Data Scientists as Game Changers in Big Data Environments

Akemi Takeoka Chatfield
School of Information Systems and Technology
University of Wollongong
Wollongong, Australia
Email: akemi@uow.edu.au

Vivian Najem Shlemoon
School of Information Systems and Technology
University of Wollongong
Wollongong, Australia
Email: vns983@uowmail.edu.au

Wilbur Redublado
School of Information Systems and Technology
University of Wollongong
Wollongong, Australia
Email: wr803@uowmail.edu.au

Faizur Rahman
School of Information Systems and Technology
University of Wollongong
Wollongong, Australia
Email: mrahman@uowmail.edu.au

Abstract

The potential power of big data to generate insights and create new forms of value in the ways which transform organizations and society has been observed by big data-driven organizations and big data experts. Despite the recent sensational declaration of a data scientist as "the sexiest job of the 21st century", however, there is a lack of published rigorous studies of what a data scientist is, and what job skills this hottest job title may require. In order to address this gap, we systematically examine relevant source material to extract key definitions and categorize them towards understanding emergent roles and skills of data scientists. We conclude that the current lack of clear skills specifications and the growing skills shortage are key barriers to realizing the potential benefits of big data in organizations, through hiring data scientists and leveraging their game changer roles, indicating important educational implications for our IS field.

Keywords

Big data, data scientist, game changer roles, big data analytics, data scientist skills.

INTRODUCTION

The Internet has transformed the way we collect and share data, having made data collection and data sharing processes much easier (Cukier and Mayer-Schönberger 2013) and big data access in real time to make managerial decision-making more agile (McAfee and Brynjolfsson 2014). This transformative power of the Internet has greatly accelerated the explosion of data and further digitization of government and business organizations. "The rise of big data" is, however, relatively new (Cukier and Mayer-Schönberger 2013, p. 28). While its impact is not yet as visible as that of the Internet, big data can transform how the organization and society process data and change the way we think about the world (Cukier and Mayer-Schönberger 2013; Mayer-Schönberger and Cukier 2014; Davenport 2014).

Importantly, the rise of big data is more noticeable among nations which have promoted the development of sustainable knowledge economy. The growing knowledge economy has created new jobs in the professions, science, and technology over the traditional jobs in the mining, accommodation, agriculture and construction industries (Trounson 2014). The global trend for push for knowledge economy seems to further stimulate the explosion of big data of different types generated by click data from growing e-commerce websites, social media

channels, sensors connected to the Internet (so called "the Internet of things"), and surveillance images and data for monitoring customer service quality of outsourced call center operations, among others.

While there are various definitions of big data, there is no rigorous definition of big data (Mayer-Schönberger and Cukier 2014). However, for our purpose in this paper we adopt a forward-looking definition which is relevant and applicable to different big data projects. Mayer-Schönberger and Cukier (2014, p. 6), authors of *Big Data*, argue: "big data refers to things one can do at a large scale that cannot be done at a smaller one, to extract new insights or create new forms of value, in ways that change markets, organizations, the relationship between citizens and governments, and more." While big data is thus viewed as providing the organization with the great potential to generate new insights or to create new forms of value, it is clear that big data are "things one can do at a large scale", indicating that big data as a new breed of organizational resource needs to be enacted by human actors to realize, capture, and deliver its full potential benefits to the key stakeholders in the organization.

In response to the accelerating big data environments, the demand for a "data scientist" has emerged and has been sharply increased in marketplace. The demand comes from small big data-driven start-ups as well as from large companies under the global competitive pressures for innovation and quality in service and product offers. In fact, leading big data experts have sensationally declared a data scientist as the sexiest job in the 21st century" in their recent *Harvard Business Review* article (Davenport and Patil 2012, p. 70). Despite the assertion that the job of a data scientist is "the hottest job title", however, there is no rigorous definition of a data scientist in the literature. In this study, therefore, we examine two inter-related research questions: (1) What is a data scientist? And (2) What are the job skills required of data scientists to fulfil their roles in the enfolding big data environments? In this study we examine these two research questions through a rigorous examination of relevant academic literatures, non-profit organization reports such as the US National Science Foundation (NSF) and the Association for Computing Machinery (ACM), and the websites and industry reports of leading big data-driven organizations.

The structure of this paper is organized as follows: the next section describe our research methods for collecting relevant journals and review articles as well as relevant data from websites and industry reports to answer the questions. The third section presents a review of the literature on big data environments, data science, and data scientists. In the fourth section, we present our key findings. In the fifth section, we discuss the managerial and educational implications of our key findings that may influence our IS curriculum design and IS education in the future and may guide the growing trend for the deployment of data scientists in various big data analytics projects in the emergent big data environments. We also present our conclusions, including the research limitations of this paper and the future research directions for empirical research on expected roles and skills of data scientists.

