SMU Data Science Review

Volume 2 | Number 3 Article 4

2019

A Data Science Approach to Defining a Data Scientist

Andy Ho
Southern Methodist University, atho@mail.smu.edu

An Nguyen
Southern Methodist University, anguyen2@mail.smu.edu

Jodi L. Pafford
Southern Methodist University, jodi.pafford@gmail.com

Robert Slater
SMU, rslater@mail.smu.edu

Follow this and additional works at: https://scholar.smu.edu/datasciencereview

 Part of the Computer Sciences Commons, and the Other Sociology Commons

Recommended Citation

Ho, Andy; Nguyen, An; Pafford, Jodi L.; and Slater, Robert (2019) "A Data Science Approach to Defining a Data Scientist," *SMU Data Science Review*: Vol. 2: No. 3, Article 4.
Available at: https://scholar.smu.edu/datasciencereview/vol2/iss3/4

This Article is brought to you for free and open access by SMU Scholar. It has been accepted for inclusion in SMU Data Science Review by an authorized administrator of SMU Scholar. For more information, please visit http://digitalrepository.smu.edu.

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, anguyen2, jpafford, rslater}@smu.edu
https://definingadatascientist.blogspot.com/

Abstract. In this paper, we present a common definition and list of skills for a Data Scientist using online job postings. The overlap and ambiguity of various roles such as data scientist, data engineer, data analyst, software engineer, database administrator, and statistician motivate the problem. To arrive at a single Data Scientist definition, we collect over 8,000 job postings from Indeed.com for the six job titles. Each corpus contains text on job qualifications, skills, responsibilities, educational preferences, and requirements. Our data science methodology and analysis rendered the single definition of a data scientist: A data scientist codes, collaborates, and communicates – transforming data into insights using techniques in statistics, analytics, and machine learning. A secondary finding confirms the hypothesis that Data Scientist job postings' features overlap with the other five job titles – explaining the absence of a common definition. Our conclusion is the application of data science algorithms and techniques on the job postings shows the most similar roles to a Data Scientist, provides a single definition of a Data Scientist, and generates the top features of a Data Scientist.

Keywords: Natural Language Processing · web scraping · Neural Networks.

1 Introduction

On a given day, 500 million tweets are sent, 294 billion emails delivered [8] and 4 petabytes of data created on Facebook [9]. Finding skilled individuals to transform data into insights is a necessity. The newness of the field has led to the creation of the data scientist role – which is still an evolving title. With no common definition, organizations and individuals are forced to create their own definition and list of skills leading to ambiguity and incorrect resource fit.

There is no common definition or list of skill sets for a Data Scientist. This becomes evident when the 'What is' or 'Who is' a Data Scientist is in question or when a job search (on multiple job sites) returns job title postings different from the search term. According to the Oxford Dictionary, 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 [1]. This definition is incorrect and/or outdated since data scientist are required to do more today.

Today's data scientist must possess the abilities to collect, clean, extract, transform, and load the data. In addition, apply statistical, analytical, and machine learning techniques to draw insights from the data. And most importantly, a data scientist must be able to communicate the findings in both written and spoken form.

Our approach is scraping job postings from an online job site, pre-process the data, explore the data, then transform the data into high-dimensional vectors to cluster, classify, and analyze. The application of natural language processing (NLP), universal sentence encoder (USE), and machine learning (ML) techniques led to quantifiable visualizations, results, and features.

The results 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 (Table 1).

Table 1.

1. Data Scientists are most associated with Data Analyst and Statistician
2. Data Scientists at minimum have programming/coding expertise in Python or R
3. Data Scientists often collaborate with other data scientists in a team environment
4. Data Scientists must be able to communicate their approach, findings, and insights
5. Data Scientists have a background or knowledge of statistics, analytics, and machine learning

The remainder of this paper is organized as follows. In section 2, we provide a brief overview on the general principles and techniques used in NLP, ML, USE for this research. In section 3, information on the data source, collection process, and content. In section 4, a detailed guide on the complete data science process performed on data. In section 5, a related published paper that was helpful for our research. In section 6, we show the results from the analysis. In section 7, we provide a few ethical issues of our approach and findings. And lastly, concluding this research in section 8.

2 Machine Learning Explained

2.1 Selecting the Optimal Machine Learning Approach

Machine learning (ML) is when computers and programs learn from the data it is provided. Data scientists select the best techniques and algorithms to model

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, Naïve 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.

the data. There are many ML algorithms, all of which can be stratified into 3 classifications, described in Table 2.

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

2.2 The Application of Natural Language Processing

Natural Language Processing (NLP) is a subfield of machine learning that focuses on how to program computers to process and analyze large amounts of human/natural language data. NLP incorporates both speech and written language. For this research, only written text is used to solve the problem. Using NLP, we are able to web scrape job postings from multiple sources to collect our data set to perform lexical, syntactic, and semantic analysis. Throughout this paper, Python packages specializing in NLP are leveraged to obtain text and document similarity scores of the job postings for each job title searched. The python libraries of Beautiful Soup, Natural Language ToolKit (NLKT), and spaCy are used to gather and clean the text.

