1. Data, Data Types, and Big Data

Data have been frequently discussed along with two other concepts: information and knowledge. However, the concept of data seems less ambiguous than that of information or knowledge. Data are considered to be symbols, or raw facts which have not yet been processed (Ackoff, 1989; Coronel, Morris, & Rob, 2012, p.5). In contrast, information has a number of different definitions, such as information as a thing, something informative, a process, or equivalent to knowledge (Buckland, 1991; Losee, 1997; Saracevic, 1999; Madden, 2000). Knowledge also does not have an agreed-upon definition. Davenport and Prusak (1998) viewed knowledge as “a fluid mix of framed experience, values, contextual information, and expert insight that provides a framework for evaluating and incorporating new experiences and information.” (Davenport & Prusak, 1998, p.5).

The three concepts of data, information, and knowledge are related. They are products of human intellectual activity at different cognitive levels. Data usually need to be processed and organized to become informative for human beings, while knowledge is obtained by making use of data and information through experiences. There is a fine line between data and information: when we talk about digital information or digital data, they are sometimes used interchangeably.

To facilitate our discussion, we follow the idea that data are symbols or raw facts that are stored on electronic media. They can be produced or collected manually by humans or automatically by instruments. The following are examples of data in different forms and formats:

- **Scientific data.** Scientific data are those collected by scientists or scientific instruments during observations or experiments. Examples of scientific data include astronomical data collected by telescopes, patients’ vitals collected by heartbeat monitors, and laboratory data collected by biologists or physicists;
- **Transaction data.** Transaction data are those collected by companies to record their business activities, such as sales records, product specifics, and customer profiles;
- **Data that is of public interest.** Data that reflect environmental status, human activities and social change, such as geographical data, and weather forecasts.
• Web data. Web data are documents, webpages, images, videos published or released on the Internet. These data also record human activities and knowledge. They can be raw facts for certain tasks or decision making processes;
• Social media data. Social media data are those generated by the Internet or mobile users, such as tweets or postings on social media sites, or comments to others’ postings.

Data can also be classified based on different facets depending on different perspectives. For examples, scientific data can be numeric, textual, or graphical. Or, we can categorize data as structured or unstructured based on whether they have been organized. Data type is an important concept in this regard. It has been a basic concept in Computer Science - a computer programmer knows that one needs to determine the data type of a particular variable before applying appropriate operations to it. Computer languages can usually handle different types of data. For example, Python, one of the most popular computer programming languages for data analysis, contains simple data types including numerical data type, string and bytes, and collection data types such as tuple, list, set, and dictionary. Data might need to be changed to a different type in order to be processed properly.

Most data we talked in this chapter are actually digital data in electronic form. Data are generated every day and at every moment. One of the characteristics of data is that it can be copied easily and precisely, or transferred from one medium to another with great speed and accuracy. However, that is not always true, especially when we start to deal with big data. Big data refer to digital data with three Vs: volume, variety, and velocity (Laney, 2011). Volume refers to the situation that data grows at a rapid rate, variety refers to the many data types in which data exist, and velocity refers to the speed of data delivery and processing. Dealing with big data requires theories, methods, technologies, and tools, which lead to the emergence of Data Science, a new discipline that focuses on data processing, which will be discussed in the remaining sections.

Data are important resources for organizations and individuals who use them to make decisions. When big data becomes publicly available, no one can ignore the huge potential impact of these data on business, education, and personal lives. Therefore, Data Science as a new discipline has drawn great attention from governments, industry, and educational institutions.

2. Data Science and Data Scientists

Dr. William S. Cleveland formally defined Data Science in 2001 as it is used today in his article “Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics.” (Cleveland, 2001). Since then, the concept of Data Science has been further developed and has been increasingly linked to data analytics and big
data. Companies have realized the value of Data Science and its usefulness in discovering knowledge and helping with decision-making. The need for Data Science workers, or data scientists, has increased tremendously in recent years (Leopold, 2017).

2.1 Defining Data Science: Different Perspectives

Providing a precise definition for a discipline is important, especially for students who are eager to acquire important knowledge and skills for future job seeking. This section summarizes the definitions of Data Science and analyzes the features of this discipline. We also provide our own definition based on our understanding of this important field.

The following are prominent Data Science definitions we found in the literature:

- Data science as an expansion of the technical areas of the field of statistics. (Cleveland, 2001).
- “Data science involves principles, processes, and techniques for understanding phenomena via the (automated) analysis of data.” (Provost & Fawcett, 2013)
- “Data science is the study of the generalizable extraction of knowledge from data.” (DhaR, 2013)
- “An extension of information systems research.” (Agarwal & Dhar, 2014)
- “This coupling of scientific discovery and practice involves the collection, management, processing, analysis, visualization, and interpretation of vast amounts of heterogeneous data associated with a diverse array of scientific, translational, and interdisciplinary applications.” (Donoho, 2015)
- “Data science is now widely accepted as the fourth mode of scientific discovery, on par with theory, physical experimentation and computational analysis. Techniques based on Big Data are showing promise not only in scientific research, but also in education, health, policy, and business.” (Michigan Institute for Data Science, 2017)
- “Data science enables the creation of data products.” (Loukides, 2011, p.1)
- “Data science is the study of where information comes from, what it represents, and how it can be turned into a valuable resource in the creation of business and IT strategies.” (Banafa, 2014)
- “Data Science is not only a synthetic concept to unify statistics, data analysis and their related methods but also comprises its results. It includes three phases, design for data, collection of data, and analysis on data.” (Hayashi, 1998,p.41)
- “Data science is a combination of statistics, computer science, and information design” (Shum et al., 2013)
- “Data science is a new trans-disciplinary field that builds on and synthesizes a number of relevant disciplines and bodies of knowledge, including statistics, informatics, computing, communication, management, and sociology, to study data following data thinking” (Cao, 2017)
The above definitions reflect current understanding of this discipline. Together they help to describe the characteristics of Data Science, which can be summarized into the following aspects:

