graduate students should have to be employable in the Big Data field. We are also able to examine what software skills and tools are currently being taught in Universities.

2.1.2 The graduate questionnaire

The graduate questionnaire (see Appendix D.2) has been designed for graduates who work in the Data Analytics workforce. The questions were mostly modified versions of those from the academics' survey because we wanted to identify the similarities and differences between *what academics think (and do)* with *what graduates think and do* in respect to the use of software tools and required skills. The graduate questionnaire addresses the following topics:

- General information (questions 1 – 4): These were the same questions as in academic questionnaire that enquired about gender, age, nationality, and language to enable to understand participants' background.
- Educational information (questions 5 – 8): These questions included the participants' highest education and the discipline along with when they completed their highest degree.
- Workplace information (questions 9 and 12 – 17): The participants' workplace i.e. the name and the operational area of their companies, as well as their working experience in the Big Data area will enable us to identify similarities and differences between different employers. We also asked questions about whether the Big Data is seen important and whether they use Big Data (and/or Big Data analysis) in their organization.
- Graduates employed in Big Data / Data Analytics roles: (questions 10 – 11 and 22): We investigated the number of new graduates hired each year and disciplines studied by these new graduates. We also asked their opinions on which university degree programs should include data analysis in their degree programs.
- Area of expertise (questions 18 - 19): These were modified versions of similar questions in the academic survey. The questions asked the respondents to give a rating to the importance of various generic skills / types of expertise and whether the respondents had the skill / expertise themselves. The answers to these questions will be compared with the answers from the academics, so that the differences can be identified.
- Software tools (questions 20 - 21): These questions were also modified versions of similar questions in the academics survey to reflect the perspective of graduates working in industry. The questions asked respondents to give a rating to the importance of particular software tools and whether these tools are used in their own organization. By comparing the answers from graduates to the answers from academics, we can identify the software skills that might be added into the university curriculum.

2.2 Participants

We had two participant groups. The first was the academics and the second was the graduate employees in the data analytics workforce. We sent an invitation email (see Appendix E.1 and E.2) to possible participants. The first page of the survey was the participant information and consent form (see Appendix F) which asked for their consent to participate in this online survey. We also used "snowball sampling" where we asked the invited participants to pass the survey link to their colleagues who may be willing to complete the survey(s). The rationale for using snowball sampling was to get a larger sample and to reach respondents we would otherwise have no way to know about or of contacting. These are some of the advantages of using snowball sampling. However, the disadvantage of using snowball sampling is the people selected that way might be a biased sample.

2.2.1 Academics

We collected the academic email addresses from 39 Australian and 8 New Zealand universities' websites in April - May 2014. The target group was the academics in Statistics, Computer Science, Actuarial Science, Information Systems, Information Technology, Mathematics, and Marketing disciplines. First, we searched for the study areas in our target disciplines in each University website. Then, we sought for the email addresses of academic staff members in those areas. In total, there were 127 potential university departments in Australian Universities and 31 for New Zealand Universities. Moreover, there were 5 more university departments from snowball sampling in the United Kingdom, Austria, Canada, and the United States. Please see Appendix G.1, G.2, and G.3 for the list of potential university departments in Australian Universities, in New Zealand Universities, and in other country Universities, respectively.

From 163 university departments sampled, there were 62 university departments that responded. Therefore the proportion of the university departments surveyed was 38%. We assume that an academic from a particular department knows what their departments do so they are representative of their departments.

2.2.2 Graduate employees

The email list of prospective participants for the graduate survey was gathered from three universities: Macquarie University (Australia), University of Western Sydney (Australia), and

Chulalongkorn University (Thailand). The graduates in our list included people who graduated between 2003 and 2014 with one of the following qualifications: Bachelor degree, Postgraduate Certificate, Postgraduate Diploma, and/or Master degree. Current Master's degree students were also included. The target disciplines were: Statistics, Computer Science, Actuarial Science, Information Technology, and Mathematics. These prospective respondents are the types of graduates who get hired to work in the data analytics workplace. In addition, we also sent the invitation email to workers in industry partners using personal connections, as well as via the snowball sampling method. We sent out 497 invitations to the graduates and 72 of them responded to our survey (14.49%). The prospective graduate participants' universities and degree programs are provided in Appendix H.

2.3 Job advertisements

In addition to conducting the two online surveys, we also analysed the knowledge and skills mentioned in recent "data scientist" job advertisements posted in StatsJobs which is the online specialist recruitment operation (http://www.statsjobs.com/) in April – September 2014. We will be presenting the requirements identified from these advertisements by classifying the statistical techniques, supplementary skills, programming skills, and non-technical skills, in the results section of this thesis. The samples of Data scientist's job advertisements are provided in Appendix I.

3. Results

From 163 university departments in 52 target universities, there were 87 academic participants from 62 university departments in 37 universities (3 respondents did not identify their university departments). Their ages were between 25 and 79 with an average age of 48 years. 69 percent of them were males. There was one respondent who declined to provide his demographic background. The academics and the university departments we wanted to reach was a small population and we invited all of them to participate. We obtained a reasonably good response rate in terms of the university departments.

There were 72 graduate participants who graduated from 33 different Universities, and currently work in 53 different companies (10 respondents did not identify their work place). The participants of the graduate survey were younger in average with the average age at 33 years and a range between 21 and 62 years. Some of them completed their university degree a long time ago before data science existed. So these people would have learned their big data skills on the job rather than from a university degree program. The population of graduates working in big data is difficult to access and there is no simple way to identify them, unlike the academics. Accordingly we used snowball sampling for this population. Random sampling would not have been feasible to do. This is a descriptive study giving a snapshot of the situation in Australia at the time of the study.

There were more male participants (60%) than female participants (40%). This pattern is aligned with the overall Australian workforce (Australian Bureau of Statistics, 2014a; Australian Bureau of Statistics, 2014b).

3.1 The gaps between what Australian universities offer and what industry and government need in terms of graduates trained in the relevant skills of Big Data

3.1.1 Technical skills

In questions 15 and 16 of the academic survey, we asked respondents their opinion about the type of expertise required for graduate students to be employed in the Big Data field and whether or not students acquire expertise in this area during their studies. We compared these academics' answers with the responses to questions 18 and 19 of the graduate survey that also asked graduates

their opinions about the type of expertise required to work with Big Data and whether they have this expertise or not. In analyzing the responses to these questions, we coded all missing values as a Not Applicable (N/A) answer.

The results for the importance ratings and the extent of expertise are shown in Appendix J.1 and J.2. Both academics (86%) and graduates (82%) consistently rated Statistical Analysis and Statistical Software Skills as the most necessary type of expertise required for working in the Big Data field (Agree and Strongly Agree combined). Almost half (47%) of the academic respondents stated that graduates had already been trained in these skills at University and 68% of graduates agreed that they had these skills.

According to Figure 1 below, the ratings of graduates and of academics were close to each other suggesting consistency between academics and graduates views on required expertise. Statistical Learning had been rated by graduates as the second most important technical skill required to work with Big Data, followed by Data Mining, Business Analysis, and Programming. While academics rankings of the expertise required for graduate students to be employed in the Big Data field were Data Mining, Statistical Learning, Programming, and then Mathematics in descending order. It is interesting that Business Analysis had the largest difference in average score between the ratings of graduates (4.03) and of academics (3.49). In contrast, Mathematics had slightly lower rating by graduates than by academics. The ratings assigned to Marketing and Artificial Intelligence were neutral in both surveys. The only category of expertise that was rated as not significant to work in Big Data workforce was Accounting. Its rating was much lower than the average score of the Statistical Analysis and Statistical Software Skills.

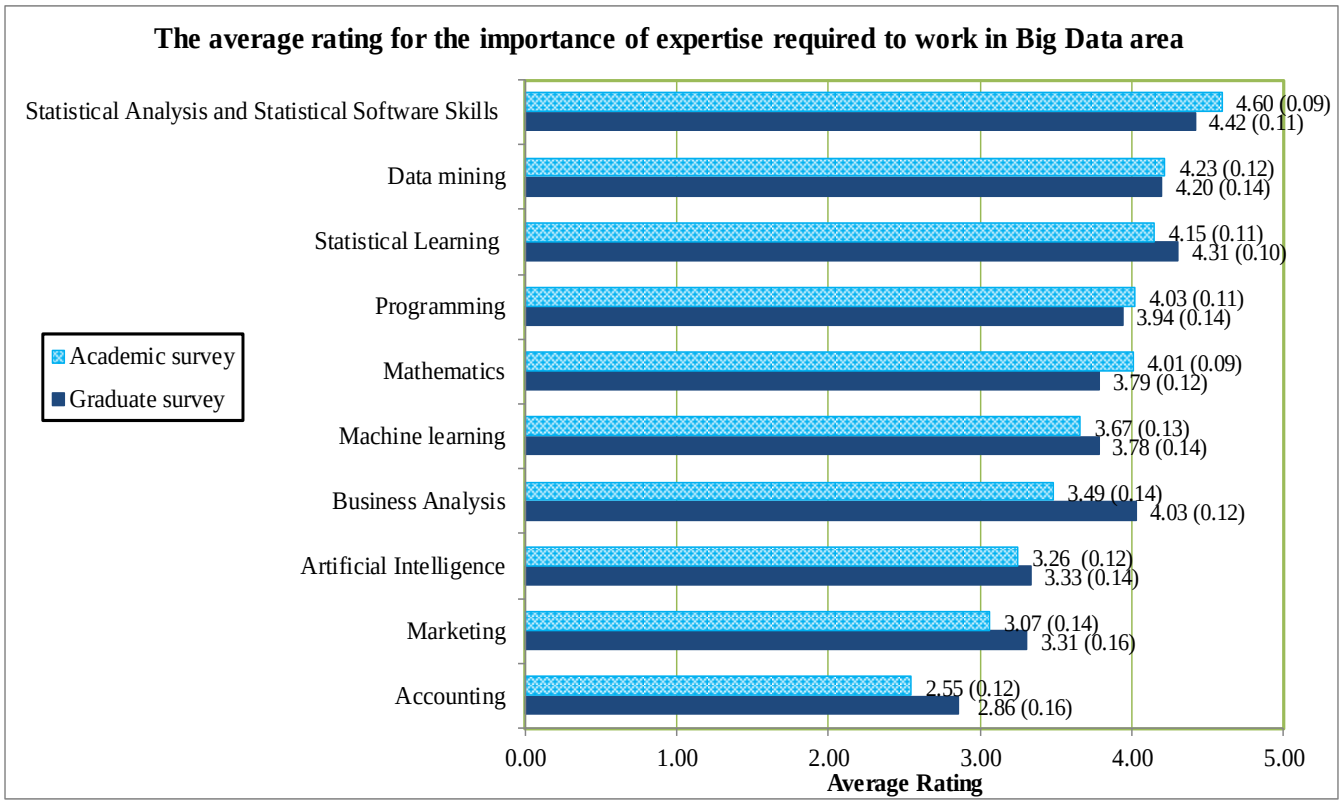


Figure 1: The average rating for the importance of expertise that is required to work with Big Data by academics (question 15) and graduates (question 18) based on 5 point Likert scale

In Figure 2 below we present results about the importance of various types of expertise as measured by the proportion of the participants who believe the graduates in the workforce have the skill (graduates) or they acquired the skill (academics). Statistical Analysis and Statistical Software Skills had the highest percentage both by graduates and academics as the required expertise to work in Big Data workforce. The Programming, Machine Learning, Marketing, and Accounting had slightly higher proportions of academics' perceived expertise for their graduates than graduates who had these skills. Within these four skills, only the Programming skill showed high percentages at 69% for academics and 63% for graduates. Over 66% of graduates had skills in Statistical learning, Mathematics, Data Mining, and Business Analysis. While more than 50% of academics stated there were courses teaching these skills. There were large differences between them. According to Figure 1, these four skills were considered to be necessary skills to work in Big Data workforce. So these skills should be given higher priority to prepare graduates who might work in Big Data analysis. In contrast, the Artificial Intelligence skill was much more likely to be considered necessary for working in Big Data by academics than by graduates. This might indicate a mismatch between what university courses cover and what graduates need to know for working in the Big Data area.

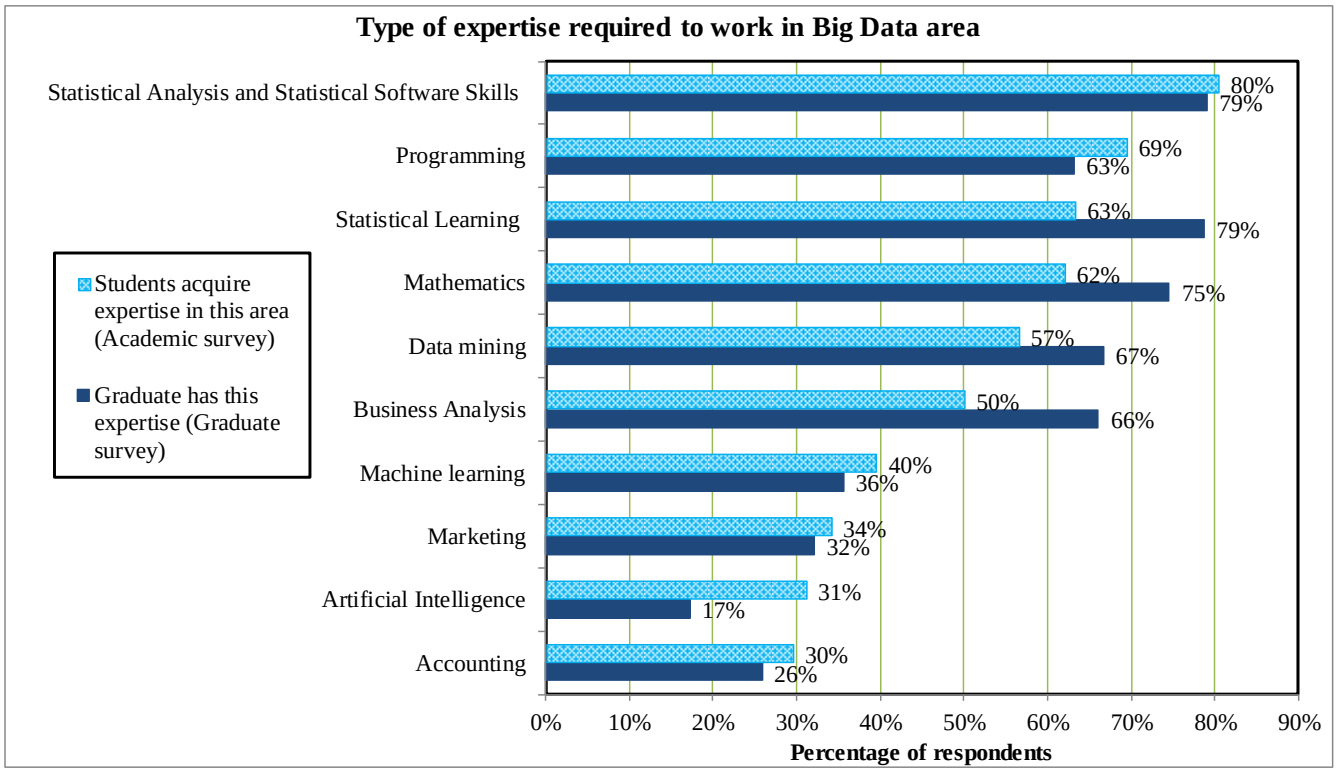


Figure 2: Perceptions of Academics (question 16) and Graduates (question 19) about expertise of the workforce working in Big Data measured as the proportion of the sample agreeing they have the skill (for the graduates) or acquired the skill (for the academics).

The results for the skills grouped by academics' perceived expertise for their graduates and whether graduates use it or not can be seen in Table 1. On average, the Statistical Analysis and Statistical Software Skills were rated highly necessary (4.93) to work in Big Data field by academics whose departments were providing a course in it while these skills were rated lower (3.50) by academics from other departments. The mean score in most skills for both academics and graduates were higher for academics' perceived expertise for their graduates and graduates who had this expertise than no expertise group, except for Artificial Intelligence in the graduate survey. There were big difference of average rating between academics that their departments teach Marketing, Business Analysis, Artificial Intelligence, and Accounting and academics that their departments did not teach these skills. Accounting also had large difference of average score between graduates who had this expertise themselves and not.

Type of expertise	Academic									Graduate								
	Yes			No			N/A			Yes			No			N/A		
	Average Rating	Standard Error	Sample Size	Average Rating	Standard Error	Sample Size	Average Rating	Standard Error	Sample Size	Average Rating	Standard Error	Sample Size	Average Rating	Standard Error	Sample Size	Average Rating	Standard Error	Sample Size
Statistical Analysis and Statistical Software Skills	4.93	0.04	40	3.50	0.34	10	4.53	0.15	32	4.50	0.12	48	4.08	0.36	12	4.43	0.30	7
Statistical Learning	4.35	0.16	31	4.06	0.21	16	4.00	0.19	31	4.37	0.12	46	4.08	0.19	12	4.29	0.36	7
Mathematics	4.13	0.16	30	3.74	0.13	19	4.06	0.16	33	3.79	0.14	42	3.64	0.31	14	4.20	0.37	5
Data mining	4.56	0.12	25	4.30	0.21	20	3.94	0.21	35	4.27	0.19	37	4.00	0.26	19	4.33	0.29	9
Machine learning	4.41	0.17	17	3.58	0.22	24	3.35	0.20	34	4.05	0.22	20	3.58	0.19	36	4.00	0.38	8
Artificial Intelligence	4.14	0.25	14	2.87	0.16	30	3.24	0.18	34	3.00	0.47	9	3.30	0.16	46	3.88	0.30	8
Programming	4.25	0.12	32	3.93	0.34	15	3.84	0.20	32	4.03	0.21	36	3.68	0.29	19	4.09	0.25	11
Marketing	4.36	0.17	14	2.54	0.19	28	2.97	0.19	32	3.88	0.26	16	3.03	0.21	36	3.44	0.38	9
Business Analysis	4.19	0.20	21	2.76	0.22	21	3.50	0.20	34	4.24	0.15	37	3.61	0.24	18	4.00	0.33	9
Accounting	3.17	0.30	12	2.20	0.15	30	2.65	0.18	34	3.71	0.27	14	2.50	0.18	40	3.11	0.45	9

Table 1: The average rating and standard deviation for the importance of type of expertise required to work with Big Data by academics (question 15) and graduates (question 18) based on 5 point Likert scale. The results are shown for academics (question 16) separately by whether or not they believe graduates have that skill and for graduates (question 19) separately by whether or not they claim to have the skill.