2.3 Feature Vector Creation using Universal Sentence Encoder

Google's TensorFlow HUB (TF Hub) is an open source library for advanced Machine Learning and for numerical computation. In addition, the library contains an arsenal of algorithms for deep learning, digit classification, image recognition, word/sentence embeddings, recurrent neural networks, and for this paper, natural language processing. The Universal Sentence Encoder (USE), which uses TensorFlow library, encodes text into feature vectors for the purpose of text classification, semantic similarity, and other natural language tasks. At the core, USE produces sentence embeddings for transfer learning [5] and is made publicly available on TF Hub. On a high level, a corpus of text is fed into the encoder

and a 512-dimensional vector is output for semantic retrieval purposes. The vectors are then used to form clusters and classification on prediction accuracy. An example is below in Fig. 1.

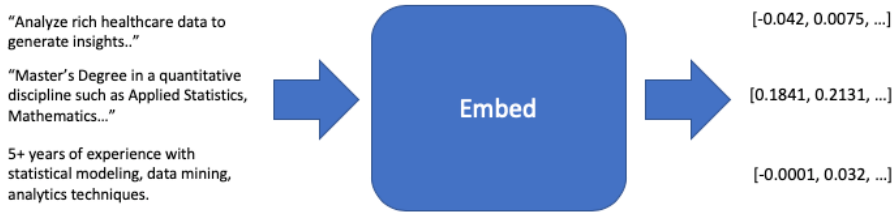

Fig. 1. Classification using a simple binary text classifier

3 Data Set Collection

Data is collected from Indeed.com. Indeed.com is the number one job site in the world with over 250 million unique visitors every month [2]. Indeed.com 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. Indeed.com does not provide job posting data for downloading; it must be web scraped. Web scraping, also referred to as web harvesting or web data extraction, is a process commonly used by Data Scientists to get data from the World Wide Web directly from the respective website [6]. Web Scraping for this project is done using the python library, Beautiful Soup (beautifulsoup4). Beautiful Soup allows us to scrape Indeed.com for information on several job titles and extract the information. "Beautiful Soup sits atop an HTML or XML parser, providing Pythonic idioms for iterating, searching, and modifying the parse tree [2]." The following data was gather directly from the search page(s) of Indeed.com: Individual job posing web link, Job Title, Location, and Salary (if available). Once 'Location' was pulled from the search page(s), location was parsed into City, State, Zip Code, and Country for further analysis. Each individual job website link was parsed to pull information about each job posting. Qualifications, skills, responsibilities, education, and requirements were pulled from the website by searching for those with bold HTML tags with bullet points following each bold word. The full text of the job description was also pulled into one column of the data set.

3.1 Data Set Contents

Data is scraped from Indeed.com for the following 6 job titles: Data Scientist, Data Analyst, Data Engineer, Database Administrator, Software Engineer, and

Statistician. Each job title is search in the following 16 cities: Atlanta, GA, Austin, TX, Bellevue, WA, Boston, MA, Chicago, IL, Cupertino, CA, Dallas, TX, Denver, CO, Houston, TX, Los Angeles, CA, Mountain View, CA, New York, NY, Pittsburgh, PA, Seattle, WA, San Francisco, CA, and Washington, DC. We attempted to pull 300 postings from each city and job title. By searching this criterion, we have 96 different possible combinations and a potential of over 28,000 job posts. However, many cities did not pull 300 job postings and some of our cities were so close to each other geographically that duplicates are found. After duplicates are eliminated, our total data set contains 8,738 unique job postings. A distribution of all the job postings pulled can be seen in Fig. 2, 3, and 4 below.

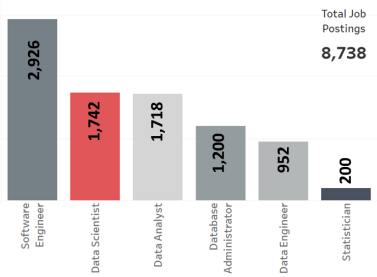

Fig. 2. Count of Job Postings

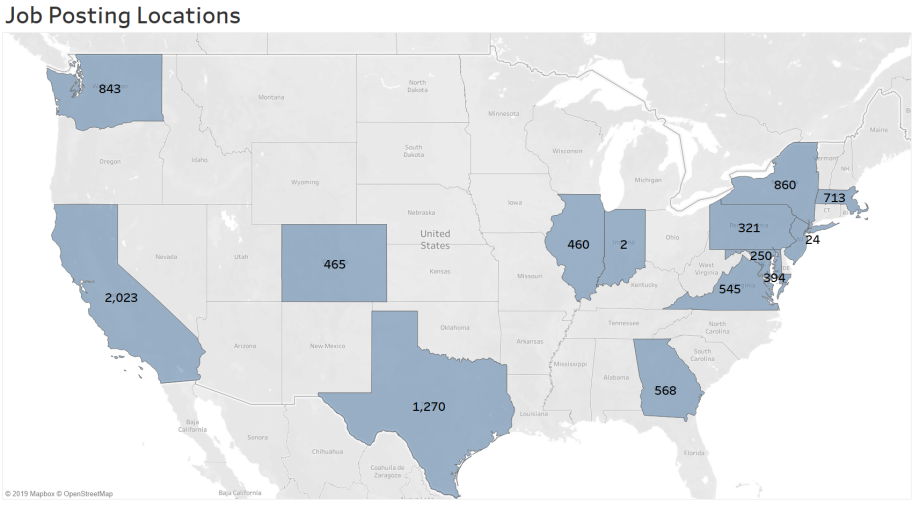

Fig. 3. Distribution of job postings across the United States

Indeed.com Job Posting Web-Scrape

Search Title	State CA	TX	MA	NY	IL	GA	WA	DC	CO	VA	MD	PA	NJ	IN	Grand Total
Data Scientist	517	170	188	207	123	72	185	69	54	96	29	30	2		1,742
Data Analyst	430	218	194	187	134	133	115	97	82	58	35	34	1		1,718
Data Engineer	98	143		187	110	54	136	32	60	94	17	15	6		952
Database Admin	182	168	74	46	82	86	82	110	42	186	92	38	10	2	1,200
Software Engineer	758	560	216	224		212	313	68	221	94	61	199			2,926
Statistician	38	11	41	9	11	11	12	18	6	17	16	5	5		200
Grand Total	2,023	1,270	713	860	460	568	843	394	465	545	250	321	24	2	8,738

