NEED FOR B.S. DATA SCIENCE DEGREE PROGRAM IN THE PHILIPPINES

Rudy H. Tan, Ph.D. * and Lourdes A. Tan, M Stat**

ABSTRACT

Graduates of the B.S. Statistics program easily find employment because their training in statistical thinking is useful in any field of application. However, they are seldom hired for the position of statistician. During the last few years, there has been a massive explosion of big data outpacing the statistical tools to analyse it. Big Data is no longer just a collection of numbers and categorical variables. It now includes emails, tweets, GPS locations, images, etc. in big data, statistical inference may no longer apply since the concept of population and sample is lost. The requisite expertise needed to handle big data goes beyond having a B.S. Statistics degree. Knowledge of computer science, mathematics, statistics, machine learning, data mining, and data visualization is needed. Thus, there is a need to introduce a B.S. Data Science degree program in schools offering statistics and IT courses. It is estimated that the U.S.A. alone will need 200,000 data scientists by 2018. This paper will propose a curriculum for B.S Data Science.

1. Introduction

Aren't We Data Science? This question is the title of an article published in AMSTAT News (July 2013). Dr. M. Davidian, President of the American Statistical Association and Professor of Statistics at North Carolina State University, wrote the article "to serve as a metaphor for the disconnect between statistics and data science." She and other statisticians were unaware that the National Consortium for Data Science (NCDS) just opened on the campus of the University of North Carolina at Chapel Hill.

Singapore, Malaysia, Indonesia, and Thailand are already preparing to offer Data Science degree programs in 2016. Is the Philippines ready? It has the most number of schools in Southeast Asia offering degree programs in statistics. Eleven schools offer B.S. Statistics; four B.S. Mathematics with a major in Statistics; and one B.S. Statistics with a major in Actuarial Science. These 17 schools are in NCR (4), Luzon (3), Visayas (4), and Mindanao (6). Tawi-Tawi in Sulu is the furthest island with a school offering B.S. Statistics under the Mindanao State University. Most of these schools adopted the B.S. Statistics curriculum of the University of the Philippines Statistical Center (UPSC).

The UPSC is the first school in Southeast Asia to offer an undergraduate degree program in statistics. In 1953, the United Nations Development Program (UNDP) under a bilateral agreement with the Philippine government established the UPSC to improve the Philippine Statistical System by training government statisticians. The Bachelor of Statistics degree program started in 1964. It was replaced by B.S. Statistics in the Academic Year 1967-1968. In 1998, the name of the UPSC was changed to the U.P. School of Statistics (UPSS).
Most of the early graduates of B.S. Statistics entered the government service, and some became faculty members of UPSC. Many found employment in manufacturing, banking, and

*Retired UPSC Professor of Statistics and UNILAB Pharmaceutical Statistician
**Retired SQCS Specialist, San Miguel Corporation

other industries because their training in statistical thinking is useful in any field of application. However, graduates still have to undergo on- the-job training to gain domain knowledge. For example, a new graduate of B.S. Statistics has to train at least one year in a drug company to become a pharmaceutical statistician.

Are the graduates of statistics degree programs from the 17 schools in the Philippines prepared to face the challenges of a rapidly growing data during the last five years? Data has grown so big in volume, velocity, and variety, exceeding the storage capacity of existing database systems and making statistical analysis inadequate. Google and Facebook collect unimaginable amounts of data every second. Satellites orbiting the earth, cameras on electric posts at every corner of the street, and monitoring devices, including sensors placed on human bodies, continuously transmit data. The age of "Big Data" and a new science of data have arrived. This paper will propose a B.S. Data Science curriculum in these 17 schools with existing B.S. Computer Science or Information Technology degree program.

2. The Age of Big Data

The "Age of Big Data," which started in 2007, spawned data driven technological and business innovations in the 21st century. However, the history of data predates the development of computer and statistics by several thousands of years. Two factors were crucial in the evolution of data: storage devices and tools to extract information for decision making. The amount of data generated has been rapidly growing since the beginning of the 20th century, but reached epic proportions only during the last 10 years.

