Systematic Study of Big Data Science and Analytics Programs

Dr. Huanmei Wu, Indiana University-Purdue University, Indianapolis

Chair of the Department of BioHealth Informatics. Research on data management and data analytics, applying to informaticcs, especially biomedical informatics and other interdisciplinary fields.

©American Society for Engineering Education, 2017

Systematic Study of Data Science and Analytics Programs

Abstract— Rapid advances in information technologies have led to the generation of massive data sets, especially in life science and biomedical informatics. These data sets are valuable assets and in great needs to be analyzed. However, there is a shortage of workforce for big data analysis. Education innovations are required to empower students with the skills and technologies for large dataset analysis. Over the last few years, there is a high demand for new programs in data science and analytics (DSA). We has performed a systematic study of the existing DSA programs in the US by checking the detailed information about the degree programs, the program competencies, the curriculum designs, the expected learning outcomes, program sizes, professional careers, and other related information. There are more than 70 DSA programs offered in the US. This study provides guidance on DSA related program development and curriculum design. It also provides the potential trainees in DSA with the current market needs and the required knowledge for their future career.

1. Introduction: There is rapid growth massive data generation and data collections from many fields, such as web-based social informatics, search engines, satellite data, health sciences, biology, and biomedical informatics. The data is not huge in size but also complicated with various data types and uncertainties. It is an important and challenge to extract valuable information from the large data volume. Applied *Data Science and Analytics (DSA)* has been emerging developing discipline, which applies modern, data-driven analytical methods over massive data to application oriented problem-solving and decision making based on evident-based data analytical results. For example, DSA can be applied to predictive modeling, marketing analysis, data mining, natural language processing, and other data driven applications.

Various DSA programs have been offered in the US at different degree levels, from BS to MS and PhD programs or some certificate programs. Due to its multidisciplinary nature, different approaches and strategies for DSA education have been proposed. Especially, there is a great advocate for conveying the statistical, visualization, computational and information technology skills for data analytics to meet the challenges of big data sciences.

However, as it is an evolving field, there is no concrete guidelines on program competencies, curriculum design, learning outcomes, and career preparations for DSA programs. This study has performed analysis of the existing DSA programs and job market study. The results will provide guidance in DSA program developments and improvements, as well as professional career readiness.

2. Overviews of the DSA programs in the US: Various DSA programs exist in the US. For individual program, there is only limited program information available on the website. However, there is no common resource to introduce the programs. One potential resource is the Data Science 101 which contains 101 universities providing some types of programs in data science [1]. The other source is the KDnuggets.com, which gives a list of universities providing various analytical degrees

[3]. Our team has visited the individual program website of 103 DSA programs from 68 universities and performed systematic study of these programs. Some of the findings are briefly summarized below.

2.1. Academic Structures of DSA Programs: First**,** the academic schools, which offered some DSA programs, are studied. It was observed that a DSA program could reside in different academic schools. The top five schools are the School of Business (41.2%), the School of Arts and Science (13.7%), the informatics schools (which includes Informatics and Computing, Information and Computing Sciences, and Information Technology, combined to 13.7%), the School of Engineering and Applied Science (11.8%), the School of Management (6%).

Usually, the DSA program is a program within a department. The most common department names are Department of Informatics, Department of Computer Science, or Department of Engineering. In addition, a DSA can be an interdisciplinary program through collaborations with several departments. One example is the *Business Analytics* program, which is usually built by collaboration from Department of Computer Science and Department of Statistics.

The names of the DSA degree programs are differently from various universities. Most popularly DSA program degree names are Business Analytics (27%), Data Science (17%), Data Analytics (21%), Marketing Analytics (6%), and Statistics (6%).

The types of the DSA degree programs include formal degrees, such as BS, MS, and PhD programs and certificate programs. Among them, Master programs (MS) in DSA is the most popular degrees. Figure 1 below illustrated the numbers among the 68 universities which has the different programs. Usually, a university is not limited to only one DSA degree program, which is demonstrated in Figure 2. Many universities have two or more degree programs. A few even have three or four types of DSA programs. The distributions of the various DSA degree programs and the numbers of universities are manifested in Table 1. It showed that the most popular combinations of DSA degree programs among the universities are MS + certificate programs, followed by the MS + PhD programs, and BS + MS programs.

