RIGHT HIRES!

Talent Forecasting

The Right People, The Right Skills, The Right Jobs

Team 52 Final Report
Julia Barnhart, Brennen Chadburn, Jonathan McKim, David Pilkington, Mike Ryder

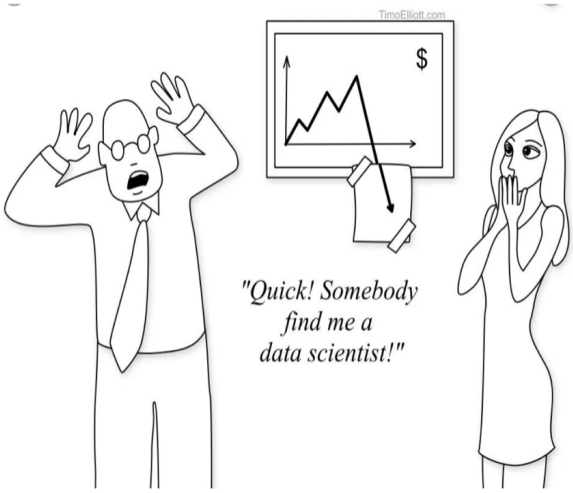
Right Hires!

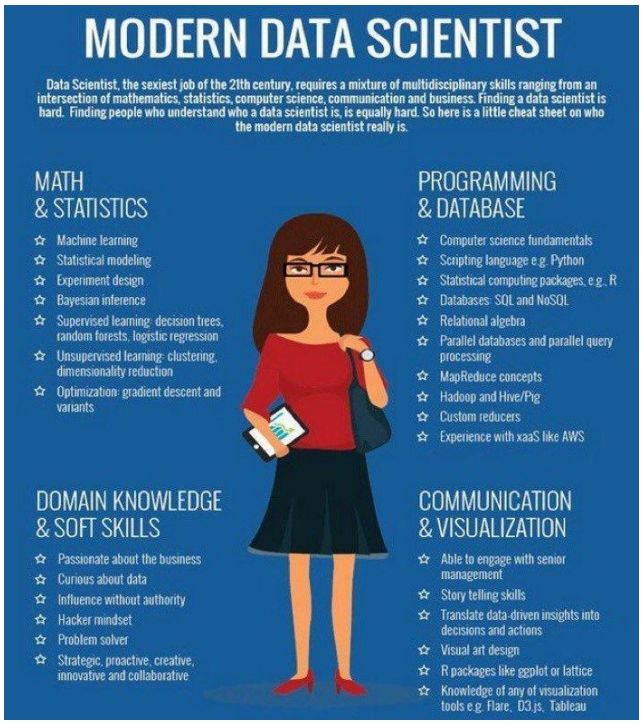
Table of Contents

The Opportunity

Market Overview

When Data Scientist was coined “the sexiest job in the 21st century” (Davenport et al., 2012), the race to find, hire, and train data scientists began. Seven years later (2019) even with record numbers of analytics talent graduating from universities (Krensky et al., 2017), the relative immaturity of the field is causing organizations to struggle to close the talent gap.

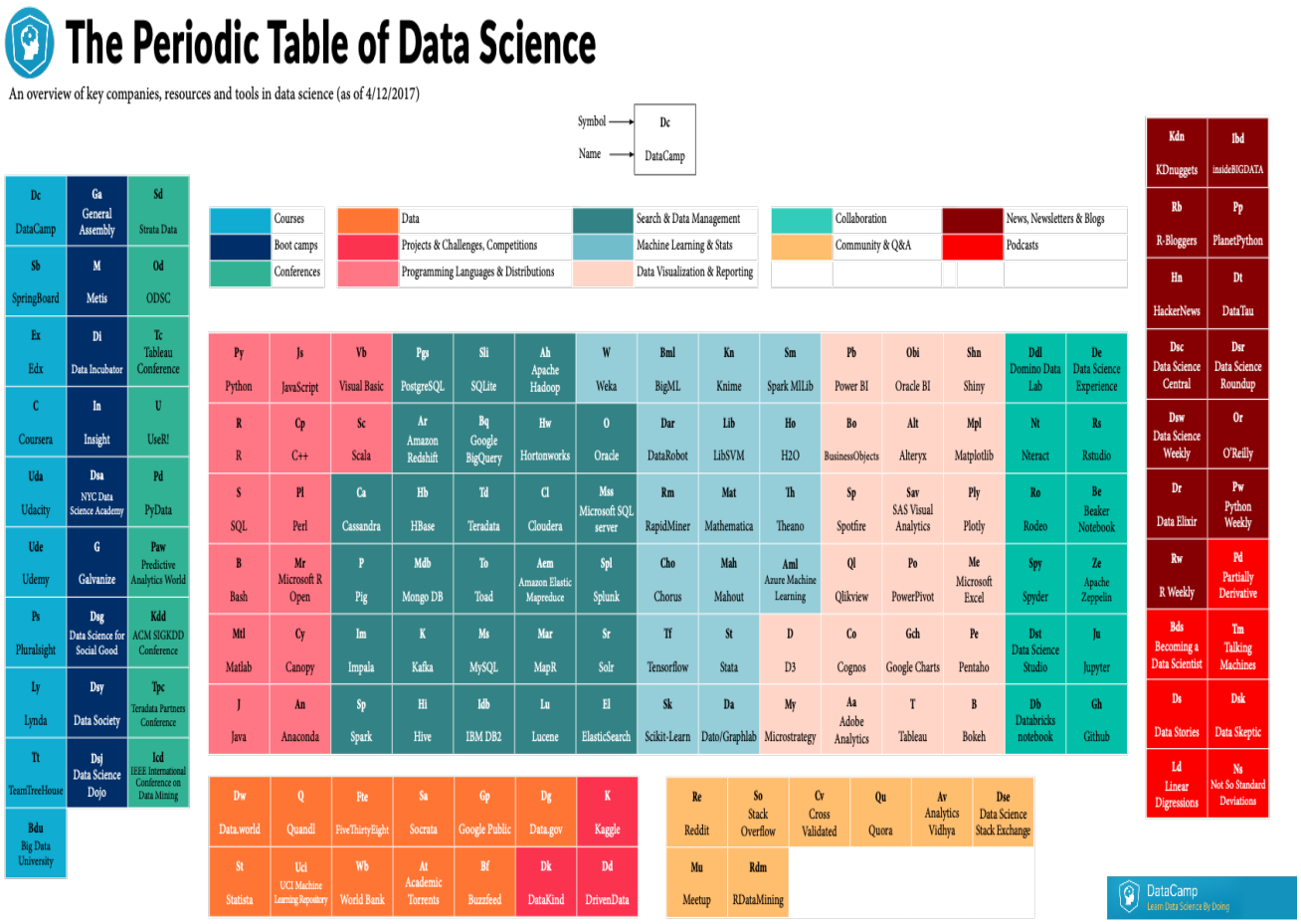




https://thedatascientist.com/data-science-considered-own-discipline/

Almost 30% of senior leaders cite finding talent as their most significant managerial challenge (Keller et al., 2017). Hiring companies don’t have a clear understanding of the breadth and depth of skills needed, both technical and nontechnical, to launch, grow, and sustain a successful data science team. There is also a mismatch in data science skills to organizational-specific program or project tasks; this includes expanding the data science team’s capabilities to include other key data, hardware, and software support roles (Linden et al., 2018). In general, organizations have difficulty in

Data Science talent management and correctly upskilling the current organizational workforce to narrow the talent gap internally. The figure below shows just how complex this specialization can be (and technology and methods are continually being released).



https://s3.amazonaws.com/assets.datacamp.com/blog_assets/Data-Science-Periodic-Table.pdf

This includes understanding the secondary education landscape for certificate, undergraduate, and graduate programs in Data Science and related fields. Additionally, recruiters are having problems sourcing suitable candidates to provide to these organizations. Technical leaders have further difficulties deciding whether to build, buy, or outsource certain data science and machine-learning capabilities (Krensky et al., 2018).

Even when candidates can be found, 83% of companies report a skills gaps in their data science organizations. From that group, nearly 75% attribute those gaps as limiting factors to deliver and support future growth (ATD, 2019) and adoption of the newest methods (like artificial intelligence).

In Search of the Data Science Unicorn

Through a synthesis of the above industry research and deep personal team experience as hiring analytics managers, data science students, data analysts, and data scientists, the business landscape in Figure 1 was defined. This landscape was used to articulate RightHires goals and offerings to help alleviate problems in the current market for Data Science talent.

Figure 1: Business Landscape

The Competition

As part of the process to define the RightHires product, the team reviewed the market for possible competitors and found two complementary offerings.

The first competitor is an offering from People Insights built for Walt Disney Company (and could be retrofitted for others). The People Insights product is generally inward focused and provides a company with insights on retention problems (exiting employees). For example, what are the characteristics of those that might leave and the probability they will leave. This information can help Disney change pay, work environments, flexibility to lower probabilities (where possible), save money, and increase employee satisfaction. Jobs covered are well defined and stable.