When we compared each skill's importance by academics' and graduates' disciplines (see Table 2 and Table 3), we found that academics in business and statistics disciplines rated highly the importance of Statistical Analysis and Statistical Software Skills. Academics and graduates in the computing discipline rated the importance of the Statistical skills lower than academics and graduates from other disciplines did. Data mining was rated more highly by academics in computing and marketing than by academics from other areas. Graduates from the statistics discipline gave a higher rating for Data mining skills than did graduates from other disciplines and rated it almost as important as skills in statistical analysis / statistical software, whereas graduates in marketing discipline rated Data mining skills lowest. Artificial Intelligence was rated lower by academics in the business discipline than by other academics but it was rated higher by graduates in business than by graduates from other disciplines. Machine learning was rated by people in computing discipline highly in the academic survey, whereas it had the lowest rating in the graduate survey (computer graduates). Interestingly, Programming was rated more highly by people in the statistics discipline than people in computing disciplines, both by academics and graduates. For Marketing skills, the academics in marketing and the graduates from business discipline gave the highest mean score. In industry, the term Business Analyst means a person who can translate the needs of business into a language that computer programmers can understand. These people bridge the communications gap between users and programmers. Business Analysis skill was rated more highly by graduates in all disciplines than by academics' view with the biggest gap between the

ratings of graduates and of academics being in the statistics discipline Accounting skills had the lowest ratings overall for both groups of participants.

Type of expertise	Academic														
	Statistics			Computing			Business			Marketing			Other		
	Average Rating	Standard Error	Sample Size	Average Rating	Standard Error	Sample Size	Average Rating	Standard Error	Sample Size	Average Rating	Standard Error	Sample Size	Average Rating	Standard Error	Sample Size
Statistical Analysis and Statistical Software Skills	4.90	0.10	20	4.34	0.16	29	5.00	0.00	10	4.50	0.31	14	4.44	0.24	9
Statistical Learning	4.50	0.17	18	3.93	0.18	28	4.44	0.24	9	4.21	0.32	14	3.78	0.32	9
Mathematics	4.10	0.23	20	3.97	0.16	29	4.20	0.20	10	3.86	0.21	14	4.00	0.17	9
Data mining	3.95	0.26	19	4.48	0.14	29	3.89	0.35	9	4.50	0.31	14	3.89	0.42	9
Machine learning	3.82	0.30	17	3.97	0.18	29	3.14	0.55	7	3.31	0.35	13	3.33	0.17	9
Artificial Intelligence	2.88	0.27	17	3.72	0.17	29	2.78	0.43	9	3.00	0.26	14	3.33	0.24	9
Programming	4.50	0.19	18	4.21	0.13	29	3.67	0.47	9	3.36	0.36	14	3.89	0.20	9
Marketing	2.38	0.31	16	2.90	0.17	29	3.14	0.51	7	4.31	0.21	13	3.00	0.33	9
Business Analysis	3.00	0.32	17	3.32	0.19	28	4.22	0.40	9	4.31	0.26	13	3.00	0.33	9
Accounting	2.06	0.30	16	2.54	0.17	28	2.56	0.41	9	3.00	0.23	14	2.78	0.28	9

Table 2: The average rating and standard deviation for the importance of type of expertise required to work with Big Data by academics (question 15) based on a 5 point Likert scale, grouped by academics' discipline

Type of expertise	Graduate														
	Statistics			Computing			Business			Marketing			Other		
	Average Rating	Standard Error	Sample Size	Average Rating	Standard Error	Sample Size	Average Rating	Standard Error	Sample Size	Average Rating	Standard Error	Sample Size	Average Rating	Standard Error	Sample Size
Statistical Analysis and Statistical Software Skills	4.78	0.10	18	4.13	0.27	16	4.39	0.26	18	4.20	0.37	5	4.40	0.22	10
Statistical Learning	4.56	0.15	18	3.88	0.29	16	4.47	0.15	17	4.25	0.48	4	4.30	0.15	10
Mathematics	4.06	0.19	18	3.36	0.29	14	4.13	0.20	16	3.50	0.50	4	3.44	0.34	9
Data mining	4.76	0.11	17	3.88	0.34	16	4.18	0.27	17	3.40	0.68	5	4.20	0.29	10
Machine learning	4.06	0.26	18	3.50	0.26	16	3.94	0.28	16	4.00	0.58	4	3.40	0.34	10
Artificial Intelligence	3.28	0.28	18	3.38	0.34	16	3.63	0.29	16	3.25	0.25	4	2.89	0.20	9
Programming	4.22	0.25	18	3.75	0.37	16	4.06	0.29	17	3.60	0.40	5	3.70	0.30	10
Marketing	3.18	0.23	17	3.13	0.36	15	3.73	0.32	15	3.25	0.75	4	3.20	0.44	10
Business Analysis	4.22	0.22	18	3.50	0.30	16	4.24	0.18	17	5.00	0.00	3	3.90	0.28	10
Accounting	2.50	0.29	18	2.69	0.22	16	3.53	0.31	15	2.50	0.65	4	2.90	0.50	10

Table 3: The average rating and standard deviation for the importance of type of expertise required to work with Big Data by graduates (question 18) based on a 5 point Likert scale, grouped by graduates' discipline

The average ratings for importance of areas of expertise in the graduate survey classified by graduates' company operating area (see Table 4) were quite similar for data analytics and finance groups except for the skills in machine learning, programming and marketing. The interesting difference was the data analytics group had higher ratings for Machine Learning and Programming, while the finance group showed greater scores for Marketing and Accounting.

Type of expertise	Graduate								
	Data analytics			Finance			Other		
	Average Rating	Standard Error	Sample Size	Average Rating	Standard Error	Sample Size	Average Rating	Standard Error	Sample Size
Statistical Analysis and Statistical Software Skills	4.53	0.27	15	4.56	0.17	18	4.29	0.16	34
Statistical Learning	4.62	0.14	13	4.50	0.15	18	4.09	0.16	34
Mathematics	4.00	0.23	13	3.89	0.19	19	3.62	0.20	29
Data mining	4.38	0.33	13	4.17	0.27	18	4.15	0.18	34
Machine learning	4.15	0.22	13	3.72	0.29	18	3.67	0.19	33
Artificial Intelligence	3.33	0.36	12	3.22	0.27	18	3.39	0.19	33
Programming	4.50	0.20	14	3.83	0.26	18	3.76	0.22	34
Marketing	3.08	0.34	12	3.56	0.26	18	3.26	0.24	31
Business Analysis	4.23	0.23	13	4.17	0.19	18	3.88	0.20	33
Accounting	2.54	0.43	13	3.12	0.22	17	2.85	0.22	33

Table 4: The average rating and standard deviation for the importance of type of expertise that is required to work with Big Data by graduates (question 18) based on 5 point Likert scale, grouped by graduates' company operating area

3.1.2 Software skills

Questions 17 and 18 of the academic survey asked respondents' opinions about the necessary software tools for Big Data analysis and whether the software tools are used in teaching Big Data analysis or not. We compared these academics' answers with the responses to questions 20 and 21 of the graduate survey that also asked graduates' opinions about the necessary software tools for Big Data analysis and whether or not the software tools are used in their company for Big Data analysis. In analyzing the responses to these questions, we coded all missing values as a Not Applicable (N/A) answer.

The results for the importance ratings and the usage of software tools are shown in Appendix K.1 and K.2. R programming had a higher rating by academics than by graduates and SQL had a higher rating by graduates than by academics. R programming and SQL were software tools with the highest proportions of usage by academics and graduates, respectively.

The importance of software skills and differences between academics and graduates is shown in Figure 3. The graduate survey had higher ratings than the academic survey for most software tools, with big differences in ratings for IBM SPSS Modeler, Base SAS, noSQL, and SQL. For the software tools R, Python, MapReduce, and Matlab, academics' ratings were higher than graduates' ratings. Hadoop showed the similar results for both groups. R programming (4.24) was the most important software skill in academic survey, followed by Python (3.89). However, the importance of R programming (3.77) and Python (3.51) were much lower in the graduate survey. The most

important software skill in graduate's opinion was SQL (4.18). WinBUGS was seen to be an unimportant software tool for Big Data analysis.

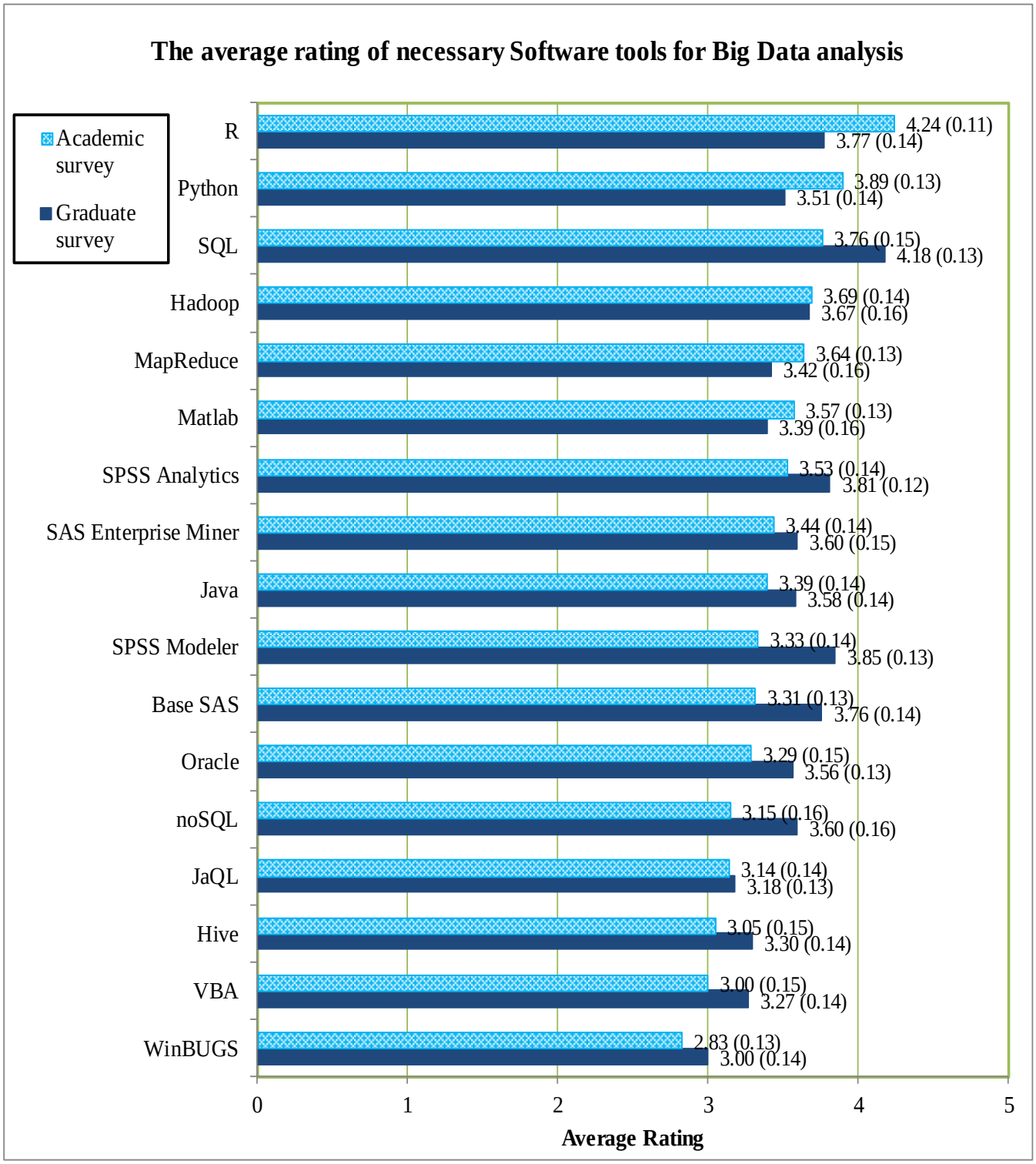


Figure 3: The average rating for the importance of software tools for Big Data analysis by academics (question 17) and graduates (question 20)

As can be seen in Figure 4, the large gap in the percentage of software tools usage between academics and graduates were present for most software tools, except for Oracle and IBM SPSS Modeler. The most common software tool used in teaching Big Data analysis was R programming (70%), whereas graduates use R programing in their company only half of that percentage (35%). Approximately 40% of academics use SQL, Matlab, Python, and IBM SPSS Analytics in their courses. However, the proportion of SQL and IBM SPSS Analytics, as well as Java and VBA usage was much higher at graduates' workplaces than taught by academics at their courses. It might be beneficial for universities to consider including these four software tools in their academic curriculum, especially in Computing, Statistics and Mathematics disciplines. On the other hand, Matlab and Python were seen as more important by academics than by graduates.

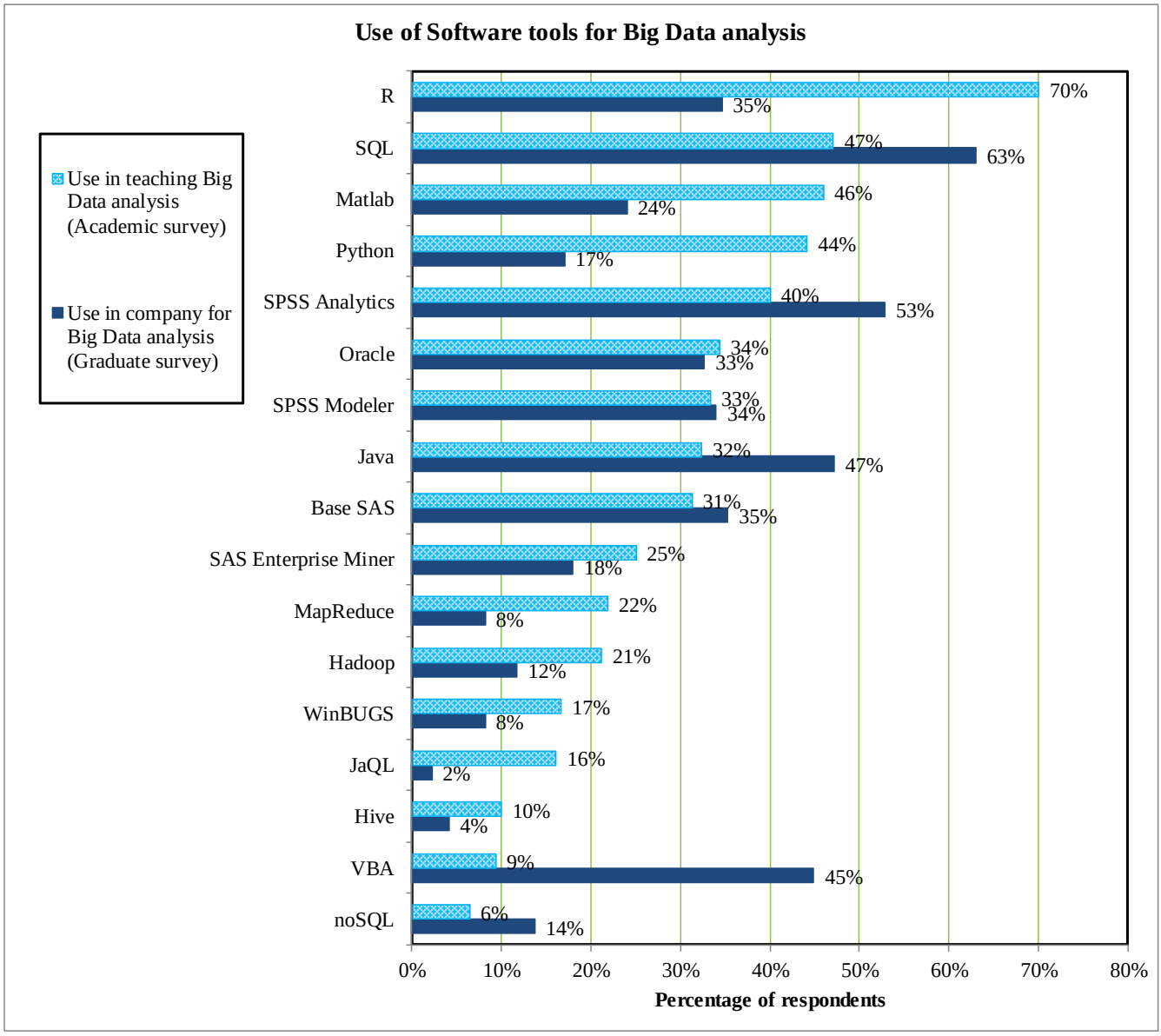


Figure 4: The extent of usage of software tools by academics in their units/courses (question 18) and by graduates in their workplace (question 21). (Please note that missing values were omitted.)

In Table 5 we present the average ratings by graduates for different software tools according to whether or not the graduates used them in their workplace and by academics according to whether or not they teach students these software skills. When we compared the importance of each software tool use (see Table 5), we found that most of them were seen as more important for graduates than by academics, except for MapReduce, Hive, and Java. MapReduce was rated the most important by academics, while graduates chose SAS Enterprise Miner as the most important software tool. MapReduce had the biggest gap in average score between academics (used in teaching) (4.50) and graduates used in working (3.33). The second most important software tools were Java and SQL for academics and graduates, respectively. R programming was the third most important software for both parties. The importance of IBM SPSS Analytics and Matlab rated by academics and graduates who use those software were similar. WinBUGS that was considered as not important software to work in Big Data area became important for graduates who use it in their companies.