Fig. 4. Count of Job Titles by State

4 Methodology from a Bird's Eye View

The high-level overview of the data science approach to solving the problem: Unstructured data from Indeed.com was web scraped, pre-processed to extract relevant text and transformed into feature vectors using USE. The next phase was a continual cycle of clustering and classification for both exploratory and insight purposes. Once an acceptable accuracy score was obtained, the clusters were reviewed to find closes neighbors to the data scientist job postings – closest neighbors being statistician and data analyst. The final step was to find the features of importance for each of the three roles by reverse engineering the results. The process is shown below in Fig. 5.

4.1 Exploratory Data Analysis

Indeed.com allows companies to post jobs with their own HTML code beyond the generic required information. This means that some companies included salary information, company logo, and/or company rating while others did not. This also means that most of the Job Description Summary sections are different base on how the company posted the job (bold with bullets, one single paragraph, and a myriad of other variation in between). Additionally, running the entire job description text through our analysis means that we include items like company information and non-discrimination clauses. After examining the data, we determine that enough of the job postings have bullet points within the text. This creates a uniform way to pull information for analysis that excludes information we do not want. After removing job postings who did not utilize the bold words, "Education", "Qualifications", "Requirements", "Responsibilities", and/or "Skills" and bullet point information below, our data set shrunk to a total 4,156 job postings. The final count of jobs used can be found in Fig. 6. It is interesting to note that all of our final counts ended up with a total of two different digits within each total (i.e. 1,311 uses the digits 1 and 3). This does not have any effect on the outcome, but we found it interesting. Fig. 7 shows the breakdown of those jobs throughout the United States. Table 8 shows the location breakdown by job title searched.

