A Data Science Approach to Defining a Data Scientist

Andy Ho[1], An Nguyen[1], Jodi Pafford[1], and Dr. Robert Slater[1]

Master of Science in Data Science, Southern Methodist University, Dallas TX 75275 USA {atho,angyuen2, jpafford, rslater}@smu.edu

Abstract. In this paper, we present a novel approach for the problem of not having a common definition and list of skillsets for a Data Scientist. As it relates to the field of data science, adoption has spanned nearly all industries and disciplines. This has increased the necessity for a wide range of technical skills, soft skills, experiences, and education to work on applications within data science (i.e. Artificial Intelligence, Machine Learning, Big Data Analytics, Data Visualization, etc.). The result is an overlap and ambiguity of various roles such as data scientist, data engineer, data architect, data analyst, research scientist, and research analyst. To solve the problem, we collected numerous job postings from multiple sources that share similar attributes and job descriptions to form our data set. Natural Language Processing (NLP) and advanced Machine Learning (ML) was performed to identify and visualize common patterns of job descriptions/summaries, experiences, skills, educational preferences to arrive at our key findings on the definition and list of skillsets of a data scientist. A secondary finding verified various industries (and even within the same industry) experience this same problem. Our conclusion is a Data Scientist is an individual who discovers valuable insights from large amounts of data possessing programming skills, well versed in statistics and mathematics, and can adequately visualize and communicate the findings.

1 Introduction

Over 2.5 quintillion bytes of data is created every single day and the pace is expected to accelerate with the growth of the Internet of Things (IoT), a society more dependent on data, and more businesses making data-driven decisions. On any given day, 500 million tweets are sent, 294 billion emails delivered, 4 petabytes of data created on Facebook, and IoT products such as driverless cars, wearables, and smart cities will push the daily amount collected to 463 exabytes. This expansion of data has created the need for individuals (i.e. Data Scientists) who are trained, skilled, and educated in the field (and application) of Data Science.

There is no common definition and a list of skillsets for a Data Scientist. This becomes evident when the question of What is a Data Scientist or What does a Data Scientist do is presented or when a job search is ran on multiple job sites (i.e. Indeed, Monster, ZipRecruiter). According to Dictionary.com, a data scientist is a person employed to analyze and interpret complex digital data, such as the usage of statistics of a website, especially in order to assist a business in its decision-making. But we postulate that there is more to a data scientist, and a common definition along with a list of skills can be reached using data science techniques.

To apply data science to this problem, our solution was to gather data, clean the data, model and analyze to find patterns and/or classifications within the data. NLP and ML was utilized to arrive at our solution. The end result is visualizations and explanations of our findings.

Since there is no public data sets available for this problem, we web scraped job postings listed under Data Scientist and other similar roles from multiple job sites to form our training data set. Features such as job title, job description/summary, experience and skill requirements, and location was collected. NLP was performed to read, decipher, understand, and make sense of each job posting. Unsupervised ML within NLP was performed to find document and text similarities between each job posting. Multiple comparative analytics provided additional results and insights to arrive at a common definition and list of skillsets of a data scientist.

The main conclusions are the definition and list of skills of data scientist varies from one job posting to another. We observed that multiple job sites have different algorithms presenting roles that seem like a data scientist but in fact an entirely different role. In applying data science techniques to this problem, the following are a few key findings:

1. Data Scientists at minimum have programming expertise in Python or R
2. Data Scientists have a background or knowledge of mathematics or statistics
3. Data Scientists have different domain expertise
4. Data Scientists, in general, possess degree(s) in computer science, mathematics, or some related science field

In addition to the Abstract and Introduction section, this paper is organized in the following manner: Tutorial Sections to educate readers on the general principles and techniques used in NLP and ML for this research; a Data Set section explaining how data was collected, features selected and visualization from the data exploration analysis; a Methods section on the data science techniques and algorithms performed; a Related Work section discussion other studies already completed; a Results section listing and explaining our experiment; an Analysis section on the results; an Ethics section on the possible ethical ramification of our findings; and a Conclusions section listings multiple conclusions from our analysis.