In contrast, the RightHires offering has a market (external) focus in a quickly evolving industry (Data Science) where jobs and their requirements vary as new technologies become available. The RightHires offering serves multiple internal and external facing customers including HR, VCs, executives, new hire candidates, and students. And the product is used early in the hiring process to define a job for a successful outcome. Overall, the RightHires product is targeted to make the hire, find a job, find education process more

efficient (which should help with retention issues when the right people are hired for the right job). The offering takes a country wide view to help employers and candidates no matter where they are based in the U.S. (other countries targeted for future). Last, the RightHires offering will be subscription based, so constantly adding new value for multiple audiences will be important to keep subscriber revenue growing.

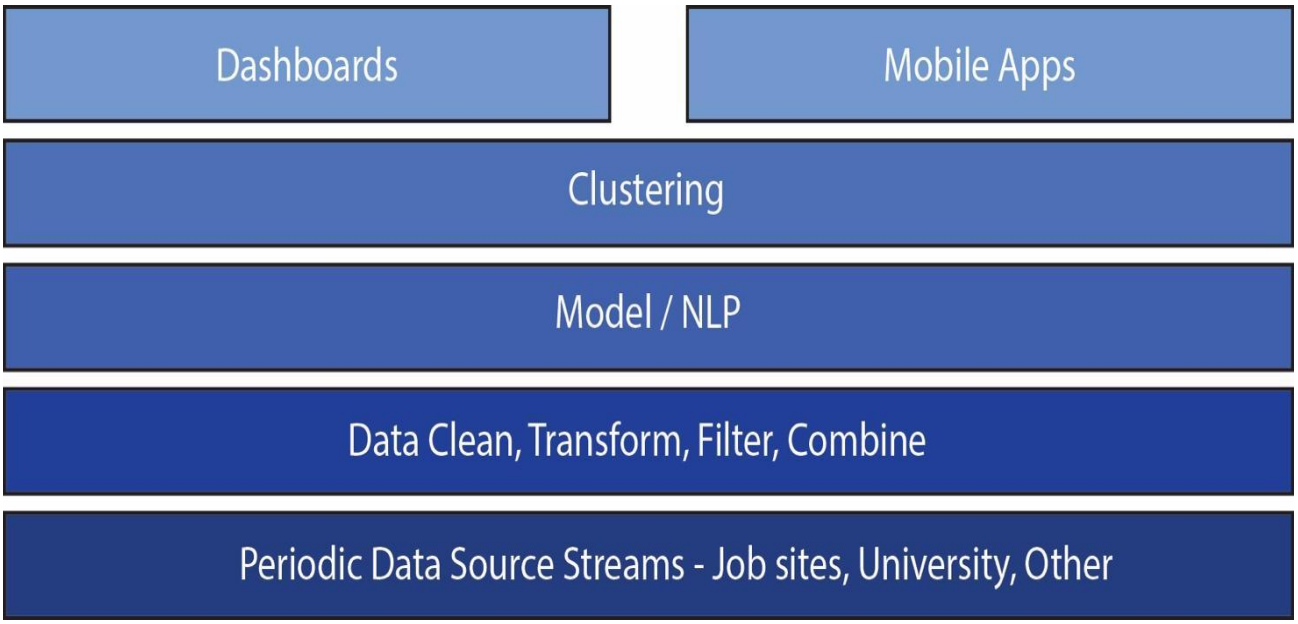
The second competitor Aspiring Minds (www.aspiringminds.com/) is an AI driven tool to help screen candidates once a job is posted (driven by RightHires) and before a candidate becomes an employee (and serviced by People Insights). Aspiring Minds provides on-line job simulations and coding evaluations as well as video assessments. AI-powered interviewing technology assesses candidate domain knowledge, language and personality traits including motivation and job fit without human bias. The AI also scores facial expressions, speech and content to offer HR and hiring managers scoring to aid in evaluation of candidates.

While all three products don't overlap and are not direct competitors, each also doesn't solve the full life cycle of all our collective users. With that in mind, a partnership should be considered to offer customers a full suite of functionality as shown in the figure on the last page. Workflow would move in the order shown in that figure, that is customers begin with RightHires, moves to Aspiring Minds, and once someone is hired, they leverage People Insights. Part of the partnership should include the trading of data (interfaces) or to share a collective data store so our common customers won't need to enter data twice. This will also reduce data errors and make the whole a larger value than any single product in the trio. As a fall back, interfacing should be done in a component way in case any company wants to leave the trio, or they cease to do business.

Proposed Solution

RightHires will periodically scan U.S. job market postings and use Natural Language Processing (NLP) and clustering to determine the most prevalent data science related job titles, demand by region, groups of related jobs, equivalent titles, the most important skills per job, and universities that offer master level data science programs. Results will be available for easy access through dynamic dashboards and mobile applications to help organizations attract relevant job applicants with the required skill set and keep up with a rapidly evolving Data Science job market. Figure 2 provides a “layers view” of our architecture.

Figure 2: High Level Solution Layers



Having the capability of comparing job titles, descriptions, and requirements to current market norms will drastically reduce the time and costs associated with building a Data Science team and reduce the problems highlighted in Figure 1. The remainder of this paper describes RightHires methodology as well as projects findings, conclusions, and plans for future releases.

Right Hires!

Goals and Deliverables

Project goals remain constant since project inception. Figure 3 describes major goals (left) and the strategy used to fulfill them (right). All goals have been completed.

Figure 3: Goals, Deliverables

Build base models (for Data Science) that can be expanded for other industries	• Scrape and purchase data from leading sources including Indeed, Twitter, and Master of Science in Data Science offerings sites. • Test various NLP and Clustering methods to find best one for targeted jobs.
Identify top Data Science job titles, alternate titles, most prevalent skills for those jobs.	• Leverage iterative cycles of NLP and clustering to determine the optimal mix of skills to jobs and their alternatives. • Make those findings available through 2 Uis below
Determine where the jobs and universities are that support Data Science.	• Combine model and clustering results from job sources with that from university programs to provide views for both audiences of the planned user interfaces.
Create interactive dashboards targeted for corporate needs.	• Deliver multiple and self-service dashboards that corporate, head hunting, and VC firms would find instrumental to help determine the best way and location to find data science resources.
Create a mobile App targeted for students and candidates that can also be leveraged by corporate users.	• Deliver one mobile application with multiple sections to help those looking for jobs or schools to determine the skills they need, where jobs and schools are, and what these jobs are called in locations of interest.
Define opportunities to expand the platform.	• To fit project in allocated time, team needed to scope features. Future expansion ideas will be included in the final report to make the executives aware of additional value opportunities.

Progress to Date

At Week 9, the RightHires team continues to track on schedule following the original 10-week project plan (Figure 4). The team continues to leverage an Agile methodology with members leading focus areas, collaborating during 3 weekly working sessions over Google Hangouts and with members helping across disciplines as needed to stay on schedule.

Figure 4: Project Status

Project deliverables and work are safely stored on Google Drive and available for external audit. All major deliverables required to complete data sourcing, data prep, filtering, and transforming data to make our data “gold copy” (all modeling and user interfaces leverage) are done.

Right Hires!

All NLP modeling and clustering is complete including testing multiple alternative methods in the search for additional improvement/value over the week 6 versions. The work done has identified 3 top clusters of job titles and the top 200 job description words found in each group (cluster) of jobs.

Throughout this iterative (NLP/clustering) process titles not of interest were filtered, stop words (both generic English and job description specific) were removed (as they added no value), and alternative job titles following an ontology developed by the team were leveraged to replace fewer common titles to those more prevalent.

All dashboards and the mobile application are now complete with a sampling contained in this report and a full walk through of all features planned for the CEO demo. The technology framework outlined in Week 6 remains constant.

The team has also specified a Google cloud infrastructure architecture that helped handle data and processing needs for the nearly 40 million job description words across the 3 top clusters. This infrastructure will become more important (and scalable up) in the future as RightHires expands into other industries and job families.

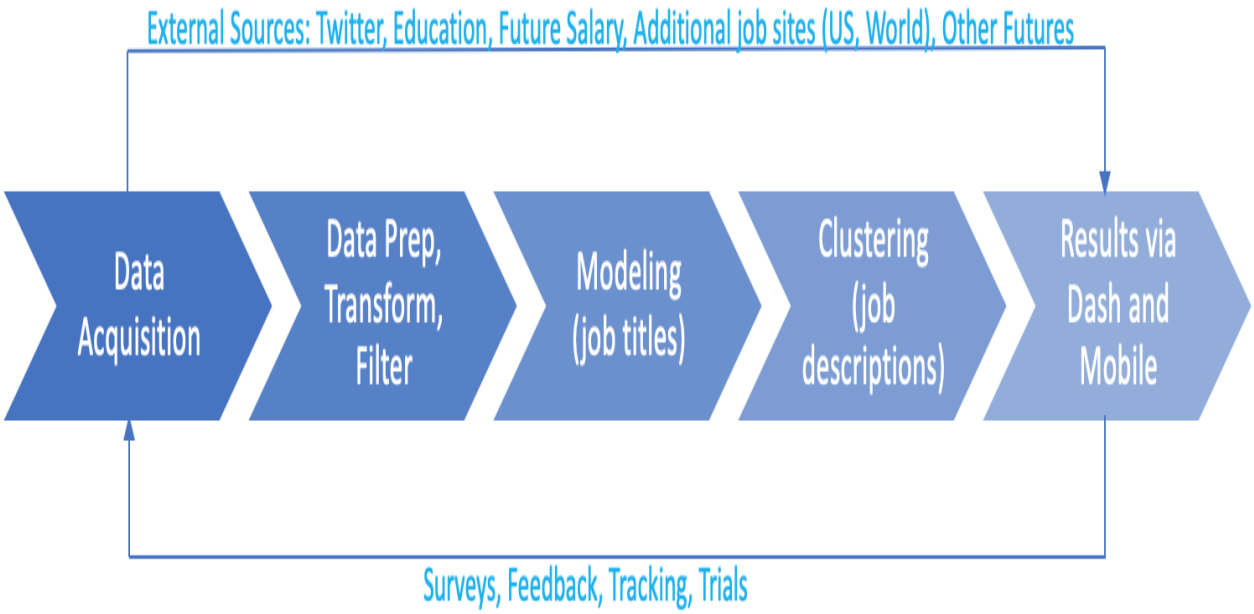
The last deliverable now in progress is to create and present the presentation our final findings to the CEO. The team is planning to make a video of the presentation so it can be viewed on demand in the future and to support executive and funding reviews of this project and ideas contained at the end of this report for new expansions.