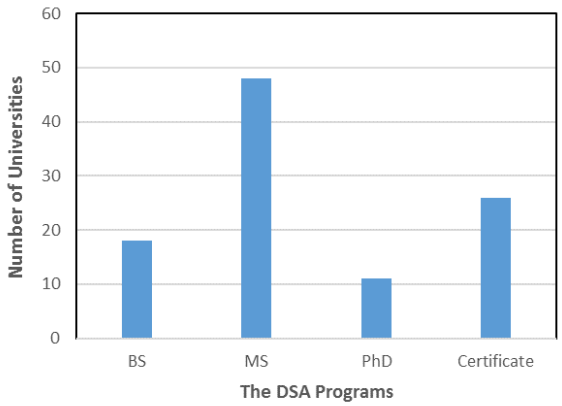

Figure 1. The number of universities for each DSA degree program.

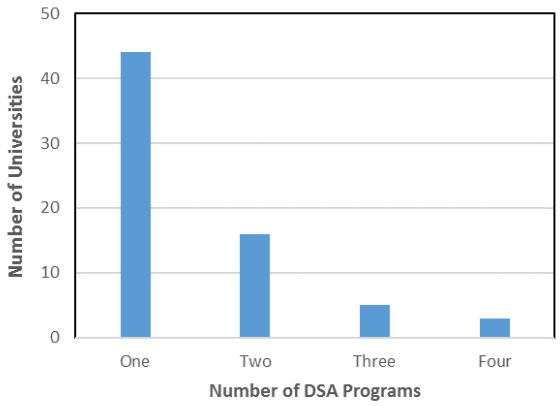

Figure 2. The number of universities and the number of DSA programs.

Table 1: The number of Universities, which offered different combinations of DSA degree programs.

Degrees	BS+ MS	BS+ PhD	BS+ C	BS+ MS+ PhD	BS+ MS+ C	BS+ PhD +C	BS+MS + PhD+C	MS+ PhD	MS +C	MS+ PhD +C	PhD +C
# of programs	8	6	5	5	3	4	3	9	16	4	5

** C stands for certificate program in DSA.*

For the DSA program sizes, the numbers of faculty and students are also studied. Some of the universities do not have the specific faculty and/or student information on their program websites. For the 41 programs with the faculty information listed, the number of faculty in the DSA programs is demonstrated in Figure 3. Most of the programs list all the faculty members together, including the core DSA faculty and the adjunct faculty. Most programs (63%) have 20 or fewer faculty members. There are 14 universities have faculty members between 21 and 40. Most DSA programs do not list their student information on the website. Only a limited number of programs have their current graduate students listed. Based on the limited information of 15 DSA programs, the average number of BS, MS and PhD students for the DSA programs are 47 students per programs.

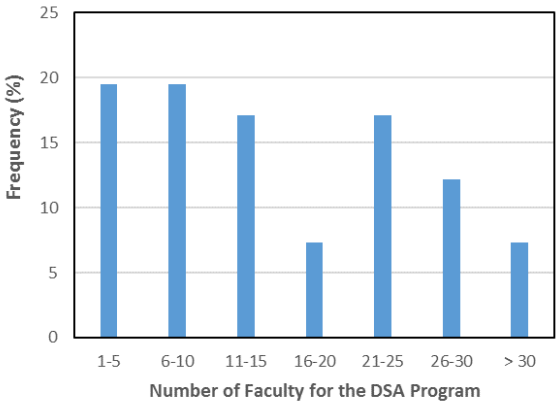

Figure 3. The DSA program faculty sizes

2.2. Curriculum and program competencies: Because of excessive advancement in technology, there are many tools and techniques are emerging in this field. It is a great challenge for trainees to stay up to date and mould with the changing technologies. The DSA programs are expected from trainees that they would be skilled in different programs under DSA like data mining, data science, data handling, and some programming languages as well as with basic database technologies. A trainee needs to know good understanding of DBMS and cloud computing emerging with the knowledge of the field of the data like marketing, informatics, business, medical, etc.,

Due to the success achieved collectively by the data mining and statistical analysis community, statistics based data analytics continues to be an active area in DSA [5, 6]. Other new data analytics techniques explore and leverage unique data characteristics, from sequential/temporal mining and spatial mining to data mining for high-speed data streams and sensor data. Unlike the structured data that can repeatedly be handled through an RDBMS, semi-structured data may call for ad hoc and one-time extraction, parsing, processing, indexing, and analytics in a scalable and distributed MapReduce or the Hadoop environment.