- The center of Data Science is data. Especially, big data becomes the subject that is investigated;
- The purpose of Data Science is to obtain information or knowledge from data. The information will help to make better decisions, and the knowledge may help an organization, a state, a country, or the whole of humanity to better understand the development or change of the nature or the society;
- Data science is a multi-disciplinary field which has applied theories and technologies from a number of disciplines and areas such as mathematics, statistics, computer sciences, information systems, and Information Science. It is expected that Data Science will bring change to impact these disciplines.

We define Data Science as:

Data Science is an interdisciplinary field that explores scientific methodology and computational technology about data, including data management, access, analysis, and evaluation for the benefits of human beings.

In our definition, data is at the center of Data Science. Any activities around data should be considered within the scope of this field. In other words, Data Science not only deals with data mining or data analysis, but also explores technologies and approaches for other important activities including data management, data access, and evaluation. We avoid to list related disciplines in the definition because we believe Data Science has impact and promises new opportunities to every scientific and industry field.

As an emerging discipline, Data Science is in still in its evolving stage. More research and investigation are needed to understand Data Science problems, challenges, theories, and methodologies. Interesting questions in Data Science to explore can be: what are its other major concepts in addition to data? What are the problems Data Science should attempt to tackle? And do we have enough theories and frameworks to guide future practice in Data Science? As specified by Provost and Fawcett (2013), in order for Data Science to flourish as a separate field, we should think beyond the specific algorithms, techniques, and tools that we are currently using and work on developing the core principles and concepts that underlie related techniques.

2.2 Most Related Disciplines and Fields for Data Science
Data science has been considered as an interdisciplinary field since its emergence. It was originally developed within the statistics and mathematics community (Cao, 2017). At its current stage, Data Science is more closely related to several disciplines than others, which is evident from the definitions in the literature provided in the previous section. Even though we believe Data Science is relevant to every discipline, some disciplines or fields are more related to it than others. These disciplines include Mathematics (especially Probability and Statistics), Information Science, Computer Science, Knowledge Management, Management Information Systems, and Decision Science. The Data Science programs at the doctoral and Master’s levels (see table 3, and table 4) confirm the above list, as most of the Data Science programs were established in colleges, schools, or departments of Information Science, Computer Science, Statistics, and Business Management.

Many studies have explored the connections between Data Science and Mathematics, Computer Science, and Management Information Systems (Cleveland, 2001; Shum, et al, 2013; Agarwal & Dhar, 2014). Mathematics including statistics provides numerical reasoning and methodologies for data analysis and processing. Some mathematic concepts and theorems such as those in Calculus, Linear Algebra, and statistics needed to be mastered by potential data science workers; Computer Science provides programming languages, database systems, and algorithms that a data science worker needs to manipulate numerical and textual data; Management Information Systems, on the other hand, have explored data mining and data analytics for years.

However, we found that the connection of Data Science with Information Science was largely neglected which may negatively impact the development of Data Science as a discipline. The impression of the public on Information Science is actually Library Science, which has focused more on classic information resource management, library management, and reference services in libraries. Even though Information Science has been closely related to Library Science in history, its landscape has dramatically changed since the advent of the Internet. Information Science can definitely provide useful theories and guidance to the development of Data Science. For example, one of the important concepts to explore in Data Science may be the data life cycle. There are many ways to describe and depict data life cycle (Chang, 2012). For data management, the data life cycle can include creating data, processing data, analyzing data, presenting data, giving access to data, and re-using data (Boston University Libraries, n.d.). Or it can be a series of processes to create, store, use, share, archive, and destroy data (Spirion, n.d.). Explanations of the different stages included in the life cycle can be different. But some processes, such as archiving, storing, and accessing have been well explored for information in Information Science.

Data Science is also closely related to Knowledge Management. Hawamdeh (2003) describes the field of Knowledge Management and its relationship with other disciplines for the purpose of educating knowledge professionals (p. 168). He considered Knowledge Management a multidisciplinary subject. Knowledge
management professionals need to acquire information technology and Information Science skills, which serve as foundation for higher-level knowledge work such as knowledge management and sharing. We expand on his framework by including an additional step, shown in Figure 1, called data processing, analysis, and processing, between the original step 2: Information acquisition and content management and Step 3: Information and knowledge sharing. In this consideration, Information Science, Data Science, and Knowledge Management are closely related and overlapping disciplines or areas. Typically, Data Science centers on activities in step 2, but should also include activities and tasks in steps 2, 4 and 5.