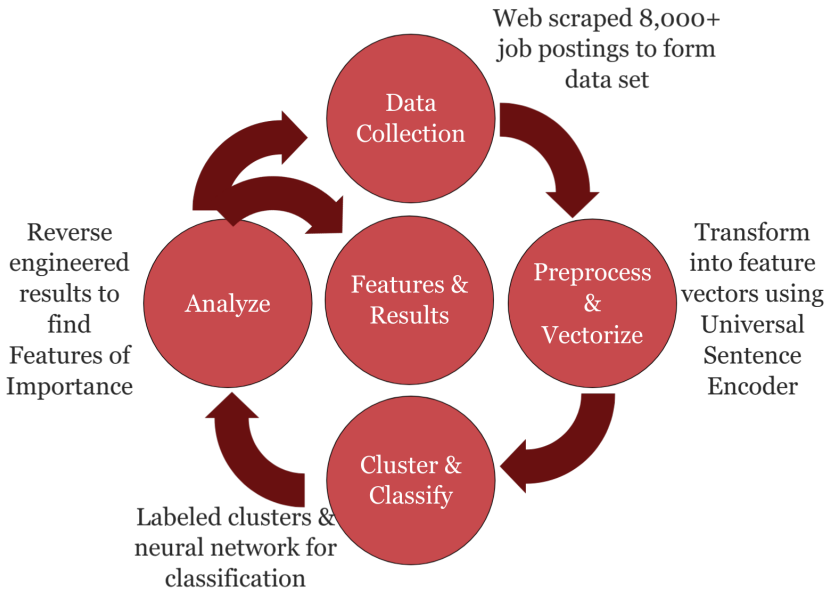

Fig. 5. Process followed for analysis

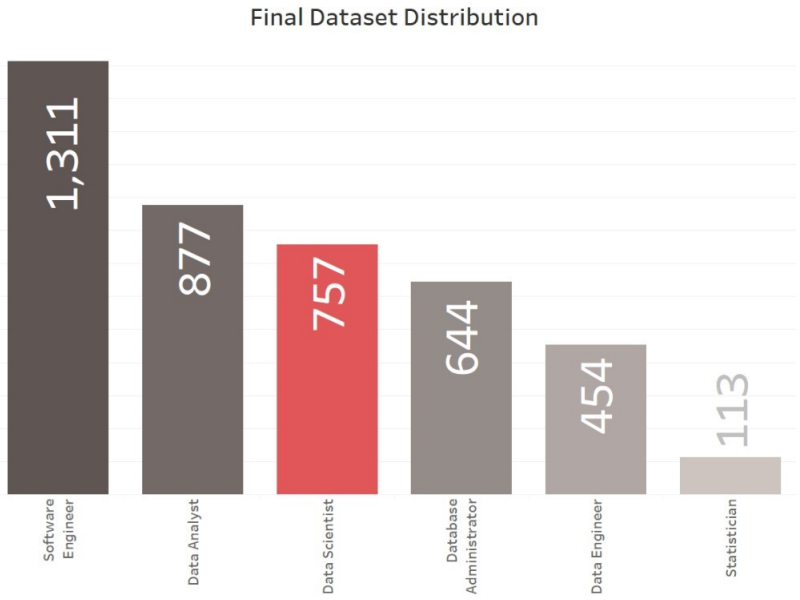

Fig. 6. Count of Jobs in the Final Data Set

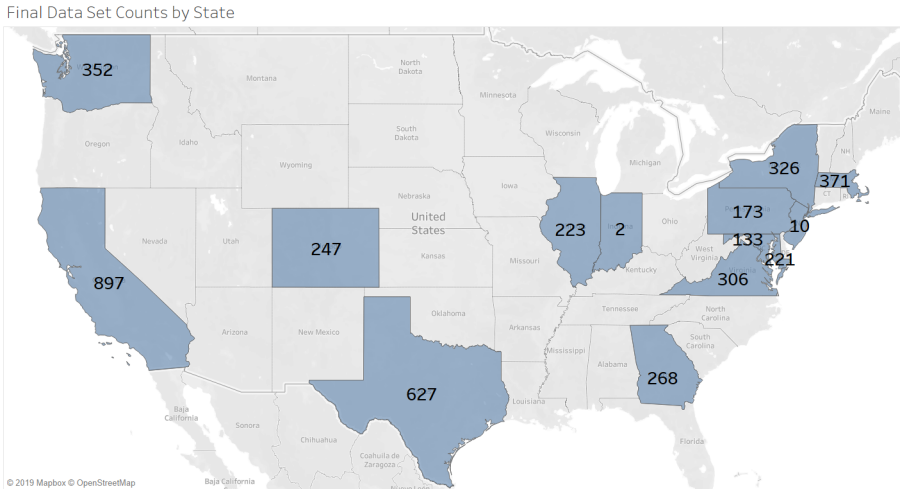

Fig. 7. Final Data Set Counts by State

Job Posting Used in Analysis

Search Title	State														
	CA	CO	DC	GA	IL	IN	MA	MD	NJ	NY	PA	TX	VA	WA	Grand Total
Data Scientist	205	28	43	42	51		83	8	1	74	14	92	54	62	**757**
Data Analyst	207	47	52	67	73		99	21		78	18	122	31	62	**877**
Data Engineer	54	31	22	31	50			7	3	78	7	79	44	48	**454**
Database Administrator	90	18	68	32	42	2	54	60	2	20	20	72	126	38	**644**
Software Engineer	326	121	25	93			108	26		70	110	255	43	134	**1,311**
Statistician	15	2	11	3	7		27	11	4	6	4	7	8	8	**113**
Grand Total	**897**	**247**	**221**	**268**	**223**	**2**	**371**	**133**	**10**	**326**	**173**	**627**	**306**	**352**	**4,156**

Fig. 8. Distribution of Job Titles Throughout the States

4.2 USE for Sentiment Retrieval – 512 Dimensional Feature Vector Transformation

Once the data was collected, it needed to be transformed into something that could be quantified and analyzed. Tensorflow-Hub's Universal Sentence Encoder (USE) transformed each of the unique job posting corpus' into numerical vectors. USE transforms our text by finding semantically similar sentences (job postings) and places it into a 512-dimensional feature vectors for clustering and classification. Tensorflow-hub's USE is pretrained with a deep averaging network (DAN) encoder and ready for our use 'out of the box'.

4.3 Cluster Analysis on Key Terms

The first step of analysis is exploratory to find a ground truth which is deciding whether there is a clear market definition of a data scientist. Using the scikit-learn library [7], principal component analysis was performed to reduce the 512 features vector to two dimensions. Each job description is plotted with the centroid of the clusters highlighted in red (Fig. 9). Judging from the initial clusters for each job title, the centroid for each job posting is located closer to the coordinates (0,0) but no decisive/clear insights is drawn.

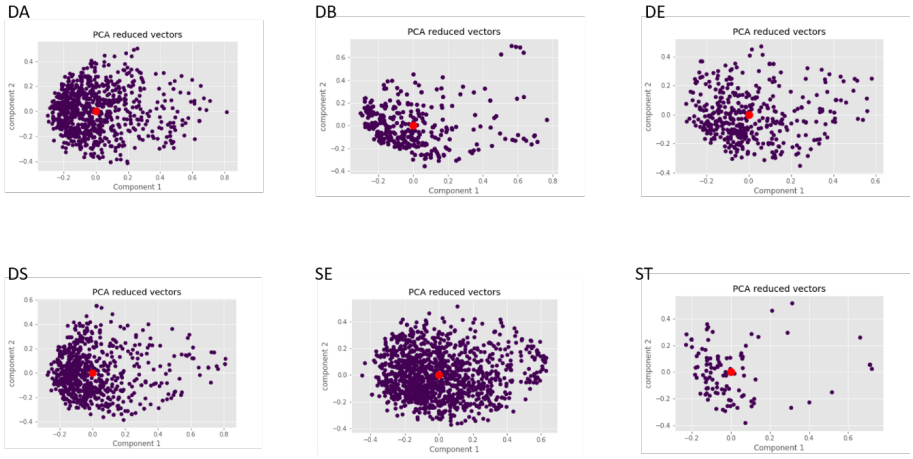

Fig. 9. PCA visualization of each job description as one cluster.