The earliest known form of data is just a simple tally on a baboon bone discovered in 1960. It has been dated to the Upper Paleolithic Era between 18,000 and 20,000 BCE. The Ishango Bone is the first known data storage device used by humans to keep a record of counts. In 3800 BCE, the king of Babylon ordered a tally of people, livestock, and commodities every six or seven years. This was the earliest known census of population, and the data were preserved on clay tablets. Ancient Egypt had one of the first national censuses in 3340 and 3050 BCE, which were documented on papyrus and inscribed on monuments. The most sophisticated census of population in the ancient world was undertaken by the Roman Empire from 443 BCE until the third century CE.

During the middle ages, the state collected data on people and wealth of the nation primarily to determine its preparedness to go to war. In the 16th century, data collected by the state was called "statistik." This was the origin of the word "statistics," derived from the Latin word "status," meaning "Political State." A person knowledgeable about the affairs of the state was called a *statist* or a statesman. Thus, the practice of statistics was considered a "politician art." The misuse of statistics by politicians popularized the phrase "lies, damned lies, and statistics," attributed to the British Prime Minister Disraeli.

In the 17th century, statistics developed into the science of data collection, organization, presentation, analysis, and interpretation. Descriptive statistics are the results of data analysis and usually presented in a table or chart. John Graunt's 1662 Observations on the Bills of Mortality is often cited as the first descriptive statistics. A vast amount of data was presented in a few tables to provide comprehensive information in a nutshell. His statistical analysis of

mortality data in London provided insights into the causes of death in the 17th century. Thus, in the plural sense of the word, statistics and data are often used interchangeably to mean a collection of "facts and figures."

Modern statistics is the science of data analysis and decision making in the face of uncertainty. Its framework in probability theory was laid down in the late 19th and early 20th century. It is divided into descriptive statistics and statistical inference, which are complementary to each other. Descriptive statistics summarizes data either numerically or graphically to facilitate interpretation. Statistical inference generalizes the information from a sample data to the population using probability theory. The introduction of the concepts of correlation and regression by Sir Francis Galton (1889), chi-square test by Karl Pearson (1900), Student's t-test by Gosset (1908), and experimental design by R.A. Fisher (1935) established modern statistics as a distinct scientific field rather than a branch of mathematics. Although the needs of business and research organizations are new concepts and procedures for analyzing large datasets generated in the information age, the development of modern statistics continues along mathematical statistics,

In 1973, the introduction of electronic calculators brought significant changes in statistical computing. The students at the U.P. Statistical Center used to spend about an hour to calculate the square root of a number using a Friden electro-mechanical calculator. The time was reduced to a few minutes after the introduction of a simple electronic calculator, and a few seconds with the press of a button on a scientific calculator with a square-root function. The arrival of personal computers and statistical software packages like SPSS, SAS, and BMDP led to the advancement of computational statistics and data analysis. In 1980, the U.P. Statistical Center established the Statistical Computing and Consulting Laboratory.

The amount of data has grown so big since 1997 when Google introduced its search engine to crawl the Internet. It was predicted that the worldwide web will increase in size 10-fold each year, creating a deluge of data to store with no way to analyze it. After 10 years, the term "Big Data" became a new buzz word. Data analytics have to be developed to discover information and provide new insights into a problem. Unlike statistical data that are structured or semi-structured, big data can be either structured or unstructured, numeric or non-numeric, including images, audio, video, and text. It is characterized by **V**olume, **V**elocity, and **V**ariety (3V's). The amount of data in the digital universe is estimated to be 2.7 zettabytes (10^9 terabytes). Facebook alone already generates more than 500 terabytes of data every day.

Statisticians have been analyzing very large datasets since the introduction of the mainframe computer. Today, data storage is no longer a problem, and many statistical software packages can now handle large datasets. Aren't these large datasets already big data? Is there a need for a new science for big data? Isn't Data Science just a rebranding of Statistics? What are the differences between Statistics and Data Science? Should schools in the Philippines start offering B.S. Data Science?