METHODOLOGY

In this research we have addressed the two research questions on roles and skills of data scientists in the enfolding big data environments. In order to answer these questions, we first conducted a systematic review of the existing literature. A systematic search accumulates a relatively complete census of relevant literature (Webster and Watson 2002). Webster and Watson (2002) recommend a normative approach which has three structured steps to determine the source material for review: (1) the major contributions are likely to be in the leading journals; (2) going backward by reviewing the citations for the articles identified in step 1 to determine prior articles we should consider, and finally (3) going forward by using any journal database to identify articles identified in the previous two steps. In academic practice, the usefulness of a piece of research often is evaluated by its uptake by other researchers and not by the fact it has been published. In contrast, there is an alternative and perhaps less time-consuming approach to identify the source material for review: "One, albeit very imperfect, way of looking at this is the use of citation counts" (Ginieis et al. 2012, p. 1).

First, we used "data scientist" as the primary keyword for our search strategy. Second, we identify six major academic databases through which we search highly cited journal articles on data scientists: SCOPUS, Web of Science, IEEE, Springer, Science Direct, and ProQuest Central. Although there exist the differences in search engines in use across the databases, we consistently employed the following generic query strategy: (Title OR Abstract) CONTAINS ("data scientist") AND (Publication Year) = (2005-2014) AND (Publication Type) = (Journal Article). This search strategy discovered 99 published journal articles as shown in Figure 1 and Figure 2. We have excluded journal articles written in other languages than English.

Although systematic search was employed, as suggested by Webster and Watson (2002) and Ginieis et al. (2012), our initial search results showed insufficient literature on the roles and skills of a generic data scientist roles and its corresponding professional and skills requirements. We hold that this lack of the academic literature may well be a function of the enfolding big data environments as being relatively new (Cukier and Mayer-Schönberger 2013) and that academic research has not caught up with enfolding big data business practices in

some of big data-driven companies. Subsequently, we expanded our search to include non-profit organization reports, such as the US National Science Foundation (NSF) and Association for Computing Machinery (ACM), and the websites and industry reports of leading big data-driven organizations. In this way, we augment our academic literatures with non-academics information sources to capture essential information relatively comprehensively, despite the newness of big data environments in general and data science issues under investigation.

For our second-phase expanded search, we used Google Scholar with a generic keyword query on "data scientist". Although Google scholar search results yielded some redundant articles, it provided us with relevant and useful websites which included known big data industry players and their publications on a data scientist as they viewed: namely, IBM, SAS, Gartner, Fortune, Forbes, Microsoft, and The Data Warehousing Institute, among others.

LITERATURE REVIEW

The Rise of Big Data and the Enfolding Big Data Environments

The rise of big data in the Internet age has transforming business environments into big data environments which are characterized by volume, velocity, and variety (McAfee and Brynjolfsson 2014) and the increasing power and importance of predictive big data analytics in these enfolding big data environments (Hayashi 2014). According to Davenport (2014, p. 1, *italics* added for emphasis), big data is defined dominantly not only by its size but also by its lack of structure: "Big data refers to data that is too big to fit on a single server, too unstructured to fit into a row-and-column database, or too continuously flowing to fit into a static data warehouse. While its size receives all the attention, *the most difficult aspect of big data really involves its lack of structure*." While there are other similarly insightful definitions of big data, there is no rigorous definition of big data in prior research (Mayer-Schönberger and Cukier 2014).

However, for our purpose in this paper, we adopt a forward-looking and functional definition which is relevant and applicable to different big data projects. Mayer-Schönberger and Cukier (2014, p. 6), authors of *Big Data*, argue: "big data refers to things one can do at a large scale that cannot be done at a smaller one, to extract new insights or create new forms of value, in ways that change markets, organizations, the relationship between citizens and governments, and more." While we do not dismiss other similar or competing definitions of big data in the literature, our central focus of this paper is on developing a better understanding of emergent roles of data scientists who operate in big data environments. In consequence, we do not further discuss these definitions in this paper.

Data Science

While the terms big data and data scientist are relatively new, the term "data science" is not new at all. In fact, the conceptions of data science have been extensively examined in diverse academic studies since early 1970s. However, in 1996, for the first time, the importance of "data science" as a field of study was officially recognized and included in the title of the academic conference "*Data Science*, Classification, and Related methods" (International Federation of Classification Societies 1996, *italics* added for emphasis). The number of studies articulated the evolving field of data science, for example, a publication entitled: "Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics" (Cleveland 2001). A new journal, *Data Science Journal*, was also launched (Committee on Data for Science and Technology 2002).

A report entitled: "Long-lived Digital Data Collections: Enabling Research and Education in the 21st Century," which aimed to develop the career track for data scientists in order to ensure that the research enterprise contains a necessary number of high-quality data scientists (National Science Board 2005). Another report by Joint Information Systems Committee entitled: "The Skills, Role & Career Structure of Data Scientists & Curators: Assessment of Current Practice & Future Needs" examined the role and career development of data scientists (Joint Information Systems Committee 2008). Borne et al. (2009) suggested that training the next generation to deal with data is needed in their paper: "The Revolution in Astronomy Education: Data Science for the Masses". In the same year, Drew Conway came up with Data Science Venn Diagram (Conway 2010), which shows that data science exists at the intersection of three different domains of knowledge and skills: math & statistics knowledge, substantive expertise, and hacking Skills. Similarly, data science has been construed as a combination of problem solving, data analysis, and computer hacking (Smith 2011).