Software tool	Academic									Graduate								
	Yes			No			N/A			Yes			No			N/A		
	Average Rating	Standard Error	Sample Size	Average Rating	Standard Error	Sample Size	Average Rating	Standard Error	Sample Size	Average Rating	Standard Error	Sample Size	Average Rating	Standard Error	Sample Size	Average Rating	Standard Error	Sample Size
Base SAS	3.89	0.31	9	3.28	0.25	18	3.15	0.17	27	4.47	0.19	15	3.56	0.17	25	3.20	0.36	10
SAS Enterprise Miner	4.14	0.26	7	3.35	0.23	20	3.32	0.20	28	4.78	0.15	9	3.38	0.17	32	3.27	0.36	11
SPSS Analytics	4.07	0.22	14	3.33	0.30	15	3.36	0.19	28	4.08	0.16	26	3.58	0.19	19	3.56	0.34	9
SPSS Modeler	3.80	0.36	10	3.19	0.26	16	3.25	0.19	28	4.33	0.16	15	3.77	0.17	26	3.36	0.34	11
WinBUGS	3.20	0.66	5	2.71	0.19	17	2.83	0.17	24	4.00	0.41	4	2.85	0.13	33	3.11	0.42	9
Matlab	4.07	0.23	15	3.31	0.25	16	3.44	0.18	27	4.09	0.39	11	3.10	0.17	30	3.50	0.45	10
R	4.33	0.18	27	3.67	0.33	9	4.33	0.13	27	4.50	0.17	18	3.31	0.17	26	3.67	0.41	9
Java	4.36	0.20	11	2.76	0.29	17	3.39	0.14	28	3.91	0.18	23	3.29	0.24	24	3.50	0.27	8
VBA	3.00	0.00	2	2.83	0.27	18	3.13	0.17	24	3.43	0.23	21	3.26	0.20	23	2.88	0.30	8
Hadoop	3.83	0.48	6	3.75	0.24	20	3.59	0.18	22	4.40	0.40	5	3.63	0.19	32	3.44	0.34	9
MapReduce	4.50	0.34	6	3.53	0.24	17	3.50	0.16	24	3.33	1.20	3	3.45	0.17	33	3.33	0.41	9
noSQL	4.00	0.00	1	3.00	0.24	19	3.24	0.22	21	4.43	0.30	7	3.48	0.20	31	3.33	0.41	9
SQL	3.71	0.24	14	3.25	0.35	12	4.00	0.20	28	4.60	0.15	30	3.74	0.23	19	3.63	0.26	8
JaQL	3.80	0.37	5	2.87	0.26	15	3.17	0.18	23	4.00	0.00	1	3.33	0.13	30	2.50	0.33	8
Hive	4.00	0.58	3	2.75	0.19	16	3.14	0.21	22	3.50	0.50	2	3.39	0.14	33	2.89	0.42	9
Oracle	3.64	0.20	11	2.83	0.32	12	3.35	0.21	26	4.18	0.21	17	3.41	0.16	29	2.89	0.26	9
Python	3.73	0.21	15	3.50	0.31	14	4.19	0.17	27	4.29	0.29	7	3.37	0.15	30	3.40	0.37	10

Table 5: The average rating and standard deviation for the importance of software tools for Big Data analysis by academics (question 17) and graduates (question 20) based on 5 point Likert scale, grouped by academics' (question 18) and graduates' (question 21) use of the software

The necessity of software tools for Big Data analysis identified by academics and graduates by their disciplines (see Table 6 and Table 7) indicated that Base SAS, IBM SPSS Analytics, IBM SPSS Modeler, SQL, and Oracle were rated higher by graduates than academics in all disciplines.

Only R programming was identified as more important by academics from all disciplines than by graduates. In the marketing discipline, the academics rated many software tools as more important than did the graduates from marketing discipline. SQL was seen as more important by graduates in the business discipline than by the academics in the same discipline, while a similar pattern is observed for NoSQL in the computing discipline. Opposite patterns were observed for R, NoSQL, Hadoop, and Python in marketing discipline (i.e. academics rated higher than graduates). Please see Appendix M for the number of participants in each discipline.

Software tool	Academic														
	Statistics			Computing			Business			Marketing			Other		
	Average Rating	Standard Error	Sample Size	Average Rating	Standard Error	Sample Size	Average Rating	Standard Error	Sample Size	Average Rating	Standard Error	Sample Size	Average Rating	Standard Error	Sample Size
Base SAS	3.47	0.26	15	3.11	0.20	19	3.17	0.31	6	3.88	0.35	8	3.00	0.52	6
SAS Enterprise Miner	3.43	0.25	14	3.25	0.22	20	3.71	0.42	7	3.78	0.36	9	3.20	0.66	5
SPSS Analytics	3.43	0.29	14	3.50	0.21	20	3.67	0.37	9	3.78	0.40	9	3.20	0.58	5
SPSS Modeler	3.21	0.28	14	3.15	0.18	20	3.50	0.50	6	3.78	0.40	9	3.40	0.68	5
WinBUGS	3.07	0.34	14	2.63	0.13	16	3.00	0.32	5	2.83	0.31	6	2.60	0.40	5
Matlab	3.59	0.29	17	3.47	0.18	19	3.86	0.26	7	3.43	0.37	7	3.63	0.46	8
R	4.32	0.23	19	4.27	0.13	22	4.00	0.38	8	4.43	0.37	7	4.00	0.31	7
Java	3.31	0.28	16	3.64	0.19	22	2.57	0.37	7	4.00	0.26	6	3.00	0.55	5
VBA	3.36	0.31	11	2.69	0.22	16	2.83	0.48	6	3.50	0.22	6	2.80	0.49	5
Hadoop	3.75	0.28	12	3.82	0.21	17	2.83	0.48	6	4.00	0.45	6	3.71	0.29	7
MapReduce	3.82	0.26	11	3.94	0.20	17	2.67	0.42	6	3.33	0.33	6	3.71	0.29	7
noSQL	3.13	0.40	8	3.13	0.18	16	2.67	0.42	6	4.20	0.49	5	2.83	0.48	6
SQL	3.85	0.36	13	4.00	0.16	20	2.83	0.54	6	3.83	0.40	6	3.67	0.41	9
JaQL	3.40	0.37	10	3.13	0.15	16	2.67	0.42	6	3.67	0.42	6	2.60	0.40	5
Hive	3.38	0.46	8	2.94	0.14	16	2.67	0.42	6	3.67	0.42	6	2.60	0.40	5
Oracle	3.22	0.43	9	3.45	0.21	20	2.86	0.40	7	3.86	0.26	7	2.67	0.33	6
Python	4.27	0.21	15	3.94	0.22	18	3.00	0.49	7	4.00	0.27	8	3.75	0.31	8

Table 6: The average rating and standard deviation for the importance of software tools for Big Data analysis by academics (question 17) based on 5 point Likert scale, grouped by academics' disciplines

Software tool	Graduate														
	Statistics			Computing			Business			Marketing			Other		
	Average Rating	Standard Error	Sample Size	Average Rating	Standard Error	Sample Size	Average Rating	Standard Error	Sample Size	Average Rating	Standard Error	Sample Size	Average Rating	Standard Error	Sample Size
Base SAS	4.21	0.21	14	3.42	0.29	12	3.69	0.33	13	4.33	0.67	3	3.38	0.26	8
SAS Enterprise Miner	4.00	0.22	16	3.33	0.28	12	3.58	0.34	12	3.00	0.82	4	3.50	0.42	8
SPSS Analytics	3.71	0.22	17	3.58	0.23	12	4.15	0.22	13	4.00	0.58	3	3.78	0.32	9
SPSS Modeler	4.06	0.21	16	3.45	0.21	11	3.92	0.29	12	4.33	0.67	3	3.70	0.33	10
WinBUGS	2.93	0.25	14	3.00	0.21	12	3.00	0.36	11	3.00	0.00	3	3.17	0.48	6
Matlab	3.31	0.25	16	4.00	0.32	13	3.08	0.38	13	3.00	0.00	3	3.17	0.48	6
R	4.13	0.22	16	4.00	0.26	14	3.33	0.38	12	3.00	0.00	3	3.63	0.32	8
Java	3.63	0.31	16	3.60	0.24	15	3.15	0.25	13	4.00	0.45	5	4.00	0.26	6
VBA	3.24	0.28	17	3.00	0.23	13	3.33	0.28	12	3.75	0.48	4	3.50	0.43	6
Hadoop	4.00	0.28	14	3.77	0.30	13	3.00	0.33	9	3.00	0.00	3	4.00	0.38	7
MapReduce	3.62	0.27	13	3.75	0.28	12	2.73	0.38	11	3.00	0.00	3	3.83	0.40	6
noSQL	3.50	0.33	14	4.15	0.22	13	3.09	0.39	11	3.00	0.00	3	3.83	0.40	6
SQL	4.27	0.23	15	4.20	0.28	15	4.13	0.27	15	4.00	0.58	4	4.13	0.30	8
JaQL	3.11	0.26	9	3.38	0.18	13	2.90	0.38	10	3.33	0.33	3	3.25	0.25	4
Hive	3.62	0.27	13	3.31	0.21	13	2.80	0.33	10	3.00	0.00	3	3.60	0.40	5
Oracle	3.67	0.25	15	3.57	0.23	14	3.29	0.30	14	4.00	0.58	4	3.63	0.26	8
Python	3.69	0.26	13	3.80	0.22	15	3.09	0.31	11	3.00	0.00	3	3.40	0.40	5

Table 7: The average rating and standard deviation for the importance of software tools for Big Data analysis by graduates (question 20) based on 5 point Likert scale, grouped by graduates' disciplines

The average rating for the importance of software tools by graduates' company operating area (see Table 8) were higher for graduates in data analytics than in finance. The exception was for SAS Enterprise Miner, IBM SPSS Modeler, and MapReduce. The highly important software tool ratings by graduates in data analytics were SQL, Base SAS, Oracle, and IBM SPSS Analytics. For Oracle and JaQL, the average score of graduates in finance were a lot lower than data analytics group.

Software tool	Graduate								
	Data analytics			Finance			Other		
	Average Rating	Standard Error	Sample Size	Average Rating	Standard Error	Sample Size	Average Rating	Standard Error	Sample Size
Base SAS	4.33	0.29	9	3.88	0.22	16	3.48	0.21	25
SAS Enterprise Miner	3.56	0.44	9	3.81	0.23	16	3.48	0.22	27
SPSS Analytics	4.00	0.30	10	3.81	0.19	16	3.75	0.18	28
SPSS Modeler	3.89	0.31	9	3.94	0.19	16	3.78	0.19	27
WinBUGS	3.22	0.15	9	3.00	0.25	13	2.92	0.22	24
Matlab	3.50	0.22	10	3.13	0.29	16	3.52	0.27	25
R	3.80	0.29	10	3.69	0.27	16	3.81	0.21	27
Java	3.92	0.36	12	3.44	0.24	16	3.52	0.18	27
VBA	3.70	0.26	10	3.06	0.23	16	3.23	0.21	26
Hadoop	3.88	0.35	8	3.43	0.29	14	3.75	0.22	24
MapReduce	3.33	0.41	9	3.43	0.29	14	3.45	0.23	22
noSQL	3.80	0.33	10	3.14	0.31	14	3.78	0.23	23
SQL	4.44	0.29	9	4.06	0.22	18	4.17	0.19	30
JaQL	3.71	0.36	7	2.92	0.19	12	3.15	0.18	20
Hive	3.50	0.27	8	3.43	0.29	14	3.14	0.18	22
Oracle	4.11	0.35	9	3.12	0.21	17	3.66	0.17	29
Python	3.89	0.31	9	3.50	0.25	14	3.38	0.19	24

Table 8: The average rating and standard deviation for the importance of software tools for Big Data analysis by graduates (question 20) based on 5 point Likert scale, grouped by graduates' company operating area

3.2 The technical and software skills required for Big Data professionals

When we consider the results shown in Appendix J.2, the graduates' survey shows that the most important types of expertise area required to work with Big Data were Statistical Analysis and Statistical Software Skills at 82% (56% Strongly Agree and 26% Agree). This was followed by Statistical Learning, and Data Mining at 81%, and 74%, respectively, then came Machine learning, Mathematics, Artificial Intelligence, and Marketing. Programming and Business Analysis were at the same importance rating (67% agree or strongly agree). The least important technical skill required in this area was Accounting (23%). Other skills suggested included Communication, Project management, Business understanding, Critical thinking, Creative thinking, Data design, and Engineering.

SQL was the most important software tools for Big Data analysis rated by graduates at 62% with 24% Agree and 38% Strongly Agree (see Appendix K.2). Then followed by IBM SPSS Analytics, IBM SPSS Modeler, R programming, Java, and Base SAS, SAS Exterprise Miner at

important rate over 40%. The importance of software tools at 30% level were Oracle, Hadoop, noSQL, and Matlab. Python, VBA, and MapReduce had the important level at 20%. The importance of Hive, JaQL, and WinBUGS were at 10% level. The other software tools suggested by graduates were Apache Mahout, STATA, AWS Cloud, Data presentation software as Tableau or D3, Julia, Google Analytics, C#, C++, PHP, Scala, Microsoft Access, Microsoft Excel, Data Conversion System, Prophet, MiniTab, and Bloomberg Terminal.

On the other hand, by analyzing the data scientist's job advertisements we found that there is wide variation in the supplementary skills required by employers to work in these kind of jobs. The most often mentioned skills/expertise were Big Data statistics, computing skills, and statistical techniques. The list of all skills listed to be required for a data scientist gathered from job advertisements can be seen in Appendix L. The important statistical techniques required to work as a data scientist can be divided into Data Mining techniques and Multivariate Statistical techniques. Data Mining techniques included the knowledge of Random Forests, Bayesian networks, Neural networks, Support Vector Machines, Decision Trees, Machine Learning algorithms, K-Nearest Neighbours, Affinity Analysis, Classifier and predictor development, Pattern detection, Natural Language Processing (Text mining), and Sentiment Analysis (Opinion mining). Multivariate Statistical techniques included the knowledge of Heuristics, Logistic Regression, Linear Regression, Generalised Linear Models, Regression analysis, Multivariate analysis, Cluster analysis, Data Munging, Experimental design (sampling, confidence), Time series analysis, and Computer Vision.

There are various software tools listed in the job advertisements. Statistical software tools included R, SAS, Matlab, IBM SPSS Analytics, and Stata. Computing programming tools included Python, Java, and C. The required computer skills for data scientist include Lisp, and Haskell. In some advertisements, the knowledge of mapping software tools e.g. ArcGIS is also listed as required knowledge. The knowledge of data warehouse back-end tools such SQL, NoSQL, Oracle, Sybase, DB2, and Teradata was also required. The newly developed analysis tools which can deal with large datasets in Big Data such as Hadoop (Hive/ Impala/ Pig), Mahout, PostgreSQL, AWS (EC2/ EMR/ S3), MongoDB, and Neo4J – Cypher were part of the requirements of the advertisements.

The important non-technical skills required from data scientist advertisements are interpersonal skills, communication skills, leadership skills, ability to work as a team member, and problem solving skills. These skills are considered *soft skills* which are rarely included in the tertiary curriculum in disciplines like mathematics, computing, actuarial studies and statistics but highly desired by employers.

3.3 The opinions of Academics in disciplines relevant to Big Data

Most academic respondents (72%) agreed that Big Data (data analytics) was important to their work and their department's work with 25% rating it as very important and 47% rating it as somewhat important. Only 9% rated Big Data as neither important nor unimportant and 15% of them rating Big Data as unimportant to their work (9% somewhat unimportant; 6% very unimportant).

When we looked at the open-ended questions, please beware only 70 out of 87 of the participants answered the open-ended question. They proposed their reasons for why Big Data was important or not important to their work. The most significant reasons for their rating of the importance of Big Data related to the ability to handle data management, retrieval, processing, transfer, and storage. Big Data generates lots of new insights and knowledge. It is used in High-performance computing (HPC), Service-orientated computing, Geospatial mapping, and the Virtual Worlds Business projects. Big Data is also widely used in analyzing huge volumes of data in areas such as Astronomy, Geoscience, Marketing, Hyperspectral imaging, Genomics and Proteomics. In addition, academics in the marketing discipline focused their attention on big data's implications for the digital and social media age, for example, extracting insights and information from customer data to make appropriate marketing and product design decisions or to develop appropriate social media campaigns.

Their opinion is that Big Data is a very important emerging area. It is regarded as a hot topic. Big Data is a growing field. Growth in demand from industry was the second major reason for the significance of Big Data. There is increasing market demand for data scientists. Their view is that this new area has the potential to attract many students. In response to this, many Universities are keen on developing units and programs relevant to Big Data. Some universities have planned to launch new degrees in data science in 2015. Examples of such new degree programs are: the Bachelor of Computer Science (Data Science) and Master of Business Analytics at Deakin University, Master of Computer Science specialising in Big Data Management and Master of Information Technology specialising in Big Data Management at RMIT University, and Masters of Business Data Science at University of Otago (New Zealand).

Another crucial reason for the importance of Big Data for academics was that it is an active research area with good prospects for funding. Some universities have organized research centers in this area such as PRaDA (Pattern Recognition and Data Analytics) at Deakin University, and Big Data Infrastruture – Big Data at RMIT University. Some of them felt that Big Data was more

important to their research effort than to teaching. However, they pointed out that the main limiting factor was the availability of data from industry partners. Industry Partnerships are highly desirable in order to have access to useful Big Data.

On the other hand, some respondents were uncertain about Big Data long-term value or popularity. Their departments have been slow to realize the importance of Big Data. They also questioned its enduring core skills. Some felt that Big Data was just a new "buzzword" to describe what they have always done. Other participants disagreed, mentioning that "the old hacks had no idea what Big Data was about" and they only thought that it was IT departments' thing.

Some academics' view was that Big Data was unimportant to their work, was not their department's area of expertise and was not supported by their senior management. Some were concerned about cost. Some thought that few people were interested in large data sets, but that most people undertake small sample studies.