2 Artificial Intelligence Explained

To understand the data science techniques applied to our research, we want to first provide a baseline understanding for Artificial Intelligence (AI). This term has been associated to robots functioning on their own or a global network of computers rising up against mankind. Though that may be the case in the distant future, present AI is the development of advanced software programs to perform tasks that would normally require human intelligence. Alan Turings paper established the fundamental goal and vision for AI [8]. Much has developed since Turings endeavor to simulate human intelligence in machines, Table 1 shows a few examples of applied AI today:

Table 1.

1. Translation between languages (i.e. Google Translate)
2. Facial Recognition (i.e. Auto Tag on Facebook)
3. Virtual Assistant (i.e. Amazon Alexa)

The data science techniques applied will be in the realm of Narrow AI where machine learning and natural language processing is used to perform tasks in data collection, analysis, and modeling.

2.1 The Importance of Machine Learning to AI

In short, ML is when computers learn from the data it is provided. Without ML, AI could not provide insights, decisions, and/or actions for the examples of applied AI. There are many ML algorithms, all of which can be stratified into 3 classifications, described in Table 2.

Table 2. Machine Learning Classifications.

Classification	Description
Supervised Learning (SL)	Labeled input/output data is fed into an algorithm multiple times to arrive at a pattern for prediction. Algorithm examples could be Linear/Logistic Regression, K-Nearest Neighbor, Nave Bayes, Decision Tree.
Unsupervised Learning (UL)	Labeled input is fed into an algorithm multiple times to form clusters for the unlabeled output data. A new input is then added to predict which cluster it is associated with. Algorithm examples could be K-means clustering, Principal Component Analysis.
Reinforcement Learning (RL)	A reward base learning where feedback is provided on the output to improve the prediction accuracy.

In this paper, we will apply both SL and UL algorithms to find clusters and patterns for the data collected from natural language processing.

2.2 The Application of Natural Language Processing

NLP is a subfield of AI that focuses on how to program computers to process and analyze large amounts of human/natural language data. Both speech and written language is included in NLP. For this paper, only written text will be used to solve the problem. Using NLP, we were able to web scrape job postings from multiple sources to create our data set to perform lexical, syntactic, and semantic analysis. Throughout this paper, Python packages specializing in NLP will be leveraged to get text and document similarity scores of the job postings for each job title searched.

3 Data Set Creation

Data is collected from Indeed.com. Indeed is the number 1 job site in the world with over 250 million unique visitors every month[6]. Indeed gives users free access to complete job-seeking tasks such as searching for jobs, posting resumes, and researching companies. Globally, indeed.com has 9.8 jobs adds posted every second. Data is not readily available for download from indeed.com, so it must be scraped. Scraping data from the internet is a process commonly used by Data Scientist to get the information needed when not already available in an acceptable format. Table 3 illustrates the data that was gather during the scraping process.

Table 3. Data Set Description.

Data Set	Variables
Job Search Page	Link
Job Search Page	Job Title
Job Search Page	Company Name
Job Search Page	Location
Job Search Page	Summary
Job Search Page	Salary (if available)
Job Description Page	Full Text divided by ”jobSectionHeader”

3.1 Exploratory Data Analysis

The data contains 500 job descriptions (100 from each job title).

4 Methodology

4.1 Document Sentiment Comparitive Analysis

Document sentiment is being achieved by NLP.

4.2 Cluster Analysis on Key Terms

ML is used to assist in the cluster analysis.

5 Related Work

In a paper from the University of Rome, Mauro et al. presented a classification of job roles and skills in the area of Big Data Analytics. The researchers used web scraping to retrieve job postings from many prominent websites. Natural language processing was then applied to this dataset to discover four essential job groups, most frequent bigrams appearing in the job titles: Business Analysts, Data Scientists, Developers and System Mangers. Then using the Latent Dirichlet Allocation, LDA, classification techniques the authors clusters skills into 9 topics that were generated by human interpretation of the skills. The 9 topics are: Cloud, Coding, Database management, Architecture, Project Management, Systems Management, Distributed Computing, Analytics, Business impact. Finally, the job skill sets are mapped to job roles by a measure of the extent at which each skill set is represented within each job post description[1]. (see Fig. 1)