It is noteworthy that high-level technical knowledge and skills has been commonly identified (e.g. hacking skills and computer hacking) together with high-level analytical skills normally required of academic researchers and scientists. In summary, it appears that the evolution of data science concepts corresponds to the diffusion process of innovations in computer technologies and computer-based information systems at the organizational level. It

also seems that it has gained significant attention of academics and managers in such areas as database marketing (BusinessWeek 1994) and data mining for extracting information from large structured databases, as data output has significantly increased in correlation with the increasing processing power of computers in the organization.

Growing Interest in a Data Scientist

The concept of the data scientist has gained significant attention in academic and managerial areas during the last few years. Of the six databases we examined with the search strategy described in our Methodology section, we found a total of 99 published articles published in the literature over nine years from 2005 to 2014. Figure 1 below shows a bar graph for an overall trend in academic interest in a data scientist. The bar graph shows the frequency distribution of these published data scientist studies across the six databases. Springer (33) leads the number of publications, which is followed by SCOPUS (27), ProQuest Central (25), Web of Science (9), IEEE (3) and Science Direct (2). Interestingly, despite the relative importance of high-level technical and computing knowledge and skills for a data scientist (Conway 2010; Smith 2011), the IEEE and Science Direct databases which traditionally include more technical than social science journals have the lowest numbers.

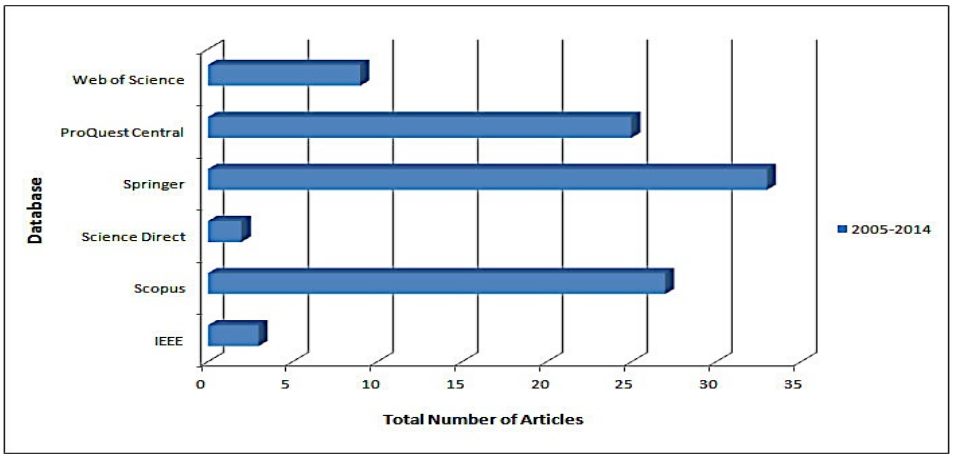

Figure 1: Overall trend across major academic databases 2005 – 2014

In comparison, Figure 2 below shows time series graph for dynamically changing trends over the same period across the six databases. The graph shows that overall academic research interest in a data scientist are on the rise, showing the six databases are publishing more studies in the recent years (2010-2012) vis-à-vis the earlier year (2005), even though conceptions and definitions of a data scientist are still new and emerging in the literature.

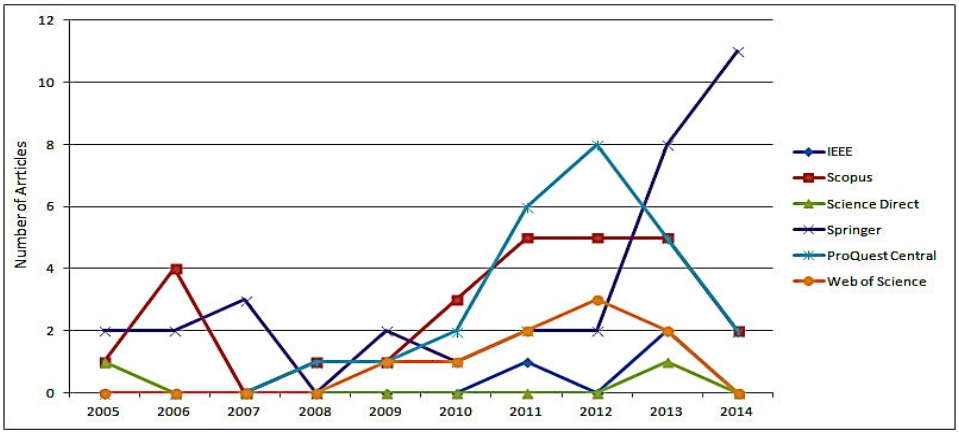

Figure 2: Yearly trend 2005 - 2014

Several researchers have investigated issues related to the job of a data scientist (National Science Board 2005; Swan and Brown 2008; Loukides 2010; Granville 2014) and its impact on business domain (Davenport and Patil 2012; Harris et al. 2013; Mohanty et al. 2013; IBM Website 2014). Despite the growing academic interest in a data scientist, however, no prior research has rigorously and systematically studied roles and skills of a data scientist in the organization in the enfolding big data environments organizations. In the next section, we will present our key results on existing definitions of a data scientist and job skills required of a data scientist we extracted from the diverse source material.