The expected program outcomes are based on the major competencies and major DSA areas, combining the different DSA programs. Upon completion of the graduate degree programs, a student is expected to possess the following DSA capabilities:

- Learn the fundamentals of data analytics and the data science pipeline

- Explain the main concepts, models, technologies, and services of cloud computing, the reasons for the shift to this model, and its advantages and disadvantages
- Explain the core challenges of cloud computing deployments, including public, private, and community clouds, regarding privacy, security, and interoperability.
- Demonstrate and compare the use of cloud storage vendor offerings, such as Amazon S3, Microsoft Azure, OpenStack, and Hadoop distributed file system.
- Develop, install, and configure cloud-computing applications under software-as-a-service principles, employing cloud computing frameworks and libraries.
- Apply the MapReduce programming model to data analytics and enhance its performance by redesigning the system architecture (e.g., provisioning and cluster configurations
- Develop competency in the Python programming language and some data-related Python libraries
- Store and access data from a variety of sources including traditional relational databases, NoSQL data stores, and other web-based sources
- Master basic software engineering practices and understand how they enable reproducible and scalable data analyses
- Learn how to scope the resources required for a data science project
- Apply statistical methods, regression techniques, and machine learning algorithms to make sense out of data sets both large and small
- Know what analyses are possible given a particular data set, including both the state of the art of the field and inherent limitations
- Fluently speak to disparate groups within an organization, from management to the IT director, to implement data analytics solution.

3. Careers for DSA Students: We performed detailed investigations on the job market for DSA in the US. The summary information below is from two major job postings portal sites: indeed.com and monster.com.

3.1. Job opportunities and requirements: The job market for DSA students is promising. The DSA graduates can work in various related fields with varying titles. Sample industrial job titles include data analyst, data scientist, data analytics engineer, business analytics specialist, product analyst, marketing analytics specialist, analytics consultant, research data analyst, statistical analyst, analytics technologist, visualization developer. Other than these famous job titles are reporting specialist, promotional analyst, quality improvement analyst, computational scientist, visualization developer, merchandise reporting & analyst, data mining specialist, etc., there are also openings in academia as postdoctoral fellows, research scientist, and research professor. The top-ranked job titles for DSA job openings from Indeed.com are illustrated in Figure 4, based on a search of a keyword of “Applied Data Analytics” in December 2015 from Indeed.com.

The salary package for DSA starts from $40,000+ and goes up to more than $120,000. With more experience in the field, one can go higher over the salary ladder. Based on more than 8000 job postings in DSA only from Indeed.com, the distributions of job opportunities and the salary ranges are illustrated in Figure 5. DSA job positions for the salary range between$40,000 and $60,000 have the most openings by itself (>3500), which are the entry-level jobs for DSA.

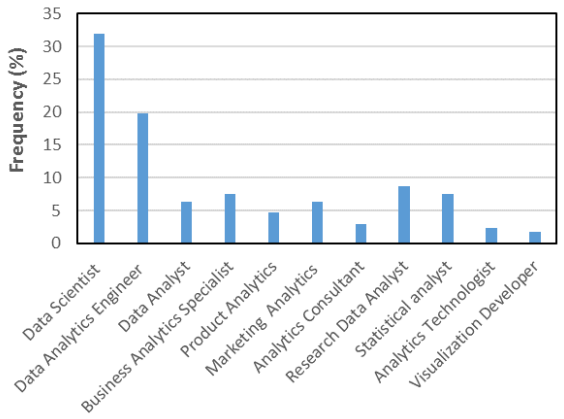

Figure 4. The job postings for different job titles.

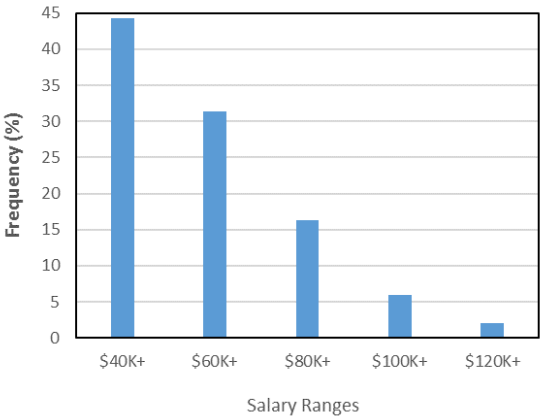

Figure 5. The salary ranges and number of job postings for DSA professionals

The combined job postings for intermediate level in DSA, salary range between $60,000 and $80,000, have about 2500 job postings. The more advanced level jobs for DSA, with a salary range between $80,000 and $100,000, have more than 1200 opening.

