DECISION SCIENCES INSTITUTE

Big Data Analytics and Data Science Undergraduate Degree Programs

(Full Paper Submission)

Cheryl Aasheim
Georgia Southern University
caasheim@georgiasouthern.edu

Susan Williams
Georgia Southern University
rebstock@georgiasouthern.edu

Paige Rutner
Georgia Southern University
prutner@georgiasouthern.edu

Adrian Gardiner
Georgia Southern University
agardine@georgiasouthern.edu

ABSTRACT

With a strong emerging demand for employees who are skilled and knowledgeable in big data and analytics, some universities have started to offer programs in big data, analytics, and data science. This paper presents a review of such endeavors within the United States at the undergraduate level. These types of programs are still in the embryonic stage, with only 21 from a sample of 86 universities offering related curricula. Results indicate that business analytics and data analytics related programs are more commonly associated with business schools, while data science programs are more commonly associated with computer science academic units.

KEYWORDS: Big data, Analytics, Data science, Curriculum

INTRODUCTION

Big data is a relatively new, hot topic that will only increase in importance over time (Davenport & Patil, 2012; Collett, 2011; Preimesberger, 2011). This topic and the terms *big data, data analytics,* and *data science* have arisen as organizations have begun to collect data sets that exceed the capacity of traditional data storage systems and analytical tools (Provost & Fawcett, 2013). There are many sectors in the U.S. — including e-government and politics, science and technology, smart health and well-being, and security and public safety — that could make significant advances by applying analytic techniques on the large quantity of available unstructured data (Chen, Chiang & Storey, 2012). Yet, in the 'era of big data,' potential advances in these fields may be hindered by the scarcity of people with the unique skill set (math, statistics, probability, programming, and business skills) needed to effectively deal with the complexity of this task (Davenport & Patil, 2012). In this respect, a much-quoted report from McKinsey & Company (Manyika, Chui, Brown, Bughin, Dobbs, Roxburgh & Byers, 2011, p. 3) on big data implications for business offers an ominous warning that: "*The United States alone* *faces a shortage of 140,000 to 190,000 people with deep analytical skills as well as 1.5 million managers and analysts to analyze big data and make decisions based on their findings."*

If academicians wish to heed this warning, higher education needs to quickly respond to the rise of big data and offer adequate relevant educational opportunities. To this end, this paper is motivated by concerns over the state of higher education in big data and data science and presents a review of existing undergraduate degree programs. In doing so, this paper also reviews the current structure of relevant curricula, and the positioning of such programs amongst various academic units. As there is no extant academic map describing these programs, this paper also adds to the literature by presenting an overview of the current state of big data, data science, and analytics academic offerings.

LITERATURE REVIEW

The following literature review provides a broad introduction to big data and highlights the significance of big data as an emerging and independent field of study. The literature review begins with a review of the history of big data and analytics as well as a definition of the terms *big data*, *data science* and *analytics*. This is followed by a section on the current state of the industry with a closer look at legal and ethical issues in the subsequent section. Novel applications of big data are then presented followed by a section on the overall implications for academia.

The Rise of Big Data and Analytics

In 2001, Laney (2001) identified a need for new data management practices that would address fundamental changes in the "volume, velocity, and variety" of data – arising largely at that time from the explosion of e-commerce applications. Laney's "3 V's" became a *de facto* definition for the term "big data" – with "volume" signifying the rapid increase in the amount of data available, "velocity" signifying that much of this data arrives at or near real-time, and "variety" signifying the plethora of incompatible formats and inconsistent semantics across data arriving from different sources. Laney contended that these conditions would push traditional data management practices to their limits, eventually rendering typical database technologies ineffective for storing and managing "big data".

Since that time, the concept of "big data" has broadened to include an even wider variety of unstructured data, as well as the tools and techniques required to analyze this data. Chen et al. (2012) view the evolution of business intelligence and analytics (BI&A) as moving through three distinct phases: BI&A 1.0 – which focuses on the storage and analysis of structured data, using relational database management system (RDBMS), data warehousing (DW), extract, transform and load (ETL) and online analytical processing (OLAP) technologies, along with traditional data mining and statistical analysis; BI&A 2.0 – which includes web-based technologies and unstructured data from "various online social media such as forums, online groups, web blogs, social net-working sites, social multimedia sites (for photos and videos), and even virtual worlds and social games" (p. 1167); and BI&A 3.0 – which centers on mobile and sensor-based technologies (e.g., the "Internet of Things"), with applications becoming location and context-aware.

In terms of defining 'big data,' Chen et al. (2012, p. 1166) offer this definition: *"… the data sets and analytical techniques in applications that are so large (from terabytes to exabytes) and* *complex (from sensor to social media data) that they require advanced and unique data storage, management, analysis, and visualization technologies*." Moore, Eyestone, and Coddington (2013, p. 1) present a similar definition of big data as being: "… *vast amounts of diverse data, both structured and unstructured, that organizations can quickly access and analyze using innovative new tools that help pinpoint opportunities to better manage and improve value."* Therefore, attempts to define big data have emphasized: (1) the size, variability and complexity of the data; (2) the techniques and technologies required for analysis; and (3) that the goal of analysis is to improve decision-making and add value. Leveraging these definitions, big data can be defined as the: "*data sets that are so large that they require advanced and unique data storage, management, analysis, and visualization technologies to pinpoint opportunities for improving decision making and adding value.*" The terms *data analytics* and *data science* refer to the analytical techniques required to extract useful information from big data (Chen et al., 2012; Provost & Fawcett, 2013).

The State of the Industry

Given the media buzz associated with big data, a number of white papers and surveys have been published by leading consulting groups and vendors. These publications have attempted to identify trends, needs and current industry practices with respect to big data.