3. The New Science of Data

The new field of Data Science has its root in data analysis. Unfortunately, statistical data analysis has not been developed to face the challenges of big data, since modern statistics continues to emphasize theory rather than applications. In a 1962 paper "The Future of Data

Analysis," John Tukey brought into question the relationship between statistics and data analysis. He was the first prominent statistician to suggest that data analysis should "take on the characteristics of science rather than those of mathematics." In 1977 Tukey published *Exploratory Data Analysis*, which is now a classic textbook in Data Science.

The first known reference to the term "data science" was in the 2001 paper entitled "Data Science: An Action Plan to Expand the Field of Statistics" by William Cleveland, a statistician at Bell Labs in the U.S.A. He introduced data science as an independent discipline, extending the field of statistics to incorporate "advances in computing with data." However, it took nearly a decade for interest in data science to build up starting in 2011 when the U.S. economy was beginning to recover from a severe recession. The job title "data scientist," which first appeared on Facebook, helped popularize data science. Some academic statisticians think that it is just a sexed up term for statistician, and data science is a rebranding of statistics. Most of the early data scientists who worked at Google, Amazon, Linkedin, Facebook, and other data-driven business enterprises were applied statisticians with advanced computer experience.

There is no accepted definition of data science. Based on several definitions given by different authors, it is clear that data science is a systematic study of digital data using statistical techniques and applications of computer science. Its goal is to make sense of vast amounts of dynamic data and extract information that will lead to new knowledge that can provide actionable insights for decision makers. In addition to advanced computer skills and in-depth knowledge of statistics, a data scientist must acquire domain knowledge to interpret data-driven insights. Like any field of science in its infancy, data science will continue to develop and play an important role in data-driven technological innovations and business transformations in the 21st century. We are already witnessing disruptive technologies that will change forever how we work and play just like the personal computer in the 1980s that displaced the typewriter. The key to successful innovation is data. Although innovation is often triggered by a brilliant insight, data science is needed to develop a concept into a new technology.

Some schools with B.S. Statistics offer analytic courses like data mining and predictive analytics. However, the students still lack the computer skills, and data science knowledge needed to do big data analytics in the real world. Storing, accessing, organizing, processing, analyzing, and using rapidly growing and ever-changing massive amounts of data will require new data platforms and analytical tools not taught in B.S. Statistics courses. Data science is more than statistics. The differences between statistics and data science can best be explained by how statistical data and big data are stored, processed, and analyzed.

Statistical data is organized in a structured table of observations (rows) and variables (columns) for analysis by a program like R, SPSS, or SAS. The data values are numbers or characters and can also grow extremely large like big data, but are generally homogeneous and typically do not require real-time analysis. Big data can be structured or unstructured, messy and varied. According to the 2015 O'Riley Data Science salary survey, Apache Scala has replaced R as the choice for big data real-time statistical computing.

Statistical data is typically stored and processed in a Relational Database Management System (RDBMS) like Oracle or DB2. Apache Hadoop, is widely used for storage and parallel processing of big data across clusters of computers. It is not a database but an open-source software for distributed file system.

Statistical analysis of a very large data stored in a single database can result in very slow computing time, and might even crash the computer. Big data is stored in several servers to increase processing time by using analytic software running simultaneously in parallel. If there is a hardware failure, the job is automatically redirected to continue the distributed computing.

To test a hypothesis, a statistician must take a sample and make inference about the population based on the result of statistical analysis. Big data can be analyzed and explored in its entirety to discover patterns and correlations. The concept of sample and population may no longer apply. Is this the demise of statistical inference? The answer is NO! There are still reasons to take a random sample to examine more closely the elements of the population.

Statistical data in databases must first be cleaned or scrubbed before processing to ensure that the data values are correct and consistent with similar data in different storage. Manual cleaning of large datasets is a very tedious process. However, there are tools that can help shape the data for analysis. These data wrangling tools can automatically reformat, merge, and filter datasets. Messy big data can be cleaned as needed and analyzed in real-time as it is generated.