RESULTS

Existing Definitions of a Data Scientist

As we discussed in the Methodology section, we have adopted a normative approach which has three structured steps to determine the source material for review (Webster and Watson 2002). While our systematic search based on this approach has found 99 articles published in the six academic databases (for detailed discussion, see the Literature Review section), we did not find too many academic definitions of a data scientist in these articles. Subsequently, we expanded our search to include non-academic source material from industry websites.

A careful review of the different types of source material has identified the following 6 academic and 18 industry website definitions of what a data scientist is. Table1 lists first academic definitions of a data scientist and then industry definitions. Table 1 content is sorted by the Document's publication year in descending order.

Table 1. Collected Definitions of Data Scientist

Academic Document	Page	Definition: Data scientist is…
Granville (2014)	73	"not statisticians, nor data analysts, nor computer scientists, nor software engineers, nor business analysts. They have some knowledge in each of these areas but also some outside of these areas."
Viaene (2013)	16	"aren't super heroes with all the qualities and knowledge to make a data science project successful."
Dhar (2013)	1	"requires an integrated skill set spanning mathematics, machine learning, artificial intelligence, statistics, databases, and optimization, along with a deep understanding of the craft of problem formulation *to engineer effective solutions*."
Mohanty et al. (2013)	252	"the practitioners of the analytics models *solving business problems*. They incorporate advanced analytical approaches using sophisticated analytics and data visualization tools *to discover patterns in data*. In many cases, these practitioners work with well-established analytics techniques such as logistic regression methods, clustering methods, and classification methods *to draw insights from data*. These practitioners have *deep understanding of the business domain and apply that effectively to analyse data and deliver the outcomes in a business understandable intuitive manner* through advanced data visualization tools."
Davenport and Patil (2012)	73	"the people who understand *how to fish out answers to important business questions* from today's tsunami of unstructured information."
Choudhury (2008)	217	"these individuals, who are few in number at the moment, possess domain-specific knowledge and data management expertise. They act as the human interface between the library and the eScience projects."
Industry Document	**Page**	**Definition: Data scientist is…**
IBM Website (2014)	1	"somebody who is inquisitive, who can stare at data and *spot trends*. It's almost like a Renaissance individual who really wants to learn and *bring change to an organization.*"
Microsoft Website (2013)	1	"so companies need to do a lot with their data: gather, collate, store, transform, clean, analyse, explore, visualise, share and discover. The people who help organisations do this are data scientists. They *turn data into products, insights and stories by adding value to raw information*."
Harris et al. (2013) (Accenture Institute for High Performance)	3	"not only about data crunching. It's about *understanding the business challenge, creating some valuable actionable insights to the data, and communicating their findings to the business*."
	3	"the most common term for the often PhD-level experts who operate at the frontier of analytics, where data sets are so large and the data so messy that less-skilled analysts using traditional tools cannot *make sense of them*. But they are more precisely described as data engineer-scientist-manager-teachers."

Davenport (2012) (SAS)	1	"are hybrids of technologists and quantitative analysts."
Cooper (2012) (Adam Cooper Cetis Blogs)	1	"someone who wants *to know what the question should be*; embodies a combination of curiosity, data gathering skills, statistical and modelling expertise and strong communication skills. Brobst argues that the working environment for a data scientist should allow them *to self-provision data*, rather than having to rely on what is formally supported in the organisation, to enable them to be inquisitive and creative."
Laney (2012) (Gartner Website)	1	"expected to work more in teams, have a comfort and experience with 'big data' sets, and are skilled at communication. They also frequently require experience in machine learning, computing and algorithms, and are required to have a PhD nearly twice as often as statisticians."
Press (2012) (Forbes Website)	1	"an engineer who employs the scientific method and applies data-discovery tools *to find new insights in data*."
Lev-Ram (2011) (Fortune Website)	8	"*helps companies make sense of the massive streams of digital information they collect every day, everything from internally generated sales reports to customer tweets*. The gig which requires the specialist to capture, sort, and figure out what data are relevant is one part statistician, one part forensic scientist, and one part hacker."
Woods (2011a) (Forbes Website)	1	"someone who *can obtain, scrub, explore, model and interpret data*, blending hacking, statistics and machine learning. Data scientists not only are adept at *working with data, but appreciate data itself as a first-class product*."
Woods (2011b) (Forbes Website)	1	"examples of those rare professionals who bring in talents from a lot of different areas"
	1	"analytically-minded, statistically and mathematically sophisticated data engineers who *can infer insights into business and other complex systems out of large quantities of data*."
Woods (2011c) (Forbes Website)	1	"they are half hacker, half analyst, they use data *to build products and find insights*."
Woods (2011d) (Forbes Website)	1	"when most people think of a data scientist, they think of a statistician, a guy with 'analyst' in his title."
	1	"Or, someone who works in IT and manages the data warehouses. To do these jobs, you certainly needed programming skills; you probably needed advanced statistics skills, or some combination of those skills."
Loukides (2010) (An O'Reilly Radar Report)	8	"combine entrepreneurship with patience, *the willingness to build data products incrementally, the ability to explore, and the ability to iterate over a solution*. They are inherently interdisciplinary. They can tackle all aspects of a problem, from initial data collection and data conditioning to drawing conclusions."
Swan and Brown (2008) (Key Perspectives Report to JISC)	1	"people who work where the research is carried out – or, in the case of data centre personnel, in close collaboration with the creators of the data – and may be involved in creative enquiry and analysis, enabling others to work with digital data, and developments in data base technology."
National Science Board (2005)	17	"the information and computer scientists, database and software engineers and programmers, disciplinary experts, curators and expert annotators, librarians, archivists, and others, who are crucial to the successful management of a digital data collection."