Eckerson (2011) attributed the need for big data analytics to a number of factors. These include: (1) changing data types, (2) hardware advances, and (3) a return to in-sourcing data-critical business processes. Eckerson identified advances such as Hadoop and MapReduce, as well as BI tools such as PowerPivot, and analytical technologies such as Massively Parallel Processing (MPP) and Complex Event Processing (CEP) as key components of the next-generation BI architecture. He proposed that success with big data requires the "right culture, people, organization, architecture and technology" (p. 11), and recommended the use of a bottom-up architecture that encourages ad-hoc data exploration, rather than one that is top-down and report-driven.

In addition to the "3 V's" – volume, velocity, and variety – associated with big data, some authors have suggested that a fourth "V" – value – is critical for organizations seeking to benefit from this trend (Lycett, 2013). In addition to collecting and storing immense volumes of data, organizations must also use that data to gain insights into their processes and to make determinations as to how to best allocate scarce resources. The ability to derive value from this type of data-driven decision making is emerging as a key factor which sets apart high-performing organizations from their competitors (Lavalle, Lesser, Shockley, Hopkins & Krushwitz, 2011). An industry survey found that "top-performing organizations use analytics five times more than lower performers" (Lavalle et al, 2011, p 23). Another industry study found that "the more companies characterized themselves as data-driven, the better they performed on objective measures of financial and operational results" (McAfee & Brynjolfsson, 2012, p. 64). Big data has been identified as relevant to leaders in all sectors and a key basis for competition and growth for individual firms (Manyika et al., 2011). Manyika et al. (2011) suggest that the use of big data enables productivity growth, particularly in sectors such as computer/electronic products, finance, insurance and government.

A study commissioned by IBM Global Business Services (Balboni, Finch, Reese & Shockley, 2013) surveyed 900 business and IT executives in 70 countries to examine how organizations are leveraging big data to create value. They identified nine attributes of organizations that have successful analytic initiatives, with several of these attributes corresponding to those identified by Eckerson (2011). Balboni et al. (2013) found that success in big data initiatives requires the right culture, people (executive sponsors and analytic expertise), organization (funding and organizational confidence), architecture and technology (platform and data), understanding the sources of value generation within the organization, and an ability to evaluate the impact of actions and decisions on business outcomes. They also found some evidence of the emergence of Chen et al.'s (2012) BI&A 3.0, in that some leaders in big data are actively implementing mobile and cloud-based hardware solutions, as well as free-form text and voice analytics.

Several of these studies, along with others in the popular press, identify general trends such as the explosive growth in stored data and the increasing availability of cheap storage (Collett, 2011; Tallon, 2013; Wittman, 2013), while others discuss important legal and ethical issues associated with big data (Nash, 2012a; Poole, 2013; Tallon, 2013; Wigan and Clarke, 2013). There are additional challenges that will need to be addressed, such as privacy, security, intellectual property, liability, adoption of new technologies/techniques, and organizational change (Manyika et al, 2011).

In all, the message from industry serves to reinforce the view that the big data 'era' has arrived, and will have significant impacts on modern organizations, as well as changing the required skills of knowledge-worker employees.

Legal and Ethical Issues Surrounding Big Data

The ability to link information from multiple sources, especially when that information is collected by different organizations, or for different purposes, raises significant legal and ethical questions. Privacy issues, in particular, take on a new dimension. Nash (2012a) reported that *Equifax* is creating analytics products from 800 billion business and consumer records. In addition to establishing credit scores, it uses these records to identify "non-obvious" relationships which are then used to reduce risk and improve marketing for its business customers. This raises the question as to whether data collected for one purpose should be used for another, as well as the potential consequences of combining information from different data sets collected for different purposes.

Wigan and Clarke (2013) identify similar concerns regarding the "exploitation" of data, and call for the development of new legal structures and business practices to protect the rights of individuals. The authors define big data as, "not only to specific, large datasets, but also to data collections that consolidate many datasets from multiple sources, and even to the techniques used to manage and analyze the data" (p. 46). The authors point out that such collections of data may be of poor quality and facilitate *dataveillance* — a term introduced in 1988, which "offers a more economical method for monitoring individuals than physical and electronic surveillance" (p. 48). Moreover, Wigan and Clarke offer several examples, including consumer profile databases developed by "data brokers," the use of loyalty cards, personal data volunteered through social media, as well as flows of data streaming from sensors, smart meters and aerial surveillance. Along these lines, Poole (2013) cites several cases where privacy issues related to big data arise. For example: (1) Amazon knows what consumers read, what they highlight, and where they give up on reading a book – which may lead to the development of a system that automatically writes books that consumers are statistically inclined to like; (2) Apple keeps Siri inquiries for two years; (3) the ability to analyze what someone “likes” on Facebook may make it possible to discern someone’s sexual orientation or history of drug use.

At the very least, big data raises the ante with respect to privacy issues, particularly when used to identify unusual (and perhaps, questionable) combinations of data (Tallon, 2013). A survey of 974 adults found that 92% believe there should be a law forcing websites and advertising companies to delete stored information about an individual — yet only 50% bother to read the privacy policies of those companies, and only 30% were able to correctly answer five true/false questions about online privacy (Nash, 2012a).

Tallon (2013) identified a need for data governance practices that balance the risks, costs and benefits of big data applications. Nash (2012a) suggests the biggest problem in achieving that balance is putting a value on a company’s data, particularly given that data has a limited useful life. Moreover, Tallon suggests that there is a “time value” of information over that useful life, which means that valuing information assets is a constantly moving target. This forces managers to continually assess whether the costs and risks to retain data justify the benefit that may be derived from that data in the future. From this standpoint, “storage investments are the premiums required to manage risk” (p. 33).

Big Data Applications

The applications of big data span across virtually every industry and business function (Manyika et al., 2011). Several of the applications cited here fall into the categories of market intelligence, fraud detection and financial risk. Much of the work in analytics research aims to identify new algorithms or novel combinations of algorithms, methods and techniques to improve predictions. Other research examines the uses of such tools to help address business problems and capitalize on business opportunities.