4. Data Science Job Trends

Job demand for data scientists started in 2011 when the U.S. economy was recovering from a severe recession. Figure 1 shows the U.S. job trends for data scientist and statistician from 2006 to 2015. Statisticians steadily dominated the analytics job market until 2011 but begin losing dominance when data-driven companies started hiring data scientists.

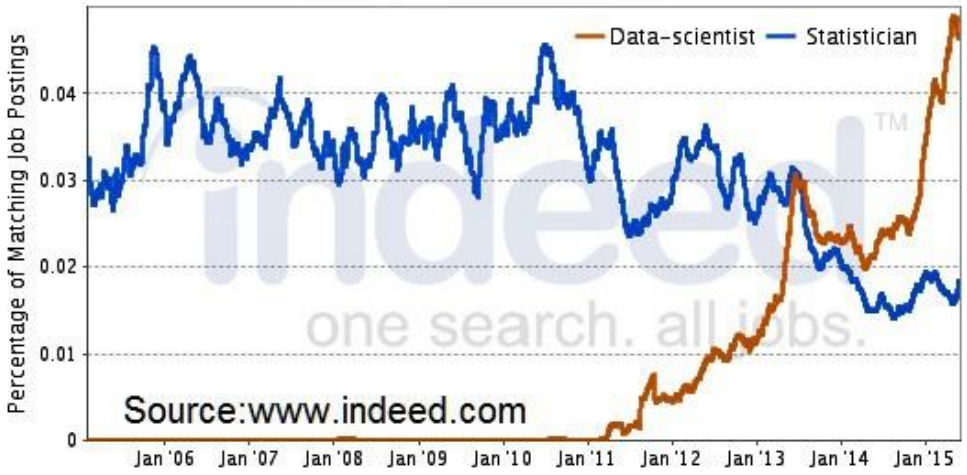

Figure1. Job trends for data scientist and statisticians from January 2006 to January 2015. The vertical axis is percentage of matching job postings.

Figures 1a and 1b show the upward linear job trend for data scientist and the downward job trend for statistician from 2012 to 2016. On May 27, 2016, the percentage of matching job posting is 0.05439% for data scientist compared with only 0.0147% for statistician. Data scientist posted a 1567% job growth compared with -43% for statistician over a 5-year period. The trends forebode a dim future for graduates of statistics. Companies will prefer to employ data scientists since they have skills in computer science and advanced knowledge of

statistical analysis.

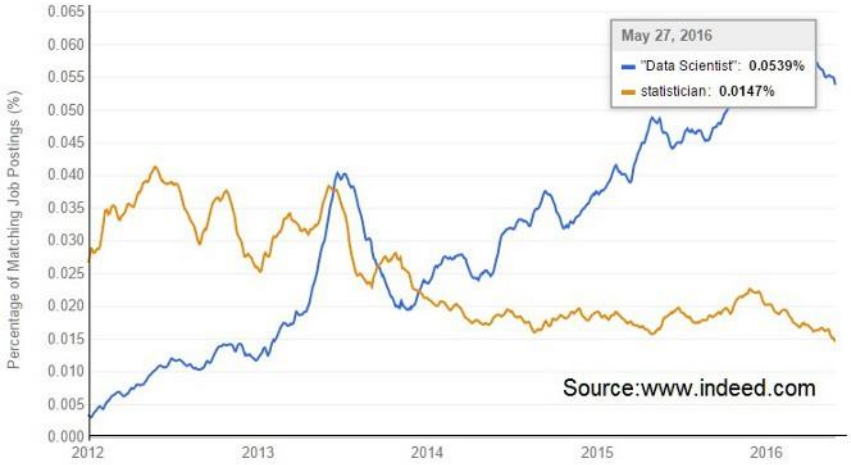

Figure 2a. Job trends for data scientist and statisticians from January 2012 to 2016. The vertical axis is percentage of matching job postings.

Figure 2a. Job trends for data scientist and statisticians from January 2012 to 2016. The vertical axis is percentage Growth.