The more advanced jobs, such as directors and managers, requires extensive experience in the fields, with salary more than $100,000. However, the job postings are getting less, with 492 postings in the $100,000+ category and 173 postings in the $120,000+ category. Thus, the majority of careers for DSA graduates are in the entry, intermediate, and advanced levels, which are 91% of all the job postings. Many job postings specifically state that a graduate degree in DSA, data science or related fields is required, although there are opportunities for Bachelor's degree holders. In addition, job seekers with PhD degrees have more openings. Many job postings explicitly indicate that PhD degree is preferred over MS and BS degrees.

Regarding the working experiences, a minimum of 3 years working experience on average was required, even for most of the entry level jobs. The required years of experience vary according to the levels of jobs. Usually, a higher paid job required more experience.

3.2. Skill requirements: Detailed analysis of the skill requirements for different job openings has also been performed. Since the focus of this work is to guide the education program, the skills requirements for the entry level jobs have been summarized below, which is more suitable for fresh graduates. Similar analytical results can be performed on intermediate and advanced jobs.

The required skills for randomly selected 100 entry levels jobs are illustrated in Figure 7. The Figure showed that the most required skills for DSA jobs are programming skills, analytical software tools, data mining and machine learning, database managements, statistical analysis, predictive analytics, marketing skills, large-scale analytics and others. More information about the required skills in DSA is summarized below.

> *Programming skills*: It was required for most DSA jobs, the most popular programming languages or script languages are R, Perl, Python, Pig, Hive, Java, C++ and Matlab. At least one of these programming languages will be listed in the job requirements. Python

programming language (Scikit-Learn), and a suite of programs for interacting with high-throughput sequencing data (Samtools).

Analytical software tools: Data analytics scope is very vast as it contains data science, various type of data analysis, data mining, visualization, etc., there are several tools used in different areas. The range of software tools started from simple excel sheet to typical analytical packages to sophisticated business suites. Here I mention some famous tools which are very important in this field. Knowledge of data analysis platform like Cloudera is as important as other tools. A command like a utility like Linux is also playing an important role.

Data mining and machine learning: For example, hierarchical clustering and decision trees are commonly required techniques. Data mining tools - RAPID MINER, orange, RAPID ANALYTICS, WEKA, KEEL, Mahout, etc.

Database management: Databases, including traditional relational database systems (e.g., SQL Server or Oracle), NoSQL databases (e.g., MongoDB), big data analytics (e.g., Vertica), and other big data databases (e.g., TCGA) are commonly required knowledge for potential DSA job seekers. Especially, for big data management or analysis, it is a widely open field and is in great demands.

Statistical software systems: such as SPSS and SAS. Also, statistical analysis using R or Python is highly demanded.

Marketing skills: Marketing and Sales Analytics including revenue/discounting analysis, customer survival and churn analysis, and business case/profitability analysis support.

Some additional skills required for the jobs are good communication skills and writing skills, managerial skills, good team player, the ability to multi-task and work independently. Strong problem-solving skills, project management skills, Strong facilitation and collaboration skills are also expected from data analytics specialist.

It can also be observed that there are some overlaps between Figure 6 and Figure 7 (the competencies and the skill requirements). In fact, the more overlap, the better, as it will show that our program competencies meet the job market for bioinformatics.

Figure 6. The major competencies of DSA programs

Figure 7. The required skills for entry level jobs in DSA.

C. Employers for DSA: Based on the job postings, the top 15 companies who have the most job postings in DSA are summarized in Table 2. It can be observed that the leading employer is Leidos, which is a biomedical research Incorporate. It develops and applies advanced technologies for translational research in cancer and AIDS treatments. The second is Center for Autism and Related Disorders. It is among the largest Autism treatment organizations in the world. Other top companies on various type of data analysis companies based on healthcare, air force, transportation, information technology and laboratory.

Another way to investigate the DSA job market is the locations of the job postings, which is summarized in Figure 8, for the top 14 locations that have the most job postings. The blue coloured areas which are about east coast New York (23%). The red coloured areas are on the west coast in San Francisco, California which is 15% of whole locations. Other 62% is divided into a various region of job posting like Chicago, Seattle, Washington, Atlanta, Boston and others.

Table 2: Top Companies with most jobs postings.