Approach

Figure 5 illustrates major steps for the project and information flows between them. The majority of these flows are internal to RightHires with results delivered to customers through dashboards and a mobile application. Findings and progress to date for all steps are provided in the “Findings” section.

Figure 5 also shows the flow of information from left to right (as it’s more refined and approaches the user interfaces). Several external paths of data (or feedback data) are also shown that support the two-way conversational nature of some of the mobile app functions (survey, trial features, real time Twitter tag views, ability for users to select what they view in real time).

Figure 5: Major Components, Steps, and Data Flows



Technology Leveraged

The team has leveraged the technologies in Figure 6 to support EDA, NLP, clustering, dashboards and mobile application. The team has also tested, and architected cloud-based infrastructure that will be capable to handle large data sizes and model loads as RightHires evolves in the future.

Figure 6: Technology and Tools Leveraged

Right Hires!

Data Sources

The majority of data for this project was sourced from the Indeed job site (indeed.com) and comprised of 3.125 million individual job postings for 2018. The data comprises all jobs that were listed on the Indeed from January 1 to December 31, 2018. This data set was purchased directly by the project team through Data Stock (https://datastock.shop/), which offers web crawled datasets from a variety of industries on a one time or subscription basis.

Buying data allowed the team to streamline the ingestion process and reduce task completion time by not having to scrape data from the web directly. It also provided an overview of one complete calendar year, including any accompanying seasonal trends and business cycles affecting job listings. The raw data was provided by the vendor XML in format and then processed by the team and stored in JSON format for modeling.

While listings posted on Indeed do not represent the entire population of job postings available, it is a main source of jobs and a good indicator of job trends. Other competitors in the space include Glassdoor and LinkedIn (and others) which could become additional data sources for RightHires in the future.

A second data set used was gathered by web-scraping a Master's in Data Science website (mastersindatascience.com).

This included a list of graduate programs in Data Science and similar fields across the continental U.S.. This data will be used to provide regional level supply and demand labor needs.

The last data set used was trending information related to Data Science top job titles identified by the teams NLP and clustering work. Trending data was gathered from Twitter and is targeted to be available through our mobile application.

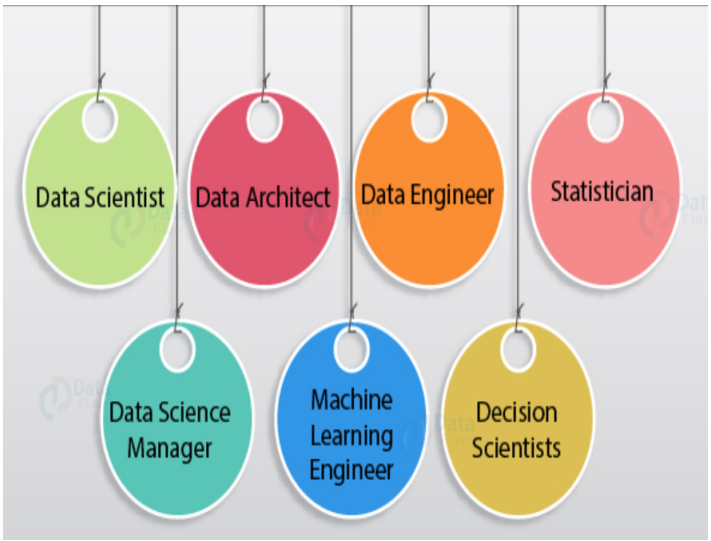
While all three data sources were gathered once for the project, RightHires plans to periodically gather and process this data to provide customers with the latest data to make the best decisions.

During the project, the team also began to set up a data store to initially work with the mobile application so RightHires could store responses and activity data from our users. Initially, survey information will be stored for the launch of RightHires. In time, more data about how our customers interact with RightHires will be stored and mined by Marketing.

Right Hires!

Modeling Strategy

Before modeling began, a taxonomy of job titles was created by scanning job titles for “Data” and “Data Science” postings from the Indeed data. These keywords were then used to conduct follow-on searches to identify related job titles. The process was then repeated several times with the new job titles creating a larger corpus of postings the team targeted for the project. In the future, RightHires can expand its offerings by expanding the list of jobs to include more industries and roles outside of Data Science.



Our modeling approach was to leverage two technologies, NLP and clustering. These technologies were used to create a taxonomy of job titles, clusters of similar job titles that can be serviced in a similar way, and identification of alternative job titles to those in the clusters (to simplify hiring and job posting for customers).

With these clusters (or groups) the team then extracted the most common terms (e.g. experience, requirements found in a job description) for each job title. The list of terms will be used by our RightHires customers to build a more complete job posting reflective of the collective industry knowledge of what experience roles in Data Science require.

To accomplish the task, the team used the Term Frequency – Inverse Document Frequency (TF-IDF) methodology to identify useful bigrams (a pair of consecutive letters, syllables, or words) and trigrams (a group of three consecutive letters, syllables, or words) that were used to identify a longer list of relevant job titles. The process was repeated many times resulting in a corpus of the most popular job titles in Data Science.

Figure 7 provides a 10,000 foot definition of what TF-IDF is since this was key to the work done on the project.

Figure 7: TF-IDF High Level Definition

TF-IDF is an approach that weighs the frequency of a term based on how frequently it appears in other the entire corpus (i.e. all documents). It is used to identify relevant words for a given document and has been used in search algorithms for years. For instance, if someone googles "where are the Himalayas?" TF-IDF is used to identify articles containing "Himalayas" as the word is relatively infrequent in the corpus.

What is TF-IDF?

10,000 foot view

TF-IDF is used in document classifications in the similar way. For our purposes, we use the TF-IDF weight for each word which provide inputs to whatever algorithm we use for clustering. When using K-means for instance, we simply compare the distance of documents based on their TF-IDF scores for each team

Next, clustering methods were employed to investigate relationships between job titles and look for clusters of jobs with potentially similar job requirements (either indicating a class of similar jobs, identifying job families, finding additional alternative jobs, or uncovering a base of job requirements many data science job descriptions should contain).

Multiple clustering methods were used along with a range of cluster sizes to obtain the optimal number of distinct (and fairly equal size) clusters to service the Data Science jobs targeted. The "Topic Modeling" section provides additional detail for how this was done, alternatives attempted, and a resulting top 3 cluster model used for the project.

Right Hires!

User Interfaces Offered

The primary delivery of value to our customers is through two user interfaces (dashboards and mobile application). While these interfaces draw from common model results, both also draw from other data sources to offer more value to the audiences as outlined in Table 1.

Table 1: RightHires User Interfaces

Dashboards		Mobile Application	
	Targeted to HR, VC, headhunters, and executives to help determine locations of Data Science centers		Targeted to job seekers, current and future students.
Multiple functionality available within each interface. Leverages model and other sourced data. Mobile application will gather/store survey data to help RightHires determine demand for new futures.			
Used to answer questions like: What are the top job titles, their alternatives, and the skills that should be in each job? Where are the centers of jobs like mine, the schools where students are likely to come from, and the growth of key Data Science jobs over time across the U.S.?		Used to answer questions like: What are key skills for jobs I'm considering? Where are the jobs I'm interested in? What locations have the most Data Science (DS) jobs? What are the family of jobs in Data Science? What are hot DS trends I should be skilled in?	