Data scientist is listed as the best paying job for 2016, by CareerCast.com, an online career search and information site. Forbes magazine named it the hottest job in 2016 with a median annual salary of $120K. In 2011, a McKinsey study predicted that by 2018 the number of data science jobs in the United States alone will exceed 490,000, but there will be fewer than 200,000 available data scientists to fill these positions. Globally, demand for data scientists is projected to exceed supply by more than 50 percent by 2018. Search engine and social network companies like Microsoft, Google, Facebook, and Linkedin employ the most number of data scientists. Based on the O'Reiley 2015 Data Science Salary Survey, more than 50% of data scientists are in the U.S. and concentrated in California. The United Kingdom and India have the

highest numbers of data scientists after the U.S. In Southeast Asia, Singapore and Malaysia are gearing up to attract and train data scientists to meet the needs of multinational companies in the region. In the Philippines, ads for data scientist job are occasionally posted by Business Process Outsourcing (BPO) companies.

Based on the Philippine Standard Occupation Classification (PSOC), B.S. Statistics graduates can find low-paying jobs in government as assistant statistician, statistical processor, statistical researcher, statistics billing assistant, and statistical aide. Those with at least a master's degree can be appointed as a statistician. In the U.S. most companies will hire only a Ph.D. graduate as a statistician, whereas a B.S. Statistics graduate will most likely find work as a computer programmer. In 2013, the Department of Labor and Employment (DOLE) listed statistician as the 10th highest paying job in the Philippines. However, according to the National Statistical Coordination Board (NSCB), the salary of Philippine government statisticians is among the lowest in the ASEAN region.

5. Proposed B.S. Data Science Curriculum

During the last few years, there has been a proliferation of online courses on data science. The most popular is Coursera, a MOOC (Massive Open Online Course) platform taught in partnership with John Hopkins University in the U.S.A. It was originally "absolutely" free, but now it will cost about P20,000 to get a Data Science Specialization Certificate. Microsoft is also offering an online Professional Degree in Data Science. A graduate of statistics or computer science will most likely benefit by enrolling in online courses to enhance one's job application portfolio. However, a campus experience is still the best for students entering college.

Many universities worldwide now have degree programs in data science. Among the schools in Southeast Asia that will offer data science degree programs in 2016 are the National Singapore University and the seven institutes of higher learning in Malaysia. In the Philippines, the Commission on National Education has accredited the University of the Philippines, Ateneo de Manila, De La Salle University, and the University of Santo Tomas to offer business analytics specialization tracks. It should be noted though that a one-semester course in business analytics does not make a data scientist.

To address the expected need for data scientists in the Philippines, schools with existing B.S. Statistics and B.S. Computer Science should consider offering a B.S. Data Science. The objective of the B.S. Data Science is to produce data analysts who will eventually become data scientists after gaining experience working in a data-driven organization. The label "scientist" is a high calling, and requires domain expertise to provide insights from data to discover new knowledge.

The authors reviewed the B.S. Statistics and B.S. Computer Science degree programs of several schools in the Philippines as well as online and campus-based data science programs of schools in the U.S. Listed below are the courses recommended for inclusion in the proposed 4-year B.S. Data Science degree program. Students in their senior years will undertake a capstone project via an internship program with participating business enterprises or research institutions to develop their domain knowledge.

1. Statistics Courses for Data Science

INTRODUCTION TO STATISTICAL METHODS (3 Units)
Data and statistics. Elements of probability. Population and sample. Descriptive statistics. Estimation and Tests of hypotheses. Correlation and regression analysis. Nonparametric statistics. Bayesian inference. Design of experiments.

EXPLORATORY DATA ANALYSIS (3 Units)
Numeric displays and schematic summaries. Re-expressing data. Comparing several batches of data. Straightening curves and plots. Plotting relationships and straightening out plots. Smoothing sequences. Making and using two-way tables. Advanced fits and three way-fits. Looking at shapes of distribution. Robust and resistant measures. Direct assessment by Jackknifing.