Figure 1. Data Science in the Context of Knowledge Process (Adapted from Hawamdeh (2003, p. 168))

2.3 Data Scientists: the Professions of Doing Data Science

The increasing interest in Data Science calls for more Data Science workers. Several reports indicated that data scientists, which is very frequently called the “sexiest job of the 21st century” (Davenport & Patil, 2012), are in high demand. Mckinsey and Company (2011) pointed to a fact that the U.S. is facing a shortage of 140,000 to 190,000 analytical and managerial skills necessary to handle business processes related to critical decision making and big data. Additionally, companies in other sectors have started recruiting data scientists as a crucial component to conduct their business strategy. For example, Greylock Partners, a leading venture capital firm in Silicon Valley that was an early investor in such well-known companies such as Facebook and LinkedIn, is now showing more concern about expanding its team-based portfolio and recruiting more talented people in the Data Science and Analytics fields (Davenport & Patil, 2012).
A data scientist is usually described as “a hybrid of data hacker, analyst, communicator, and trusted adviser” (Davenport & Patil, 2012, p. 73). Data scientists’ jobs “involve data collecting, cleaning, formatting, building statistical and mathematical models, and communicate professionally with their partners and clients” (Carter & Sholler, 2016). According to Stodder (2015), data scientists begin by analyzing business problems, and then providing strategic and meaningful insights about probable actions and future plans. This strategy involves identifying business drivers, establishing a creative team which is characterized by excellent technical and communication skills, and using the latest analytics and visualization techniques (Stodder, 2015). Provost and Fawcett (2013) specified that successful data scientists “must be able to view business problems from a data perspective.”

The literature also describes knowledge and skills that are necessary for data scientists. For example, DhaR (2013) believed that “a data scientist 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.” He considered four basic types of courses that contribute to the knowledge and skills for a qualified data scientist: Machine Learning, Computer Science (data structure, algorithms, and systems), correlation and causation, and problem formulation.

Data scientist can be a general term for people working in Data Science. Different titles can be found from job postings and literature for related positions. It is important to realize that data scientists may vary in their background, experiences, skills, and analytical workflows. In semi-structured interviews of 35 data analysts, Kandel, Paepcke, Hellerstein, and Heer (2012) indicated that analysts fall into three categories which are hacker, scripter, and application user, each with different levels of technical skills and tasks that can be performed. They, however, generally share five common tasks that support daily insights, which are discovery, wrangling, profiling, modeling, and reporting.

In order to verify what we have learned from the literature and to understand the increase in the demand for data scientists (Leopold, 2017), we conducted a small-scale job analysis. The analysis is reported in the next section.

3. Data Science and Data Analytics Jobs: An Analysis

One of the goals of this chapter is to explore how higher education could produce qualified Data Science workers, or data scientists. We therefore conducted a preliminary job analysis to understand current job markets for Data Science workers. This section will present the purposes of the analysis, the research questions, the process of data collection and analysis, and the results.
3.1 Purposes of Analysis and Research Questions

This analysis aimed to understand the current job market for data scientists. Such understanding would provide us with guidelines to designing Data Science courses and programs at undergraduate and graduate levels. Specifically, we would like to find answers to the following questions:

(1) What characterizes the employers who hire data scientists?
(2) What were the job titles used when employers are looking for qualified Data Science workers?
(3) What are the general qualifications, knowledge, tools, or skills required by employers?

As mentioned earlier, our analysis also served as a validation of knowledge, skills, and competences for qualified data scientists addressed in the literature.

3.2 Data Collection

We recruited the students of graduate-level Information Science class at the University of North Texas to collect the data. INFO 5717 is a course in which students learn to build web-based database systems. The students were assigned to five teams with three members in each team. They were asked to collect, process, and upload at least 60 job postings pertaining to data analysts, information management specialists, data scientists, and business intelligence analysts to a MySQL database.

Prior to data collection, we provided a list of metadata elements to guide the process. Students were required to collect information for at least the following elements for each job posting:

- Title of the Job
- Start date and end date for the position
- Name of the employer
- Salary
- URL of the job posting at the employer’s website
- Position requirements
- Position responsibilities

Students were instructed to collect the above data on the Internet using methods they considered appropriate. The data was collected from job aggregator websites such as Linkedin.com, indeed.com, dice.com, and glassdoor.com in November 2016. Additionally, some of the teams also collected data pertaining to the employers such as the URL of the official employer website, contact email, and the mailing address. The data was initially saved as text files.
We also used this project as an opportunity to train students for future Data Science related skills. Therefore we asked students to perform a series of tasks to process the collected data: each student team was required to design a job database using MySQL, containing one or two database tables to store the job posting and the employer information. Using HTML and PHP scripts, students created simple web entry forms to insert the job and employer information into their databases. As part of their course requirements, students developed web pages to manipulate or send queries to retrieve information from the data. They also developed authentication and session-control mechanisms to secure and use their database. In total the teams collected 364 job postings.

3.3 Data Cleanup and Integration

After students’ classwork, we exported the job and employer data from five databases to Excel spreadsheets for analysis. We found that even though we gave students metadata specifications, the database tables from each team were not completely aligned – some had more elements, or the same elements were called different names. Employer information was more problematic because we did not specify what exactly information should be collected about employers except that we needed to have the name and URL of the employer for each job posting. Furthermore, students might have collected the same job posting (duplication) as the teams worked independently of each other. Some of the collected records had some missing values such as job location and salaries because the information pertaining to the corresponding job was not stated clearly in the original announcements.

We conducted manual data cleanup to fix the above data issues. Some minor cases of missing salary values were handled by using online popular salaries estimate tools such as the job-search engine indeed.com (Doherty, 2010). The manual data cleanup was possible because we had only 364 job postings. After removing the duplicates and consolidating the data elements, we obtained 298 usable observations for analysis.