Deriving greater market intelligence and understanding consumer behavior are ripe areas for big data applications. Examples include efforts to determine consumer sentiments from web forum postings (Chen, 2010) or to determine how to retain cellular phone customers upon expiration of their contracts (Provost & Fawcett, 2013). Improving customer recommendations is another potential application of big data analytics. Sahoo, Singh, and Mukhopadhyay (2012) modeled the problem of making personalized recommendations when user preferences are changing. Using real world data sets and simulations, the authors showed that when user preferences are changing, their model outperformed alternative approaches. Analysis of consumer behavior obtained through e-commerce sites can lead to new product development such as Netflix’s move to expand streaming content and their development of the House of Cards television series (Lycett, 2013). Customer loyalty cards represent another rich source of big data which can help with marketing efforts. Collett (2011) cited two examples of organizations using big data to discern buyer behavior. They first matched TV ads in a home with actual buying behavior in retail stores, using loyalty card data, to identify correlations between ads and purchases. Another organization used a customer loyalty database of purchase history data for more than 190 million grocery shoppers to help major consumer goods manufacturers and supermarket chains predict buying behavior and the potential level of new product interest. Similar techniques have also been used to understand the information that customers report about themselves. Park, Huh, and Han (2012) examined individuals’ social circles and communication patterns within their circles to develop a framework for determining the accuracy of self-reported customer profiles. A combination of methods was used, including query processing, statistical inference, social network analysis, and user profiling. On a dataset consisting of more than 20 million mobile call transactions, their model consistently outperformed alternative approaches.

Prevention and detection of fraud is an ongoing concern for businesses of all types. Nash (2012b) reported on a study conducted by the Association of Certified Fraud Examiners that examined the role of IT in fraud prevention, especially that associated with theft of proprietary data or trade secrets. IT controls and big data analytics were used to help identify suspicious transactions, thereby facilitating fraud detection. Visualization and analytical techniques can help to identify anomalies that might be caused fraudulent credit card transactions or insurance claims (Griffin, 2012). Researchers are attempting to develop techniques to improve fraud prevention and detection. Abbasi, Albrecht, Vance and Hansen (2012) examined financial fraud using external, publicly available data. These researchers designed and developed a BI-based meta-learning framework for enhanced fraud detection and demonstrated that this approach improved performance over existing methods.

There are numerous other examples of the applications of big data and associated analytical techniques in a wide variety of contexts. For example, Chau and Xu (2012) applied analytics to the task of deriving business intelligence from published blogs using text, web and network analytics. They explored a method / approach to mining blogs and demonstrated how this method could be used to determine how bloggers interact with one another, who are the most influential bloggers, and whether there is a relationship between blogger types and content. In the area of financial services, Hu, Zhao, Hua and Wong (2012) examined methods and tools for modeling systematic risks of bank failures. They developed an approach that incorporates a network view of modern portfolio theory into a BI algorithm to analyze risk at the individual bank level, thereby assisting central banks in managing those risks. Lau, Liao, Wong and Chiu (2012) examined the success/failure rate of cross-border mergers and acquisitions. The authors suggest that one reason for failures is overlooking the nonfinancial aspects involved (such as sociocultural and political issues) – much of which is available as qualitative data. They proposed the design of a “due diligence scorecard” that gathers intelligence from the web and leverages it in the mergers and acquisitions decision making process.

A full exploration of the breadth of applications of big data and analytics is beyond the scope of this paper. In addition to the examples discussed above, there are many others exploring the use and implications of big data in areas such as health care (Schouten, 2013), education (Picciano, 2012), and government (Kim, Trimi, & Chung, 2014). However, this diversity underscores an important point with respect to the knowledge and skills for budding data scientists and analysts. An understanding of the context within which analytical tools will be used is critical to understanding the appropriate use and application of those tools. The following section will discuss these issues in greater detail.

Implications for Academia

As mentioned previously, the McKinsey & Company report (Manyika et al., 2011) predicts that by 2018 the United States will face a shortage of 140,000 to 190,000 people with deep analytical skills, a gap of 50-60% relative to supply, as well as a deficit of 1.5 million managers and analysts who can effectively analyze big data for purposes of decision making. Davenport and Patil (2012) similarly suggest that there is scarcity of people with the unique skill set (math, statistics, probability, programming, and business skills) to deal with big data.

In addition to technical skills, the literature also identifies a set of “soft skills” that future data scientists will need to possess. Beyond technology skills focused on statistics, computer science and mathematics, data scientists will also need business acumen sufficient to understand the nuances of the their organization’s needs and identify areas where big data will be most likely to create value (Kelly, 2014; Provost & Fawcett, 2013). Data scientists will also need strong teamwork and communication skills in order to build and convey compelling business cases and stories through big data visualization (Kelly, 2014; Van Dyke, 2013).

Other soft skills found include a level of curiosity and inquisitiveness, self-motivation, and an aptitude for problem-solving. As a case in point, Marshall (2012) describes a qualitative researcher’s participation on a big data project to identify “interesting” tweets from Twitter feed. The account of this experience highlights the need for understanding the data well enough to spot a “spike in the graph” and identify the questions that need to be asked to determine if it really is a meaningful finding. Similarly, Marshall suggests that data scientists need to be able to identify “ancillary datasets” (for example, population growth rates from an external source when studying college enrollment data), determine how to interpret them, and decide if they are trustworthy and whether they were collected in a way that did not violate privacy. In other words, for quality analysis and insight generation, there is a need to understand the greater context and surrounding events, as they relate to the datasets being examined.