Note: Italics are added for emphasis.

These 24 definitions listed in Table 1 show different conceptions of roles of a data scientist in the enfolding different big data environments. However, almost all definitions seem to describe *who* a data scientist is (*person* or *personal attributes*), *what* a data scientist produces and delivers (*product*), and *how* (*process* of producing desired outcomes). Of a total of 24 definitions listed in Table 2 bellow, we have highlighted by italics what a data scientist produces and delivers in the organization. The findings on data scientist definitions will be discussed in the next section.

Table 2. Common Attributes of Data Scientist

Attributes	IBM Website (2014)	Cooper (2012)	Davenport (2012)	Davenport and Patil (2012)	Dhar (2013)	Granvill (2014)	Harris et al. (2013)	Laney (2012)	Loukides (2010)	Microsoft Website (2013)	Mohanty et al. (2013)	National Science Board (2005)	Swan and Brown (2008)	Vangelova (2014)	Provost and Fawcett (2013)	Press (2012)	Lev-Ram (2011)	**Total**
Entrepreneurship/Business domain knowledge			x	x			x	x	x		x		x		x			8
Computer scientist			x	x		x						x		x		x	x	7
Effective communication		x	x	x	x	x	x					x						7
Creating valuable actionable insight			x	x			x		x	x	x	x						7
Curious	x	x		x					x				x		x			6
Statistical and modelling		x			x	x							x	x	x			6
Data visualisation			x	x								x	x		x			5
Mathematics			x	x	x			x						x				5
Data management		x			x	x						x						4
Analytical			x			x							x		x			4
Software engineer			x	x		x						x						4
PhD qualification			x	x				x					x					4
Programmer												x				x		2
Understanding business challenges							x				x							2
Machine learning					x			x										2
Economics				x									x					2
Technologist & quantitative analysts			x															1
Outside of IT knowledge						x												1
Works in a team								x										1
Experience with big data sets								x										1
Optimisation					x													1
Interdisciplinary									x									1
Artificial intelligence					x													1

DISCUSSION

This paper has raised the two research questions: (1) What is a data scientist? And (2) What are the job skills required of data scientists to fulfil their roles in the enfolding big data environments?

Definitions of a Data Scientist

Of the twenty-four definitions we collected from the diverse source material, one of more comprehensive academic definitions of what a data scientist produces and delivers to the organization in a manner he/she adds value is found in Mohanty et al. (2013, p. 252): Data scientists are "the practitioners of the analytics models *solving business problems*. They incorporate advanced analytical approaches using sophisticated analytics and data visualization tools *to discover patterns in data*.

In many cases, these practitioners work with well-established analytics techniques such as logistic regression methods, clustering methods, and classification methods *to draw insights from data*. These practitioners have deep understanding of the business domain and apply that effectively *to analyse data and deliver the outcomes in a business understandable intuitive manner* through advanced data visualization tools." This definition underscores the role of a data scientist as someone who solves business problems by discovering patterns or trends in data, drawing insights from data and communicating these big data-driven insights to business decision makers in a manner that they can understand the insights from big data. Other academic definitions (Dhar 2013; Davenport and Patil 2012) also underscore the important role of a data scientist in engineering business solutions and generating answers to important business questions. While the all academic definitions of a data scientist refer to the term big data, the messy and unstructured nature of big data has not been sufficiently articulated except in one study (Davenport and Patil 2012).

In contrast, two of more insightful industry definitions of a data scientist do mention the nature of big data. According to Lev-Ram (2011) cited on (Fortune Website, p. 8): A data scientist "*helps companies make sense of the massive streams of digital information they collect every day, everything from internally generated sales reports to customer tweets*." In this definition, the diversity of internal structured big data such as sales reports and external unstructured bid data such as customer-generated tweets from the company's social media platforms is mentioned, which is critically important. It is because big data is characterized by volume, velocity, and variety (McAfee and Brynjolfsson 2014). It is also because according to Davenport (2014, p. 1, *Italics* added) "*the most difficult aspect of big data really involves its lack of structure*."