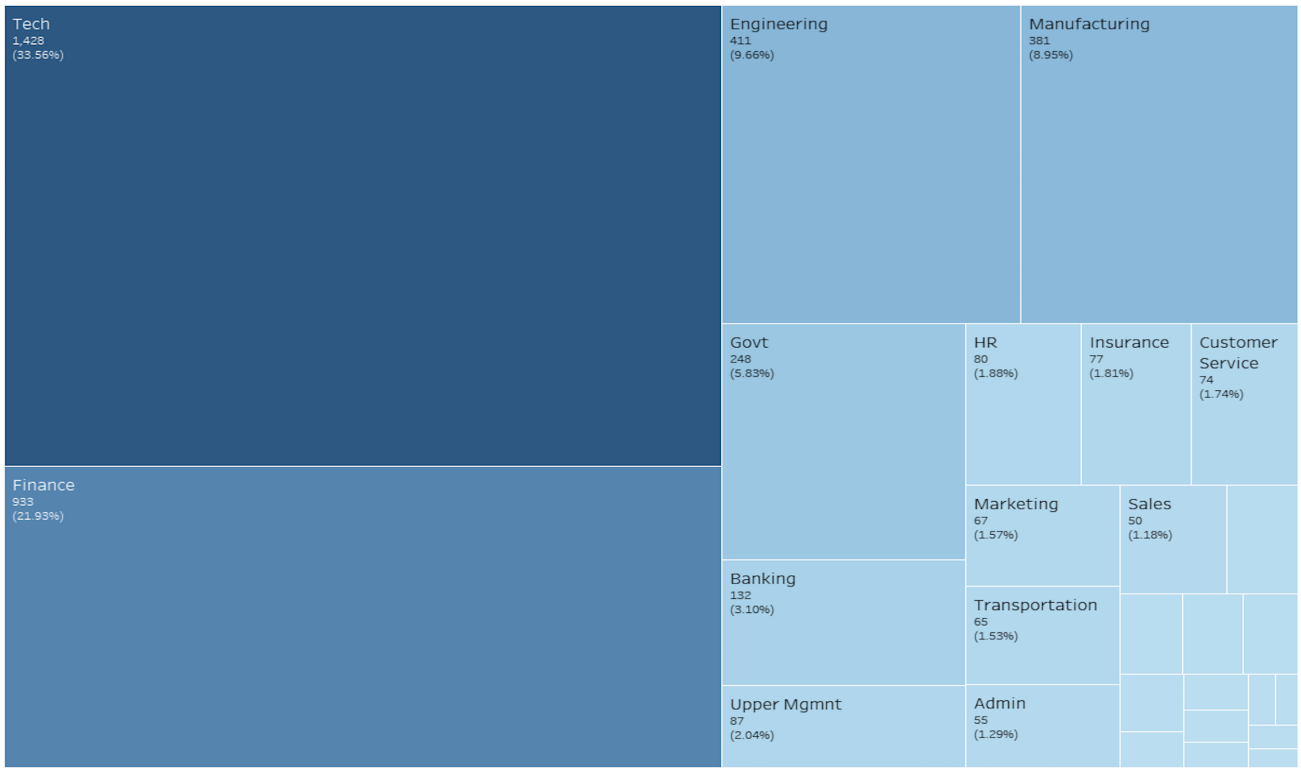
Findings

Exploratory Data Analysis (EDA)

Of the 3.125 million job postings sourced from Indeed, RightHires only needed a tiny fraction for this project (less than 1% of those related to Data Science kinds of jobs). This represented ~3,000 job postings, spread over 1,100 unique companies and 25 industries.

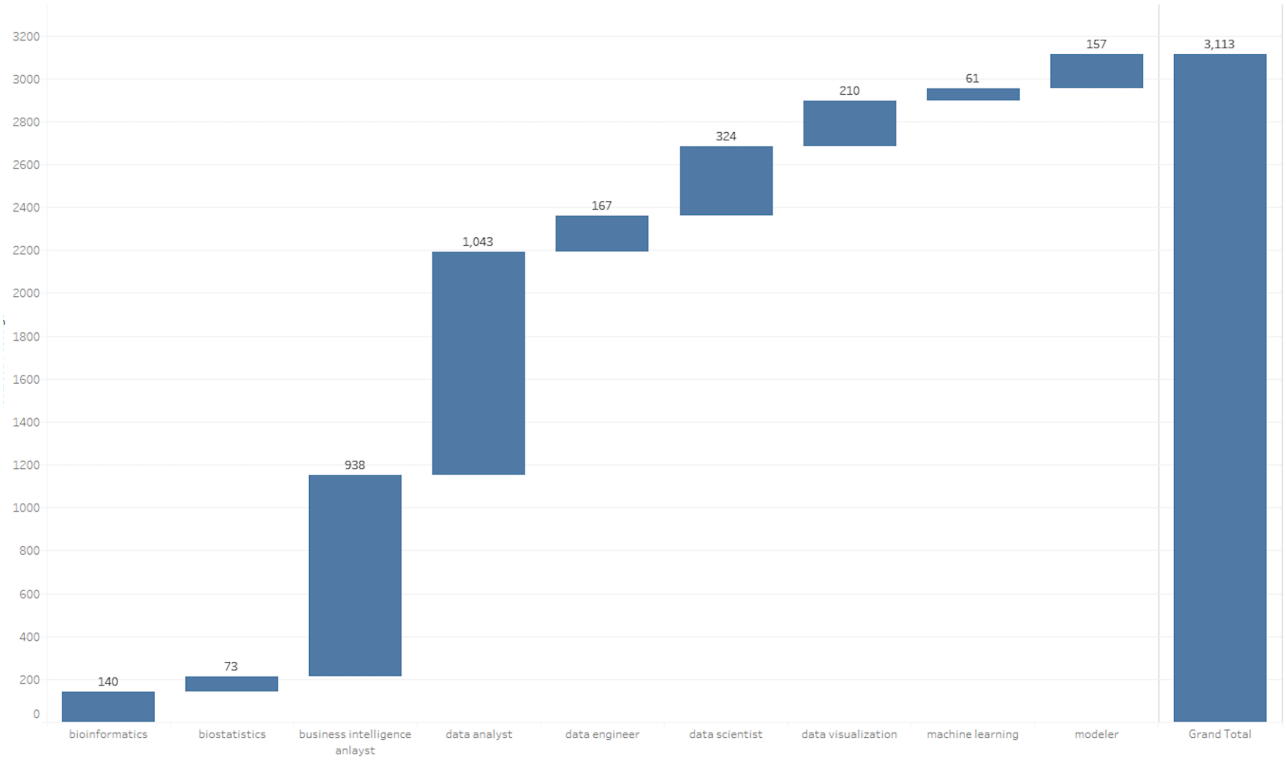
The team leveraged its user interface technology early to obtain visualizations to help with EDA. Plots shown in the next several figures were generated by our dashboard team and are candidates for inclusion in our dashboard product offering. Some of the top industries related to jobs found are illustrated in Figure 8. These tree map visualizations helped the team understand the number of postings by industry relative to the whole dataset. For example, technology, finance, and manufacturing are leading employers.

Figure 8: Industries Represented – Indeed Data for Data Science



While there were many job titles found in the data, the team focused on those related to “Data” and “Data Science”. This uncovered job titles like those shown in Figure 9. The plot shows the relative size of postings in 2018 for the job titled targeted by the team. For example, BI, Analyst, Data Engineer and Scientist were all top roles. This was a logical finding as many of these roles also perform some aspect of data science.