KMeans is used for this clustering. KMeans attempts to classify n number of samples into k number of clusters. The samples are assigned to the clusters with the nearest mean. The KMeans algorithm minimizes a criterion known as inertia, the sum of the squared distances of each sample from the centroid. The results were used to form the baseline model when comparing to the neural network clusters. The method of initialization is 'k-means++', the number of times the algorithm runs is set at 10 with 300 maximum number of iterations per run and relative tolerance set at 0.0001. The data is not modified and the "elkan" algorithm was used. A total of seven data sets and a range of 1 through 30 clusters separately informs the optimal number of clusters for the data sets. Each job title is clustered individually and all six titles concatenated into one data set is clustered. The number of clusters for each analysis was determine by locating the elbow from plotting the sum of the squared distances of each sample to their closest cluster center (Fig. 10 and 11).

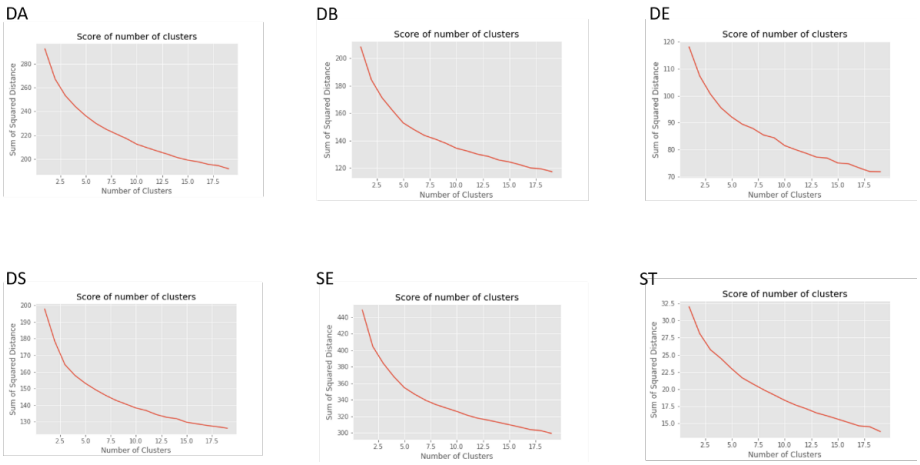

Fig. 10. Sum of Squared Distance for each job title with different number of clusters.

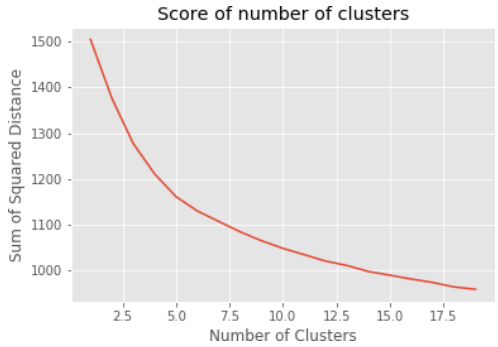

Fig. 11. Sum of Squared Distance for all the job titles combined.

From the sum of squared plots, 3 clusters for the individual job titles and 4 clusters when all job titles are combined into one data set is chosen. This is an incorrect conclusion as it is known that there should only be one cluster for each individual job titles and 6 clusters for the combined data sets (Fig. 12 and 13). From these clusters, the ground truth was established – there is market confusion and/or ambiguity for the role of data scientist.

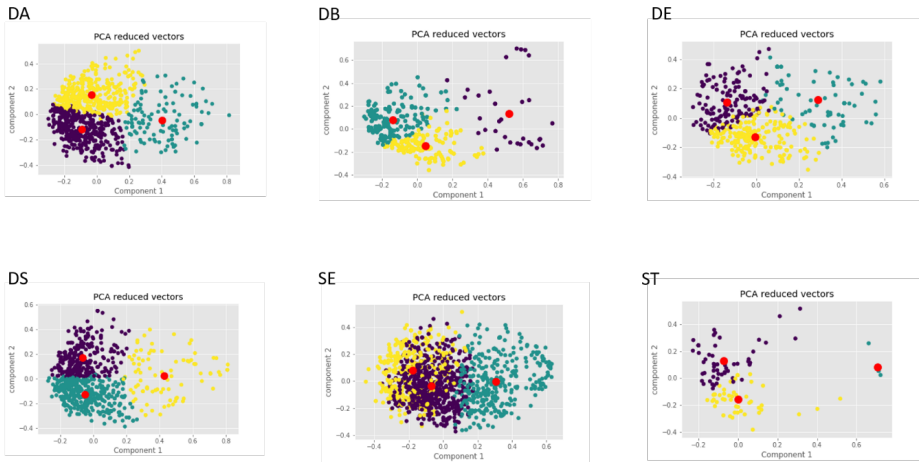

Fig. 12. PCA visualization of each job description as three clusters.

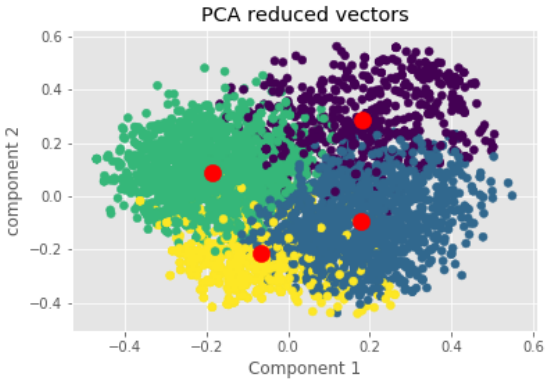

Fig. 13. PCA visualization of the combination of all six job titles with 4 clusters

Figure 14, below, is the KMeans clustering of the full data set when 2 clusters is specified. Because this is unsupervised clustering an arbitrary label of 0 and 1 is assigned by the algorithm to each vector. These labels can be compared to the original data sets. 0 is found to contain more vectors from the DS job description and 1 contains all the other job description, non-DS. After determining the label, the accuracy of the model is determined by the percentage of vectors correctly labeled. The accuracy of the binary KMeans clustering algorithm is 55.37% (Fig. 14) - not useful for conclusion.