4. Conclusions

In this thesis we have examined the perceptions of graduates working in the new field of Big Data Analytics and the perceptions of academics working in disciplines producing graduates likely to work in that field. A new hybrid type of career, the "data scientist" has emerged recently. Our interest is in what areas of expertise are required for data scientists and what software tools do they need to know. Some academics and graduates in statistics regard "data science" as applied statistics whereas academics and graduates from computer science may see it as a set of new software skills and tools belonging in the realm of computer science, not statistics. We have examined the differences between what academics think and what graduates in industry think about the areas of expertise and the types of software tools required for working in Big Data analytics / data science. This research can help to inform university academics about the design of the curriculum for degrees in Statistics, Computing, Mathematics, and Data Science.

Both the graduates and the academics rated expertise in Statistical Analysis and Statistical Software as the most important type of expertise, followed by Statistical Learning and Data Mining. Expertise in Accounting was rated as the least important area of expertise by both academics and graduates. Machine Learning and Artificial Intelligence were given similar ratings by academics and graduates and the rating was somewhere between "neutral" and "agree". Expertise in Programming was rated more highly than this and the ratings by the academics and by the graduates were similar for Programming. Expertise in Mathematics was rated lower than expertise in Programming by both groups. Business Analysis had a higher rating by graduates than by academics.

A big range of different software tools are used in industry and in academia for Big Data analysis. Some software tools may be more suited to teaching than to industry practice. This may partly explain the differences between the results for academics and for graduates in the average importance ratings and level of usage of different software tools. The R statistical software package is open source and widely used in academia. It has a higher rating and higher level of usage according to academics than according to graduates in industry. For SQL the opposite is true. Some software tools have very low levels of usage in industry. For instance, WinBUGS, Hive, JaQL, and MapReduce had less than 10% of the graduates saying it is used in their workplaces. When graduates join the workplace they should be capable of learning new software tools. It may not be necessary for university degrees to provide training in all of the software tools we have identified as being used in industry.

The interesting gaps between academics' and graduates' perceptions of expertise required to work in Big Data Analytics are as follows:

- Business Analysis had mean score much lower in academics than in graduates with the biggest gap in statistics discipline. This might mean that Business Analysis was more important in the Big Data workplace than academics thought it was.
- Other big differences were in the importance of Marketing, Artificial Intelligence, and Accounting. The academics that their departments teach these skills had much higher important rate than the academics that their departments did not teach these skills.
- The academics in business and statistics discipline gave extremely high importance to the Statistical Analysis and Statistical Software Skills.
- The academics and graduates in computing gave lower importance rates to the Statistical skills than other disciplines.
- In marketing discipline, academics gave much higher importance rates to Data mining than graduates.
- The importance of Artificial Intelligence in business discipline was the lowest in academic survey while it was the highest in graduate survey.
- Programming skill had higher importance rating by people in statistics discipline than people in computing disciplines.

In helping universities designs data science programs, we need to know what skills are important for students to know. The Statistical Analysis and Statistical Software Skills were obviously the most necessary type of expertise required for working in the Big Data field. The other types of expertise that should be given priority in teaching were Statistical Learning, Data Mining, Business Analysis, Programming, Mathematics, and Machine Learning. There was a big difference between academics' and graduates' perception of importance of Business Analysis.

The noteworthy differences between the results from the academic and graduate surveys about the importance of software tools were as follows.

- It was clear that most software tools were rated more highly by graduates than by academics, except R programming, Python, MapReduce, and Matlab where academics' average ratings were higher.
- IBM SPSS Modeler, Base SAS, noSQL, and SQL had much lower importance rating by academics than by graduates.

- Most software tools had a big gap in the proportion of usage between academics and graduates use in their work, except for Oracle and IBM SPSS Modeler. This points to a mismatch in software tools used in education and in industry for Big Data analysis.
- MapReduce had a higher average rating by academics than by graduates working in Big Data analysis.
- Base SAS, IBM SPSS Analytics, IBM SPSS Modeler, SQL, and Oracle had higher important ratings by graduates than by academics in all disciplines.
- R programming was identified as more important by academics from all disciplines than by graduates.
- The academics in the marketing discipline rated many software tools as more important than graduates did in that same discipline.
- SQL was seen as more important by the graduates in the business discipline than by academics in the same discipline, while a similar pattern is observed for NoSQL in computing discipline. Opposite patterns were observed for R, NoSQL, Hadoop, and Python in marketing discipline (i.e. academics rated higher than graduates).

Data Science degrees should include SQL as the most necessary software skill for Big Data analysis. It might be beneficial for universities to consider including the software tools identified in this study, exclude WinBUGS, into their academic curricula, especially in the Computing, Statistics and Mathematics disciplines. WinBUGS may be regarded as an unimportant software tool for Big Data analysis.

As in any study, this study had limitations such as variability in respondents' interpretation of the terms used, academics' knowledge of the needs of industry, and graduate's awareness of developments in both industry and academia. Participants may have very different levels of knowledge and experience and areas of application of Big Data techniques.

We approached the entire population of relevant academics and university departments and we were able to approach a reasonable number of graduates via snowball sampling. This was a descriptive study with the idea of describing the current the situation which will possibly suggest more research questions to be investigated in follow up studies.

This study suggests further research. We have surveyed academic departments from around Australia and New Zealand and while we obtained a reasonable response rate, it would have been good if we had been able to interview some of these academics from these departments and obtain more in depth understanding of their experiences and opinions about the issues involved in

education for careers in Big Data analytics. Likewise it would have been good to interview some of the graduates working in the area and interview their employers too. We obtained a reasonable number of responses from the graduates. We used snowball sampling to contact these people as other methods to contact them were not practicable. Employers / industry representatives are often too busy to make themselves available or do not want to divulge much information about their business. Such interview based research is time consuming and labour intensive and this could not be done in the time frame available for completing the thesis. By their nature, online surveys will not always get a response from a random sample of the population we want to sample from. We plan to do a follow up study and interview some of these people.

Another set of issues worthy of study is the extent to which industry is ahead of academia in the development of methods and software for Big Data analysis, and how universities need to respond to this and form partnerships with industry. This was beyond the scope of this thesis. The nature of statistics practice in industry and government is changing in response to the availability of new software tools. This poses a challenge for educational institutions providing education for statisticians, in particular in terms of the balance between teaching statistical theory and teaching computer skills and software. It is a rapidly changing area.

Appendices

A. Glossary

ArcGIS	An application of Geographic Information System (GIS)
AWS	Amazon Web Services (a cloud computing service)
BUGS	Bayesian inference Using Gibbs Sampling
C	A programming language
C#	A programming language
C++	A programming language
Cloudera	A Big Data database provider
Coursera	A MOOC provider
CRM	Customer relationship management
CSV	Comma Separated Value
Cypher	A Neo4j query language
D3	Data-Driven Documents (a JavaScript library for manipulating documents based on data)
Datameer	A Big Data analytics company
DB	Database
DB2	A database server
EC2	Amazon Elastic Compute Cloud (a web service that provides resizable compute capacity in the cloud)
EMR	Amazon Elastic MapReduce (a web service to quickly and cost-effectively process vast amounts of data)
Gartner	An information technology research and advisory company
Hadoop	Apache Hadoop (an open-source software framework for reliable, scalable, and distributed computing)
Haskell	An advanced purely-functional programming language
Hive	Data warehouse software facilitates querying and managing large datasets residing in distributed storage
Hortonworks	A computer software company, focuses on Hadoop
HPC	High-performance computing
IBM	International Business Machines (a global technology and innovation company)
Impala	Cloudera Impala (an open-source, interactive SQL for Hadoop)
Intel	A silicon innovation, develops technologies and products company
IT	Information Technology

JaQL	Primarily for a query language for JavaScript Object Notation (JSON) (IBM Institute, n.d.)
Java	a programming language and computing platform
JavaScript	A programming language
JSON	JavaScript Object Notation (a lightweight data interchange format)
Julia	A high-level, high-performance dynamic programming language for technical computing
Linux	An operating system
Lisp	LISt Processing (a high-level programming language)
Mahout	Apache Mahout (a library of scalable machine-learning algorithms, implemented on top of Apache Hadoop)
MapReduce	A programming paradigm that allows for massive scalability of servers in a Hadoop cluster
Matlab	Matrix laboratory (a high-level language and interactive environment for numerical computation, visualization, and programming)
MiniTab	A statistical software
MongoDB	From "hu**mongo**us" and database (an open-source document database, classified as a NoSQL database)
MOOC	Massive Open Online Course
MySQL	My Structured Query Language (an open-source relational database management system)
Neo4j	An open-source graph database, implemented in Java
NoSQL	Not only SQL (a database computer language)
Oracle	A software or a global technology and innovation company
PBT	Prescient Business Technologies
Petabytes	1000^5 bytes
PHP	Hypertext Preprocessor (an open-source scripting language)
Pig	Apache Pig (a high-level platform for creating MapReducee programs used with Hadoop)
PostgreSQL	An open-source object-relational database system
Prophet	A statistical and sequence analysis software
PwC	Pricewaterhouse Coopers (professional services firms)
Python	A free software programming language and environment for statistical computing and graphics
Quantium	A developed data, analytics, and software tools company

R	R programming (a free software programming language and environment for statistical computing and graphics)
S3	Amazon S3 (a cloud storage for the Internet)
SaaS	Software as a Service (delivering applications over the Internet)
SAS	Statistical Analysis System (a software or a global technology and innovation company)
Scala	A programming language
SPSS	A statistical analysis software
SQL	Structured Query Language (a database computer language)
Stata	A data analysis and statistical software package
StatsJobs	An online specialist recruitment operation
Sybase	A software to manage, analyze, and mobilize information, using relational databases, analytics and data warehousing solutions and mobile-application development platforms
Tableau	A computer software of interactive data visualization products focused on business intelligence
Terabytes	1000^4 bytes
TSV	Tab Separated Value
Udacity	A MOOC provider
Unix	An operating system
VBA	Visual Basic for Applications (a programming language)
WinBUGS	A part of the BUGS project (a flexible statistical software for the Bayesian analysis of complex statistical models using Markov chain Monte Carlo methods
XML	Extensible Markup Language
Zettabytes	1000^7 bytes

B. Industry websites about Big Data

IT Corporation	Link
Oracle	http://www.oracle.com/us/technologies/big-data/index.html
Intel	http://www.intel.com.au/BigData
IBM	http://www.ibm.com/DB2BigData
IBM	http://www.ibmbigdatahub.com
Quantium	http://www.quantium.com.au/harnessing-value-big-data
Cloudera	http://www.cloudera.com/content/cloudera/en/training/courses.html
Hortonworks	http://hortonworks.com/blog/stinger-phase-2-the-journey-to-100x-faster-hive/
Datameer	http://www.datameer.com/product/big-data.html
Gartner	http://www.gartner.com/technology/topics/big-data.jsp

C. University degree programs relating to Big Data

University	School	Degree program
Australia		
University of South Australia	School of Information Technology & Mathematical Sciences	Master of Data Science degree, Graduate Diploma of Data Science, Graduate Certificate in Data Science
Macquarie University	Faculty of Science	Master of Data Science degree, Graduate Diploma of Data Science, Graduate Certificate in Data Science
Deakin University	School of Information Technology (Bachelor), School of Information and Business Analytics (Master)	Bachelor of Computer Science (Major in Data Science), Master of Business Analytics, Graduate Diploma of Business Analytics
University of Western Sydney	School of Computing, Engineering and Mathematics	Master of Information and Communications Technology (Specialisation in Distributed Computing)
University of Technology Sydney	Faculty of Engineering and Information Technology	Masters of Data Science and Innovation, Graduate Diploma in Data Science and Innovation, Graduate Certificate in Data Science and Innovation
RMIT University	School of Computer Science and Information Technology	Bachelor of Computer Science (Major in Big Data), Master of Information Technology (Specialisation in Big Data Management), Master of Computer Science (Specialisation in Big Data Management)
New Zealand		
University of Auckland	Faculty of Science	Master of Professional Studies in Data Science
University of Otago	Otago Business School	Master of Business Data Science

University	School	Degree program
United States		
Arizona State University	W.P. Carey School of Business	Master of Science in Business Analytics
Bentley University	Graduate School of Business	Master of Science in Marketing Analytics, Master of Business Analytics
Carnegie Mellon University	Heinz College	Master of Information Systems Management (Concentration in Business Intelligence and Data Analytics), Master of Science in Information Technology (Concentration in Business Intelligence and Data Analytics)
Columbia University	The Fu Foundation School of Engineering and Applied Science	Masters of Science in Computer Science
DePaul University	College of Computing and Digital Media	Master of Science in Predictive Analytics
Drexel University	LeBow College of Business	Master of Science in Business Analytics
Indiana University, Bloomington	Kelley School of Business	Master of Business Administration (Major in Business Analytics)
Louisiana State University	E.J. Ourso College of Business	Master of Science in Analytics
Massachusetts Institute of Technology	The MIT Sloan School of Management	Master of Business Administration
Michigan State University	Broad College of Business	Master of Science in Business Analytics
New York University	Stern School of Business	Master of Science in Business Analytics
North Carolina State University	Institute for Advanced Analytics	Master of Science In Analytics
Northwestern University	School of Continuing Studies (MSPA) and McCormick School of Engineering and Applied Science (MSiA)	Online Master of Science in Predictive Analytics, Master of Science in Analytics
Purdue University	Krannert School of Management	Master of Business Administration (Concentration in Business Analytics)
Rutgers University	Graduate School, Professional Science Masters Programs	Master of Business and Science (Concentration in Analytics – Discovery Informatics and Data Sciences)
Stanford University	School of Engineering, Computer Science Department	Master of Science In Computer Science (Specialisation in Information Management and Analytics)
University of California, Berkeley	College of Engineering (M.Eng), School of Information (MIDS)	Master of Engineering (concentration in Data Science), Master of Information and Data Science
University of Cincinnati	Lindner College of Business	Master of Science in Business Analytics
University of Connecticut	School of Business, Department of Operations and Information Management	Master of Science in Business Analytics and Project Management
University of Maryland	Robert H. Smith School of Business	Master of Science in Business for Marketing Analytics
University of San Francisco	School of Management and the College of Arts and Sciences	Master of Science in Analytics
University of Tennessee	College of Business Administration, Statistics, Operations and Management Science	Master of Science in Business Analytics, Professional MBA with Business Analytics
Villanova University	School of Business	Master of Science in Analytics

D.Survey questionnaire

D.1 Academic Survey

Academic Survey

Research Project: Big data and its implications for the statistics profession and statistics education

The study is being conducted by Miss Busayasachee Puang-Ngern under the supervision of Dr Ayse Bilgin, Department of Statistics and Dr Timothy Kyng, Department of Applied Finance and Actuarial Studies, Macquarie University. It will form the thesis component for a Master of Research degree.

The purpose of this study is to investigate the implications of "big data" for the statistics profession, statistics education and society. For more information about the scope and purpose of the study please see the information form.

Big data analysis is a new multidisciplinary area. People working in the field include graduates from mathematics, statistics, computer science, actuarial studies, marketing and others.

Question 1: What gender are you? (1) Female (2) Male

Question 2: What is your year of birth?

Question 3: In which country were you born?

1) Australia (2) Other (please specify)

Question 4:

a. Do you speak language(s) other than English at home?
   (1) Yes (2) No

b. If yes, which language(s) do you speak? _____

Question 5: Which university do you work for?

Question 6: In which faculty do you work?

Question 7: In which department do you work?

(1) Computer Science (2) Statistics (3) Mathematics
(4) Actuarial Science (5) Marketing (6) Other, please specify:

Question 8: How many students graduate with degrees in disciplines taught by your department each year (approximately)?

Question 9: How long (years) have you worked for your current university?

Question 10: How long (years) have you worked in big data area?

Question 11: Are there any degree programs in your department involved with any of the following areas: big data / data analytics / data science / machine learning / statistical learning / data mining / statistical analysis?

(1) Yes, please specify: (2) No

Question 12:

a. Are there any subjects (also called "units") offered by your department that cover the above areas broadly included in "big data"?

(1) Yes, please specify: (2) No

b. If yes, in which area is that subject involved with big data?

(1) Statistical Analysis (2) Statistical Learning (3) Mathematics

(4) Data mining (5) Machine learning (6) Programming

(7) Marketing (8) Business Analysis

(9) Business Development (10) Other (please specify)

c. If yes, how many students enroll in that subject/unit each year (approximately)?

Question 13: How important is big data (data analytics) to your work, your department does? Please give a rating:

(1) Very Unimportant

(2) Somewhat Unimportant

(3) Neither Important nor Unimportant

(4) Somewhat Important

(5) Very Important

Question 14: Why or why not is big data important to your department? Please briefly explain.

Question 15: In your opinion, which type of expertise is required for graduate students to be employed in the big data field?

Please indicate your level of agreement with the following statements:

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree	N/A	**Question 16:** Do students acquire expertise in this area?
Statistical Analysis and Statistical Software Skills	(1)	(2)	(3)	(4)	(5)	(6)	(1) Yes (2) No
Statistical Learning	(1)	(2)	(3)	(4)	(5)	(6)	(1) Yes (2) No
Mathematics	(1)	(2)	(3)	(4)	(5)	(6)	(1) Yes (2) No
Data mining	(1)	(2)	(3)	(4)	(5)	(6)	(1) Yes (2) No
Machine learning	(1)	(2)	(3)	(4)	(5)	(6)	(1) Yes (2) No
Artificial Intelligence	(1)	(2)	(3)	(4)	(5)	(6)	(1) Yes (2) No
Programming	(1)	(2)	(3)	(4)	(5)	(6)	(1) Yes (2) No
Marketing	(1)	(2)	(3)	(4)	(5)	(6)	(1) Yes (2) No
Business Analysis	(1)	(2)	(3)	(4)	(5)	(6)	(1) Yes (2) No
Accounting	(1)	(2)	(3)	(4)	(5)	(6)	(1) Yes (2) No
Other (please specify)	(1)	(2)	(3)	(4)	(5)	(6)	(1) Yes (2) No

Question 17: In your opinion, which software tools are necessary for big data analysis? Please indicate your rating of their importance.