3.4 Tools for Data Analysis

For this small dataset, we mainly used Microsoft Excel to sort the different elements, to generate frequency counts and percentages, and to produce diagrams. Additionally, SAS Enterprise Miner (SASEM) 14.1 was utilized to perform cluster analysis (SAS Institute Inc, 2014) on qualifications and required experience as presented in the job postings. We also wrote a simple program using the R language to extract the most frequent keywords and produced a visual representation for the results called a word cloud (Heimerl, Lohmann, Lange, & Ertl, 2014). Manual data transformation and adjustment were performed in order to use these tools for the purpose of answering the research questions proposed in Section 3.1.
3.5 Results and Discussion

In this section, we present the results of our analysis and answer our research questions:

3.5.1 Characteristics of the Employers

From the 298 valid job postings, 236 company names were identified by the students. Among them we found some well-known companies such as Apple, Blue Cross and Blue Shield, Capital One, Bank of America, Ericsson, Google, IBM, Infosys, Nike, Samsung, Sprint, Verizon, Walmart, Wells Fargo, and Xerox. These employers covered different industrial fields such as Information Technologies, Healthcare, Insurance, Finance, Retail, Biotechnology, Business services, Manufacturing, and Media sectors.

With respect to job postings’ geographical concentration, Texas leads the United States with more than 24% of the jobs posting, next comes Illinois and California with 17% and 15% respectively (Figure 2). The collected data also showed that both graduates and undergraduates were possible candidates for Data Science jobs. The preferred levels of experience ranged from two to more than 10 years. These results portrayed an image of the job market for Data Science.

![Figure 2. Job Distributions Among States](image-url)
3.5.2 Job Titles

A frequency analysis of the 298 job postings showed that the top job titles used by the employers were: data scientist, data analyst, business intelligence analyst or intelligence analyst, information management specialist/analyst, and data engineer. For data scientist and data analysts, the job title could specify different levels, such as lead or principal, senior, junior, or intern. Some job titles, such as data analytics scientist and Data Science quantitative analyst, were a mixture of data analyst, data science, or data scientist. Other titles included big data architect, data integrity specialist, data informatics scientist, marketing analyst, MIS specialist, and project delivery specialist and consultant on data analytics.

<table>
<thead>
<tr>
<th>Job Titles</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Scientist</td>
<td>95 (31.9%)</td>
</tr>
<tr>
<td>Data Analyst</td>
<td>95 (31.9%)</td>
</tr>
<tr>
<td>Business Intelligence Analyst (Intelligence Analyst)</td>
<td>50 (16.8%)</td>
</tr>
<tr>
<td>Information Management Specialist/Analyst</td>
<td>13 (4.4%)</td>
</tr>
<tr>
<td>Data Engineer</td>
<td>10 (3.3%)</td>
</tr>
<tr>
<td>Research Analyst</td>
<td>5 (1.7%)</td>
</tr>
<tr>
<td>Other</td>
<td>30 (10%)</td>
</tr>
<tr>
<td>Total</td>
<td>298</td>
</tr>
</tbody>
</table>

3.5.3 Word Cloud and Clusters on Qualifications and Responsibilities

We used the R programming language to visually summarize the most frequent terms occurred in two fields: qualifications and responsibilities. The word cloud in Figure 3 shows that terms like “experience”, “business”, “analyst”, “analysis”, “management”, and “skills” stood out for their frequent occurrences, as depicted in Figure 1. This was supported by many field researchers, career practitioners, and job posters who confirmed that effective data scientists and analysts need to have a substantial level of managerial and business experience in addition to sophisticated skills in programming, development, design, and statistics (Davenport & Patil, 2012; Kandel, Paepcke, Hellerstein, & Heer, 2012).
We also conducted cluster analysis on these two fields. The purpose of cluster analysis was to identify concepts that are implied in the dataset. The process starts by exploring the textual data content, using the results to group the data into meaningful clusters, and reporting the essential concepts found in these clusters (SAS Institute Inc, 2014). Using the SASEM cluster analysis function, which is based on the mutual information weighting method, we obtained five clusters as presented in Table 2. According to the terms in each cluster, we could name Cluster 1 as project management; Cluster 2 as machine learning and algorithmic skills; Cluster 3 as statistical models and business analytics; Cluster 4 as database management & systems support, and Cluster 5 as communication skills. Among them, Cluster 2 is the largest cluster with a relative frequency of 47% among all the terms in these two fields.

Table 2: The Clusters of the Job Postings

<table>
<thead>
<tr>
<th>Cluster ID</th>
<th>Cluster Name</th>
<th>Descriptive Terms</th>
<th>Frequency (Percentage)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Project Management</td>
<td>stakeholder, project, presentation, recommendation, communicate, plan, understand, fast-pace, requirement, manage, team multiple, issue, improvement</td>
<td>46 (15%)</td>
</tr>
<tr>
<td>2</td>
<td>Machine Learning and</td>
<td>statistical, technique, machine, model, test, learning, system, report, create,</td>
<td>144 (47%)</td>
</tr>
<tr>
<td></td>
<td>Algorithmic skills</td>
<td>Design, requirement multiple, implement</td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>-------------------</td>
<td>-----------------------------------------</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Statistic models and business analytics</td>
<td>sale, market, customer analytics, help analytic, build, insight, service, opportunity, engineer key, solution, team</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Database management &amp; systems support</td>
<td>request, assign, perform, process, procedure information, maintain, database, improvement, relate, assist, support, issue, service</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Communication skills</td>
<td>ability, experience, skill strong, fast-pace highly, demonstrate, communication, write multiple, service analytical, customer information</td>
<td></td>
</tr>
</tbody>
</table>

---

Figure 4. The Concepts Link Map of the Job Postings

The Text Filter Node functionality in SAS enabled us to invoke a filter viewer to explore interactive relationships among terms (SAS Institute Inc, 2014). To preview the concepts that are highly correlated with the term *data*, a concepts link map was generated, as depicted in Figure 4. The five clusters in Table 2 become the six nodes (two notes on Analysis/Analyze) with links to other outer notes.