Company	**Jobs**
Leidos	244
Center for Autism and Related Disorders	159
Johns Hopkins University	59
MetLife	41
Applied Memetics LLC	40
Autism Home Support	39
Maxim Healthcare Services	35
People's Care, Inc	31
The Johns Hopkins Applied Physics Laboratory	27
Total Spectrum	27
Department of the Air Force	27
Metropolitan Transportation Authority	26
Booz Allen Hamilton	26
General Dynamics Information Technology	25
City of Houston, TX	24

4. Applications of the information: The information learned from this study provides great guidance for the redesign of the curriculum, help for the students to evaluated their potential DSA skills, and promote of hands-on experience for DSA students.

Various DSA courses are offered at our institute. The information from this project helps the DSA to be a separate master's program and following the learning outcomes that can take care while developing this program. It helps to eliminate redundant contents and develop new course modules to cover the top required skills. Modular courses of 1 or 2- credits have been designed to cover specifically of one or two required skills. This will guarantee

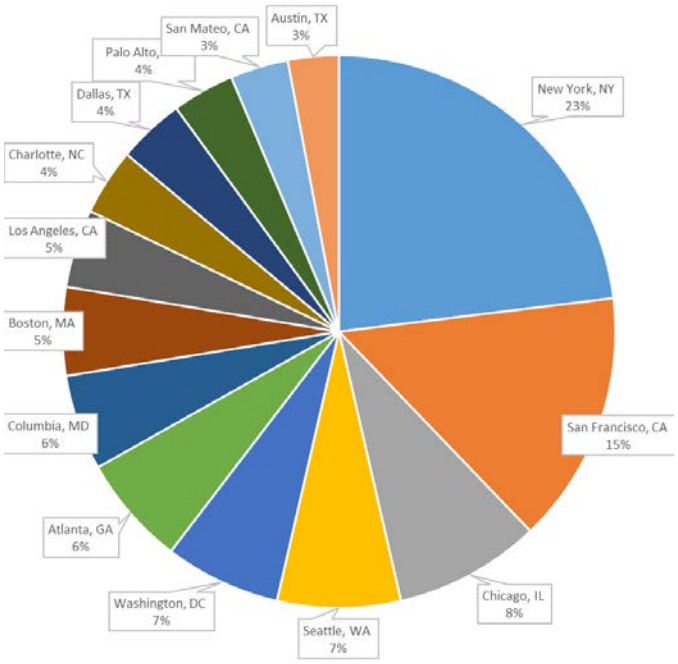

Figure 8. The geographical distribution of the job market in DSA for the top 14 locations

that our students are career ready upon graduation. For example, for the DSA students with a limited background in a programming language, the programming courses (C++, R, Python, Matlab, and Java) are taught in their first semester entering the programs. This will make sure they are ready to work on DSA projects from the beginning, or same can apply for data management system.

In addition, all the students in DSA program will be provided with the opportunities to have hands-on experiences, starting from the beginning when they entering the programs. All full-time students are provided with partial scholarships and hourly research assistantship. With the acceptance of the scholarships, the students are expected to work on DSA projects upon entering the programs. The student records showed that this is a great way to increase the students' experiences and to improve the portfolios of the students when they are ready for jobs.

5. Conclusions: This project has studied the DSA education in the United States and investigated the potential employments of the DSA students. The current state of the arts of the DSA programs nationwide has been summarized. The DSA program competencies and the career skill requirements are described in details. Overall, the DSA is a promising field with high demands, better payments, and interesting projects.

References:

1. 23 Great Schools with Master's Programs in Data Science, Master's in Data Science. http://www.mastersindatascience.org/schools/23-great-schools-with-masters-programs-in-data-science/
2. Jason Davies, wordcloud, https://www.jasondavies.com/wordcloud/#
3. KD nuggets, Certificates and Certification in Analytics, Data Mining, and Data Science, http://www.kdnuggets.com/education/analytics-data-mining-certificates.html
4. Ryan Swanstrom, College with Data Science Degree, Data Science 101, http://101.datascience.community/2012/04/09/colleges-with-data-science-degrees/
5. Chen, Hsinchun, Roger HL Chiang, and Veda C. Storey. "Business intelligence and analytics: From big data to big impact." MIS quarterly 36.4 (2012): 1165-1188.
6. Ferguson, Rebecca. "Learning Analytics: drivers, developments and challenges." TD Tecnologie Didattiche 22.3 (2014): 138-147.
7. Ferguson, Rebecca. "Learning analytics: drivers, developments and challenges." International Journal of Technology Enhanced Learning 4.5-6 (2012): 304-317.
8. Gaurav Vohra, 10 most popular analytic tools in business, Analytics Training, http://analyticstraining.com/2011/10-most-popular-analytic-tools-in-business/