Similarly, Harris et al. (2012, p. 3) posted on the Accenture Institute for High Performance website identify the messy nature of big data: "the most common term for the often PhD-level experts who operate at the frontier of analytics, where data sets are so large and the data so messy that less-skilled analysts using traditional tools cannot *make sense of them*." Moreover, some of the definitions include desirable personal attributes of a data scientist. For example, Anjul Bhambhri, vice president of big data products at IBM defined "data scientist" as "somebody who is inquisitive, who can stare at data and spot trends. It's almost like a Renaissance individual who really wants to learn and bring change to an organization" (IBM Website 2014).

Skills Required of a Data Scientist

As Table 2 in our Results section shows, we found that out of the 23 attributes listed in Column 1, six attributes had been identified in six or more articles. Therefore, we hold that these six attributes are more commonly perceived by academic researchers as important for the job of a data scientist. We can identify an initial set of job skills required of a data scientist. A set of six common attributes are: (1) Entrepreneurship and business domain knowledge, (2) Computer scientist, (3) Effective Communication skills, (4) Create valuable and actionable insights, (5) Inquisitive and curious, and (6) Statistics and modeling.

Based on a comprehensive systematic review of the relevant source material, it is clear that, while there is rather a very small set of academic definitions was uncovered, we can identify the six common attributes of desired skills required of a data scientist in the organization operating in the enfolding big data environments. Furthermore, we can categorize these attributes into three categories: *Person* (or personal attributes which a data scientist brings to the organization as a person), *Process* (or analytical and technical attributes required to perform expected roles of a data scientist in the organization), and *Product* (or a combination of diverse attributes required of a data scientist to produce desired outcomes that create value to the organization). A set of skills required of a data scientist is shown in Figure 3 below.

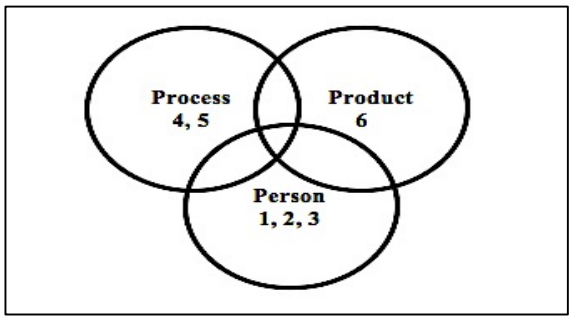

Figure 3: Data scientist skills requirements

Currently there is a serious demand for data scientists and it is predicted that the job of a data scientist will grow enormously in demand during the next few years. In their *Harvard Business Review* article "The Sexiest Job of the 21st Century," Davenport and Patil (2012) mentioned that no university program has yet been designed to produce data scientists. More recently, however, form 2012 to 2014, many universities have launched a variety of postgraduate programs in data science, although some of these programs are not named data science. For example, in 2012 "Data Science 101 - learning to become a data scientist" website posted a list of "185" universities with data science degrees.

According to *McKinsey Global Institute Report* (2011) and Accenture Institute for High Performance website (2013), data science talents are in high demand and this has been driven by the success of the major Internet companies such as Google, LinkedIn, Facebook and Amazon, all of which have made their marks by using data creatively to produce big data-driven products (Manyika et al. 2011; Harris et al. 2013). Accenture Institute for High Performance website (2013) also indicated that in the United States alone, it is expected to create around 400,000 new data science jobs between 2010 and 2015 (Harris et al. 2013). However, it is likely to produce only about 140,000 qualified graduates to meet the market demand (Harris et al. 2013). At the same time, McKinsey

Global Institute Report (2011) indicated that by 2018 as much as 140,000-190,000 deep analytical talent positions and 1.5 million data-aware managers in the United States alone will be needed to create business value from big data which can transform the way we work and think (Manyika et al. 2011).

This skills shortage exists because data scientists require a scarce combination of diverse skills as Tables 1 and 2 clearly showed. Patil in his Building Data Science Teams Report (2011) mentioned some roles of data scientists can fall in decision sciences and business intelligence fields. He also listed some skills that data scientists should have: technical expertise, curiosity, storytelling and cleverness. Surprisingly, he indicated that data scientists do not need to have a background in computer science. This contradicts with our finding on the importance of computer science skills in Table 2. In contrast, based on Interviews with Professionals Using Science in the Workplace, Luba Vangelova indicated that most data scientists have a background in computer science, mathematics, statistics, or one of the natural or social sciences that relies on quantitative methods (Vangelova 2014).

Despite the recent sensational declaration of a data scientist as "the sexiest job of the 21st century" (Davenport and Patil 2012), however, there has been no rigorous definition of what a data scientist is, and what job skill requirements this hottest job title may need remain unclear. In this paper we examined comprehensively the diverse source material to extract definitions, analyze and categorize them. By so doing and developing a simply classification scheme for further advancing our knowledge of emergent roles and skills required of data scientists in the enfolding big data environments, we have shown a step towards reducing the gap in the literature and developing a better understanding of roles and skills of data scientists in the big data environments. However, our research has some limitations, which include the need for empirical validation of our key findings through primary data collection. Our future research includes a quantitative research to validate the proposed critical skills of data scientists.