It is clear that universities must develop degree programs that will produce graduates to fill job roles such as (1) data scientists who can apply a variety of tools and techniques to capture, manage and identify patterns in structured and unstructured data; (2) business analysts who can apply analytics to gain business insights that lead to improved decision making; as well as (3) information security specialists and software engineers who can protect the data resources and continue the development of new software tools. Although academia is working to develop relevant programs, the supply of appropriately skilled graduates still lags behind the demand (Dumbill, Liddy, Stanton, Mueller, & Farnham, 2013). Given that big data and big data analytics is a multi-disciplinary field, workers within this field will require a variety of talents, knowledge and skills. Understandably, a variety of degree programs will be needed.

METHODOLOGY

Based on the review of the literature, skills that are important in the areas of big data, data science and analytics include math, statistics and probability, data mining, visualization techniques, programming, problem-solving, knowledge of technologies and techniques for data capture, data storage, and data management, an understanding of “unstructured” data and data “quality”, familiarity with hardware, platforms and architectures like Hadoop, MPP, CEP, and cloud computing. In addition, an understanding of ethical considerations, especially privacy, and data governance policies, are important. Business acumen and decision making skills, communication skills, and personal traits such as curiosity and persistence are also cited as being important. The purpose of this study is to understand the current state of academia as it relates to big data and analytics programs at the undergraduate level and to specifically examine programs in light of the skills identified in the literature.

To this end, the authors identified a list of universities that might be expected to offer programs or concentrations in data science or analytics. Web sites of these universities were searched to determine where in the university such programs were housed as well as to understand the content of these programs in light of the aforementioned skills.

Universities Included within the Sample

The first stage of data collection was to identify institutions that might be expected to offer either courses or programs in the area of big data, data science, or data analytics. Due to relatively recent development of this area of study, no existing comprehensive lists of such institutions were found. Therefore it was necessary to compile an initial list from various sources with the primary focus on programs associated with business and/or computing disciplines such as computer science, information systems, and information technology. Although there is wide interest in big data in areas as diverse as history (see, e.g., Hoffman, 2013) and biology (see, e.g., Aronova, Baker, & Oreskes, 2010), those disciplines were considered to be outside the scope of this project.

To compile the list, rankings from sources such as *US News and World Report*, *Bloomberg BusinessWeek*, and several others were reviewed; see Table 1 for a full list of the sources consulted. Where possible, the entire published list was retrieved. In cases where the published list included more than 25 schools (such as *Bloomberg BusinessWeek*) only the top 25 schools were included. In the case of *US News and World Report's* undergraduate rankings, only the top 10 schools on each list were publicly available and therefore included in the compilation. Additionally, institutions recognized for their graduate degrees in relevant areas were incorporated based on the rationale that a school might seek to leverage its faculty expertise in the area of big data or analytics by offering courses or programs at the undergraduate level. After compiling the full list and eliminating duplicates, the initial list contained 76 universities.

Recognizing that degree programs are continually changing and new offerings are emerging on an ongoing basis, the initial list of institutions was supplemented by conducting a web search using the terms "business analytics major" and "data science undergraduate major." This search yielded an additional 10 institutions, many of which were drawn from a listing compiled on a blog entitled Data Science 101 (Swanstrom, 2012), which was the first item returned on the search for "data science undergraduate major". The final list of universities included in this study is shown in Table 2. Although the authors cannot guarantee that this is a comprehensive listing of all programs in the area of big data or data science, it is sufficient for this initial research effort.

Table 1: Sources Used to Compile the List of Universities

SOURCE	LIST	NUMBER OF SCHOOLS	URL
US News and World Report	Best Undergraduate Business Schools	10 (of 388 on full list)	http://colleges.usnews.rankingsandreviews.com/best-colleges/rankings/business
US News and World Report	Management Information Systems Rankings	10 (of 18 on full list)	http://colleges.usnews.rankingsandreviews.com/best-colleges/rankings/business-management-

			information-systems
US News and World Report	Quantitative Analysis Rankings	10 (of 14 on full list)	http://colleges.usnews.rankingsandreviews.com/best-colleges/rankings/business-quantitative-analysis
Bloomberg BusinessWeek	Undergraduate Business School Programs	25 (of 124 on full list)	http://www.businessweek.com/bschools/rankings#5
Education Portal	Top Information Technology Schools	3	http://education-portal.com/article_directory/Top_Information_Technology_Schools.html
Tech Republic	Top 10 U.S. College Programs for IT	10	http://www.bit.vt.edu/careers/TechRepublicITProgramsReview-Final.pdf
College Atlas	Best Information Systems Colleges of 2012-2013	10	http://www.collegeatlas.org/information-systems-college-rankings.html
NetworkWorld	Top 20 Colleges for Computer Science Majors, based on earning potential	20	http://www.networkworld.com/news/2013/091713-computer-science-college-ranking-273910.html
Remzi Arpaci-Dusseau University of Wisconsin, Madison	Google-based Ranking for Computer Science and Engineering Departments	10	http://pages.cs.wisc.edu/~remzi/rank.html
US News and World Report	Best Graduate Schools for Computer Science	25 (of 177 on full list)	http://grad-schools.usnews.rankingsandreviews.com/best-graduate-schools/top-science-schools/computer-science-rankings
InformationWeek	Big Data Analytics Master's Degrees: Top 20 Programs	20	http://www.informationweek.com/big-data/big-data-analytics/big-data-analytics-masters-degrees-20-top-programs/d/d-id/1108042?