Figure 9: Job Title Groups – Indeed Data for Data Science

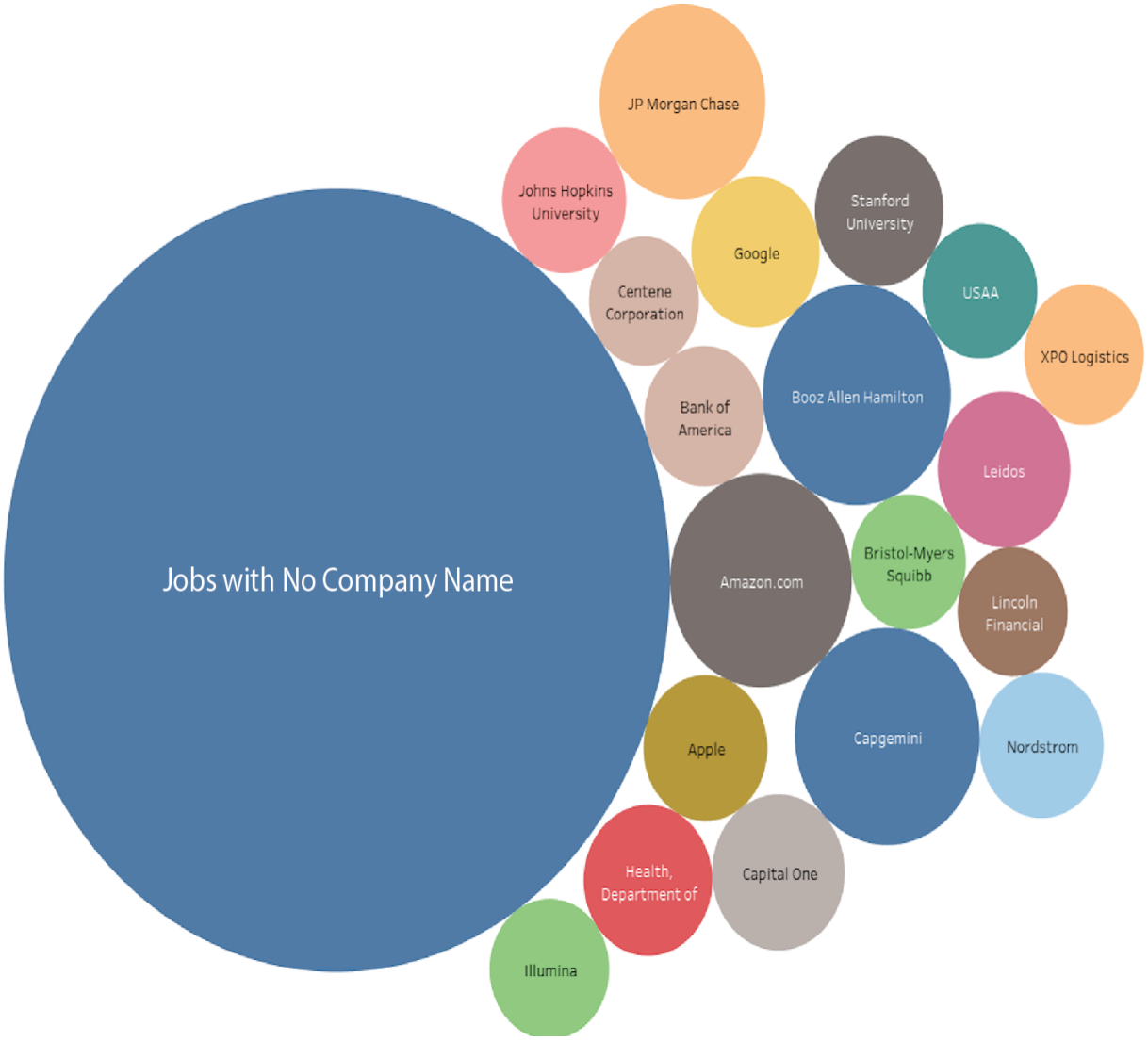


Jobs found were spread across 6,200 hiring companies, Figure 10 shows companies offering the most jobs in 2018. The team found many jobs in Healthcare and Finance, which may represent a higher than normal investment by these sectors to ramp up Data Science efforts (especially Healthcare which is generally behind other industries in new technology adoption). This same pattern was later found when a wide range if clusters were tested. During those tests, distinct Healthcare and Finance clusters

were identified with jobs requiring both Data Science and industry specific experience.

Figure 10 also shows a significant number of job postings not related to a company. After more research, the team verified this wasn't a missing data problem. Rather, these represented confidential searches, promotional listings, and general fishing for candidates by headhunters or companies to fill files for future openings.

Figure 10: Top Hiring Companies – 2018



The last macro data characteristic the team dived into was to understand the pattern of job postings, to look for peak/average volumes, and to determine if there is a seasonality (e.g. is there a good time to look for these roles?). Figure 11 shows job volume over 2018 was relatively constant at ~25K/month except for a spike in late March. That spike may represent the time it takes for new fiscal year funds to be released and HR departments to begin posting jobs. The spike may also be due to data scraping issues, as the team became aware of those kinds of issues after purchasing the data.

In general, no seasonality or cycles were detected within the data. A problem was discovered by the team with the vendor data feed resulting in no postings from January to late March. However, the team determined it had enough data to continue as is.

Figure 11: Job Posting Volume Over 2018

Overall, data was generally clean and complete except for salary information. Data came to the team in the form of variable text-based fields for job title, location, and job description. There were no fields that needed cleaning nor numeric transformations.

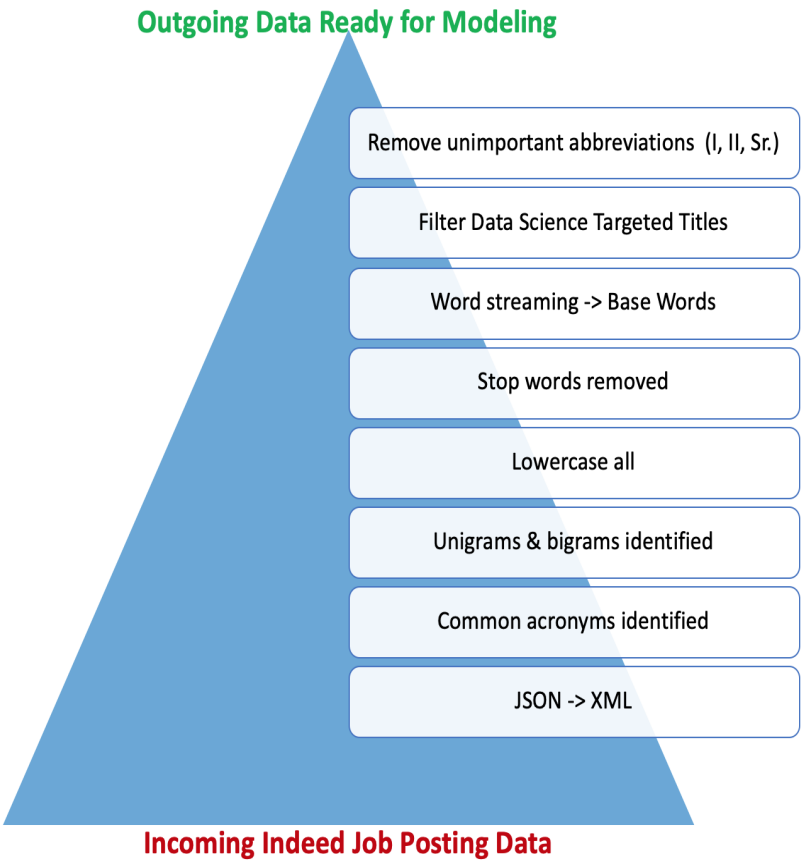
One field of interest and not initially available was salary. In a minority of postings, Indeed provides salary ranges based on their insight or ranges from hiring companies. However, since most postings didn't include salary; the team decided to search for better salary source in the future based on the roles targeted.

Data Preparation

Most of the work with the data came in the form of data preparation. This work included format transformations required by model and clustering tools, iterations of filtering, removal of unimportant words (called stop words in NLP), and other simplification strategies including the combining of like jobs. The remainder of this section provides details of the steps completed to ready data for modeling. Figure 12 provides an overview of the work done to take raw incoming Indeed job postings and get it to a state ready for modeling.

The initial raw data set (or corpus), was provided as a JSON formatted file containing job postings written in English. Initially, all data-related job postings were selected to keep the initial refined data set broad. Text was extensively pre-processed in order to make it useful for further analysis. For example, data was cleaned, common acronyms written out, punctuation removed, and common bigrams

Figure 12: Data Preparation Steps



(two words found together) were combined (e.g. “power bi” to “powerbi”). All text was changed to lowercase and generic (“a”, “the”, “an”, “to”, etc.) and job

description stop words were removed that didn't add semantic value to the modeling process.

Preparation also included stemming or iterating through each word and returning the base of that word (for simplification). For example, "playing", "plays", and "played" were all stemmed to "play". Lastly, we required that the text or title included data related terms such as "data scientist", "data science", "analyst", "data engineer", "big data", "data analyst", "machine learning" and a long list of other potentially related terms. This helped further filter postings to those RightHires was targeting for a 10-week project.

Additionally, some acronym-mapping was implemented. This included mapping common tokens such as "svc" to "service" to make the language more consistent. Tokens such as "I" and "II" (that indicate the level of a position, as in "Data Scientist II") were also removed from job titles to help the team get to (common) base jobs.

With the above complete, a review of the jobs remaining found positions related to "Data Analyst" and "Business Systems Analyst" had the most records. These were followed by "Data Scientist" and "Data Engineer" positions. Jobs like "Artificial Intelligence Developer" represented the least number of jobs. The team concluded this was due to the newness of the role in 2018, or that AI work is a component of work several roles cover. For example, those doing Machine Learning work also do some AI.

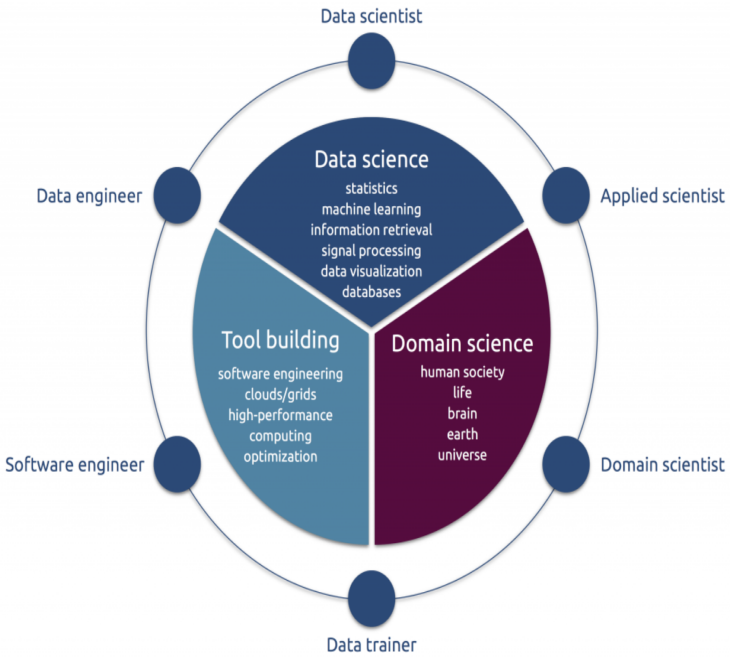
Selecting Data Science Job Titles for Modeling

To help scope the project for 10 weeks, the top 200 job titles were reviewed and evaluated by the team. A list of job titles was identified that didn't fit within the scope of data science titles (e.g. "electrical engineer", "software developer", "web developer", "java developer", "etl developer") and filtered out. Next, another group of 200 titles were reviewed that might be related to

data science and those not relevant removed (like DBA, system administrator). Other jobs like “Data Engineer” were kept.