In conclusion, we hold that the potential of big data in the organization to transform the business, the government and relationships between the government and citizens will require a critical mass of data scientists as game changers. The lack of clear understanding of roles and skills of data scientists in the enfolding big data environments and the current level of the growing skills shortage of data scientists are the key issues in hindering the organization from realizing the full potential of big data. The dynamically changing big data environments indicate the urgent need for forward-looking plans for urgently assessing the existing information systems programs design and development guidelines to meet the market demand.

REFERENCES

Borne, K.D., Jacoby, S., Carney, K., Connolly, A., Eastman, T., Raddick, M.J., Tyson, J.A., and Wallin, J. 2009. "The Revolution in Astronomy Education: Data Science for the Masses," *Joint Information Systems Committee (JISC)*, Retrieved 3 July, 2014, from http://arxiv.org/pdf/0909.3895v1.pdf

BusinessWeek, 1994. "Database Marketing", Retrieved 15 July, 2014, from http://www.businessweek.com/stories/1994-09-04/database-marketing

Choudhury, G.S. 2008. "Case Study in Data Curation at Johns Hopkins University," *Library Trends* (57:2), pp 211-220.

Cleveland, W.S. 2001. "Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics", *International Statistical Review/Revue Internationale de Statistique* (69:1), pp 21-26.

Committee on Data for Science and Technology (CODATA), 2002. Retrieved 15 July, 2014, from http://www.codata.org/dsj/index.html

Conway, D. 2010, "The Data Science Venn Diagram," Retrieved 15 July, 2014, from http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

Cooper, C. 2012. "Analytics and Big Data - Reflections from the Teradata Universe Conference 2012, Retrieved 20 July, 2014, from http://blogs.cetis.ac.uk/adam/2012/04/27/analytics-and-big-data-reflections-from-the-teradata-universe-conference-2012/

Cukier, K., and Mayer-Schönberger, V. 2013. "The Rise of Big Data: How It's Changing the Way We Think About the World," *Foreign Affairs*, (92:3), pp. 28-40.

Data Science 101 - Learning to Become a Data Scientist Website, 2014. "Colleges with Data Science Degrees," Retrieved 10 July, 2014, from http://datascience101.wordpress.com/2012/04/09/colleges-with-data-science-degrees/

Davenport, T. H. 2014. *Big Data @Work Dispelling the Myths, Uncovering the Opportunities*, Harvard Business Review Press, Boston.

Davenport, T.H. 2012. "Help wanted: Data scientist", *SAS*, Retrieved 6 August, 2014, from http://www.sas.com/en_us/news/sascom/2012q4/data-scientist.html

Davenport, T.H., and Patil, D.J. 2012. "Data Scientist: the Sexiest Job of the 21st Century," *Harvard Business Review* (90:10), pp 70-76.

Dhar, V. 2013. "Data Science and Prediction", *Association for Computing Machinery*, Retrieved 5 August, 2014, from http://cacm.acm.org/magazines/2013/12/169933-data-science-and-prediction/fulltext

Ginieis, M., Sánchez-Rebull, M.V., and Campa-Planas, F. 2012. "The Academic Journal Literature on Air Transport: Analysis Using Systematic Literature Review Methodology", *Journal of Air Transport Management*, (19:1), pp 31-35.

Granville, V. 2014. *Developing Analytic Talent: Becoming a Data Scientist*, John Wiley and Sons, Incorporated, US.

Harris, J.G., Shetterley, N., Alter, A.E., and Schnell, K. 2013. "The Team Solution to the Data Scientist Shortage," *Accenture Institute for High Performance*, Retrieved 11 July, 2014, from http://www.accenture.com/SiteCollectionDocuments/PDF/Accenture-Team-Solution-Data-Scientist-Shortage.pdf

Hayashi, A. M. 2014. "Thriving in a Big Data World," *Sloan Management Review*, (55:2), pp 35-39.

IBM Website, 2014. "What is a Data Scientist". Retrieved 20 July, 2014, from http://www.01.ibm.com/software/data/infosphere/data-scientist/

International Federation of Classification Societies (IFCS), 1996. "Data Science, Classification, and Related Methods," *Proceedings of the Fifth Conference of the International Federation of Classification Societies (IFCS-96), Kobe, Japan, March 27–30, 1996.*

Joint Information Systems Committee (JISC), 2008. "Skills, Role & Career Structure of Data Scientists & Curators: Assessment of Current Practice & Future Needs", Retrieved 8 July, 2014, from http://www.jisc.ac.uk/whatwedo/programmes/digitalrepositories2007/dataskillscareers.aspx

Laney, D. 2012. "Emerging Role of the Data Scientist and the Art of Data Science", *Gartner Website*, Retrieved 16 July 2014, from http://blogs.gartner.com/doug-laney/defining-and-differentiating-the-role-of-the-data-scientist/

Lev-Ram, M. 2011. "Data Scientist: The Hot New Gig in Tech", *Fortune Websit*e, Retrieved 30 July, 2014, from http://fortune.com/2011/09/06/data-scientist-the-hot-new-gig-in-tech/

Loukides, M. 2010. "What is Data Science?," *An O'Reilly Radar Report*, Retrieved 26 August, 2014, from http://origin-www2.informatica.com/INFA_Resources/ar_strata2011_what-is-data-science_1781.pdf

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., and Byers, A.H. 2011. "Big Data: The Next Frontier for Innovation, Competition, and Productivity," *McKinsey Global Institute Report*, Retrieved 5 July, 2014, from http://bit.ly/McKinseyBigDataReport

Mayer-Schönberger, V., and Cukier, K. 2014. Big Data, Houghton Mufflin Harcourt, New York.