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree	N/A	**Question 18:** Which software tools are used in teaching big data analysis?
Base SAS	(1)	(2)	(3)	(4)	(5)	(6)	(1)
SAS Enterprise Miner	(1)	(2)	(3)	(4)	(5)	(6)	(2)
SPSS Analytics	(1)	(2)	(3)	(4)	(5)	(6)	(3)
SPSS Modeler	(1)	(2)	(3)	(4)	(5)	(6)	(4)
WinBUGS	(1)	(2)	(3)	(4)	(5)	(6)	(5)
Matlab	(1)	(2)	(3)	(4)	(5)	(6)	(6)
R	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Java	(1)	(2)	(3)	(4)	(5)	(6)	(8)
VBA	(1)	(2)	(3)	(4)	(5)	(6)	(9)
Hadoop	(1)	(2)	(3)	(4)	(5)	(6)	(10)
MapReduce	(1)	(2)	(3)	(4)	(5)	(6)	(11)
noSQL	(1)	(2)	(3)	(4)	(5)	(6)	(12)
SQL	(1)	(2)	(3)	(4)	(5)	(6)	(13)
JaQL	(1)	(2)	(3)	(4)	(5)	(6)	(14)
Hive	(1)	(2)	(3)	(4)	(5)	(6)	(15)
Oracle	(1)	(2)	(3)	(4)	(5)	(6)	(16)
Python	(1)	(2)	(3)	(4)	(5)	(6)	(17)
Other (please specify)	(1)	(2)	(3)	(4)	(5)	(6)	(18)

Question 19: In your opinion, how could your courses be improved in relation to big data?

Question 20: What content should be covered?

Question 21: Are you willing to be interviewed as part of this research?

(1) Yes, please provide your contact detail (phone, email address): (2) No

The results will be made available on Macquarie University library website.

If you are interested to get this web address, please provide your email address:

Thank you for your time to complete this survey.

D.2 Graduate Survey

Graduate Survey

Research Project: Big data and its implications for the statistics profession and statistics education

The study is being conducted by Miss Busayasachee Puang-Ngern under the supervision of Dr Ayse Bilgin, Department of Statistics and Dr Timothy Kyng, Department of Applied Finance and Actuarial Studies, Macquarie University. It will form the thesis component for a Master of Research degree.

The purpose of this study is to investigate the implications of "big data" for the statistics profession, statistics education and society. For more information about the scope and purpose of the study please see the information form.

Big data analysis is a new multidisciplinary area. People working in the field include graduates from mathematics, statistics, computer science, actuarial studies, marketing and others.

Question 1: What gender are you? (1) Female (2) Male

Question 2: What is your year of birth?

Question 3: In which country were you born?

1) Australia (2) Other (please specify)

Question 4:

c. Do you speak language(s) other than English at home?
   (1) Yes (2) No

d. If yes, which language(s) do you speak? _____

Question 5:

a. Which University did you attend to achieve your highest qualification?

b. If it is not in Australia, please tell us which country:_____

Question 6: What is your highest qualification?

(1) No university degree (2) Bachelor (3) Master (4) PhD

(5) Other (please specify)

Question 7: In which year did you complete your highest qualification?

Question 8: In which discipline was your highest qualification?

(1) Computer Science (2) Statistics (3) Mathematics

(4) Actuarial Science (5) Marketing (6) Other, please specify:

Question 9: Which company do you work for?

Question 10: How many graduate students does your company employ each year (approximately)?

Question 11: In which disciplines these new graduates come from?

(1) Computer Science (2) Statistics (3) Mathematics

(4) Actuarial Science (5) Marketing (6) Other, please specify:

Question 12: In which area does your company operate?

(1) Data analytics (2) Insurance (3) Banking

(4) Retail (5) Business Consulting (6) Internet services

(7) Utilities (i.e. gas, electricity) (8) Other (please specify)

Question 13: How long (years) have you worked for your current employer?

Question 14: How long (years) have you worked in big data area?

Question 15: For what purposes does your organisation use big data?

(1) Business Intelligence

(2) Understanding and modelling customer behaviour

(3) Targeted-marketing of products to the customers

(4) Understanding and Optimizing Business Processes

(5) Improving Security and Law Enforcement

(6) Product design/development

(7) Detect fraudulent transactions

(8) Conducting risk assessments for credit and insurance,

(9) Other (please specify)

Question 16: How important is big data (data analytics) to your work, your company or organisation does?

Please give a rating:

(1) Very Unimportant

(2) Somewhat Unimportant

(3) Neither Important nor Unimportant

(4) Somewhat Important

(5) Very Important

Question 17: Why or why not is big data important to your work, your company or organisation? Please briefly explain.

Question 18: In your opinion, which type of expertise is required to work with big data?

Please indicate your level of agreement with the following statements:

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree	N/A	**Question 19:** Do you have this expertise yourself?
Statistical Analysis and Statistical Software Skills	(1)	(2)	(3)	(4)	(5)	(6)	(1) Yes (2) No
Statistical Learning	(1)	(2)	(3)	(4)	(5)	(6)	(1) Yes (2) No
Mathematics	(1)	(2)	(3)	(4)	(5)	(6)	(1) Yes (2) No
Data mining	(1)	(2)	(3)	(4)	(5)	(6)	(1) Yes (2) No
Machine learning	(1)	(2)	(3)	(4)	(5)	(6)	(1) Yes (2) No
Artificial Intelligence	(1)	(2)	(3)	(4)	(5)	(6)	(1) Yes (2) No
Programming	(1)	(2)	(3)	(4)	(5)	(6)	(1) Yes (2) No
Marketing	(1)	(2)	(3)	(4)	(5)	(6)	(1) Yes (2) No
Business Analysis	(1)	(2)	(3)	(4)	(5)	(6)	(1) Yes (2) No
Accounting	(1)	(2)	(3)	(4)	(5)	(6)	(1) Yes (2) No
Other (please specify)	(1)	(2)	(3)	(4)	(5)	(6)	(1) Yes (2) No

Question 20: In your opinion, which software tools are necessary for big data analysis?

Please indicate your rating of their importance and whether your organisation uses them.

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree	N/A	**Question 21:** Which software tools are used in your company (organization) for big data analysis?
Base SAS	(1)	(2)	(3)	(4)	(5)	(6)	(1)
SAS Enterprise Miner	(1)	(2)	(3)	(4)	(5)	(6)	(2)
SPSS Analytics	(1)	(2)	(3)	(4)	(5)	(6)	(3)
SPSS Modeler	(1)	(2)	(3)	(4)	(5)	(6)	(4)
WinBUGS	(1)	(2)	(3)	(4)	(5)	(6)	(5)
Matlab	(1)	(2)	(3)	(4)	(5)	(6)	(6)
R	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Java	(1)	(2)	(3)	(4)	(5)	(6)	(8)
VBA	(1)	(2)	(3)	(4)	(5)	(6)	(9)
Hadoop	(1)	(2)	(3)	(4)	(5)	(6)	(10)
MapReduce	(1)	(2)	(3)	(4)	(5)	(6)	(11)
noSQL	(1)	(2)	(3)	(4)	(5)	(6)	(12)
SQL	(1)	(2)	(3)	(4)	(5)	(6)	(13)
JaQL	(1)	(2)	(3)	(4)	(5)	(6)	(14)
Hive	(1)	(2)	(3)	(4)	(5)	(6)	(15)
Oracle	(1)	(2)	(3)	(4)	(5)	(6)	(16)
Python	(1)	(2)	(3)	(4)	(5)	(6)	(17)
Other (please specify)	(1)	(2)	(3)	(4)	(5)	(6)	(18)

Question 22: In your opinion, which university disciplines should be including data analysis in their degree programs?

(1) Computer Science (2) Statistics (3) Mathematics

(4) Actuarial Science (5) Marketing (6) Data Science

(7) Others, please specify:

Question 23: Are you willing to be interviewed as part of this research?

(1) Yes, please provide your contact detail (phone, email address): (2) No

The results will be made available on Macquarie University library website.

If you are interested to get this web address, please provide your email address:

Thank you for your time to complete this survey.

E. Invitation Email

E.1 Academic Survey

Dear Academic staff,

You are invited to participate in a study of **"Big Data and its implications for the statistics profession and statistics education"**. The purpose of the study is to understand the needs of Industry and of Society regarding “big data” and its analysis. With the advent of the Internet, Social Media and the increasing need for the storage of and analysis of large datasets has come a range of new methods and techniques for the handling and analysis of such data. New disciplines have emerged on the boundary between mathematics, statistics and computer science. These include “data science”, “machine learning”, “statistical learning”, “data mining” and so on.

Increasingly graduates are finding employment doing “data analytics”. Graduates from a diverse range of disciplines are working in these areas. This includes graduates in computer science, statistics, mathematics, marketing and actuarial science. Many different types of corporations and government organisations are interested in collecting, analysing and being able to retrieve data from large data sets. This includes financial institutions, retailers, the police, the Roads and Traffic Authority and many others. Much of the effort is directed towards the modelling of consumer behaviour and the targeting of marketing campaigns.

This project is designed to investigate the needs of government, industry and society generally for appropriately trained analysts, and the extent to which the universities are providing the education needed in this area. The project is also designed specifically to investigate the extent to which statistical education has changed and may need to change in the future in response to this.

The study is being conducted by Miss Busayasachee Puang-Ngern (busayasachee.puang-ngern@students.mq.edu.au) who is completing Master of Research Project in the Department of Statistics at Macquarie University, Australia under the supervision of Dr Ayse Bilgin and Dr Timothy Kyng (contact details: phone (61-2) 9850 8509 or (61-2) 9850 7289 and email ayse.bilgin@mq.edu.au or timothy.kyng@mq.edu.au)

We would appreciate if you could complete this survey and also pass this survey link (https://mqedu.qualtrics.com/SE/?SID=SV_3yJclVNjvy6CZRX) on to any of your academic colleagues who may be willing to complete the survey.

E.2 Graduate Survey

Dear Graduate,

You are invited to participate in a study of **"Big Data and its implications for the statistics profession and statistics education"**. The purpose of the study is to understand the needs of Industry and of Society regarding "big data" and its analysis. With the advent of the Internet, Social Media and the increasing need for the storage of and analysis of large datasets has come a range of new methods and techniques for the handling and analysis of such data. New disciplines have emerged on the boundary between mathematics, statistics and computer science. These include "data science", "machine learning", "statistical learning", "data mining" and so on.

Increasingly graduates are finding employment doing "data analytics". Graduates from a diverse range of disciplines are working in these areas. This includes graduates in computer science, statistics, mathematics, marketing and actuarial science. Many different types of corporations and government organisations are interested in collecting, analysing and being able to retrieve data from large data sets. This includes financial institutions, retailers, the police, the Roads and Traffic Authority and many others. Much of the effort is directed towards the modelling of consumer behaviour and the targeting of marketing campaigns.

This project is designed to investigate the needs of government, industry and society generally for appropriately trained analysts, and the extent to which the universities are providing the education needed in this area. The project is also designed specifically to investigate the extent to which statistical education has changed and may need to change in the future in response to this.

The study is being conducted by Miss Busayasachee Puang-Ngern (busayasachee.puang-ngern@students.mq.edu.au) who is completing Master of Research Project in the Department of Statistics at Macquarie University under the supervision of Dr Ayse Bilgin and Dr Timothy Kyng (contact details: phone (61-2) 9850 8509 or (61-2) 9850 7289 and email ayse.bilgin@mq.edu.au or timothy.kyng@mq.edu.au)

We would appreciate if you could complete this survey and also pass this survey link (https://mqedu.qualtrics.com/SE/?SID=SV_5apM3YWkuI8mlgN) on to any recent graduates in the data analytics workforce who may be willing to complete the survey.

F. Participant Information and Consent Form

Name of Project: Big Data and its implications for the statistics profession and statistics education

You are invited to participate in a study of Big Data and its implications for the statistics profession and statistics education. The purpose of the study is to understand the needs of Industry and of Society regarding "big data" and its analysis. With the advent of the Internet, Social Media and the increasing need for the storage of and analysis of large datasets has come a range of new methods and techniques for the handling and analysis of such data. New disciplines have emerged on the boundary between mathematics, statistics and computer science. These include "data science", "machine learning", "statistical learning", "data mining" and so on.

Increasingly graduates are finding employment doing "data analytics". Graduates from a diverse range of disciplines are working in these areas. This includes graduates in computer science, statistics, mathematics, marketing and actuarial science. Many different types of corporations and government organisations are interested in collecting, analysing and being able to retrieve data from large data sets. This includes financial institutions, retailers, the police, the Roads and Traffic Authority and many others. Much of the effort is directed towards the modelling of consumer behaviour and the targeting of marketing campaigns.

This project is designed to investigate the needs of government, industry and society generally for appropriately trained analysts, and the extent to which the universities are providing the education needed in this area. The project is also designed specifically to investigate the extent to which statistical education has changed and may need to change in the future in response to this.

The study is being conducted by Miss Busayasachee Puang-Ngern who is completing Master of Research Project in the Department of Statistics under the supervision of Dr Ayse Bilgin and Dr Timothy Kyng (contact details: phone (61-2) 9850 8509 or (61-2) 9850 7289 and email ayse.bilgin@mq.edu.au or timothy.kyng@mq.edu.au)

If you decide to participate, you will be asked to complete a brief online questionnaire. The questionnaire will ask for some demographic information about you and a series of questions about IT projects and their management.

Any information or personal details gathered in the course of the study are confidential, except as required by law. No individual will be identified in any publication of the results. We will only publish aggregated results. Only the researchers will have access to the data. The results will be made available on Macquarie University library website.

Participation in this study is entirely voluntary: you are not obliged to participate and if you decide to participate, you are free to withdraw at any time without having to give a reason and without consequence.

The ethical aspects of this study have been approved by the Macquarie University Human Research Ethics Committee. If you have any complaints or reservations about any ethical aspect of your participation in this research, you may contact the Committee through the Director, Research Ethics (telephone (02) 9850 7854; email ethics@mq.edu.au). Any complaint you make will be treated in confidence and investigated, and you will be informed of the outcome.

G. The target university departments for the academic survey

G.1 Australian University

	Australian University	Number of responded
1	**Curtin University**	
	Curtin Business School, School of Economics & Finance	
	Curtin Business School, School of Information Systems	1
	Curtin Business School, School of Marketing	
	Faculty of Science & Engineering, School of Electrical Eng & Computing, Department of Computing	
	Faculty of Science & Engineering, School of Science, Department of Mathematics & Statistics	
2	**Edith Cowan University**	
	Faculty of Health, Engineering and Science, School of Computer and Security Science	
	Faculty of Business and Law, School of Business	2
3	**Murdoch University**	
	School of Engineering and Information Technology, Discipline of Information Technology	
	School of Engineering and Information Technology, Discipline of Mathematics & Statistics	
	School of Management and Governance	1
4	**University of Notre Dame**	
	School of Arts and Sciences	1
5	**University of Western Australia**	
	Business School, Discipline of Marketing	2
	Faculty of Engineering, Computing and Mathematics, School of Computer Science and Software Engineering	
	Faculty of Engineering, Computing and Mathematics, School of Mathematics and Statistics	
6	**Charles Darwin University**	
	Faculty of Engineering, Health, Science and the Environment, School of Engineering and Information Technology, Discipline of Information Technology and Mathematics	
	Faculty of Law, Education, Business and Arts, School of Business, Discipline of Management and Marketing	
7	**Flinders University**	
	Faculty of Science and Engineering, School of Computer Science, Engineering and Mathematics, Discipline of Information Technology, Software Engineering, Mathematics and Statistics	
	Faculty of Social and Behavioural Sciences, Flinders Business School	
8	**University of Adelaide**	
	Faculty of Engineering, Computer & Mathematical Sciences, School of Computer Science	
	Faculty of Engineering, Computer & Mathematical Sciences, School of Mathematical Sciences	1
	Faculty of the Professions, Business School, Discipline of Marketing	
9	**University of South Australia**	
	Business School, School of Marketing	1
	Division of Information Technology, Engineering and the Environment, School of Information Technology & Mathematical Sciences	1
10	**Australian Catholic University**	
	Faculty of Law and Business, School of Business	
11	**Queensland University of Technology**	
	QUT Business School, School of Advertising, Marketing and Public Relations	
	Science and Engineering Faculty, Discipline of Applied and Computational Mathematics	
	Science and Engineering Faculty, Discipline of Statistical Science and Mathematical Sciences	

	Australian University	Number of responded
12	**Bond University**	
	Faculty of Business, Discipline of Actuarial Science	
	Faculty of Business, Discipline of Economics and Statistics	
	Faculty of Business, Discipline of Marketing	
13	**Central Queensland University**	1
	School of Business and Law	
	School of Engineering and Technology	
14	**Griffith University**	
	Griffith Business School, Department of Marketing	
	Griffith Sciences, School of Information and Communication Technology	
15	**James Cook University**	
	Faculty of Law, Business & Creative Arts, School of Business, Discipline of Information Technology	1
	Faculty of Law, Business & Creative Arts, School of Business, Discipline of Marketing and Management	
	Faculty of Science and Engineering, School of Engineering and Physical Sciences, Discipline of Mathematics	
16	**Southern Cross University**	
	Southern Cross Business School	
17	**University of Queensland**	
	Faculty of Business, Economics and Law, School of Business, Discipline of Marketing	
	Faculty of Engineering, Architecture and Information Technology, School of Information Technology and Electrical Engineering	2
	Faculty of Science, School of Mathematics and Physics, Discipline of Mathematics and Statistics	
18	**University of Southern Queensland**	
	Faculty of Business, Education, Law and Arts, School of Management and Enterprise, Discipline of Information Systems	
	Faculty of Business, Education, Law and Arts, School of Management and Enterprise, Discipline of Marketing	
	Faculty of Health, Engineering and Sciences, School of Agricultural, Computational and Environmental Science, Discipline of Computer Science	
	Faculty of Health, Engineering and Sciences, School of Agricultural, Computational and Environmental Science, Discipline of Computing	
	Faculty of Health, Engineering and Sciences, School of Agricultural, Computational and Environmental Science, Discipline of Mathematics	
	Faculty of Health, Engineering and Sciences, School of Agricultural, Computational and Environmental Science, Discipline of Statistics	
19	**University of the Sunshine Coast**	
	Faculty of Arts and Business, School of Business, Discipline of Information Systems	
	Faculty of Arts and Business, School of Business, Discipline of Marketing	
20	**Charles Sturt University**	
	Faculty of Business, School of Computing & Mathematics	1
	Faculty of Business, School of Management and Marketing	
21	**Macquarie University**	
	Faculty of Business and Economics, Department of Applied Finance and Actuarial Studies	1
	Faculty of Business and Economics, Department of Marketing and Management	3
	Faculty of Science, Department of Computing	3
	Faculty of Science, Department of Mathematics	
	Faculty of Science, Department of Statistics	