During the process of reviewing groups of 200 titles, the team selected a list of words that were required to appear within a title to be relevant (and of these words). These included “data”, “analyst”, “data scientist”, “sql”. The team then determined (after looking at sample job postings of jobs excluded) to include “data analyst”, “business intelligence analyst”, “data scientist”, “data engineer”, “biostatistician”, “bioinformatics”, and “data visualization” (as it was found many companies still used those terms for jobs that were effectively Data Science positions). This second set allowed more relevant jobs to be included in the data for modeling. Last, the most common unigrams and bigrams were reviewed and several popular terms unrelated to data science skills were removed to further refine data for modeling.

Topic Modeling



https://www.datascience-paris-saclay.fr/data-science/

Using job descriptions related to job titles the team created a TF-IDF matrix with a TF-IDF vectorizer. This step is necessary because NLP/Machine Learning mechanisms require a numeric representation of words to operate. Functions called by this step have many “dials” (or parameters) that can be varied to control the sensitivity of what’s done.

For example, two parameters leveraged by the team were min and max “df”. The team found values of 10 and 0.9 for “min_df” and “max_df” respectively worked the best. The 10 “min_df” value removes words that aren’t included in at least 10 documents. The .90 “max_df” removes words that were in more than 90% of documents.

Once the TF-IDF matrix was created the team used K-means to create topic clusters. Then three tables were created to compare clusters while also seeing if they corresponded with the job titles evaluated (2-6 clusters were attempted). At this point the team could now see the top 10 job description words for each cluster tested. That provided insight to the team as to how jobs could be classified, if we had clusters that overlapped (not as useful), small clusters (not good), or clusters that were large, nearly equal in size, and distinct covered targeted jobs (best case and our goal).

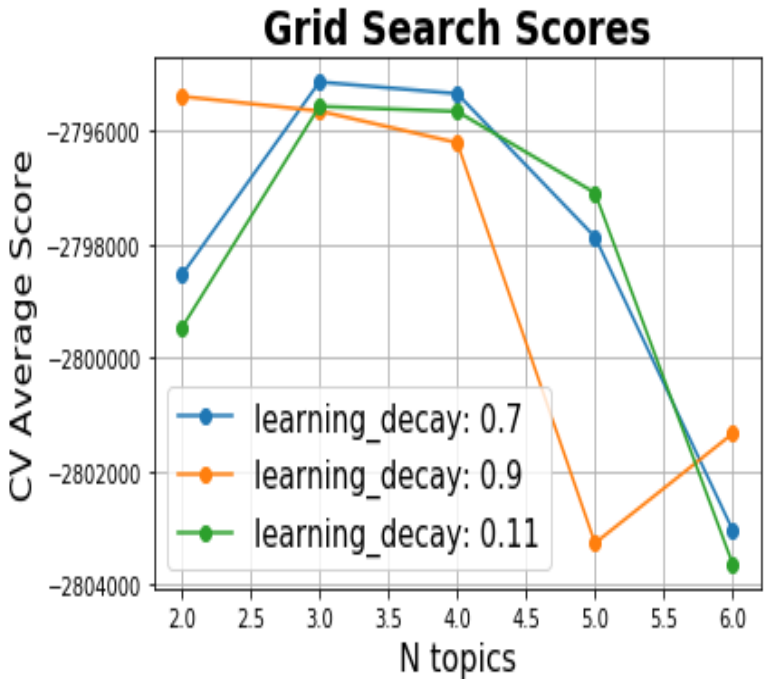
The team then followed the same steps using bigrams to determine if that would improve clusters and make them more distinct than prior clusters found with K-means. Bigrams performed somewhat poorly compared to unigrams, so the team moved forward using the original unigrams for the TF-IDF step. At this point, multidimensional scaling was conducted using t-distributed stochastic neighbor embedding (T-SNE). This was done to get a 2-dimension view of our results (see T-SNE section). This method was repeated for three and four clusters to help provide a view into the dispersion within a cluster and pick an optimal case (from 3 or 4 clusters).

To further test alternative methods and help the team verify it got to an optimal solution, hierarchical cluster analysis was tried. However, it was found that it did a poor job clustering with most jobs in one big cluster. Doc2Vec was also tested using the processed text, then K-means and T-SNE. Again, most job titles were simply clustered together in one group and therefore not useful. The team them conducted bi-clustering with spectral co-clustering. This method is not common in NLP and a long shot. The performance for this model was also poor and was abandoned.

Last, Non-negative Matrix Factorization (NMF) and Latent Dirichlet Allocation (LDA) were tested to evaluate differences in topics. LDA created a probabilistic model of how documents were grouped by latent topics. In other words, we evaluated common topics by usage of related words and determined which documents were most related to those identified topics. A grid search (Figure 13) was then used for parameter tuning. The plot below helped the team compare different numbers of topics by log likelihood scores. These scores helped the team determine the best fit for the LDA model (the higher the negative score, the better).

Figure 13: Tool to Select Optimal umber of Clusters

The LDA grid search in Figure 13 shows 3 or 4 clusters best explained the data (lowest or most negative score). This confirmed K-means findings of either a 3 or 4 clusters as optimal. Because scores for 3 or 4 clusters cases are so close, and simplicity is best, the team selected to go forward with 3 clusters.



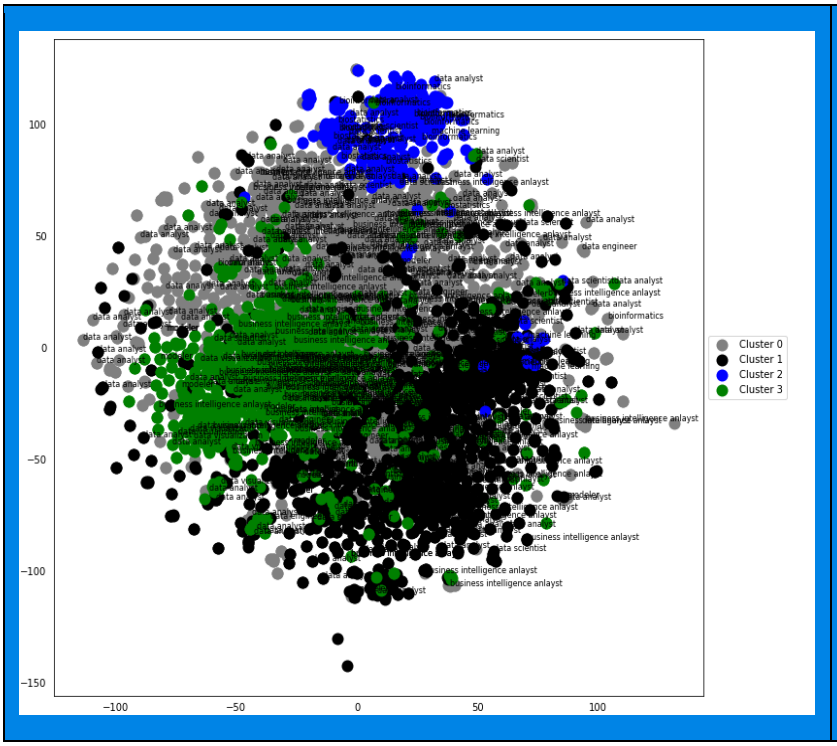
The next step involved the use of the package pyLDAavis. This allowed the team to visualize the clusters along with the relevance and frequency of (job description) terms for each. Examples of pyLDAavis plots by cluster are shown in Figures 16-19. These plots were used to name each cluster after reviewing top job description words identified for each cluster.

Right Hires!

Picking the Optimal Number of Clusters

This section provides further insight into the tools and strategies used by the team to decide between 3 or 4 clusters.

Figure 14: Four Cluster Job Title Model



Cluster	0	1	2	
job_title				
bioinformatics	7.0	2.0	157.0	
biostatistics	15.0	NaN	80.0	
business intelligence anlayst	192.0	227.0	1.0	51
data analyst	1070.0	618.0	70.0	32
data engineer	23.0	121.0	5.0	3
data scientist	25.0	290.0	19.0	
data visualization	20.0	19.0	NaN	17
machine learning	7.0	38.0	17.0	
modeler	56.0	94.0	4.0	1

Targeted job titles (left) and the number of times they occur in each of the 4 clusters (right)

Figure 14 (4-cluster) provides a graphical view of the distribution of jobs in each cluster along with a numeric distribution of key jobs across the 4 clusters. From Figure 14 we can see that not all four clusters are completely clear, so 3 clusters were then considered. While Cluster 2 (above) is the very distinct, Cluster 0 is very dispersed (not ideal).

Cluster 0 (grey): Seems to be most related to the ‘data analyst’. However, we saw that Data Analyst is fragmented into several clusters (not ideal).