The word cloud and cluster analysis showed a broader view of the terms from the job market perspective. Also, further exploration of the concepts link map revealed that data scientists need to have substantial skills in Hadoop, Hive, Pig, Python,
Machine Learning, C++ and Java programming, R, and SQL, in addition to being proficient with statistical and modeling packages such as SPSS and SASEM.

3.6 Summary of the Job Posting Analysis

This small-scale study covers many activities a data analyst may perform according to (Donoho, 2015). We realized that a Data Science worker needed to know how to organize and cleanup data, how to analyze them, and how to create data visualizations using different tools. It would be very beneficial if he or she also possesses basic research skills, such as how to collect data and how to evaluate it. Data analysis also involves team-based work of data scientists who use effective tools to build statistical and predictive models to produce business and actionable insights within their organizations, which should be also built in Data Science curriculum.

Our job analysis above provides insight into this new Data Science discipline and its career opportunities. These insights can guide educators to develop appropriate Data Science programs and courses that meet the short-term and long-term needs of the digital economy. Next we present our thoughts on Data Science education based on this job analysis and an overview of existing Data Science programs.

4. Data Science Education: Current Data Science Programs and Design Considerations

The increasing demand in Data Science workers has motivated rapid growth of Data Science related programs, as presented in this section. Universities and colleges have been actively developing new Data Science programs and courses at different levels to meet the needs of the job market on qualified Data Science workers. It is necessary to review and summarize what has been established so far in order to develop high-quality Data Science programs.

4.1 Data Science Programs Overview

As we have discussed earlier, Data Science is an interdisciplinary field. It integrates and applies knowledge and techniques in multiple disciplines and fields. Data Science programs and courses have been established to meet the needs for training effective data scientists. These programs offer a variety of courses to train students with the research and professional skills to succeed in leading edge organizations. This section summarizes existing Data Science programs and trainings.

4.1.1 PhD Programs in Data Science

According to a Data Science salary study, 88% of data scientists have advanced degrees and 46% have a Ph.D. (Burtch Works Executive Recruiting, 2017). As of
June 2017, the website MastersInDataScience.org lists 18 universities offering
doctorate-level programs in Data Science and Data Analytics, as showed in Table 3.