	Australian University	Number of responded
22	**University of New England**	
	Graduate School of Business (UNE Business School), Discipline of Management	
	School of Science and Technology, Discipline of Computer Science	
	School of Science and Technology, Discipline of Mathematics	
	School of Science and Technology, Discipline of Statistics	
23	**University of New South Wales**	
	Australian School of Business (ASB), School of Information Systems, Technology and Management	
	Australian School of Business (ASB), School of Marketing	
	Australian School of Business (ASB), School of Risk & Actuarial Studies	
	Faculty of Engineering, School of Computer Science and Engineering	2
	Faculty of Science, School of Mathematics and Statistics	3
24	**University of Newcastle**	
	Faculty of Business and Law, Newcastle Business School, Discipline of Marketing	
	Faculty of Engineering and Built Environment, School of Electrical Engineering and Computer Science	
	Faculty of Science and Information Technology, School of Mathematical and Physical Sciences	
25	**University of Sydney**	
	Business School, Discipline of Marketing	1
	Faculty of Engineering and Information Technologies, School of Information Technologies	
	Faculty of Science, School of Mathematics and Statistics	
	Faculty of Health Sciences, Sydney Centre for Aboriginal and Torres Strait Islander Statistics	1
26	**University of Technology Sydney**	
	UTS Business School, Discipline of Marketing	
	Faculty of Engineering and Information Technology	1
	Faculty of Science, School of Mathematical Sciences	1
27	**University of Western Sydney**	
	School of Business, Discipline of Marketing and International Business	
	School of Computing, Engineering and Mathematics	
28	**University of Wollongong**	
	School of Management, Operations and Marketing, Discipline of Marketing	1
	Faculty of Engineering and Information Sciences, School of Computer Science and Software Engineering (SCSSE)	
	Faculty of Engineering and Information Sciences, School of Electrical, Computer and Telecommunications Engineering (SECTE)	1
	Faculty of Engineering and Information Sciences, School of Information Systems and Technology (SISAT)	
	Faculty of Engineering and Information Sciences, School of Mathematics and Applied Statistics (SMAS)	
29	**University of Canberra**	
	Faculty of Business, Government & Law, School of Information Systems & Accounting	
	Faculty of Business, Government & Law, School of Management	
	Faculty of Education, Science, Technology & Mathematics, Discipline of Information Technology & Engineering	
	Faculty of Education, Science, Technology & Mathematics, Discipline of Mathematics & Statistics	1
30	**Australian National University**	
	ANU College of Business & Economics, Discipline of Actuarial Studies and Statistics	1
	ANU College of Business & Economics, Discipline of Business Information Systems	
	ANU College of Business & Economics, Discipline of Marketing	1
	ANU College of Engineering & Computer Science, Research School of Computer Science	2
	ANU College of Physical & Mathematical Sciences, Mathematical Sciences Institute	2

	Australian University	Number of responded
31	**Deakin University**	
	Faculty of Business and Law, School of Management and Marketing	
	Faculty of Business and Law, School of Information and Business Analytics	2
	Faculty of Science, Engineering and Built Environment, School of Information Technology	1
32	**Monash University**	
	Faculty of Business and Economics, Department of Marketing	
	Faculty of Business and Economics, Department of Econometrics and Business Statistics	1
	Faculty of Engineering, Department of Electrical and Computer Systems Engineering	
	Faculty of Information Technology	3
	Faculty of Science, School of Mathematical Sciences	1
33	**RMIT University**	
	College of Business, School of Business IT and Logistics	2
	College of Business, School of Economics, Finance and Marketing	1
	College of Science, Engineering and Health, School of Computer Science and IT	1
	College of Science, Engineering and Health, School of Mathematical and Geospatial Sciences	1
34	**Swinburne University of Technology**	
	Faculty of Business and Enterprise, Swinburne Business School, Department of Information Systems and Logistics	1
	Faculty of Business and Enterprise, Swinburne Business School, Department of Marketing, Tourism and Social Impact	1
	Faculty of Science, Engineering and Technology, School of Science, Centre for Astrophysics and Supercomputing	2
	Faculty of Science, Engineering and Technology, School of Science, Department of Mathematics	1
	Faculty of Science, Engineering and Technology, School of Software and Electrical Engineering, Centre for Advanced Internet Architectures	
	Faculty of Science, Engineering and Technology, School of Software and Electrical Engineering, Centre for Computing and Engineering Software Systems	
	Faculty of Science, Engineering and Technology, School of Software and Electrical Engineering, Department of Computer Science and Software Engineering	
35	**Federation University**	
	Faculty of Business, School of Business and Economics (Gippsland)	
	Faculty of Business, The Business School (Ballarat), Discipline of Marketing	
	Faculty of Science, School of Information Technology (Gippsland)	
	Faculty of Science, School of Science, Information Technology and Engineering (Ballarat)	2
	Research & Innovation, Research Centres, CeRDI	1
36	**University of Melbourne**	
	Faculty of Business and Economics, Department of Management and Marketing	1
	Faculty of Engineering, Melbourne School of Engineering, Department of Computing and Information Systems	1
	Faculty of Science, Department of Mathematics and Statistics	3
37	**La Trobe University**	
	Faculty of Business, Economics and Law, Discipline of Management Information Systems	
	Faculty of Business, Economics and Law, La Trobe Business School, Department of Marketing and Tourism and Hospitality	1
	Faculty of Science, Technology and Engineering, School of Engineering and Mathematical Sciences, Discipline of Computer Science, Computer Systems Engineering, Information Technology and Information Systems	
	Faculty of Science, Technology and Engineering, School of Engineering and Mathematical Sciences, Discipline of Mathematics and Statistics	1

	Australian University	Number of responded
38	**Victoria University**	
	College of Business	1
	College of Engineering & Science	1
39	**University of Tasmania**	
	Tasmanian School of Business & Economics	2
	Department of Computing and Information Systems	
	School of Mathematics and Physics	2
	Grand Total	**74**

G.2 New Zealand University

	New Zealand University	Number of responded
1	**Auckland University of Technology**	
	AUT Business School, Department of Business Information Systems	
	AUT Business School, Department of Marketing, Advertising, Retailing and Sales	
	Faculty of Design and creative technologies, School of Computer and Mathematical Sciences	1
2	**Lincoln University**	
	Faculty of Commerce, Department of Business Management, Law and Marketing, Discipline of Marketing	
	Faculty of Environment, Society and Design, Department of Applied Computing	
3	**Massey University**	
	College of Business, School of Communication, Journalism and Marketing	
	College of Sciences, School of Engineering and Advanced Technology, Discipline of Computer Science and Information Technology	
	College of Sciences, School of Engineering and Advanced Technology, Discipline of Electronics, Information & Communication Engineering	
4	**University of Auckland**	
	Business School, Department of Information Systems and Operations Management	
	Business School, Department of Marketing	
	Faculty of Science, Department of Computer Science	
	Faculty of Science, Department of Mathematics	
	Faculty of Science, Department of Statistics	1
5	**University of Canterbury**	
	College of Business and Law, School of Business and Economics, Department of Accounting and Information Systems	
	College of Business and Law, School of Business and Economics, Department of Management, Marketing and Entrepreneurship, Discipline of Marketing	
	College of Engineering, Department of Computer Science and Software Engineering	1
	College of Engineering, School of Mathematics and Statistics	
6	**University of Otago**	
	Otago Business School, Department of Information Science	
	Otago Business School, Department of Marketing	1
	Division of Sciences, Department of Computer Science	
	Division of Sciences, Department of Mathematics & Statistics	

	New Zealand University	Number of responded
7	**University of Waikato**	
	Waikato Management School, Discipline of Marketing	
	Faculty of Computing and Mathematical Sciences, Discipline of Computer Science	
	Faculty of Computing and Mathematical Sciences, Discipline of Computing and Mathematical Sciences	
	Faculty of Computing and Mathematical Sciences, Discipline of Mathematics	
	Faculty of Computing and Mathematical Sciences, Discipline of Statistics	
8	**Victoria University of Wellington**	
	Faculty of Commerce (Victoria Business School), School of Information Management	1
	Faculty of Commerce (Victoria Business School), School of Marketing and International Business	
	Faculty of Engineering, School of Engineering and Computer Science	
	Faculty of Science, School of Mathematics, Statistics and Operations Research, Discipline of Mathematics	
	Faculty of Science, School of Mathematics, Statistics and Operations Research, Discipline of Statistics and Operations Research	
	Grand Total	**5**

G.3 Other country University

	Other country University	Number of responded
1	**Aberystwyth University (United Kingdom)**	
	Institute of Mathematics, Physics and Computer Science, Department of Computer Science	1
2	**The University of Sheffield (United Kingdom)**	
	School of Mathematics and Statistics	1
3	**Klagenfurt University (Austria)**	
	Faculty of Technical Sciences, Institute of Statistics	1
4	**McMaster University (Canada)**	
	Faculty of Engineering, Department of Computing and Software	1
5	**University of California, Los Angeles (United States)**	
	Division of Physical Sciences, Department of Statistics	1
	Grand Total	**5**

H. The participants directly approached[1] for the graduate survey

	University	Course Title / Unit Title	Enrolment Year	Number of target participants
1	**Macquarie University**			
	Faculty of Science, Department of Statistics	Master of Applied Statistics	2014	93
	Faculty of Science, Department of Statistics	Postgraduate Certificate of Applied Statistics	2014	3
	Faculty of Science, Department of Statistics	Postgraduate Diploma of Applied Statistics	2014	5
	Faculty of Business and Economics, Department of Applied Finance and Actuarial Studies	Master of Actuarial Practice	2012 - 2013	15
	Faculty of Science, Department of Statistics	Data Mining	2014	29
	Faculty of Business and Economics, Department of Applied Finance and Actuarial Studies	Investment and Asset Modelling	2014	21
	Faculty of Business and Economics, Department of Applied Finance and Actuarial Studies	Investment Management	2014	106
	Faculty of Business and Economics, Department of Applied Finance and Actuarial Studies	Actuarial Control Cycle 2	2014	119
2	**University of Western Sydney**			
	School of Computing, Engineering and Mathematics	Bachelor of Mathematics and Information Technology	2003 - 2011	51
3	**Chulalongkorn University (THAILAND)**			
	Faculty of Commerce and Accountancy	Bachelor of Science in Statistics (Major in Applied Statistics, Mathematical Statistics, Insurance, and Information Technology for Business)	2008	25
4	**Other University**	Graduate working in industry partner		30
	Grand Total			**497**

[1] We did snowball sampling, so we expected these people to forward the invitation to others.

I. Sample of Data scientist's job advertisements

I.1 Data Analytics Manager, Liverpool Victoria

Posted on August 7, 2014 by evertsemeijn

LV= is the UK's largest friendly society and today, we have more than 5 million customers, of which over 1.1 million are members. We're famous for providing a first-class service to our customers and our mission is to help people look after what they love in life. We believe LV= is a great place to work and with an impressive workforce of over 5,700 employees across the UK. We're looking to the future with confidence and enthusiasm, as we continue to attract more and more customers and enjoy further success and growth.

A new team has been set up within the LV= Life & Pensions Business to act as centre for excellence in data sciences, provide thought leadership and lead best in class modelling and analytics, to add value to our existing data and analytics capability and deliver real commercial benefit in support of the Life Business performance agenda. A key responsibility for the new team is to identify opportunities and insight, which can be actioned by the business to drive financial benefits, and highlight areas of competitive advantage to LV=. The team will achieve this by building strong relationships with key stakeholders in both the Life Business and Group Functions, and optimising the use of internal and external data, including public and big data sources, through strategic collaborations with key industrial and academic partners.

The Data Scientist role within this new team will be an experienced applied mathematician/statistician with hands on experience and a track record of delivering business relevant and impactful analytics projects. The job holder will play a key role in:

- Providing thought leadership for all Life Business analytics ensuring that claims, longevity, retention, customer experience and marketing analysis use appropriate techniques to delivers real commercial benefit in support of the Life Business performance agenda.
- Building strong relationships with key stakeholders in the Pricing & Commercial team, the Protection and Retirements Business Units, Marketing, Underwriting and Claims, Life Finance and Group Finance to help drive appropriate decision making.
- Leading the development of ad-hoc and strategic analytics projects and optimising the use of various data available within the Life Business and external "big data" sources to highlight possible commercial opportunities.

About You

You will have:

- Exceptional statistical, mathematical modelling and data manipulation skills

- Modelling experience using statistical and mathematical tools such as SPSS, SQL, SAS or Matlab
- Experience of advanced statistical techniques, such as GLM and Bayesian networks
- Solid understanding of standard statistical modeling techniques (e.g. regression; multivariate, cluster analysis)
- Statistical or Actuarial qualification – desirable but not essential
- Life Insurance background and expertise – desirable but not essential
- Expertise in the development, application and implementation of predictive customer behaviour models e.g. propensity to buy and retention models – highly desirable
- Strong analytical and reporting skills
- Ability to work independently and as part of a team, pro-actively, multi-task and deliver
- Track record of commercial focus – keen focus on financial benefit of work
- Evidence of ability to form strong relationships with internal/external stakeholders
- Good written and verbal communication skills combined with strong interpersonal skills
- Proven analytical skills with the ability to be pragmatic

The Details

Want to know the details? Here's more about what you would be doing:

- drive and lead the delivery of analytics projects of emerging Life SBU experience, including mortality, morbidity, longevity and persistency.
- develop, maintain and enhance mathematical and statistical models and algorithms to best in practice standard using appropriate analysis and modelling techniques and software (e.g. Generalised Linear Modelling, Bayesian Networks, Machine Learning algorithms, etc).
- formally document the results of the annual experience analysis, recommending changes to the pricing basis as a result of emerging experience as part of the formal Annual Pricing Basis Review.
- deliver insightful regular and ad-hoc analytics reports to the Protection and Retirement Solutions business units ensuring any trends in experience any the implications are understood to improve the company's financial and competitive position, providing recommendations as appropriate.
- provide thought leadership and senior subject matter expertise and experience to support the marketing analysis team
- to proactively pursue the benefits available to the Life SBU through "big data", finding ways to drive commercial value from the MI and data available within LV= and outside the company (e.g. opportunities for new product development or areas to target growth such as digital) championing these through to implementation.

- build strong relationships with Pricing & Commercial, Financial Planning & Analysis, Marketing, Protection and Retirement business units, the MI team, Claims & Underwriting and Group Finance.
- work with IT and Life MI so as to optimise the MI available and the systems and processes in place to support its analysis.
- manage, coach and develop junior team members if appropriate and deputise for the Head of Analytics as/when required.
- ensure an attitude of continuous personal development is adopted for both self and any reports. Keep up to date with the latest developments in the data sciences through e.g. conference attendance and membership of relevant professional associations and industry events/committees.

About the Rewards

We want you to love what you do. That's why we've put together a benefits package that recognises and rewards a job well done. We'll give you a competitive pension, an annual bonus scheme, 30 days' holiday, private medical insurance through BUPA, and a flexible benefits package. There's also the option of 25% off our general insurance products, including home and car and up to 12% off our life products.

I.2 Data Analyst, Science |American International Group

AIG formed the Science team at the beginning of 2012 after recognising the power of technology, data, and computational science to transform the insurance industry. The team consists of world class business minds and scientists and has been created to drive transformational change through evidence-based decision making at the company. The team has grown extremely rapidly with headcount now over 100.

The group is highly visible and fully supported by the leadership team of AIG and has a broad and global mandate ranging from solving complex business problems to partnering with leading academics on the development of next generation modelling techniques. The group's intent is to be a centre of innovation at the company and a catalyst for change.

The Science team currently has 13 members in EMEA and the healthy pipeline of work has resulted in the need to continue to expand the team.

We currently have a vacancy for a Data Analyst.

Position Summary:

The position of Data Analyst will require outstanding data management and manipulation skills. Primary responsibilities will be to develop and manage technical processes for transforming and

loading data into data models for use by Business Intelligence and reporting tools, and for statistical and predictive modelling. These processes must be robust and "production" ready. The role will require a considerable amount of collaboration and communication with many different parts of AIG, from IT to business data users.

Organisational Structure and Interface:

The Science team is a highly matrixed organisation. While all team members have a senior supervisor, the work is managed in project teams that form, and reform, dynamically as business needs evolve. The role will report to the Science data engineering lead.

Performance Objectives:

- Contribute to the design of data models and architecture for use by Business Intelligence and reporting tools
- Design and develop data transformation and loading processes for the population and ongoing management of the data models and architecture
- Develop detailed knowledge of data structures and meanings from data sources across AIG
- Perform QC and verification of data according to specified criteria. Take the initiative in managing data problems / exceptions
- Define and implement rules-based derived fields
- Thoroughly document the data transformation processes used
- Participate in cross-functional projects

Position Requirements

The Candidate Must Have:

- Experience and in-depth understanding of relational database modelling and structures
- Extensive experience working with very large data sets (preparing data and tables for presentation of analyses), independently and as part of a team.
- Experience in handling data quality and integrity issues
- Expert in SQL and database tools (e.g., Oracle, Sybase, DB2, Teradata).
- Expert knowledge of SAS (scripting in base SAS and knowledge of SAS macros)
- Experience with software development lifecycle including testing
- Experience with business intelligence data modelling techniques and tools
- Degree qualification in relevant discipline

The Ideal Candidate Would Also Have:

- Direct experience in preparing datasets for use in statistical and predictive modeling
- An interest in statistical and predictive modelling
- Skills in other coding languages such as R, Python, C++
- Experience in data warehousing, dimensional modelling, or NoSQL databases

- Exposure to Business Intelligence and data visualisation tools

About Us

American International Group, Inc. (AIG) is a leading international insurance organization serving customers in more than 130 countries and jurisdictions. AIG companies serve commercial, institutional, and individual customers through one of the most extensive worldwide property-casualty networks of any insurer. In addition, AIG companies are leading providers of life insurance and retirement services in the United States.