Table 2: Final List of Universities

INSTITUTION	NUMBER OF TIMES INSTITUTION APPEARED ON UNDERGRADUATE LISTS	APPEARED ON GRADUATE LIST	WEB SEARCH
Arizona State University	0	*	
Bentley University	1	*	

Boston College	1		
Boston University	1		
Brigham Young University	2		
Brown University	0	*	
California Institute of Technology	0	*	
California Polytechnic State University	1		
Carnegie Mellon University	8	*	
College of Charleston			x
Columbia University	0	*	
Cornell University	4	*	
Creighton			x
DePaul University	0	*	
Drexel University	1	*	x
Emory University	1		
Fordham University	0	*	
George Mason University			x
Georgetown University	1		
Georgia Institute of Technology	3	*	
Georgia State University	1		
Harvard University	0	*	
Illinois Institute of Technology			x
Indiana University - Bloomington	5		
James Madison University	1		x
Kennesaw State University	0	*	
Louisiana State University	0	*	
Massachusetts Institute of Technology	7	*	
Miami University	1		
Michigan State University	0	*	
New York University	3	*	
North Carolina State University	0	*	
Northeastern University	1		
Northern Kentucky University			x
Northwestern University	0	*	
Old Dominion University			x
Pennsylvania State University	1		
Princeton University	1	*	
Purdue University	2	*	
Rice University	0	*	
Rutgers University	2	*	x

San Jose State University	1		
Stanford University	2	*	
Stevens Institute of Technology	0	*	
Syracuse University	1		
Temple University	1		
The Ohio State University			x
University of Arizona	2		
University of California - Berkeley	5	*	
University of California - Irvine	1		
University of California - San Diego	1	*	
University of California - Santa Barbara	1		
University of California at Los Angeles	1	*	
University of Cincinnati	0	*	x
University of Connecticut	0	*	
University of Florida	1		
University of Georgia	1		
University of Illinois - Urbana-Champaign	5	*	
University of Iowa			x
University of Kentucky			x
University of Maryland	5	*	
University of Massachusetts - Amherst	1	*	
University of Michigan - Ann Arbor	4	*	
University of Michigan - Dearborn	0	*	
University of Minnesota - Twin Cities	1		
University of North Carolina - Chapel Hill	3	*	
University of Notre Dame	2		
University of Ottawa	0	*	
University of Pennsylvania	4	*	
University of Pittsburgh	1		
University of Richmond	1		
University of Rochester			x
University of San Francisco	0	*	x
University of Southern California	0	*	
University of Tennessee	0	*	x

University of Texas - Austin	6	*	
University of Virginia	2		
University of Washington	2	*	
University of Wisconsin - Madison	0	*	
Villanova University	1		x
Virginia Commonwealth University	0	*	
Virginia Tech	2		
Wake Forest University	1		
Washington University, St. Louis	1		
Worcester Polytechnic University	1		
York University	0	*	

Web-Site Data Collection

Most universities within the United States use their web sites to deliver important services to enrolled students and the public at-large, such as e-library services (Wu, Hsieh & Chang, 2013) and access to a learning management system, such as Blackboard. Additionally, web sites are increasingly being used as a source of empirical data for research. One primary focus of research has been on web site quality – how can this construct be formulated and measured, and ultimately, how can the design of web sites be improved? Closely related to the issue of web site quality is investigating the existing quality of e-services and information available through web sites. For example, Eschenfelder and Miller (2007) investigated the quality of text-based public information provided on program-level government agency web sites. While the 'quality' and extent of services and information available from existing university websites may vary (Ojino, Mich, Ogao & Karume., 2013; Olvera-Lobo, Aguilar-Soto & Ruiz-de-Osma, 2012), it is accepted practice for universities to publish information on their websites regarding course and program offerings. Moreover, universities are incentivized to keep this information relevant and up-to-date as this information is commonly used to attract new students to a program (Anyangwe, 2012), as well as used by current students to plan their studies. For example, it is reasonable to expect that once a new program has been approved and scheduled, information about this curriculum will be quickly posted on a university web site in order to maximize the potential to attract new students. Information published on web sites available to the public has also been used in content analysis studies (see, e.g., Lee, Au & Law 2013), with some studies specifically sampling information from university web sites (see, e.g., Lee et al., 2013). In summary, information from university web sites has been featured in a number of published research studies, with the practice of sampling information from university web sites now well established.

Using the approach of gathering information from university web sites, two of the authors independently searched the public web site of each of the 86 institutions on the list to identify undergraduate programs and course offerings related to big data, analytics, or data science. This web site 'navigational' search was complemented by a Google Advanced Search in which the university's 'site or domain' was restricted to web pages associated with that institution's web domain. The aim of the Google search was to identify web pages in which the terms 'big data,' 'analytics,' or 'data science' appeared and thus provide further indicators of undergraduate programs or individual courses offered by that institution. All searches were performed during the period December 2013 – April 2014.

RESULTS

Table 3 summarizes the results of the search – including the degree title and the academic unit that houses the degree. In addition, universities that have a specific concentration or track in data science or data analytics are identified.

Table 3: Universities with Majors or Concentrations/Tracks in Data Science (DS) or Business/Data Analytics (BA/DA)

INSTITUTION	DS MAJOR	BA/DA MAJOR	TRACK	ACADEMIC UNIT*	DEGREE PROGRAM*
Arizona State University		x		Information Systems/Business	BS Business DA
Carnegie Mellon University			x	Information Systems, Humanities and Social Sciences	Content Area in Quantitative Analysis
College of Charleston	x			CS/Sciences & Math	BS in DS
Creighton		x		Business	Business Intelligence and Analytics
Drexel University		x		Business	Co-major in BS in BA
Illinois Institute of Technology			x	CS/Science	Specialization in DS
Louisiana State University			x	CS/Eng	Concentration in DS and Analytics (beginning Fall 2014)
Massachusetts Institute of Technology			x	MS/Business	Concentration in BA & OR
Miami University		x		IS & Analytics/Business	Co-major in Analytics
Northern Kentucky University	x			CS/Informatics	BS in Data Science
Old Dominion		x		BA/Business & Public Administration	BS in Business Administration - BA Major
Rutgers University		x		MS & IS/Business	Business Analytics and Information Technology
The Ohio State University		x		Stats/Arts & Sciences CS & Eng/Eng	BS in DA