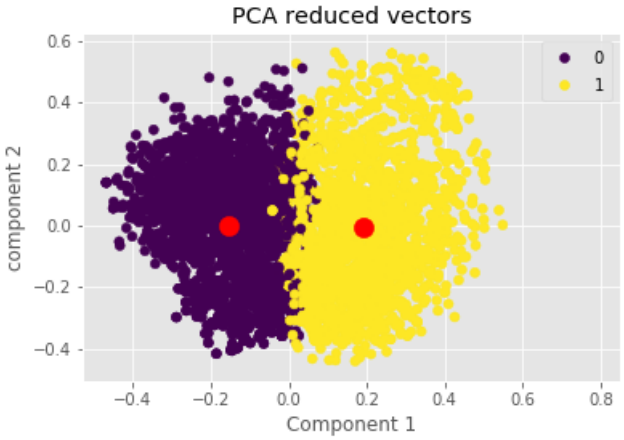

Fig. 14. PCA visualization of the combination of all six job titles with 2 clusters

The same procedure is used but with 6 clusters specified in Fig. 15 below. The Data Scientist job is mapped to group 1, dark blue. Data Analyst job is mapped to group 2, dark green. Statistician is mapped to group 4, light green. Database Administrator is mapped to group 5, yellow. Software Engineer is mapped to both group 0 and 3, purple and green. Data Engineer was not detected by the KMeans algorithm. Data Scientist and Statistician has a large overlap. The Data Analyst group is separated from the Data Scientist/Statistician group. Database Administrator is also well separated from Data Scientist/Statistician group. Software Engineer is well separated from the Data Scientist/Statistician group and the Data Analyst group but overlaps with the Database Administrator group. The algorithm could not detect the Data Engineer group. A possible explanation is that Data Engineer is too similar the Software Engineer group (refer back to figure 9). The accuracy of the model is 53.51% - not useful for conclusion.

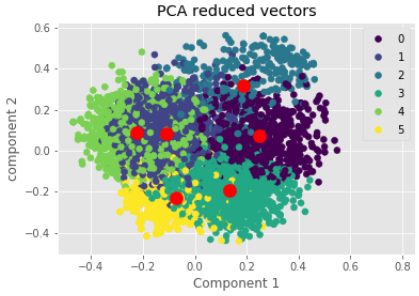

Fig. 15. PCA visualization of the combination of all six job titles with 6 clusters

To increase the model's predictive accuracy, a two layer dense neural network (NN) was used. Figure 16 is the summary of the NNs. The activation function of the input layer for both NNs is 'relu' and 256 dimensions output. The binary NN output layer's activation function is 'sigmoid' and 1-dimension output. The loss function used is 'binary_crossentropy.' To get 6 clusters the output layer's activation function is 'softmax' and 6-dimensions output. The loss function is 'sparse_categorical_crossentropy.' For both NN the 'Adam' optimizer with a learning rate of 0.001 was used. Both NNs have 15 epochs with a batch size of 75. The binary NN's, looking at DS jobs and non-DS jobs, accuracy is 92.25% (Fig. 17). While the NN of all 6 job descriptions accuracy is 89.27% (Fig. 18).

Two clusters

Model: "sequential_1"

Layer (type)	Output Shape	Param #
dense_1 (Dense)	(None, 256)	131328
dense_2 (Dense)	(None, 1)	257

Total params: 131,585
Trainable params: 131,585
Non-trainable params: 0

Six clusters

Model: "sequential_5"

Layer (type)	Output Shape	Param #
dense_9 (Dense)	(None, 512)	262656
dense_10 (Dense)	(None, 6)	3078

Total params: 265,734
Trainable params: 265,734
Non-trainable params: 0

Fig. 16. Summary of Neural Network Outcomes

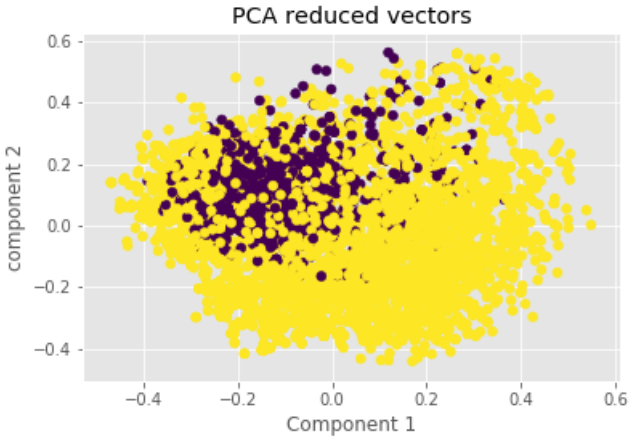

Fig. 17. PCA visualization of the combination of all six job titles with 2 classifications