I.3 Data Scientist/ Statistical Analyst, PruHealth

PruHealth is an award winning, dynamic and vibrant insurance provider, with a ground-breaking vision for the future, where individuals are enabled to succeed and are rewarded and recognised for their contribution to our business.

Our CORE PURPOSE is to make people healthier and to enhance and protect their lives. From people to products and processes, we aspire to deliver on our purpose in everything we do.

Our VISION is to be the BEST insurer in the UK

We are looking for talented individuals who are committed to living our values and delivering an award winning service to our customers.

In this role you will be required to design and develop advanced algorithms and models.

You will utilize various techniques such as Linear & Logistic Regression, Decision Trees, Random Forests, Neural Networks, K-Nearest Neighbours, Support Vector Machines, Affinity Analysis etc.

You will discover trends, patterns and stories told by the data and present them to stakeholders. You will leverage state-of-the-art data mining tools and analytical methodologies to drive improved business decisions. You will produce creative data visualizations using intuitive graphics to represent complex analytics.

You will work on analytical projects across the organisation interacting with several departments such as Marketing, Digital, Sales, Risk, Pricing & Fraud.

You will have an undergraduate degree in a numerical subject, and ideally a Masters or other advanced degree in Statistics, Mathematics or a related subject.

You will have experience of accessing, compiling and analysing data using tools such as SQL, R, SAS, SPSS, etc. You will also have experience applying statistical methodologies and building predictive and other mathematical models. Previous PMI or other insurance experience would be an advantage, as would experience with visualisation tools and epidemiological experience.

You will be highly analytical with good communication skills and able to work on your own initiative.

Working for PruHealth, you'll experience an exciting mix of creativity and innovation, within a framework of challenging objectives and a passion for delivering the best.
We think work should be fun and sociable, and we want our people to get the most out of every day. Our people are chosen for their skills, knowledge, enthusiasm and attitude but above all, their belief that anything can be achieved.

I.4 Data Scientist, GE Aviation

Data Scientists will be responsible for the development and application of advanced data driven analytics enabling services targeted at improving airline operational and maintenance efficiency. This role will be in the new JV between GE Aviation and Accenture, Taleris.
The role is to participate in the development and application of data driven analytic technologies including feature extraction, anomaly modelling, reasoning, prediction, optimisation and virtual sensing. The successful candidate will join an expanding team of data scientists, whose aim is to combine these techniques with expert domain knowledge to extract value from large sets of airline data. The role involves the exploration, prototyping, evaluation and implementation of novel analysis techniques for aircraft flight data, and refining these based on in-service experience. It will also contribute to the development of real-world systems such as on-board and off-board aircraft diagnostics, prognostics and maintenance systems, and sophisticated web-based airline information systems.

- Bachelor of Science in Engineering, Physics, Chemistry, Mathematics/ Computer Science or other qualification with demonstrable IT, engineering, scientific/ technical experience
- Some previous experience in an engineering position
- Strong degree level education or PhD in science, computing, engineering or equivalent
- Experience of advanced data-analysis methodologies
- Experience of applying analytics to real world problems
- Algorithm design, prototyping and validation experience
- Experience of statistical and Artificial Intelligence techniques
- Relevant software experience, e.g. MATLAB, R, Python, C, Java, .NET
- Database experience, including SQL Server
- Ability to effectively provide direction and training to other engineers
- Interdisciplinary communication to forge connections with diverse domain specialists
- Strong oral and written communication skills
- Strong interpersonal and leadership skills
- Capability to work in teams

- PC proficiency

I.5 Data Scientist, Bank of England

Posted on August 12, 2014 by evertsemeijn

Advanced Analytics is a new division created as part of the implementation of the Bank's recent Strategic Plan. It lies at the heart of the Bank's research infrastructure and its role is to develop our ability to exploit the information contained in all forms of relevant data. It supports all areas of the organisation, including both the analytical and support functions.

A particular focus of the team is on developing techniques to extract information from relatively unstructured and granular data sets, often described as 'Big Data'. These data sources are becoming more useful to policy makers as the amount of information they contain increases exponentially and improvements in both software and analytical techniques make rapid strides. They offer the possibility of getting a richer understanding of behaviour at a highly-disaggregated level, allowing more detailed, precise and timely insights for policy makers into key phenomena.

Central banks have only recently begun to systematically use these data sets and analytical techniques, so a crucial part of the team's work is to partner with other areas of the Bank in order to raise our understanding of the available techniques and issues involved. We also want to make optimal use of the knowledge outside the Bank and will work with academics, other policy institutions and private-sector experts on specific projects.

The use of more varied data and improved software means that researchers are able to communicate their findings in more interesting and effective ways than simply relying on old-fashioned line and bar charts. The team will work with other parts of the Bank in ensuring that our communication tool kit is state-of-the-art.

Brief Description

We need analysts who will develop and use cutting-edge techniques to extract policy-relevant information from both proprietary and publically-available data sets and add to the Bank's pool of expertise on these matters.

Detailed Description

You will:

- Work with the Advanced Analytics (AA) team, other parts of the Bank and outside organisations to identify and analyse specific projects;
- Identify data sources which can shed light on the problem being addressed;
- If the data is embedded in some other medium (eg the internet) use appropriate software to put it into a usable format;

- Ensure we properly understand the data used in these projects including the links between different data sources;
- Use those links to reach a more complete picture of the underlying situation than any one data set would allow;
- Use the appropriate IT and statistical techniques to answer policy-relevant questions pertaining to the data;
- Document the techniques and data used to enable the analysis to be replicated, either directly or in other similar contexts;
- Communicate the findings and approach used to our stakeholders in a clear, non-technical but rigorous way;
- Visualise the analysis effectively.

Job Requirements

Essential

- Excellent quantitative and technical skills demonstrated by degree or experience;
- Programming experience (eg Python);
- Strong interest in data analysis and an enquiring mind;
- Interest in policy issues;
- Excellent communication skills with the ability to clearly explain complex technical issues;
- Ability to co-ordinate work involving multiple stakeholders;
- Good relationship management skills and the ability to work comfortably with Bank of England staff at all levels;
- Flexibility and resilience – able to work under pressure in an environment where priorities may change at short notice;
- Ability to reason laterally in order to identify solutions to seemingly intractable problems;
- Strong problem solving skills – able to analyse, identify root causes and resolve issues;
- Planning and organizing skills – able to prioritise and manage potentially conflicting demands through a focus on delivery of key objectives;
- Personal accountability – a proactive responsibility for work, taking action and making decisions when necessary;
- Good team player.

Desirable

Knowledge and experience in the following areas:

- An understanding of the Bank of England's work;
- Experience validating data and maintaining data quality standards;
- Both relational and non-relational database skills (eg SQL, NoSQL);

- Data mining and machine learning techniques and software;
- Statistical/econometric/mathematical packages (eg R, Matlab).

What will I get from the role?

- An opportunity to do cutting-edge work in a field that is developing quickly and going to become increasingly important;
- Experience of working at the heart of one of the key policy institutions in the UK, collaborating with all areas of the Bank of England;
- An insight into the links between analysis, data and policy;
- You will engage in intellectually stimulating and demanding work, within a high profile and highly regarded team. In doing so you will develop a wide range of contacts across the Bank of England and outside bodies (especially academia and the Government).

I.6 Data Scientist, Cap Gemini

Posted on August 8, 2014 by evertsemeijn

We're looking to recruit a Data Scientist to join our agile data science delivery team based in London.

You'd be working in a team of delivery focused engineers and analysts specialising in delivering leading edge analytics projects across public and private sector clients, with a focus on open source and NoSQL technologies.

Main Purpose of Role

The role largely involves:-

- Working as a senior member of the analytics delivery team on a day to day basis
- Coaching other team members in advanced analytical methods
- Engaging with our clients staff to develop use cases and user stories
- Building analytical processes and code to meet client requirements

Skills and Experience

The ideal candidate should be able to demonstrate most of the following skills:

- Data Systems

o Hadoop (Hive / Impala / Pig)

o Mahout

o PostgreSQL

o AWS (EC2 / EMR / S3)

o MongoDB

o Neo4J – Cypher

- Analytical Skills

o Data Munging

o Statistical Analysis

o Experimental design (sampling, confidence)

o Classifier and predictor development

o Time series analysis

o Pattern detection

o Natural Language Processing

o Sentiment Analysis

o Computer Vision

- Supplementary Skills

o Data Visualisation

o API Exploitation

o Source control using Git

o Continuous Integration (Jenkins)

o Unit Testing

o Virtual Machines (Vagrant, Virtualbox)

- Languages

o Python (+ SciPy/Pandas/SciKitLearn)

o R

o Excel (+VBA)

o SQL

o Bash Scripting

o Clojure (+ Cascalog)

o HMTL5 (+ D3/Bootstrap/Yeoman)

Person Profile

The ideal candidate should demonstrate the following attributes:-

- Experience of delivering data analyses in any of the following domains:

o Central Government

o Consumer Products & Retail

o Telecommunications

o Energy & Utilities

• Methodologies

o Agile working practices

o Business analysis

o Workshop Management

- The Candidate should be prepared to apply for SC security clearance (UK Government)

Additional Information

- Logical career planning route up the career ladder within BIM
- Seeking individuals looking to build a career with us, who can grow into a more senior role with coaching and exposure to challenging projects

I.7 Big Data Analyst (KTP Associate)

Posted on August 26, 2014 by StatsJobs Staff

This KTP project offers a fantastic opportunity for an ambitious postgraduate to launch a career in industry with the support of company and academic mentors.

Majestic 12 Ltd surveys and maps the Internet and has created the largest commercial Link Intelligence database in the world. This Internet map is used by SEOs, New Media Specialists, Affiliate Managers and online Marketing experts for a variety of uses surrounding online prominence including Link Building, Reputation Management, Website Traffic development, Competitor analysis and News Monitoring.

Aston University has joined forces with Majestic 12 on this project to develop a suite of configurable tools to enhance interpretation of internet usage data.

We are looking for confident, credible and personable candidates with good inter-personal skills to work within Majestic 12's small team and with its collaborators and clients. You should be a self-starter with the ability to show high levels of initiative and motivation and the ability to work autonomously to agreed targets and goals. You should be able to articulate ideas and effectively interpret user requirements.

You will need a Bachelor's degree in Mathematics, a numerate science subject or a business related subject at 2.1 level or above and a Master's degree in Computing Science, Econometrics or Statistics.

You should have experience of advanced statistics techniques e.g. Bayesian analysis and high level computing skills, preferably with a knowledge of programming in 'R' (open source) and SQL software languages; data warehousing, data mining and data analytics; database operation; web application development and an understanding of digital marketing concepts and operations.

You should also bring some business analysis skills to the role e.g. analysis of digital marketing activity and market opportunities. Experience of working in a fast moving entrepreneurial environment would also be an advantage.

As part of this project you will be able to develop skills in developing and managing project and product plans for research, development, testing and roll-out. You will also develop your technical skills in data warehouse design and development, data analytics, web-based application development, real world computing design skills, knowledge of 'R' and Hadoop programming and Bayesian statistics.

The KTP Associate Development Programme will give you the management skills required to successfully deliver the project. In addition you will have a generous personal development budget for industry standard training, and the opportunity to apply for membership of the British Computer Society and Chartered Institute of Marketing. There may also be the opportunity to gain certification in Hadoop for data analysis. Alongside this you could achieve either an M.Phil or M.Res from Aston University depending on your existing qualifications and interests.

You will be employed by Aston University but will be based at the company, Majestic 12 Ltd on Birmingham Science Park Aston.

Aston University is committed to disability equality and is a Positive about Disabled People Symbol User.

Please visit our website http://www.aston.ac.uk/jobs for further information and to apply online. If you do not have access to the internet telephone 0121-204-4500 quoting **reference number: R140277.**

Interviews planned for w/c 29th September at Majestic 12 Ltd.

J. Survey results about types of expertise, their importance for industry and acquisition by graduates in their university education

J.1 Results from the academic survey

Academic survey	Question 15: In your opinion, which type of expertise is required for graduate students to be employed in the big data field?									Question 16: Do students acquire expertise in this area?		
	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree	N/A	Sample Size	Average Rating	Standard Error	Yes	No	N/A
Statistical Analysis and Statistical Software Skills	1%	2%	5%	17%	69%	6%	82	4.60	0.09	47%	11%	41%
Statistical Learning	2%	2%	14%	32%	39%	10%	78	4.15	0.11	36%	21%	44%
Mathematics	1%	2%	17%	47%	26%	6%	82	4.01	0.09	36%	22%	43%
Data mining	3%	3%	9%	29%	47%	8%	80	4.23	0.12	30%	23%	47%
Machine learning	5%	8%	21%	31%	22%	14%	75	3.67	0.13	20%	30%	51%
Artificial Intelligence	5%	15%	34%	24%	11%	10%	78	3.26	0.12	16%	36%	48%
Programming	2%	8%	7%	41%	32%	9%	79	4.03	0.11	39%	17%	44%
Marketing	9%	17%	28%	21%	10%	15%	74	3.07	0.14	17%	33%	49%
Business Analysis	5%	14%	25%	22%	22%	13%	76	3.49	0.14	25%	25%	49%
Accounting	14%	28%	33%	9%	3%	13%	76	2.55	0.12	15%	36%	49%

J.2 Results from the graduate survey

Graduate survey	Question 18: In your opinion, which type of expertise is required to work with big data?									Question 19: Do you have this expertise yourself?		
	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree	N/A	Sample Size	Average Rating	Standard Error	Yes	No	N/A
Statistical Analysis and Statistical Software Skills	3%	0%	8%	26%	56%	7%	67	4.42	0.11	68%	18%	14%
Statistical Learning	1%	1%	7%	39%	42%	10%	65	4.31	0.10	67%	18%	15%
Mathematics	1%	6%	24%	33%	21%	15%	61	3.79	0.12	61%	21%	18%
Data mining	6%	1%	10%	26%	47%	10%	65	4.20	0.14	56%	28%	17%
Machine learning	4%	6%	22%	31%	26%	11%	64	3.78	0.14	29%	53%	18%
Artificial Intelligence	7%	10%	32%	25%	14%	13%	63	3.33	0.14	14%	67%	19%
Programming	7%	3%	15%	31%	36%	8%	66	3.94	0.14	50%	29%	21%
Marketing	8%	11%	29%	18%	18%	15%	61	3.31	0.16	25%	53%	22%
Business Analysis	3%	3%	17%	33%	33%	11%	64	4.03	0.12	54%	28%	18%
Accounting	14%	19%	32%	10%	13%	13%	63	2.86	0.16	21%	60%	19%

K. Survey results of software tools

K.1 Results for the importance and the using of software tools in academic survey

Academic survey	Question 17: In your opinion, which software tools are necessary for big data analysis?									Question 18: Which software tools are used in teaching big data		
	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree	N/A	Sample Size	Average Rating	Standard Error	Yes	No	N/A
Base SAS	2%	8%	26%	18%	7%	38%	54	3.31	0.13	11%	25%	63%
SAS Enterprise Miner	2%	8%	23%	20%	10%	37%	55	3.44	0.14	9%	28%	63%
SPSS Analytics	2%	9%	17%	25%	11%	34%	57	3.53	0.14	16%	24%	60%
SPSS Modeler	2%	11%	20%	21%	8%	38%	54	3.33	0.14	13%	25%	62%
WinBUGS	3%	13%	29%	6%	2%	47%	46	2.83	0.13	6%	29%	66%
Matlab	2%	6%	21%	28%	10%	33%	58	3.57	0.13	20%	23%	57%
R	1%	1%	9%	29%	32%	28%	63	4.24	0.11	32%	14%	54%
Java	3%	7%	24%	21%	9%	36%	56	3.39	0.14	13%	26%	61%
VBA	3%	9%	25%	9%	3%	49%	44	3.00	0.15	3%	33%	63%
Hadoop	1%	3%	20%	18%	13%	45%	48	3.69	0.14	8%	30%	62%
MapReduce	1%	1%	25%	15%	11%	46%	47	3.64	0.13	8%	29%	63%
noSQL	3%	6%	23%	10%	5%	53%	41	3.15	0.16	2%	33%	64%
SQL	3%	2%	17%	22%	17%	38%	54	3.76	0.15	18%	21%	61%
JaQL	3%	3%	30%	8%	5%	51%	43	3.14	0.14	6%	30%	64%
Hive	3%	5%	30%	5%	5%	53%	41	3.05	0.15	3%	31%	66%
Oracle	3%	6%	25%	15%	7%	44%	49	3.29	0.15	13%	24%	63%
Python	1%	3%	16%	24%	20%	36%	56	3.89	0.13	17%	22%	61%

K.2 Results for the importance and the using of software tools in graduate survey