University of Iowa			x	MS/Business	Track in BA
University of Kentucky		x		Business	BBA in Analytics
University of Pennsylvania			x	Business	Concentration in Operations and Information Management with Track in BA and Decision Processes
University of Rochester	x			Interdisciplinary	DS
University of San Francisco	x			Arts & Sciences	BS in DS
University of Tennessee		x		Stats, Operations & MS/Business	BS in Business Administration - BA Major
University of Virginia			x	Business	Track in Business Analytics
Virginia Tech			x	Business IT/Business	Track in Decision Support Systems
* BA = Business Analytics, DA = Data Analytics, DS = Data Science, IS = Information Systems, CS = Computer Science, Eng = Engineering, IT = Information Technology, MS = Management Science, OR = Operations Research, Stat = Statistics					

Of the 86 universities examined, 21 offer some type of formal program in data science, data analytics or business analytics. Specifically, nine universities offer business or data analytics majors (or co-majors) and four offer data science majors. Furthermore, eight universities offer tracks, concentrations or specializations in data science or business/data analytics.

Of the nine universities with majors in analytics, eight are housed within business colleges or schools. The ninth (offered by The Ohio State University), is an interdisciplinary program overseen by the Computer Science and Engineering Department in the College of Engineering and the Statistics Department in the College of Arts and Sciences. The four data science majors are more varied in terms of the degree program and the college or school in which they are housed.

An observation of note is the correspondence of the degree program title with the academic unit offering the program. As can be seen in Table 3, data science programs are more commonly offered through computer science or interdisciplinary academic units. Business analytics and data analytics programs are more commonly offered through business-related academic units such as information systems, statistics, or operations management. To further explore the distinction between the two types of programs, curriculum requirements for programs across institutions were examined to identify commonalities and differences with a focus on the skills identified in the literature review for data scientists and analytics workers.

Table 4 describes the business/data analytics program requirements. The programs listed include those titled as data or business analytics with the exception of The Ohio State program (this program more closely resembles the data science programs and is therefore included in

Table 5 rather than in Table 4). Based on the skills identified in the literature, requirements for traditional courses in math, statistics, programming and database / data warehousing were examined and compared to the requirements for similar programs within the same institution, which were often information systems or similarly named degree programs. Credit hours with asterisks next to them indicate that the designated course is required in the traditional information systems degree or business degree as well as the analytics degree program. This distinction means that in some cases (e.g., the statistics requirements for Miami University) two numbers are reported for credit hours. The first number (with an asterisk) indicates a course requirement that overlaps with an existing program and the second number (with no asterisk) indicates additional course requirements in that area. Also, some institutions did not have existing information systems programs. If there was no information systems degree program, as indicated by universities with two asterisks next to the university name, the comparison was made to the core business requirements with respect to math and statistics. One institution, Drexel University, operates on a quarter system. Credit hours for this university were converted to a semester system equivalent by dividing by 1.5. This scaling reflects the fact that degree requirements for universities operating on a quarter system are typically about 180 hours for an undergraduate degree while those operating on a semester system are typically about 120 hours. This is a 1.5 to 1 ratio.

Table 5 describes the curriculum requirements for the four universities with data science majors. The table also includes the data analytics program at The Ohio State University as it more closely resembles the data science degrees offered at other institutions than it does other business/data analytics degrees. All credit hours are based on a semester system as all universities in Table 5 were on the semester system. Credit hours with asterisks next to them indicate that the designated course is required in the traditional computer science degree as well as the data science degree program. Universities with two asterisks have data science degree programs that are listed as interdisciplinary so a comparison to a single degree program, such as computer science, would be inappropriate.

Table 4: Analysis of Degree Programs by Credit Hours in Subject for Universities with Majors in Business/Data Analytics								
UNIVERSITY	DB/DW***	PROG	STATS	MATH	VISUAL	BIG DATA	ANALYTICS/ MODELING	DM
Arizona State University	3 (DW)		3*	3* - Brief Calc 3* - Math for Bus Majors	1.5	1.5	6	3
Creighton**	3 (DB)	3*	3*	3* - Applied Calc			4	
Drexel University	2.7* (DB)		2.7* 5.3	2.7* - Intro to Analysis II			8	
Miami University	3* (DB)	9 (optional)	4* 6	3* - Calculus 1	3	3		3
Old Dominion			6*	3* - Business Calc			9	
Rutgers University**	3 (DB)	3*	6*	3* - Calculus 1			15	
University of Kentucky**	3 (DB)		4.5*	3* - Elem Calc & Finite Math			6	3
University of Tennessee**	2 (DB)		3* 6	3* - Basic Calc			3	3
* Designated course is required in the traditional information systems degree or business degree as well as the analytics degree program ** Universities without a degree in information systems *** DB = Database, DW = Data Warehousing, DS = Data Science, PROG = Programming, VISUAL = Visualization, DM = Data Mining								

Table 5: Analysis of Degree Programs by Credit Hours in Subject for Universities with Majors in Data Science								
UNIVERSITY	DB/DW***	PROG	STATS	MATH	VISUAL	BIG DATA	DS/ ANALYTICS/ MODELING	DM
College of Charleston		13*	3* 9	3* - Discrete 3 - Calculus 2 3 - Linear Algebra			9	3
Northern Kentucky University	3* (DB) 3 (DB)	9*	3* 3	3* - Discrete 3* - Calculus C 3 - Linear Algebra	3	3	9	3
The Ohio State University**	6 (DB)	12	18	3 - Calculus 2 3 - Linear Algebra	3			3
University of Rochester**		15		3 - Discrete 3 - Calculus 2 3 - Linear Algebra				3
University of San Francisco		12*	6	3* - Discrete 3 - Calculus 2 3* - Linear Algebra	3		3	3
* Designated course is required in the traditional computer science degree as well as the data science major. ** Universities with interdisciplinary degree programs. A comparison to a single degree program would be inappropriate. *** DB = Database, DW = Data Warehousing, DS = Data Science, PROG = Programming, VISUAL = Visualization, DM = Data Mining								

DISCUSSION

This paper presents the findings from a detailed review of the curriculum requirements for data science and business/data analytics programs across the country. Data science programs are typically offered through computer science academic units or are interdisciplinary while business/data analytics programs are typically offered through business-related academic units. Further inspection of the courses required by these programs highlighted distinctions between the two types of degrees.