Table 3. PhD Programs in Data Science

<table>
<thead>
<tr>
<th>University</th>
<th>School or Department</th>
<th>Degree, Online or Campus, and URL</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Colorado Technical University</strong></td>
<td>Computer Science Department</td>
<td>Doctor of Computer Science with a concentration in Big Data Analytics, Online, <a href="http://www.coloradotech.edu/degrees/doctorates/computer-science/big-data-analytics">http://www.coloradotech.edu/degrees/doctorates/computer-science/big-data-analytics</a></td>
</tr>
<tr>
<td><strong>Indiana University Bloomington</strong></td>
<td>School of Informatics and Computing</td>
<td>PhD minor in Data Science, Campus or Online <a href="http://www.soic.indiana.edu/graduate/degrees/data-science-minor.html">http://www.soic.indiana.edu/graduate/degrees/data-science-minor.html</a></td>
</tr>
<tr>
<td><strong>Brown University</strong></td>
<td>Computer Science Department</td>
<td>PhD in Computer Science – concentration in Data Science, Campus, <a href="http://cs.brown.edu/~kraskat/phd.html">http://cs.brown.edu/~kraskat/phd.html</a></td>
</tr>
<tr>
<td><strong>Chapman University</strong></td>
<td>Schmid College of Science &amp; Technology</td>
<td>Ph.D. in Computational and Data Sciences, Campus, <a href="http://www.chapman.edu/scst/graduate/phd-computational-science.aspx">http://www.chapman.edu/scst/graduate/phd-computational-science.aspx</a></td>
</tr>
<tr>
<td><strong>Georgia State University</strong></td>
<td>Department of Computer Science</td>
<td>Ph.D. (Bioinformatics Concentration), Campus, <a href="http://cs.gsu.edu/graduate/doctor-philosophy/ph-d-bioinformatics-concentration-degree-requirements/">http://cs.gsu.edu/graduate/doctor-philosophy/ph-d-bioinformatics-concentration-degree-requirements/</a></td>
</tr>
<tr>
<td><strong>Indiana University-Purdue University-Indianapolis</strong></td>
<td>School of Informatics and Computing</td>
<td>PhD in Data Science, Campus, <a href="https://soic.iupui.edu/hcc/graduate/data-science-phd/">https://soic.iupui.edu/hcc/graduate/data-science-phd/</a></td>
</tr>
<tr>
<td><strong>Kennesaw State University</strong></td>
<td>Department of Statistics and Analytical Sciences</td>
<td>PhD in Analytics and Data Science, Campus, <a href="http://cs.kennesaw.edu/datascience/">http://cs.kennesaw.edu/datascience/</a></td>
</tr>
<tr>
<td><strong>New York University</strong></td>
<td>Tandon School of Engineering</td>
<td>PhD in Computer Science with Specialization in Visualization, Databases, and Big Data, Campus, <a href="http://engineering.nyu.edu/academics/programs/computer-science-phd">http://engineering.nyu.edu/academics/programs/computer-science-phd</a></td>
</tr>
<tr>
<td><strong>Newcastle University</strong></td>
<td>School of Computer Science</td>
<td>EPSRC CDT in Cloud Computing for Big Data, Campus, <a href="http://www.bigdata-cdt.ac.uk/">http://www.bigdata-cdt.ac.uk/</a></td>
</tr>
<tr>
<td>Institution</td>
<td>Department</td>
<td>Program Description</td>
</tr>
<tr>
<td>-------------------------------------------------</td>
<td>----------------------------------------------</td>
<td>-------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Oregon Health &amp; Science University</td>
<td>Department of Medical Informatics and Clinical Epidemiology</td>
<td>PhD in Biomedical Informatics - Bioinformatics &amp; Computational Biology Track, Clinical Informatics Track, Campus, <a href="http://www.ohsu.edu/xd/education/schools/school-of-medicine/departments/clinical-departments/dmice/educational-programs/clinical-informatics.cfm">http://www.ohsu.edu/xd/education/schools/school-of-medicine/departments/clinical-departments/dmice/educational-programs/clinical-informatics.cfm</a></td>
</tr>
<tr>
<td>University of Cincinnati</td>
<td>College of Medicine</td>
<td>PhD in Biostatistics - Big Data Track, Campus, <a href="http://med.uc.edu/eh/divisions/bio/programs/phd-biostatistics-big-data-track">http://med.uc.edu/eh/divisions/bio/programs/phd-biostatistics-big-data-track</a></td>
</tr>
<tr>
<td>University of Maryland- College Park</td>
<td>College of Information Studies</td>
<td>PhD in Information Studies – Concentration in Big Data/Data Science, Campus, <a href="https://ischool.umd.edu/phd-admissions">https://ischool.umd.edu/phd-admissions</a></td>
</tr>
<tr>
<td>University of Massachusetts-Boston</td>
<td>College of Management</td>
<td>PhD in Business Administration - information systems for data science track, Campus <a href="https://www.umb.edu/academics/cm/business_administration_phd/mis">https://www.umb.edu/academics/cm/business_administration_phd/mis</a></td>
</tr>
<tr>
<td>University of Michigan-Ann Arbor</td>
<td>Department of Statistics</td>
<td>Ph.D. in Statistics, Campus, <a href="http://lsa.umich.edu/stats/graduate-students/graduate-degree-programs/phdprograms.html">http://lsa.umich.edu/stats/graduate-students/graduate-degree-programs/phdprograms.html</a></td>
</tr>
<tr>
<td>University of Pittsburgh</td>
<td>Graduate School of Public</td>
<td>PhD in biostatistics, Campus, <a href="http://www.publichealth.pitt.edu/biostatistics/academics/phd">http://www.publichealth.pitt.edu/biostatistics/academics/phd</a></td>
</tr>
<tr>
<td>University of Washington-Seattle</td>
<td>eScience Institute</td>
<td>PhD in Big Data and Data Science, Campus, <a href="http://escience.washington.edu/education/phd/igert-data-science-phd-program/">http://escience.washington.edu/education/phd/igert-data-science-phd-program/</a></td>
</tr>
<tr>
<td>Worcester Polytechnic Institute</td>
<td>College of Arts &amp; Sciences</td>
<td>PhD in Data Sciences, Campus, <a href="https://www.wpi.edu/academics/study/data-science-phd">https://www.wpi.edu/academics/study/data-science-phd</a></td>
</tr>
</tbody>
</table>

The majority of these PhD programs require students to have undertaken technical coursework or earned a technical degree in Computer Science, Mathematics, or Statistics as a pre-requisite for entrance. These programs are on-campus and require students to be present in class to complete their degrees, only a couple of
programs offer online classes but still have the requirement of attending campus activities in order to obtain the degree.

4.1.2 Master Programs in Data Science

The same Data Science salary study also mentioned that 59% of early-career data scientists’ highest degree was a Master’s, a significant increase from 48% in 2015 (Burtch Works Executive Recruiting, 2017). As of June 2017, the site MastersInDataScience.org lists 23 schools offering Master’s programs in Data Science, as showed in Table 4.