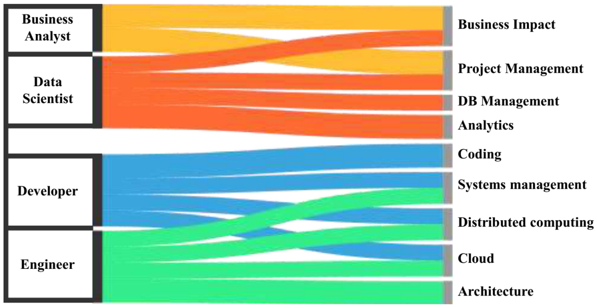
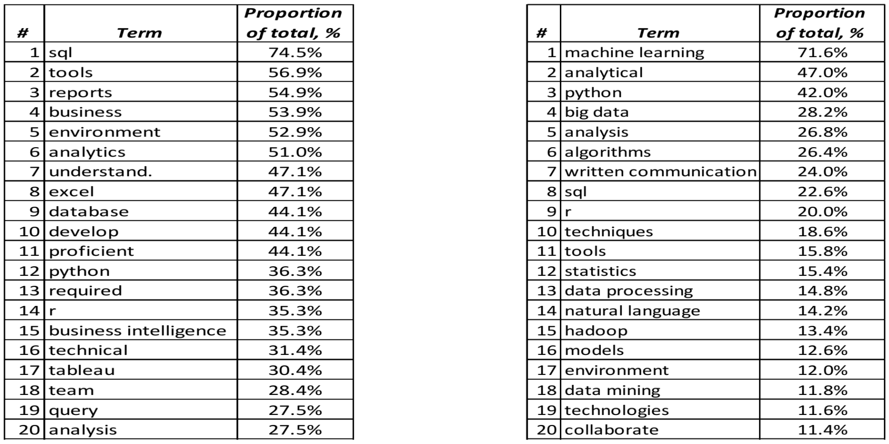


Fig. 1. Job skill sets are mapped to job roles by a measure of the extent at which each skill set is represented within each job post description.

A second group out of California State University focused entirely on the difference between Business data analytics, DBA and data science, DS. Radovilski et al. manually collected job descriptions of DBA and DS jobs from job boards. Using the job description they identify skill sets associated with Business, Analytical, Technical and Communication knowledge domains. Using text mining approaches, Document Data Matrix, Term Cloud, Singular Vector Decomposition, VARIMAX rotation and Latent Class Analysis the authors found the most frequent BDA and DS terms used, (see Fig. 2) and created a Term Cloud, (see Fig. 3)[9].

Some published papers to add to the paper at a later time. [4, 5, 2, 7, 8]



#	Term	Proportion of total, %
1	sql	74.5%
2	tools	56.9%
3	reports	54.9%
4	business	53.9%
5	environment	52.9%
6	analytics	51.0%
7	understand.	47.1%
8	excel	47.1%
9	database	44.1%
10	develop	44.1%
11	proficient	44.1%
12	python	36.3%
13	required	36.3%
14	r	35.3%
15	business intelligence	35.3%
16	technical	31.4%
17	tableau	30.4%
18	team	28.4%
19	query	27.5%
20	analysis	27.5%

#	Term	Proportion of total, %
1	machine learning	71.6%
2	analytical	47.0%
3	python	42.0%
4	big data	28.2%
5	analysis	26.8%
6	algorithms	26.4%
7	written communication	24.0%
8	sql	22.6%
9	r	20.0%
10	techniques	18.6%
11	tools	15.8%
12	statistics	15.4%
13	data processing	14.8%
14	natural language	14.2%
15	hadoop	13.4%
16	models	12.6%
17	environment	12.0%
18	data mining	11.8%
19	technologies	11.6%
20	collaborate	11.4%

Fig. 2. Job skill sets are mapped to job roles by a measure of the extent at which each skill set is represented within each job post description.