The review of skills identified in literature in the areas of big data, data science and analytics indicate that math, statistics and probability, data mining, visualization techniques, programming, problem-solving, knowledge of technologies and techniques for data capture, data storage, and data management, an understanding of "unstructured" data and data "quality", and familiarity with hardware, platforms and architectures are important. In addition, an understanding of ethical considerations, especially privacy, data governance policies, business acumen and communication skills are important.

The programs listed in Tables 4 and 5 were examined in light of a subset of the aforementioned skills and the curriculum requirements were summarized in both tables for analytics and data science programs, respectively. The subset includes math, statistics and probability, programming, database / data warehousing, data mining, visualization, analytics and big data. The distinctions between the two types of programs reflect the nature of the related academic units, with the analytics programs being housed in colleges of business and the data science programs being more closely affiliated with computer science programs or more interdisciplinary entities. Based on the comparison of these requirements, there are several specific observations that can be made.

- Analytics programs in colleges or schools of business all require a traditional database or data warehousing course, whereas data science degree programs do not all have such a requirement.
- Analytics programs do not necessarily require programming courses, whereas data science programs are typically programming intensive.
- Both types of programs require the completion of a number of statistics courses, often significantly more than what is normally required in comparable programs within the college. However, the data science programs are more statistics intensive than business/data analytics programs. The average number of credit hours for the data science programs in statistics is 10.2 as compared to 6.2 for analytics programs.
- The data science programs typically require math courses beyond those that are required for comparable programs, as well as requiring a higher level of math when compared to analytics programs.

Both types of programs require roughly an equal number of visualization, big data, data science/analytics/modeling, and data mining courses. The average number of credit hours for courses of that type for the analytics programs is 9 and for data science programs the average is 9.6. These types of courses are typically newer courses that focus on the tools and technology for handling big data. Maintaining currency is critical for technically based academic courses and programs; such new courses serve to differentiate data science and analytics programs from traditional computer science or information systems degree programs.

CONCLUSION AND LIMITATIONS

Industry demand for employees with the skills to extract value from big data is growing rapidly. As the results of this study show, academia is working to meet that demand in various ways. This paper offers a review and summary of the different types of degree programs offered to help address industry needs. As such, it may provide a roadmap for institutions wishing to enter the rapidly expanding big data field.

A limitation of this study is that it is focused on courses covering technical and statistical topics required by big data programs. Courses were not examined in sufficient depth to determine the extent to which they teach soft skills, including effective communication, team building, and conflict resolution. Nevertheless, such skills are critical to be effective in understanding the expectations of stakeholders and in communicating findings and results. Understanding the situational context in which big data is employed (e.g., business or health care) is also important. While a review of such course requirements was beyond the scope of this paper, that omission should not be taken as an implication that such skills are unimportant. Future research should incorporate a review of such courses when examining data science and business/data analytics programs.

REFERENCES

Abbasi, A., Albrecht, C., Vance, A., & Hansen, J. (2012). Metafraud: A Meta-Learning Framework for Detecting Financial Fraud. *MIS Quarterly, 36*(4), 1293-1327.

Anyangwe, E. (2012). A New Era for Attracting Students? Falmouth Creates Interactive Prospectus. *Guardian Professional*. Retrieved from http://www.theguardian.com/higher-education-network/2012/jun/07/interactive-student-prospectus-falmouth April 1, 2014.

Aronova, E., Baker, K. S., & Oreskes, N. (2010). Big Science and Big Data in Biology: From the International Geophysical Year through the International Biological Program to the Long Term Ecological Research (LTER) Network, 1957-Present. *Historical Studies in the Natural Sciences, 40*(2), 183-224.

Balboni, F., Finch, G., Reese, C. R., & Shockley, R. (2013). *Analytics: A Blueprint for Value*. IBM Global Business Services (white paper). 1-27. April 23, 2014.

Chau, M., & Xu, J. (2012). Business Intelligence in Blogs: Understanding Consumer Interactions and Communities. *MIS Quarterly, 36*(4), 1189-1216.

Chen, H. C. (2010). Business and Market Intelligence 2.0. *IEEE Intelligent Systems, 25*(1), 68-71.

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

Collett, S. (2011). Why Big Data is a Big Deal. *Computerworld.com*, 19-24. Retrieved from http://www.computerworld.com/s/article/357092/Why_Big_Data_Is_a_Big_Deal April 14, 2014.

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

Dumbill, E., Liddy, E. D., Stanton, J., Mueller, K., & Farnham, S. (2013). Educating the Next Generation of Data Scientists. *Big Data, 1*(1), 21-27.

Eckerson, W. (2011). *Big Data Analytics: Profiling the Use of Analytical Platforms in User Organizations*. SAS (white paper). 1-48. Retrieved from http://www.sas.com/content/dam/SAS/en_us/doc/research2/big-data-analytics-105425.pdf July 23, 2013.

Eschenfelder, K. R., & Miller, C. A. (2007). Examining the role of Web site information in facilitating different citizen-government relationships: A case study of state Chronic Wasting Disease Web sites. *Government Information Quarterly, 24*(1), 64-88.

Griffin, R. (2012, December). Using Big Data to Combat Enterprise Fraud. *Financial Executive,* 44-47.