<table>
<thead>
<tr>
<th>University</th>
<th>School/College</th>
<th>Degree, Online Option, URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Southern Methodist University</td>
<td>Dedman College of Humanities and Sciences, Lyle School of Engineering and Meadows School of the Arts</td>
<td>Master of Science in Data Science, online option, <a href="https://requestinfo.datascience.smu.edu/index10.html">https://requestinfo.datascience.smu.edu/index10.html</a></td>
</tr>
<tr>
<td>University of California-Berkeley</td>
<td>School of Information</td>
<td>Master of Information and Data Science, online option, <a href="https://requestinfo.datascience.berkeley.edu/index3.html">https://requestinfo.datascience.berkeley.edu/index3.html</a></td>
</tr>
<tr>
<td>Arizona State University</td>
<td>W.P. Carey School of Business</td>
<td>Master of Science in Business Analytics, online option, <a href="https://programs.wpcarey.asu.edu/masters-programs/business-analytics">https://programs.wpcarey.asu.edu/masters-programs/business-analytics</a></td>
</tr>
<tr>
<td>Carnegie Mellon University</td>
<td>School of Computer Science</td>
<td>Master of Computational Data Science, No online option, <a href="https://mcds.cs.cmu.edu/">https://mcds.cs.cmu.edu/</a></td>
</tr>
<tr>
<td>Cornell University</td>
<td>Department of Statistical Science</td>
<td>Master of Professional Studies in Applied Statistics (Option II: Data Science), No online option, <a href="http://stat.cornell.edu/academics/mps">http://stat.cornell.edu/academics/mps</a></td>
</tr>
<tr>
<td>Georgia Tech</td>
<td>College of Computing, College of Engineering, and Scheller College of Business</td>
<td>Master of Science in Analytics, No online option, <a href="https://analytics.gatech.edu/">https://analytics.gatech.edu/</a></td>
</tr>
<tr>
<td>Institution</td>
<td>Department/Program</td>
<td>Degree Offered</td>
</tr>
<tr>
<td>-----------------------------------------------------------------</td>
<td>-------------------------------------------------------------------------------------</td>
<td>---------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Illinois Institute of Technology</td>
<td>College of Science</td>
<td>Master of Data Science, online option, <a href="http://science.iit.edu/programs/graduate/master-data-science">http://science.iit.edu/programs/graduate/master-data-science</a></td>
</tr>
<tr>
<td>Indiana University, Bloomington</td>
<td>School of Informatics &amp; Computing</td>
<td>Master of Science in Data Science, online option, <a href="http://www.soic.indiana.edu/graduate/degrees/data-science/index.html">http://www.soic.indiana.edu/graduate/degrees/data-science/index.html</a></td>
</tr>
<tr>
<td>New York University</td>
<td>Center for Data Science</td>
<td>Master of Science in Data Science, No online option, <a href="http://cds.nyu.edu/academics/ms-in-data-science/">http://cds.nyu.edu/academics/ms-in-data-science/</a></td>
</tr>
<tr>
<td>North Carolina State University</td>
<td>Institute for Advanced Analytics</td>
<td>Master of Science in Analytics, No online option, <a href="http://analytics.ncsu.edu/">http://analytics.ncsu.edu/</a></td>
</tr>
<tr>
<td>Northwestern University</td>
<td>McCormick School of Engineering and Applied Science, and School of Continuing Studies</td>
<td>Master of Science in Analytics, and Master of Science in Predictive Analytics, <a href="http://www.mccormick.northwestern.edu/analytics/">http://www.mccormick.northwestern.edu/analytics/</a></td>
</tr>
<tr>
<td>Rutgers, The State University of New Jersey</td>
<td>Computer Science Department</td>
<td>Master of Science in Data Sciences, No online option, <a href="https://msds-cs.rutgers.edu/msds/aboutpage">https://msds-cs.rutgers.edu/msds/aboutpage</a></td>
</tr>
<tr>
<td>Texas A&amp;M University</td>
<td>Department of Statistics</td>
<td>Master of Science in Analytics, online option, <a href="http://analytics.stat.tamu.edu/">http://analytics.stat.tamu.edu/</a></td>
</tr>
<tr>
<td>University of California, San Diego</td>
<td>Departments of Computer Science and Engineering</td>
<td>Master of Advanced Study in Data Science and Engineering, No online option, <a href="http://jacobsschool.ucsd.edu/mas/dse/">http://jacobsschool.ucsd.edu/mas/dse/</a></td>
</tr>
<tr>
<td>University of Illinois at Urbana-Champaign</td>
<td>Department of Computer Science, Department of Statistics, School of Information Sciences and Coursera</td>
<td>Professional Master of Computer Science in Data Science, online option, <a href="https://online.illinois.edu/mcs-ds">https://online.illinois.edu/mcs-ds</a></td>
</tr>
<tr>
<td>University of Minnesota – Twin Cities</td>
<td>College of Science &amp; Engineering, College of Liberal Arts and School of Public Health</td>
<td>Master of Science in Data Science, No online option, <a href="https://datascience.umn.edu/">https://datascience.umn.edu/</a></td>
</tr>
</tbody>
</table>

18
<table>
<thead>
<tr>
<th>University of San Francisco</th>
<th>College of Arts and Sciences</th>
<th>Master of Science in Analytics, No online option, <a href="https://www.usfca.edu/arts-sciences/graduate-programs/analytics">https://www.usfca.edu/arts-sciences/graduate-programs/analytics</a></th>
</tr>
</thead>
<tbody>
<tr>
<td>University of Southern California</td>
<td>Viterbi School of Engineering</td>
<td>Master of Science in Computer Science – Data Science, online option, <a href="https://gapp.usc.edu/graduate-programs/masters/computer-science/data-science">https://gapp.usc.edu/graduate-programs/masters/computer-science/data-science</a></td>
</tr>
<tr>
<td>University of Virginia</td>
<td>Data Science Institute</td>
<td>Master of Science in Data Science, No online option, <a href="https://dsi.virginia.edu/academics">https://dsi.virginia.edu/academics</a></td>
</tr>
<tr>
<td>University of Washington - Seattle Campus</td>
<td>Interdisciplinary</td>
<td>Master of Science in Data Science, No online option, <a href="https://www.datasciencemasters.uw.edu/">https://www.datasciencemasters.uw.edu/</a></td>
</tr>
<tr>
<td>University of Wisconsin</td>
<td>Department of Statistics, and UW Extension</td>
<td>Master of Science in Statistics - Data Science, and Online Master of Science in Data Science, <a href="https://advanceyourcareer.wisc.edu/degrees/data-science/">https://advanceyourcareer.wisc.edu/degrees/data-science/</a></td>
</tr>
</tbody>
</table>

Source: http://www.mastersindatascience.org

These Master's programs have a duration from nine months to about two years, depending on curriculum requirements and the part-time or full-time status of the students. Students usually begin by taking required courses specified in the curriculum in the first semester. Once they obtain basic knowledge and become familiar with technologies in Data Science such as programming, databases, and statistics, they proceed to do practicums, application, and capstone projects. Some of the schools provide internship opportunities and practicums so students can work with partner companies in industry or government agencies during their study.