McAfee, A., and Brynjolfsson, E. 2014. "Big Data: The Management Revolution," *Harvard Business Review*, (90:10), pp 60-66, 68, 128.

Microsoft Website, 2013. "What is a data scientist?". Retrieved 6 August, 2014, from http://www.microsoft.com/en-gb/enterprise/enterprise-insights-blog/articles/what-is-a-data-scientist.aspx#fbid=J5ZhWyG9gG9

Mohanty, S, Jagadeesh, M., and Srivatsa, H. 2013. *Big Data Imperative: Enterprise Big Data Warehouse, BI Implementations and Analytics*, Apress.

National Science Board, 2005. "Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century: Chapter Two: The Elements of the Digital Data Collections Universe", Retrieved 6 August, 2014, from http://www.nsf.gov/pubs/2005/nsb0540/nsb0540_4.pdf

Naur, P. 1974. *Concise Survey of Computer Methods*, Studentlitteratur, Lund, Sweden.

Patil, D.J. 2011. "Building Data Science Teams: Data Science Teams Need People with the Skills and Curiosity to Ask the Big Questions," Retrieved 6 August, 2014, from http://radar.oreilly.com/2011/09/building-data-science-teams.html

Press, G. 2012. "Data Scientists: The Definition of Sexy", *Forbes Websit*e, Retrieved 17 July 2014, http://www.forbes.com/sites/gilpress/2012/09/27/data-scientists-the-definition-of-sexy/

Provost, F., and Fawcett, T. 2013. *Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking*, Silicon Valley Data Science, Mountain View, CA, USA.

Scott Morton, M. S. 1991. MIT90s Model, *Harvard Business Review* Press, Boston.

Smith, D. 2011. "'Data Science': What's in a name?." Retrieved 18 July, 2014, from http://blog.revolutionanalytics.com/2011/05/data-science-whats-in-a-name.html

Swan, A., and Brown, S. 2008. "The Skills, Role and Career Structure of Data Scientists and Curators: An Assessment of Current Practice and Future Needs," *Key Perspectives Report to the JISC*, Retrieved 28 July, 2014, from http://repository.jisc.ac.uk/245/3/dataskillscareersfinalreport.pdf

The Data Warehousing Institute, "Are Advanced Analytics Possible without a Data Scientist?", Retrieved 15 July 2014, http://tdwi.org/webcasts/2014/06/are-advanced-analytics-possible-without-a-data-scientist.aspx

The Data Warehousing Institute, "The Doctor Is In: The Role of the Data Scientist for Analyzing Big Data", Retrieved 15 July 2014, from http://tdwi.org/webcasts/2013/05/the-doctor-is-in-the-role-of-the-data-scientist-for-analyzing-big-data.aspx

Trounson, A. 2014. "Knowledge Economy at Risk of Falling Behind Low-Cost Rivals," *The Australian*, 6 August, Retrieved 7 August 2014, from http://www.theaustralian.com.au/higher-education/knowledge-economy-at-risk-of-falling-behind-lowcost-rivals/story-e6frgcjx-1227014554629

Vangelova, L. 2014. "Career of the Month," *The Science Teacher* (81:5), pp 70.

Viaene, S. 2013. "Data Scientists Aren't Domain Experts," *IT Professional* (15:6), pp 12-17.

Webster, J., and Watson, R.T. 2002. "Analyzing the Past to Prepare for the Future: Writing a Literature Review", MIS Quarterly (26:2), pp xiii-xxiii.

Woods, D. 2011a. LinkedIn's Daniel Tunkelang on What Is a Data Scientist?," *Forbes Website*, Retrieved 20 July, 2014, from http://www.forbes.com/sites/danwoods/2011/11/27/linkedins-monica-rogati-on-what-is-a-data-scientist/

Woods, D. 2011b. "EMC Greenplum's Steven Hillion on What Is a Data Scientist?," *Forbes Website*, Retrieved 20 July, 2014, from http://www.forbes.com/sites/danwoods/2011/10/11/emc-greenplums-steven-hillion-on-what-is-a-data-scientist/

Woods, D. 2011c. "LinkedIn's Monica Rogati on What Is A Data Scientist?," *Forbes Website*, Retrieved 20 July, 2014, from http://www.forbes.com/sites/danwoods/2011/11/27/linkedins-monica-rogati-on-what-is-a-data-scientist/

Woods, D. 2011d. "Tableau Software's Pat Hanrahan on What Is a Data Scientist?," *Forbes Website*, Retrieved 20 July, 2014, from http://www.forbes.com/sites/danwoods/2011/11/30/tableau-softwares-pat-hanrahan-on-what-is-a-data-scientist/

COPYRIGHT