Hoffman, L. (2013). Looking Back at Big Data. *Communications of the ACM, 56*(4), 21-23.

Hu, D. N., Zhao, J. L., Hua, Z. M., & Wong, M. C. S. (2012). Network-Based Modeling and Analysis of Systemic Risk in Banking Systems. *MIS Quarterly, 36*(4), 1269-1291.

Kelly, J. (2014). Big Data: Hadoop, Business Analytics and Beyond Retrieved from http://wikibon.org/wiki/v/Big_Data:_Hadoop,_Business_Analytics_and_Beyond April 14, 2014.

Kim, G. H., Trimi, S., & Chung, J. H. (2014). Big-Data Applications in the Government Sector. *Communications of the ACM, 57*(3), 78-85.

Laney, D. (2001). *3D Data Management: Controlling Data Volume, Velocity, and Variety*. White Paper. META Group, Inc. Retrieved from http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf April 23, 2014.

Lau, R. Y. K., Liao, S. S. Y., Wong, K. F., & Chiu, D. K. W. (2012). Web 2.0 Environmental Scanning and Adaptive Decision Support for Business Mergers and Acquisitions. *MIS Quarterly, 36*(4), 1239-1268.

Lavalle, S., Lesser, E., Shockley, R., Hopkins, M. S., & Kruschwitz, N. (2011). Big Data, Analytics and the Path From Insights to Value. *MIT Sloan Management Review, 52*(2), 21-32.

Lee, H., Au, N., & Law, R. (2013). Presentation Formats of Policy Statements on Hotel Websites and Privacy Concerns: A Multimedia Learning Theory Perspective. *Journal of Hospitality & Tourism Research, 37*(4), 470-489.

Lycett, M. (2013). 'Datafication': making sense of (big) data in a complex world. *European Journal of Information Systems, 22*(4), 381-386.

Manyika, J., Chui, M., Brown, Brad, Bughin, J., Richard, D., Roxburgh, C., & Byers, A. H. (2011). *Big Data: The Next Frontier for Innovation, Competition, and Productivity*. McKinsey Global Institute (white paper). 1-143. Retrieved from http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation December 31, 2013.

Marshall, C. (2012). Big Data, the Crowd and Me. *Information Services & Use, 32*(3/4), 213-224.

McAfee, A., & Brynjolfsson, E. (2012). Big Data: The Management Revolution. *Harvard Business Review, 90*(10), 60-68.

Moore, K. D., Eyestone, K., & Coddington, D. C. (2013). The Big Deal About Big Data. *HFM Magazine*. Retrieved from http://www.hfma.org/Content.aspx?id=18550 April 11, 2014.

Nash, K. S. (2012a). Big Data, Big Brother, Big Bucks. *CIO*, 30-39. Retrieved from http://www.cio.com/article/706457/Equifax_Eyes_Are_Watching_You_Big_Data_Means_Big_Brother March 8, 2013.

Nash, K. S. (2012b). Fraud Happens. *CIO*, 34-41. Retrieved from http://www.cio.com/article/702107/How_CIOs_Can_Learn_to_Catch_Insider_Crime March 8, 2013.

Ojino, R. O., Mich, L., Ogao, P., & Karume, S. M. (2013). The Quality of Kenyan University Websites: A Study for the Re-engineering of the Masinde Muliro University Website. *Journal of e-Learning and Knowledge Society, 9*, 169-176.

Olvera-Lobo, M. D., Aguilar-Soto, M., & Ruiz-de-Osma, E. (2012). Evaluation of websites for biomedical postgraduate courses in Spanish. *Transinformacao, 24*(1), 47-60.

Park, S. H., Huh, S. Y., Oh, W., & Han, S. P. (2012). A Social Network-Based Inference Model for Validating Customer Profile Data. *MIS Quarterly, 36*(4), 1217-1237.

Picciano, A. G. (2012). The Evolution of Big Data and Learning Analytics in American Higher Education. *Journal of Asynchronous Learning Networks, 16*(3), 9-20.

Poole, S. (2013, 24-30 May 2013). The Digital Panopticon. *New Statesman,* 22-25.

Preimesberger, C. (2011). Big Ideas About 'Big Data'. *eWeek,* 34-37.

Provost, F., & Fawcett, T. (2013). Data Science and its Relationship to Big Data and Data-Driven Decision Making. *Big Data, 1*(1), 51-59.

Sahoo, N., Singh, P. V., & Mukhopadhyay, T. (2012). A Hidden Markov Model for Collaborative Filtering. *MIS Quarterly, 36*(4), 1329-1356.

Schouten, P. (2013). Big Data in Health Care. *Healthcare Financial Management,* 40-42.

Swanstrom, R. (2012, last update April 2014). Colleges with Data Science Degrees. *Data Science 101* Retrieved from http://datascience101.wordpress.com/2012/04/09/colleges-with-data-science-degrees/ April 20, 2014.

Tallon, P. P. (2013). Corporate Governance of Big Data: Perspectives on Value, Risk, and Cost. *Computer, 46*(6), 32-38.

Van Dyke, M. (2013). Predictive Analytics: Pinpointing How to Best Allocate Patient Resources. *HFM Magazine*. Retrieved from https://www.hfma.org/Content.aspx?id=16069 April 11, 2014.

Wigan, M. R., & Clarke, R. (2013). Big Data's Big Unintended Consequences. *Computer, 46*(6), 46-53.

Wittman, A. (2013). Brave the Big Data Wave. *InformationWeek*. Retrieved from http://reports.informationweek.com/abstract/81/11935/Business-Intelligence-and-Information-Management/Brave-the-Big-Data-Wave.html April 11, 2014.

Wu, C. M., Hsieh, C. L., & Chang, K. L. (2013). A Model for Assessing the Service Quality of University Library Websites. *Mathematical Problems in Engineering, 2013*, 1-9.