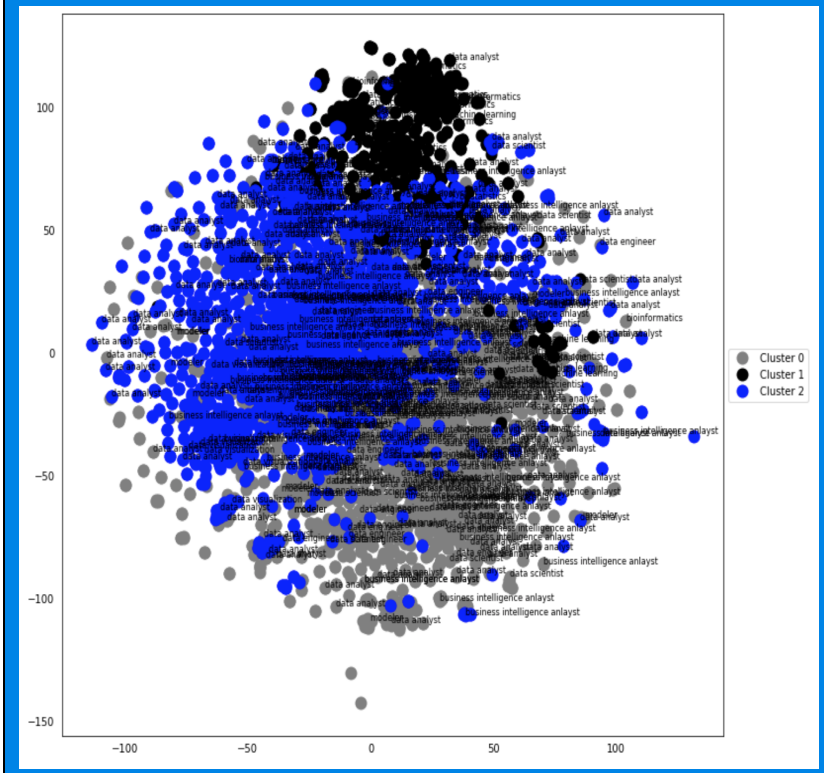
Cluster 1 (black): From the table we can see that Data Engineer and Data Scientist jobs are largely categorized here. Looking at the terms for this cluster the team confirmed job description words like “model” and “learn”.

Cluster 2 (blue): Was the most distinct cluster. We saw that bioinformatics and biostatistics are predominantly grouped here. Interestingly, in our initial explorations we left out Bioinformatics, Biostatistics and data visualizations and this cluster was still the most distinct cluster. The words (and we will see this when we conduct LDA) indicate that these roles are specific to the healthcare industry.

Cluster 3 (green): includes BI analyst and data analyst jobs. In addition to the top terms of 'BI analyst' and 'data analyst', we also saw terms like 'report' and 'dashboard' being used.

Because using 3 or 4 clusters was so close on the Grid Search Score plot, the team tried 3 clusters to see if we could simplify further. Figure 15 shows what happened when 3 clusters were used. Groupings were more distinct and with less overlap, so the team decided to go forward with 3 clusters.

Figure 15: Three Cluster Job Title Model

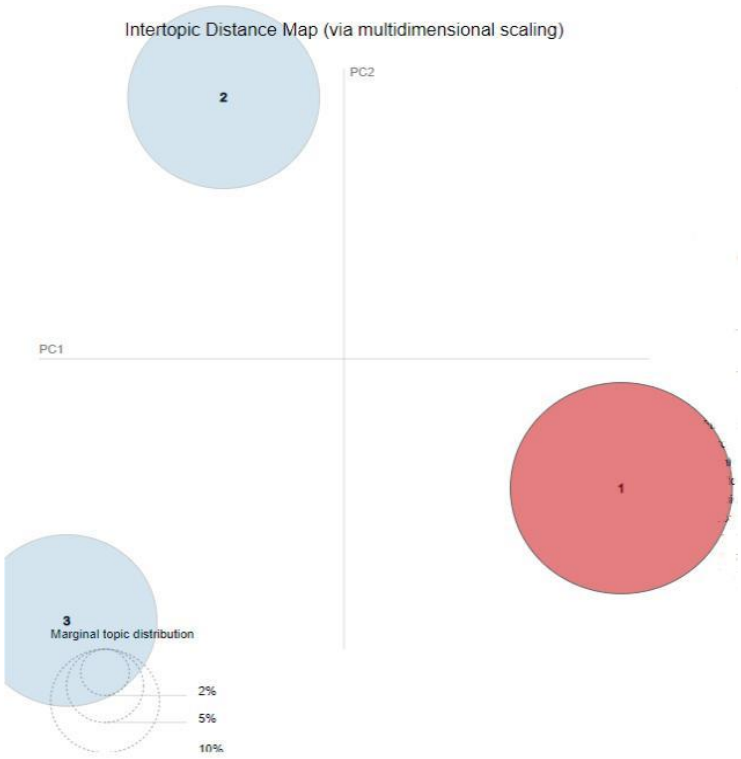


Cluster	0	1	2
job_title			
bioinformatics	7.0	153.0	2.0
biostatistics	NaN	94.0	1.0
business intelligence anlayst	280.0	56.0	602.0
data analyst	782.0	402.0	898.0
data engineer	119.0	2.0	58.0
data scientist	297.0	30.0	23.0
data visualization	39.0	3.0	180.0
machine learning	41.0	15.0	6.0
modeler	124.0	20.0	54.0

Targeted job titles (left) and the number of times they occur in each of the 3 clusters (right)

Upon inspection of job description words, the team assigned clusters with the names indicated in Figure 16. Relative cluster size is similar confirming the team obtained an optimal state.

Figure 16: Resulting Clusters



Cluster 1:
BI & Analyst

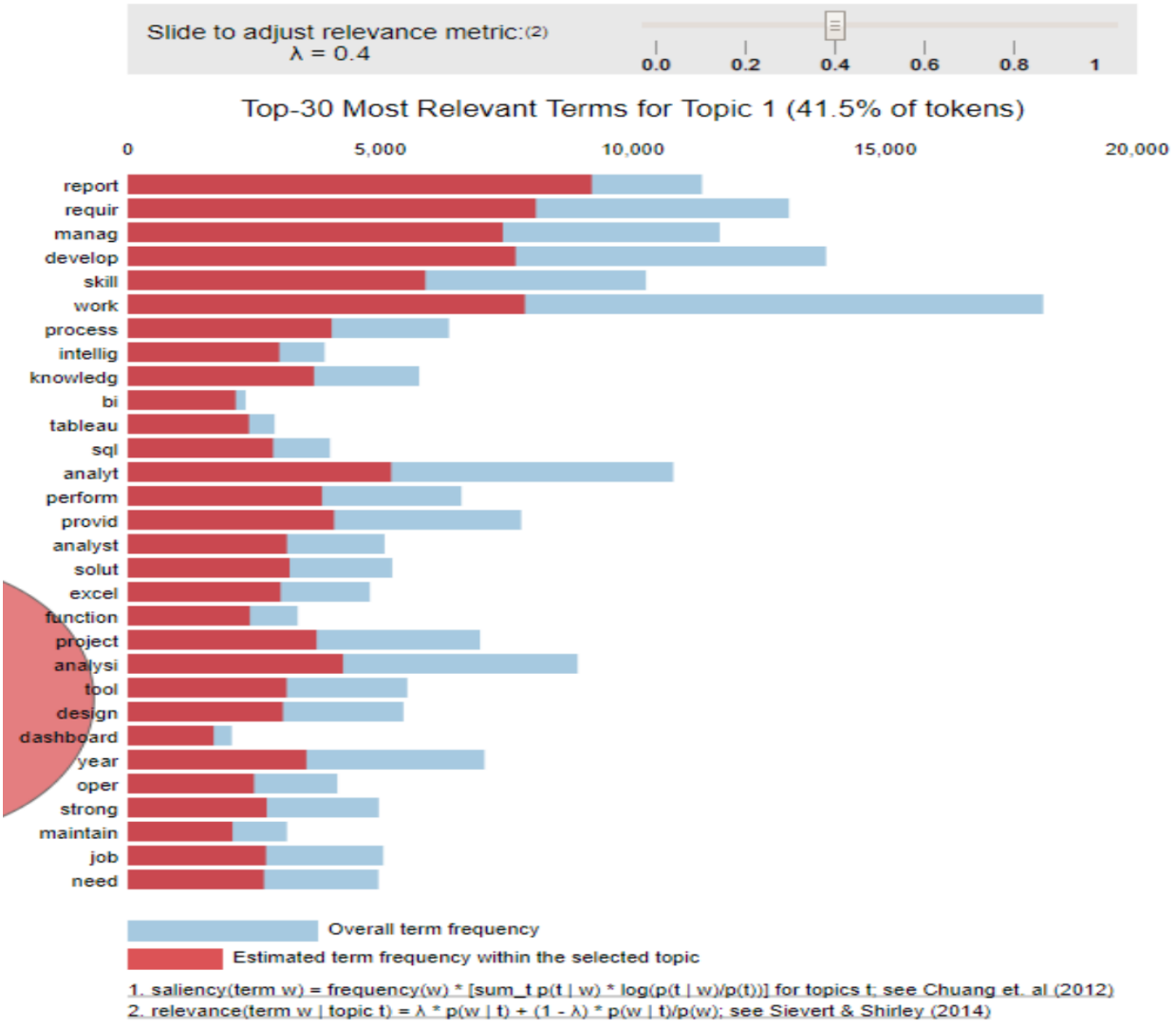
Cluster 2:
Data Science & Data Engineer

Cluster 3:
Healthcare (data science jobs in healthcare)

Further testing of the 3-cluster structure was performed to see if improvements could be made or at least verify the team reached optimal clusters. For example, K-means, LDA, and NMF were tested and all provided similar results. The pyLDAavis function was then leveraged to view top 30 job description words, verify the words were all distinct (a best case), and verify those words added together matched the names assigned to clusters.