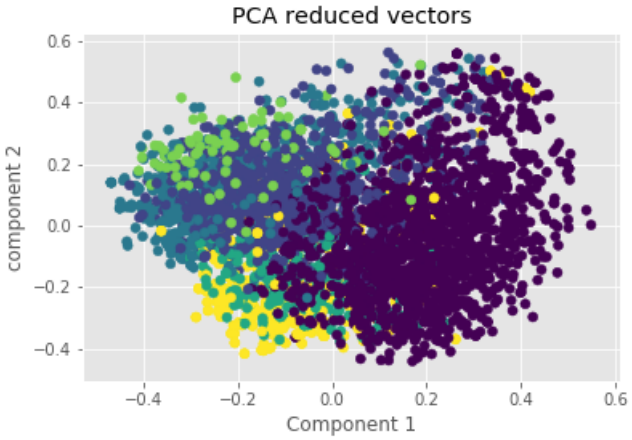

Fig. 18. PCA visualization of the combination of all six job titles with 6 classifications

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 data set 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 [3] (see Fig. 19).

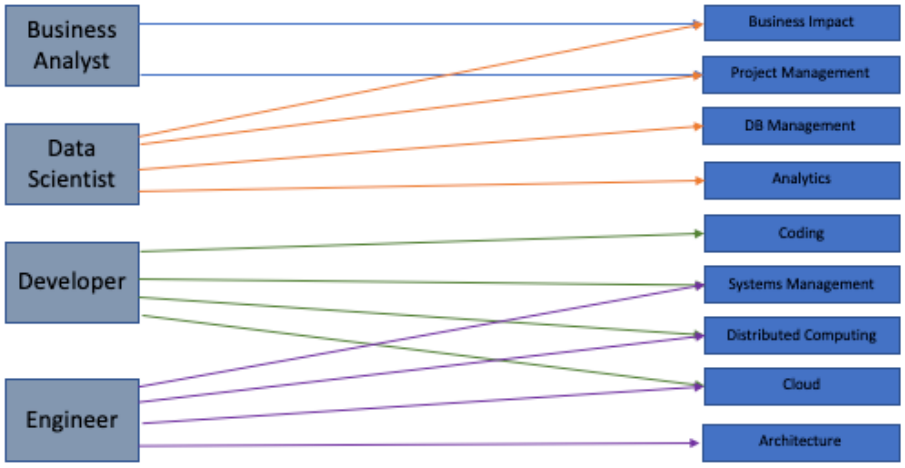

Fig. 19. Job skill sets are mapped to job roles by a measure of the extent at which each skill set is represented within each job post description.

A second group out of California State University focused entirely on the difference between Business data analytics, Database Administrator (DBA) and Data Science (DS). Radovilski et al. manually collected job descriptions of Database Administrator and Data Science 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. 20)[10].

Database Administrator

#	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%

Data Scientist

#	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. 20. Terms used in job descriptions shown as proportion of the whole.

6 Defining a Data Scientist

The final neural network model shows that Data Scientist is most closely related to Data Analyst and Statistician. With this insight, we look back at our data set job postings to reverse engineer the features of importance. All duplicate words within each posting as well as some additional stop words that would not inform our results (preferred, quality, field, work, strong, working, etc, large, experience, ability...) are removed. The top (highest word count) 25 words for all 6 job titles are below in Fig 21.

Rank	Data Analyst	Data Engineer	Data Scientist	Database Administrator	Software Engineer	Statistician
1	data	data	data	database	software	statistical
2	degree	python	science	sql	development	data
3	analysis	sql	python	data	computer	analysis
4	business	science	machine	performance	design	design
5	management	development	statistics	databases	science	statistics
6	sql	degree	computer	systems	degree	methods
7	bachelors	design	analysis	degree	engineering	communication
8	communication	engineering	degree	management	systems	research
9	analytical	systems	statistical	support	team	development
10	tools	tools	business	server	code	team
11	team	computer	models	security	java	sas
12	support	business	analytics	design	technical	projects
13	reports	technologies	sql	bachelors	programming	management
14	written	solutions	engineering	development	bachelors	programming
15	excel	technical	tools	administration	data	andor
16	reporting	spark	techniques	team	develop	written
17	information	pipelines	mathematics	oracle	communication	degree
18	analytics	big	modeling	tuning	applications	clinical
19	develop	management	team	communication	web	results
20	systems	aws	develop	procedures	technologies	project
21	science	software	development	technical	understanding	technical
22	analyze	etl	communication	recovery	tools	phd
23	development	develop	design	computer	test	relevant
24	projects	cloud	algorithms	maintain	python	analyses
25	computer	bachelors	quantitative	issues	testing	review

Fig. 21. Top 25 words found in the job postings in the final data set.

6.1 Focusing on Data Scientist, Data Analyst, and Statistician

The most common words in the job postings for Data Scientist, Data Analyst, and Statistician are used to create the diagram below in Fig 22. Data Scientist and Data Analyst have more words in common and therefore, Statistician is a smaller circle. Data Scientist, Data Analyst, and Statistician share common

words such as Research, Analysis, Statistics, Communication, and Team. Data Scientist and Data Analyst share words such as Python, Tools, Business, Support, sql. Data Scientist and Statistician do not share any top words that are not also shared with Data Analyst. A Data Scientist differs from both a Data Analyst and Statistician by including skills and tools such as Programming, Modeling, Computer Science, Engineering, and Machine Learning.