4.1.3 Graduate Certificates Programs in Data Science

Certificate programs provide a flexible way for people to study Data Science. A certificate program generally takes less time and cost less money than a program intended for earning a degree. It is an attractive option for people who are interested in pursuing a career in this field. The website MastersInDataScience.org has a list of 92 universities offering graduate certificate programs in Data Science and related specialties.
4.1.4 Massive Open Online Courses (MOOCs)

MOOCs are online courses with open (mostly free) access via the Internet. These courses are similar to university courses, but do not tend to offer academic credit. They allow students to choose their own academic path and complete the courses on their own schedule. To study Data Science or data analysis, students may choose courses from MOOCs – Coursera, for example, and construct a learning plan primarily from the following sample areas:

- **Programming languages**: The most popular languages currently used are Python and R for Data Science courses
- **Statistics and Probability**: There are broad theories that are used for making informed choices in analyzing data and for drawing conclusions about scientific truths from data.
- **Data collection**: Extracting data from the web, from APIs, from databases, and from other sources
- **Data cleaning and managing**: Manipulating and organizing the data collected to make data set useful for data analysis tasks
- **Exploratory data analysis**: Exploring data to understand the data’s underlying structure and summarizing the important characteristics of a data set
- **Data Visualization**: Using plotting systems to construct data graphics, analyzing data in graphical format, and reporting data analyze results
- **Machine Learning**: Building prediction functions and models by using supervised learning, unsupervised learning and reinforcement learning

4.1.5 Bootcamps

Bootcamp programs are non-traditional educational paths. Compared with traditional degrees, these programs are intense and have faster routes to the workplace. It is another education option for considering a career as a data scientist. Data Science Bootcamps provided by Data Science community (Datascience community, n.d.) has a list of bootcamps available for Data Science and data engineering.

In summary, Data Science programs at undergraduate, master, and doctoral level have been developed in the U.S. These programs provide a wide range of choices to students who want to obtain knowledge and skills in Data Science. Still, more Data Science programs are being developed. It maybe the time to explore the characteristics of a competitive Data Science program.

4.2 Data Science Program: An Integrated Design
Based on our understanding of Data Science and its related concepts, the knowledge and skills required for Data Science workers, and current Data Science programs, we believe that a high quality Data Science program should provide courses and training in the following areas:

- **Fundamental concepts, disciplines, and the profession.** The program should offer one or two courses to introduce the student basic concepts, disciplines, and the profession of Data Science. These courses set up solid foundation for students to learn more advanced concepts and applications. They may teach not only concepts and characteristics of data, information, knowledge, and data life cycle, but also related mathematical concepts, functions, models, and theorems. Students should develop an affection to data and willing to do Data Science;

- **Statistical analysis and research methodology.** The program should offer multiple courses to teach data collection and data analysis skills, which are usually taught in master or doctoral level research methodology classes.

- **Programming languages, algorithms, and tools.** Courses such as computer algorithms, programming languages, data mining, database design, and machine learning help students to apply computational techniques for data collecting, cleaning, analysis, prediction, and presentation.

- **Business logic and soft skills.** Students need to understand the purpose of Data Science, they need to learn how to communicate and collaborate with different business units such as sales, marketing, and production. Also, learn to effectively present data analysis results is also important.

- **Practicum or a capstone project.** Data Science students should do at least one project that integrates what they have learned to address a challenging real-world Data Science problem. The ideal case would be to have the student work in a company or an organization to participate in real Data Science projects that may help the organization to make better business decisions.

Each student can choose different courses or focus on different areas to improve their knowledge and skills based on their backgrounds and personal interests. For example, a student with business background may want to take mainly courses on data management and programming languages, while, an Information Science student should study business logic in addition to statistical analysis and data visualization through coursework or the practicum.

A course may also teach student knowledge and skills in different areas. For example, a database course may teach students data collection, cleaning, analysis and report writing through its term project. The instructor can also design projects that of research values or reflecting real industry needs.

In general, a flexible curriculum and close connections with profit and non-profit organizations provide students much flexibility and opportunities to become successful data workers.
5. Summary and Conclusions

This chapter explores the concept and characteristics of data. It reviews different perspectives from literature on Data Science and provides the authors' definition of this emerging field. The interdisciplinary nature of Data Science is discussed. We then present a preliminary analysis of about 300 job postings on data scientists and related positions. Based on the literature review and the job posting analysis, we summarize the knowledge and skills as required for different levels of Data Science workers. Furthermore, we reviewed current Data Science programs, focusing on the shared values and commonalities of these programs. This chapter concludes with a list of curriculum considerations for designing a Data Science program.

References


Datascience community (n.d.) Data science bootcamps. Available at: http://datascience.community/bootcamps.