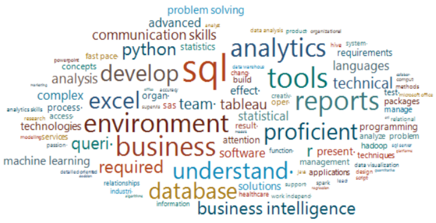


FIGURE 3. TERM CLOUD FOR BDA JOBS.

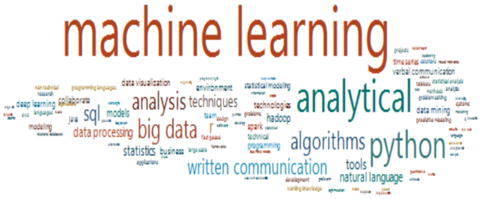


FIGURE 4. TERM CLOUD FOR DS JOBS.

Fig. 3. Job skill sets are mapped to job roles by a measure of the extent at which each skill set is represented within each job post description.

6 Results

6.1 What is a Data Scientist?

7 Analysis

7.1 Discoveries

8 Ethical Considerations

Ethics plays a role in the entire job search and interviewing process. There are many laws and regulations that oversee the process once the interviewing begins, however, there are not many laws and regulations when it comes to the job search process.

8.1 Website Usage

Scraping data from the website must be done with extreme caution. Each website is required to publish a robots.txt file that describes sections of the website that is not allowed to be scraped. Additionally, a websites terms and conditions may prevent someone from scraping. Falling outside the guidelines and/or company policies can be bad. There are criminal implications such as identity theft and hacking if information is scraped from a website without following the proper protocol. For the novice programmer, it can be easy to make this mistake. According to the 2016 lawsuit, Linkedin V. Doe Defendants, Linkedin sued 100 people who scraped their website anonymously. The lawsuit was stopped in the U.S. District Court where Judge Edward Chen ruled that LinkedIn couldnt block companies from deploying bots to scrape data from a public website. Though this was not held up in court, it speaks to the breadth of the dangers and risks of web scraping.

8.2 Job Search Ramifications

Many ethical issues related to job searches revolve around the truthful representations of jobs. Employers may try to entice more applicants by displaying the role as more desirable than it is. According to the Society for Human Resource Management (SHRM), creating fake job descriptions is a common way to get more applicants in a pool even though the role doesnt exist and is not advertised as a pool. SHRM has a code of ethics for the overall human resource profession which addresses recruiting [3].

9 Conclusion and Future Work

9.1

References

1. Andrea DeMauro, Marco Greco, M.G.P.R.: Human resources for big data professions: A systematic classification of job roles and required skill sets. Information Processing & Management **54**(5) (9 2018)
2. Ankita Srivastava, Yogesh Tiwari, H.K.: Attrition and retention of employees in bpo sector. International Journal of Computer Technology and Applications **2**(6), 3056–3065 (2011)
3. Bates, S.: Do recruiters need a code of ethics (2019), https://www.shrm.org/resourcesandtools/hr-topics/talent-acquisition/pages/do-recruiters-need-code-of-ethics.aspx
4. Brijesh Kishore Goswami, S.J.: Attrition issues and retention challenges of employees. International Journal of Scientific & Engineering Research **3**(3) (4 2012)
5. Collins Marfo Agyeman, P.V.P.P.: Employee demographic characteristics and their effects on turnover and retention in msmes. International Journal of Recent Advances in Organizational Behaviour and Decision Sciences **1**(1) (2014)
6. Indeed: (2019), https://www.indeed.com/about
7. Kaplan, J.: Artificial Intelligence: What Everyone Needs to Know. What Everyone Needs To Know®, Oxford University Press (2016), https://books.google.com/books?id=wPvmDAAAQBAJ
8. Touring: Alan Turing: His Work and Impact. Computing Machinery and Intelligence®, Gale Virtual Reference Library (GVRL) (2013)
9. Zinovy Radovilsky, Vishwanath Hegde, A.A.U.U.: Skills requirements of business data analytics and data science jobs: A comparative analysis. Journal of Supply Chain and Operations Management **16**(1) (3 2018)