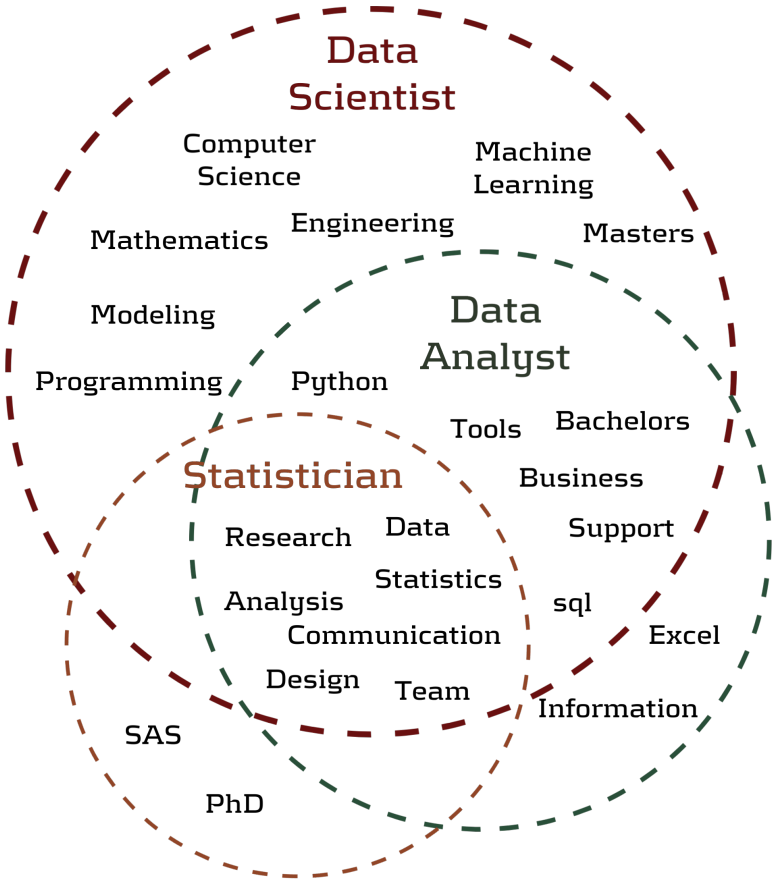

Fig. 22. Top words found in the job postings for Data Scientist, Data Analyst, and Statistician.

6.2 What is a Data Scientist?

Based on the information gathered in job postings, when an employer is looking for a Data Scientist, they look for the following:

A Data Scientist codes, communicates, and collaborates – transforming data into insights using statistical, analytical, and machine learning techniques.

7 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.

7.1 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 a website that is not allowed to be scrapped. Additionally, a website's 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 [5], 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 couldn't 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.

7.2 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 does not exist and is not advertised as a pool. SHRM has a code of ethics for the overall human resource profession which addresses recruiting [4]. Inversely, applicants could utilize algorithm or apply data science to falsify or embellish their resumes. The goal would be to trick resume tracking software into classifying the applicant as qualified, competent, and/or a fit for the open position. Not only is this misrepresentation but it prevents other qualified applicants from being interviewed.

7.3 Model Used for Profiling Candidates - aka AI bias

One shared concerned regarding Artificial Intelligence (AI) is its ability to be fair and neutral. Although there are many advantages of AI (i.e. speed and capacity to process), the software/programs are still written by humans who are bias and judgmental. The application of NLP and ML to identify features or patterns in job postings could also be used by employers to profile applicants. The risk is knowingly removing applicants based on gender, race, creed, religion, and/or sexual orientation

8 Conclusion and Future Work

The title of 'Data Scientist' is still a new concept that has and will continue to evolve as the role molds to the needs of artificial intelligence and business requirements. Our current definition is the result of the current market's view via job postings. Additional future work is required to capture a higher accuracy. In data science, the insights are only as reliable and accurate as the data itself.

For future work, additional data such as surveys, in person interviews with working data scientists, and web scraping additional websites could provide a more well-rounded (and accurate) prediction on the definition and skill set. One of the flaws of unstructured data is the possibility of selection bias during the pre-processing phase. A possible solution would be to continually collect job postings to better train the model over time.

In conclusion, the common definition and skill set of a data scientist is simplified within the 3Cs and SAM. A data scientist Codes, Collaborates, and Communicates – transforming data into insights using techniques in Statistics, Analytics, and Machine learning.

References

1. Oxford Dictionary of English. Oxford University Press, 2015 edn. (2010), https://www.lexico.com/en/definition/data_scientist
2. About indeed (9 2018), https://www.indeed.com/about
3. 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)
4. 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
5. Cer, D., Yang, Y., Kong, S., Hua, N., Limtiaco, N., John, R.S., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C., Sung, Y., Strope, B., Kurzweil, R.: Universal sentence encoder. CoRR **abs/1803.11175** (2018), http://arxiv.org/abs/1803.11175
6. E. Vargiu, M.U.: Exploiting web scraping in a collaborative filteringbased approach to web advertising. Artificial Intelligene Research **2**(1) (2013), http://dx.doi.org/10.5430/air.v2n1p44

7. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research **12**, 2825–2830 (2011)
8. Pioryshkina, E.: Big data in ecommerce: 5 remarkable use cases (10 2019), https://www.iflexion.com/blog/big-data-ecommerce
9. Smith, K.: 53 incredible facebook statistics and facts (6 2019), https://www.brandwatch.com/blog/facebook-statistics/
10. 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)