LINEAR AND NONLINEAR MODELS (3 Units)
Elements of matrix algebra. Multivariate Normal Distribution. Linear regression model. General Linear Models. Experimental Design Models. Analysis of Variance and Covariance Models. Linear Mixed Models. Nonlinear models.

MULTIVARIATE DATA ANALYSIS (3 Units)
Review of matrix algebra. Data matrix and attributes. Characterizing and displaying multivariate data. Dimensionality reduction by principal component analysis. Linear discriminant analysis and classification. Hierarchical and non-hierarchical clustering techniques.

TIME SERIES ANALYSIS AND STOCHASTIC PROCESSES (3 Units)
Time series and stochastic processes. Stationary and nonstationary time series. Time series modelling and forecasting. Discrete time versus continuous time stochastic processes. Markov, Poisson, Renewal, Brownian motion and other processes.

STATISTICAL COMPUTING (3 Units)
Statistical algorithms. Programming in R. Macro programming in VBA. Using Excel, SAS, SPSS, Minitab, Mathlab, and other software packages for data analysis.

2. Data Science Core Courses

INTRODUCTION TO DATA SCIENCE (3 Units)
What is data science? About Data and Big Data. Fundamental concepts of data science. Exploring data. Visualizing data. Analyzing data. Making predictions. Computing for data science. Evolving role of data scientists Ethics and privacy concerns. Applications.

DATA VISUALIZATION (3 Units)
What is data visualization? Types of data visualization. Visualization with big data. Visual analytics. Visualization techniques. Deployment considerations. Using software for data visualization. Reading Tufte's books: Beautiful Evidence, Visual Display of Quantitative Information, Visual Explanation, and Envisioning Information.

DATA MANAGEMENT (3 Units)
What is data management? Data integration, virtualization, streaming, quality, security, and governance. Master data management. iCloud data management. Web data management. Graph data management. Spatial data management. Big data management. Application softwares for data management.

DATA MINING (3 Units)
What is data mining? Warehousing. Basics, tasks, process, and methodologies of data mining. Operational techniques. Knowledge extraction. Web mining. Applications.

DATA ANALYTICS (3 Units)
What is data analytics? Phases of data analytics. Tools of data analytics. Techniques of data analytics. Next generation data analytics. Business analytics.

PREDICTIVE ANALYTICS (3 Units)
What is predictive analytics? Review of statistical tools for predictive analytics. Identifying and selecting the prediction variables. Setting the data for model building and validation. Preparing and organizing the data. Building the prediction model. Validating and evaluating the model performance. Using the model for predicting. Profiling based on model results. Using MS Excel and software packages. Applications.

3. Computer Science Courses for Data Science

INTRODUCTION TO COMPUTER SCIENCE (3 Units)
Evolution of computer. Computer science, information technology, and information science. Hardware and software. Types of computers. Number representation and computer arithmetic. Computer organization and architecture. Computer systems. Operating systems. Networking. Advances in computer science

PROGRAMMING LANGUAGES FOR DATA SCIENCE (3 Units)
Introduction to C++, Java, R, SQL, Apache, Python, Spark, Scala, and Azure. Software installation. Fundamental concepts. Essential programming elements. Object-oriented programming. Writing, testing and debugging a simple program. Programming project. Applications.

DATA STRUCTURES AND ALGORITHMS (3 Units)
Concepts, algorithms & applications of complex data structures: tables, trees, graphs, heaps, generalized lists, multilinked structures. Basic algorithmic techniques & analysis: sorting algorithms, hash tables, binary search trees, balanced trees.

DATABASE MANAGEMENT SYSTEMS (3 Units)
Concept of Database Management System (DBMS). Database models and architectures. Logical and physical database designs. Database implementation. Structured Query Language (SQL). Database administration and security. Client/server database. Distributed database. Hadoop Distributed File System (HDFS).

COMPUTER NETWORKING (3 Units)
Networking fundamentals. Understanding network terminology. Network models and layers. Network protocols. Concurrency. Network interconnection. Parallel and distributed computing. Networking & communication software. Emerging network technologies.

ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING (3 units)
Artificial intelligence versus machine learning. Fundamental principles of artificial intelligence. Knowledge representation & reasoning. Agents. Machine learning & neural networks. Current research applications. Google Tensor Flow Machine Learning software

CLOUD COMPUTING (3 Units)
Introduction computing paradigms. Fundamentals of cloud computing. Cloud computing management and

governance. Networking for cloud computing. Risks and security in cloud computing. Microsoft Azure. Cloudera. Applications.

HACKING AND CYBERSECURITY (3 Units)
Introduction to hacking. . Planning hacking attacks. Types of hacking attacks and techniques. Hacking tools and software. Hacking windows operating systems. Hacking mobile devices. Hacking communication systems. Hacking websites. Hacking applications. Hacking databases. Ethical hacking. Penetration testing. Cybersecurity.

6. Conclusion

Big data is now considered an economic resource. According to a special report published by the Economist, it is "becoming the new raw material of business: an economic input almost on par with capital." Data science is the new discipline that can exploit big data for its economic value to transform information into insights needed to create better services and new products in the 21st century and beyond.

Data science is not just applied statistics! It requires computer experience and domain knowledge or expertise. The scope of data science is wider than statistics, and data scientists are paid higher than statisticians. The hiring trend for data scientists is increasing in the U.S., EU, India, Singapore, and emerging economies in the ASEAN region like Malaysia and Thailand. Data science graduates will become the future high-paying Overseas Filipino Workers (OFWs).

In spite of the demands for data science graduates, the statistics degree program will not yet face extinction in the near future. However, its graduates will increasingly find it difficult to find jobs in data-driven organizations. BPO, call centers, banks, telecommunication, multinational retail companies, and other big businesses will prefer to hire B.S. Data Science graduates for their analytical needs to remain competitive.

References

Ahalt, S. (2012). Establishing a National Consortium for Data Science. Renaissance Computing Institute, University of North Carolina at Chapel Hill, USA

Davenport, T. & D.J. Patil (2012 October). Data Scientist: The Sexiest Job of the 21st Century. Harvard Business Review, pp 70-76

Davidian, M. (2013 July): Aren't We Data Science? AMSTAT News, No. 433, pp 3-5

Hand, D.J. (2014 January/February). Hand Writing: Data, data, everywhere, but let's just stop and think. Institute of Mathematical Statistics Bulletin, Vol. 43, Issue 1, p4

McAfee A. & E. Brynjolfsson (2012 October). Big Data: The Management Revolution. Harvard Business Review, pp 61-68.

Lohr, S. (2012 February 12). The Age of Big Data. Sunday Review, The New York Times.

Manyika, J., M. Chiu, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, and A.H. Byers (2011). Big Data: Next Frontier for Innovation, Computation, and Productvity. The McKinsey Global Institute, McKinsey & Company.

Press, Gil (2013 May 28). A Very Short History of Data Science. Forbes Magazine.

Rao, C.R. (1992). R.A. Fisher: The Founder of Modern Statistics. Statistical Science, Vol. 7 No. 1, pp 34-48.

Schutt, Rachel & Cathy O'Neil (2014): Doing Data Science. O' Reilly Media, Inc., CA, USA.

Stigler, S.M. (1986). The History of Statistics - The Measurement of Uncertainty before 1900. Harvard University Press, USA

The Economist (2015) Big data evolution: forging new corporate capabilities for the long term. The Economist Intelligence Unit Ltd., London

The Economist (2010): Data, data, everywhere. A special report on managing information.The Economist Group, London

Tukey, John W. (1962). The Future of Data Analysis. The Annals of Mathematical Statistics, Vol. 33, No. 1, pp 1-67.

Varberg, D.E. (1963 April) The Development of Modern Statistics. The Mathematics Teacher, Vol. 56, No. 4, pp 252-257.

Wickham, H. (2014 September). Data Science: How is it different to statistics? Institute of Mathematical Statistics Bulletin, Vol. 43, Issue 6, p7.