Graduate survey	Question 20: In your opinion, which software tools are necessary for big data analysis?									Question 21: Which software tools are used in your company		
	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree	N/A	Sample Size	Average Rating	Standard Error	Yes	No	N/A
Base SAS	1%	4%	24%	21%	19%	31%	50	3.76	0.14	25%	46%	29%
SAS Enterprise Miner	3%	6%	29%	15%	19%	28%	52	3.60	0.15	13%	57%	31%
SPSS Analytics	0%	4%	24%	29%	18%	25%	54	3.81	0.12	39%	35%	26%
SPSS Modeler	0%	6%	19%	28%	19%	28%	52	3.85	0.13	24%	46%	31%
WinBUGS	4%	10%	36%	10%	4%	36%	46	3.00	0.14	6%	63%	32%
Matlab	6%	7%	28%	15%	15%	29%	51	3.39	0.16	17%	53%	31%
R	3%	3%	24%	24%	21%	26%	53	3.77	0.14	25%	47%	28%
Java	3%	6%	28%	25%	15%	24%	55	3.58	0.14	35%	39%	26%
VBA	3%	11%	31%	19%	8%	28%	52	3.27	0.14	31%	38%	32%
Hadoop	1%	4%	28%	11%	19%	36%	46	3.67	0.16	8%	63%	29%
MapReduce	4%	3%	31%	13%	13%	38%	45	3.42	0.16	6%	63%	32%
noSQL	4%	1%	29%	13%	18%	35%	47	3.60	0.16	10%	61%	29%
SQL	1%	3%	14%	24%	38%	21%	57	4.18	0.13	47%	28%	25%
JaQL	3%	3%	33%	13%	3%	46%	39	3.18	0.13	1%	61%	38%
Hive	3%	3%	36%	13%	7%	39%	44	3.30	0.14	3%	64%	33%
Oracle	1%	4%	38%	17%	17%	24%	55	3.56	0.13	24%	49%	28%
Python	1%	4%	31%	18%	11%	35%	47	3.51	0.14	11%	54%	35%

L. Skills required for data scientist analysed from job advertisement

Supplementary Skills	Statistical Techniques	Programming / Computing languages (Software tools)
Big data science	Random Forests	R
Machine learning techniques and softwares	Bayesian networks	SAS (scripting in base SAS and knowledge of SAS macros)
Data mining techniques and softwares/tools	Neural networks	Python (+ SciPy/Pandas/SciKitLearn)
Business Intelligence	Heuristics	Matlab
Artificial Intelligence techniques	Support Vector Machines	SPSS
Data warehousing	Logistic Regression	Ruby
Statistical modeling	Linear Regression	Julia
Predictive modelling	GLM	Stata
Mathematical modeling	Decision Trees	MLWin
Dimensional modelling	Machine Learning algorithms	ArcGIS
Forecasting	K-Nearest Neighbours	Java
Experimental design	Affinity Analysis	C
Data capture	Regression analysis	C++
Data analysis	Multivariate analysis	.NET
Analytical methodologies	Cluster analysis	Excel (+VBA)
Data visualizations tools	Data Munging	Bash Scripting
Reporting tools	Experimental design (sampling, confidence)	Clojure (+ Cascalog)
Relational database modelling and structures	Classifier and predictor development	HTML5 (+ D3/Bootstrap/Yeoman)
Software development lifecycle including testing	Time series analysis	
Preparing datasets	Pattern detection	**Functional languages**
Statistical techniques	Natural Language Processing	Lisp
Analytical skills	Sentiment Analysis (Opinion mining)	Haskell
Data manipulation	Computer Vision	
Statistical analysis		**Database tools**
Quantitative skill	**Speciality skills**	SQL
Technical skill	Primary Mortgage Insurance (PMI)	NoSQL
Data management	Insurance	Oracle
User segmentation	Epidemiological	Sybase
Predictive customer behaviour models e.g. propensity to buy and retention models	Policy issues	DB2
Relational database skills	Biostatistics	Teradata
Non-relational database skills	Engineering	
API Exploitation	Physics	**Data Systems**
Source control using Git		Hadoop (Hive / Impala / Pig)
Continuous Integration (Jenkins)		Mahout
Unit Testing		PostgreSQL
Virtual Machines (Vagrant, Virtualbox)		AWS (EC2 / EMR / S3)
Teradata platform		MongoDB
Algorithm design		Neo4J – Cypher
Prototyping		
Validation		

Non-technical skills and capabilities
Interpersonal skills
Coordinate work
Ability to coordinate work involving multiple stakeholders
Ability to participate in cross-functional projects
Good relationship management skills and the ability to work comfortably with staff at all levels
Ability to work collaboratively
Communication skills
Strong oral communication skills
Strong written communication skills
Ability to clearly explain complex technical issues
Interdisciplinary communication to forge connections with diverse domain specialists
Planning and Organizational skills
Able to prioritise and manage potentially conflicting demands through a focus on delivery of key objectives
Manage workflow in accordance with project timelines
Autonomy
Ability to work independently
Initiative
Ability to work on your own initiative
Initiative in managing data problems / exceptions
Personal accountability
A proactive responsibility for work, taking action and making decisions when necessary
Leadership skills
Team work
Capability to work in teams
Business acumen
Business/commercial sense to combine with analytics to help drive recommendations
Ability to navigate a P&L and demonstrate an understanding of the commercial impact of their activities
Problem solving skill
Ability to structure a large business problem into tractable components
Ability to reason laterally in order to identify solutions to seemingly intractable problems
Able to analyse, identify root causes and resolve issues
Applied analytics
Ability to uses empirical data to support decisions
Ability to glean intelligence and behavioral inferences from data
Demonstrable history of using data to drive incremental value
Strong interest in data analysis and an enquiring mind
Flexibility and resilience
Able to work under pressure in an environment where priorities may change at short notice
Confidence
Enthusiasm
Line management skills (including coaching, training and mentoring)
Able to provide direction and training to other colleagues
Able to manage, coach and develop junior team members
Attitude of continuous personal development
Interest in policy issues

M. Number of participants in each discipline

Discipline	Academic	Graduate
Statistics	20	19
Computing	31	17
Business	11	19
Marketing	15	6
Other	10	11
Total	**87**	**72**

N.The final ethics approval letter

Dear Dr Bilgin,

RE: Ethics project entitled: "Big data and its implications for the statistics profession and statistics education"

Ref number: 5201400282

The Faculty of Science Human Research Ethics Sub-Committee has reviewed your application and granted final approval, effective 3rd June 2014. You may now commence your research. We also wish to apologise for the technical issues which delayed your approval.

The following condition of approval will need to be addressed in an email to sci.ethics@mq.edu.au :

Please indicate how you will choose the initial participants and will any of them be known to the researchers prior to beginning the research?

This research meets the requirements of the National Statement on Ethical Conduct in Human Research (2007). The National Statement is available at the following web site:

http://www.nhmrc.gov.au/_files_nhmrc/publications/attachments/e72.pdf.

The following personnel are authorised to conduct this research:

Dr Ayse Bilgn
Miss Busayasachee Puang-Ngern
Mr Timothy Kang

NB. STUDENTS: IT IS YOUR RESPONSIBILITY TO KEEP A COPY OF THIS APPROVAL EMAIL TO SUBMIT WITH YOUR THESIS.

Please note the following standard requirements of approval:

1. The approval of this project is conditional upon your continuing compliance with the National Statement on Ethical Conduct in Human Research (2007).

2. Approval will be for a period of five (5) years subject to the provision of annual reports.

Progress Report 1 Due: 3rd June 2015
Progress Report 2 Due: 3rd June 2016
Progress Report 3 Due: 3rd June 2017
Progress Report 4 Due: 3rd June 2018
Final Report Due: 3rd June 2019

NB. If you complete the work earlier than you had planned you must submit a Final Report as soon as the work is completed. If the project has been discontinued or not commenced for any reason, you are also required to submit a Final Report for the project.

Progress reports and Final Reports are available at the following website:

http://www.research.mq.edu.au/for/researchers/how_to_obtain_ethics_approval/human_research_ethics/forms

3. If the project has run for more than five (5) years you cannot renew approval for the project. You will need to complete and submit a Final Report and submit a new application for the project. (The five year limit on renewal of approvals allows the Committee to fully re-review research in an environment where legislation, guidelines and requirements are continually changing, for example, new child protection and privacy laws).

4. All amendments to the project must be reviewed and approved by the Committee before implementation. Please complete and submit a Request for Amendment Form available at the following website:

http://www.research.mq.edu.au/for/researchers/how_to_obtain_ethics_approval/human_research_ethics/forms

5. Please notify the Committee immediately in the event of any adverse effects on participants or of any unforeseen events that affect the continued ethical acceptability of the project.

6. At all times you are responsible for the ethical conduct of your research in accordance with the guidelines established by the University. This information is available at the following websites: http://www.mq.edu.au/policy/

http://www.research.mq.edu.au/for/researchers/how_to_obtain_ethics_approval/human_research_ethics/policy

If you will be applying for or have applied for internal or external funding for the above project it is your responsibility to provide the Macquarie University's Research Grants Management Assistant with a copy of this email as soon as possible. Internal and External funding agencies will not be informed that you have final approval for your project and funds will not be released until the Research Grants Management Assistant has received a copy of this email.

If you need to provide a hard copy letter of Final Approval to an external organisation as evidence that you have Final Approval, please do not hesitate to contact the Ethics Secretariat at the address below.

Please retain a copy of this email as this is your official notification of final ethics approval.

Yours sincerely,
Richie Howitt, Chair
Faculty of Science Human Research Ethics Sub-Committee
Macquarie University
NSW 2109

References

Australian Bureau of Statistics. (2014a). *4125.0 - Gender Indicators, Australia, August 2014.* Retrieved from http://www.abs.gov.au/ausstats/abs@.nsf/Lookup/4125.0main+features110August%202014

Australian Bureau of Statistics. (2014b). *Women's participation in the labour force lower than men's*. Retrieved from http://www.abs.gov.au/ausstats/abs@.nsf/Lookup/by%20Subject/4125.0~August%202014~Media%20Release~Women's%20participation%20in%20paid%20work%20lower%20than%20men's%20(Media%20Release)~10008

Batty, M. (2012). Smart cities, big data. *Environment and Planning-Part B, 39*(2), 191.

Bender, A. (2013, January 11). Skilling up: Deakin Uni's big data and analytics program. *CIO*. Retrieved from http://www.deakin.edu.au/buslaw/information-business-analytics/courses/business-analytics.php

Bentley University. (n.d.a). *Master of Business Analytics*. Retrieved from http://www.bentley.edu/graduate/ms-programs/business-analytics

Bentley University. (n.d.b). *Masters of Marketing Analytics*. Retrieved from http://www.bentley.edu/graduate/ms-programs/marketing-analytics

Billhowe. (2013, January 11). New Ph.D. Tracks in "Big Data". *eScience Institute*. Retrieved from http://escience.washington.edu/blog/new-phd-tracks-big-data

Bournemouth University. (2014). *Applied Data Analytics MSc.* Retrieved from http://courses.bournemouth.ac.uk/courses/postgraduate-degree/applied-data-analytics/msc/1975/

Boyd, D., & Crawford, K. (2011). Six provocations for big data. *A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society*. Retrieved from http://ssrn.com/abstract=1926431

Bradshaw, L. (2013, May 28). Big Data and What it Means. *U.S. Chamber of Commerce Foundation*. Retrieved from http://www.uschamberfoundation.org/library/2013/05/big-data-and-what-it-means

Brown, B., Chui, M., & Manyika, J. (2011). Are you ready for the era of 'big data'. *McKinsey Quarterly, 4*, 24-35.

Chen, H., Chiang, R. H., & Storey, V. C. (2012). Business Intelligence and Analytics: From Big Data to Big Impact. *MIS quarterly, 36*(4), 1165-1188.

Datameer. (2014). *What is Big Data?*. Retrieved from http://www.datameer.com/product/big-data.html

Degree Prospects, LLC. (2013). 23 Great Schools with Master's Programs in Data Science. *Master's in Data Science*. Retrieved from http://www.mastersindatascience.org/schools/23-great-schools-with-masters-programs-in-data-science/

Deakin University. (2014a). *Bachelor of Computer Science*. Retrieved from http://www.deakin.edu.au/__data/assets/pdf_file/0008/246824/comp-sci-flyer-2014.pdf

Deakin University. (2014b). *Master of Business Analytics*. Retrieved from http://www.deakin.edu.au/__data/assets/pdf_file/0006/43548/bus-analytics-Jun2014.pdf

Dhar, V. (2013). Data Science and Prediction. *Communications of the ACM, 56*(12), 64-73.

E-skills UK. (2013). *Big Data Analytics: An assessment of demand for labour and skills, 2012-2017*. Retrieved from http://www.photonics21.org/download/other_news/BigDataAnalytics_Report_Jan2013.pdf

Gartner. (2012, October 22). *Gartner Says Big Data Creates Big Jobs: 4.4 Million IT Jobs Globally to Support Big Data By 2015.* Retrieved from http://www.gartner.com/newsroom/id/2207915

Gilpin, L. (2014, June 13). How Intel is using IoT and big data to improve food and water security. *TechRepublic*. Retrieved from http://www.techrepublic.com/article/how-intel-is-using-iot-and-big-data-to-improve-food-and-water-security/

Glance, D. (2013, December 2). Solving Big Data's big skills shortage. *The Conversation*. Retrieved from http://theconversation.com/solving-big-datas-big-skills-shortage-20352

Hurwitz, J., Nugent, A., Halper, F., & Kaufman, M. (2013). Big data for dummies. *For Dummies*.

IBM Institute. (2013, August 14). *IBM Narrows Big Data Skills Gap By Partnering With More Than 1,000 Global Universities*. Retrieved from http://www-03.ibm.com/press/us/en/pressrelease/41733.wss

IBM Institute. (n.d.). *What is JaQL?*. Retrieved from http://www-01.ibm.com/software/data/infosphere/hadoop/jaql/

IBM Big Data & Analytics. (2013). *The Four V's of Big Data*. Retrieved from http://www.ibmbigdatahub.com/infographic/four-vs-big-data

Jeske, M., Grüner, M., & Weiß, F. (2013). Big data in logistics: a DHL perspective on how to move beyond the hype. *DHL Customer Solutions & Innovation*.

Lambert, D. (2003). What use is statistics for massive data?. *Lecture Notes-Monograph Series*, 217-228.

Lunet-Levi, T. (2013, October 24). The 3 V's of Big Data and their Technologies. *Geektime.* Retrieved from http://www.geektime.com/2013/10/24/the-3-vs-of-big-data-and-their-technologies/

Macquarie University. (2014). *Master of Data Science*. Retrieved from http://mq.edu.au/pubstatic/public/download.jsp?id=122186

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., et al. (2011). Big Data: The next frontier for innovation, competition, and Productivity. *McKinsey Global Institute*.

Marr, B. (2013). The Awesome Ways Big Data Is Used Today To Change Our World [Web log post]. Retrieved from http://www.linkedin.com/today/post/article/20131113065157-64875646-the-awesome-ways-big-data-is-used-today-to-change-our-world?trk=mp-author-card

Mayer-Schönberger, V., & Cukier, K. (2013). *Big data: A revolution that will transform how we live, work, and think*. Houghton Mifflin Harcourt.

McAfee, A., & Brynjolfsson, E. (2012). Big data: the management revolution. *Harvard business review*, *90*(10), 60-68.

Kepner et al., 2014

Kepner, J., Gadepally, V., Michaleas, P., Schear, N., Varia, M., Yerukhimovich, A., & Cunningham, R. K. (2014). Computing on masked data: a high performance method for improving big data veracity. *arXiv preprint arXiv:1406.5751*.

Knights, M. (2013, September 4). Big data or big statistics. *Raconteur Media*. Retrieved from http://raconteur.net/technology/big-data-or-big-statistics

Molinari, C. (2012, October 9). No one size fits all strategy for big data, says IBM. *BNamericas*. Retrieved from http://www.bnamericas.com/news/technology/no-one-size-fits-all-strategy-for-big-data-says-ibm

Nerney, C. (2013, August 28). Universities Expanding Big Data Analytics Courses with IBM Aid. *Data Informed*. Retrieved from http://data-informed.com/universities-expanding-big-data-analytics-courses-with-ibm-aid/

Professional Advantage. (2014). *Big Data: What is Big Data?*. Retrieved from http://www.bianalytics.com.au/qlik/big-data/?gclid=CjsKDwjwmuafBRCQ7ef6zJXhdRIkAFH2mefRxylXaHc9cVWJF5vYMAt7tjAk-rpHrYSpee47Pxh6GgL-PvD_BwE

RMIT University. (2014a). *Computing and Information Technology - Degree and Diploma Guide*. Retrieved from http://www2.rmit.edu.au/Courses/pdf/comp.pdf

RMIT University. (2014b). *Computer Science and Information Technology - Postgraduate*. Retrieved from http://www2.rmit.edu.au/Courses/pdf/pg_csit.pdf

Shaw, J. (2014). Why “Big Data” is a big deal. *Harvard magazine*. Retrieved from http://harvardmagazine.com/2014/03/why-big-data-is-a-big-deal

SINTEF. (2013, May 22). Big Data, for better or worse: 90% of world's data generated over last two years. *ScienceDaily*. Retrieved from www.sciencedaily.com/releases/2013/05/130522085217.htm

The Data Alchemists. (2013). *The 5 V’s of Big Data*. Retrieved from http://dataalchemists.com.au/2013/05/5-vs-big-data/

University of Auckland. (2013). *Master of Professional Studies in Data Science: a taught masters for industry*. Retrieved from https://cdn.auckland.ac.nz/assets/science/about/our-faculty/prospectuses-handbooks/pdfs/2013-data-science-flyer.pdf

University of Otago. (2013). *Master of Business Data Science.* Retrieved from http://www.otago.ac.nz/business/study/postgraduate/otago072318.pdf

University Of Pittsburgh. (2013). *Big Data Analytics Specialization.* Retrieved from http://www.ischool.pitt.edu/ist/degrees/specializations/big-data.php

University of South Australia. (2014). *Data Science.* Retrieved from http://www.unisa.edu.au/Global/ITEE/ITMS/DATA%20SCIENCE/Data%20Science-05-web.pdf

University of Technology Sydney. (2014). *Master of Data Science and Innovation.* Retrieved from http://www.uts.edu.au/future-students/analytics-and-data-science/master-data-science-and-innovation/about-course

University of Western Sydney. (n.d.). *Postgraduate Specialisation - Distributed Computing*. Retrieved from http://handbook.uws.edu.au/hbook/specialisation.aspx?unitset=ST3033.1

Zikopoulos, P., & Eaton, C. (2011). Understanding big data: Analytics for enterprise class hadoop and streaming data. *McGraw-Hill Osborne Media*.