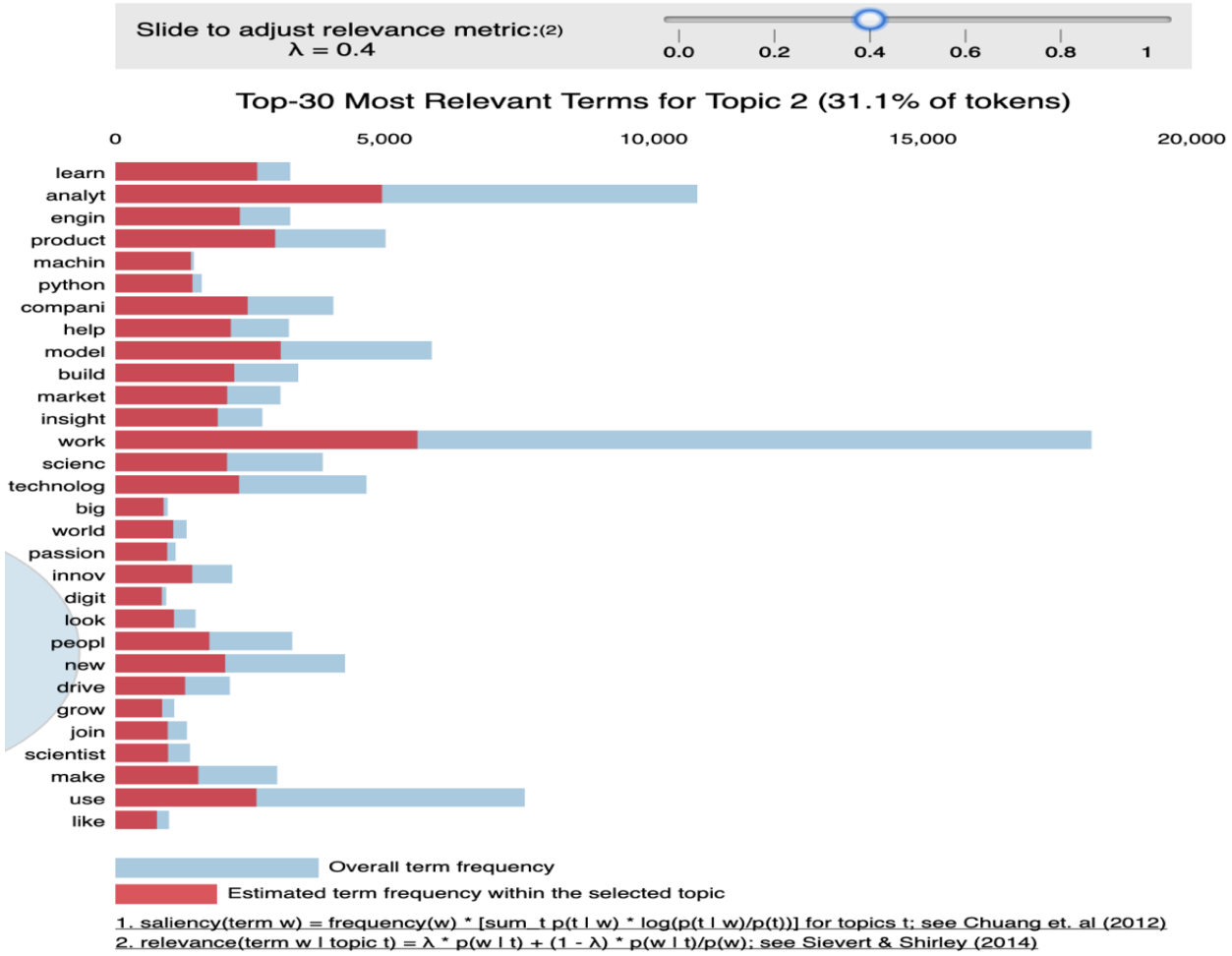
The three figures that follow show the top 30 job description words for each cluster sorted by relevance (e.g. “report” is most relevant job description word in Cluster 1 which is associated with BI and Analysts). The relevance position of the job description words shown consider the frequency of the word in that cluster (red bar) compared to the frequency of the word in the corpus (grey bar).

Figure 17: Cluster 1 (BI/Analyst) Top Job Description Words



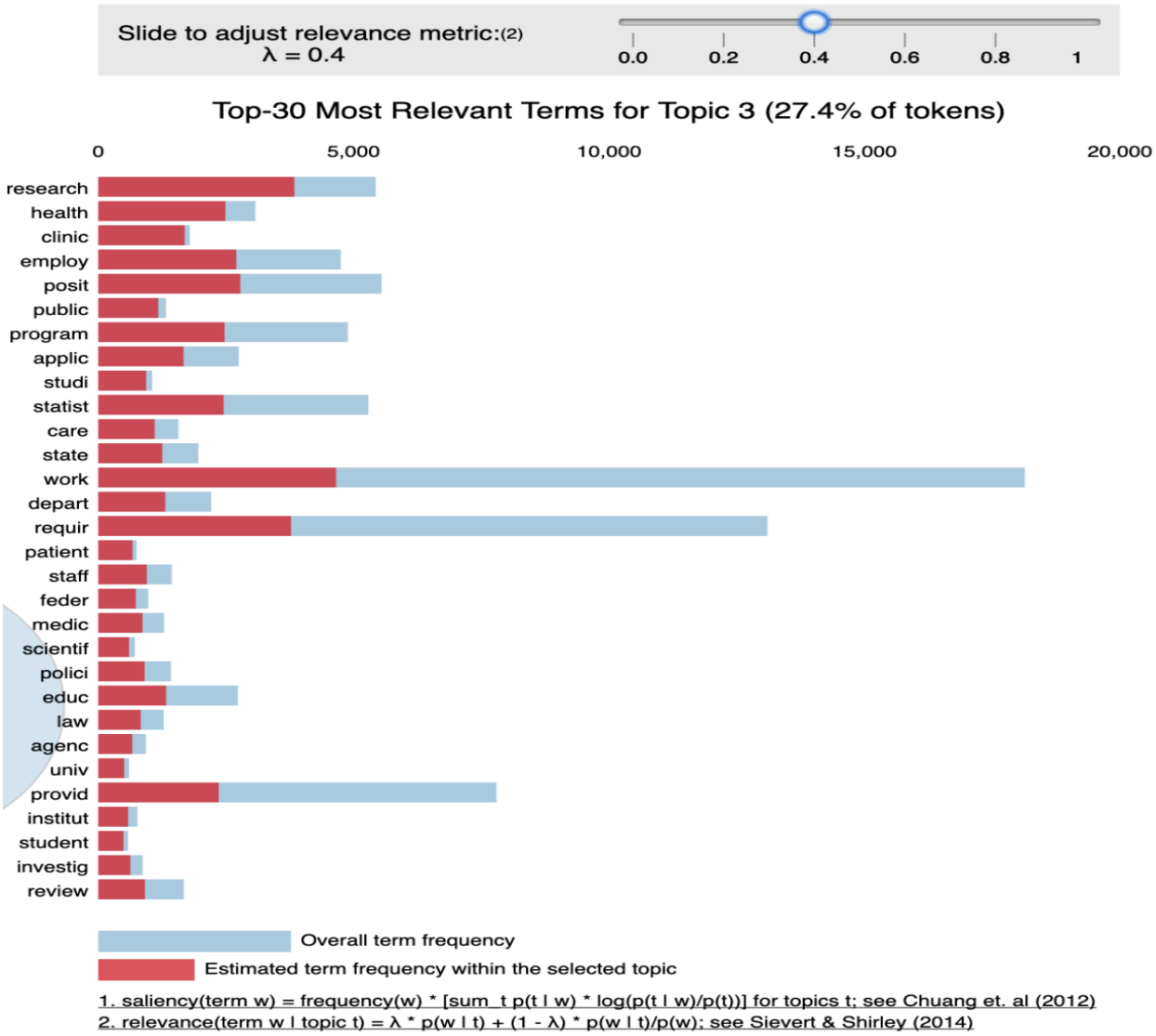
While Figures 17-19 plots are dynamic, reducing the lambda setting (right top of screens) increases the ratio of the frequency by that cluster to the frequency to all docs. A lambda in the .2 to .4 range was selected which allowed the team to identify the most relevant words to formulae a cluster identity. Said another way, the plots are listing the job description terms that uniquely define the cluster (common to this cluster and not common to all other clusters). This is important as it minimizes overlap of clusters and makes their identity to our customers clear.

Figure 18: Cluster 2 (Data Science/Data Engineer) Top Job Description Words



Right Hires!

Figure 19: Cluster 3 (Healthcare) Top Job Description Words



Right Hires!

Job Cluster Conclusions

Going forward the team used the 3-cluster model that have the characteristics shown in Table 2. These cluster are leveraged by our customers via the two user interfaces.

Table 2: Top 3 Clusters Covering RightHires Targeted Jobs

Cluster	1	2	3
Job Classification Name	**BI & Analyst**	**Data Science & Data Engineer**	**Healthcare**
Percent of all targeted jobs covered by each cluster	41.5%	31.1%	27.4%
Alternative job titles in cluster And occurrence in the corpus in parenthesis	bi analyst (304) data analyst (111) senior bi analyst (62) tableau developer (56) data engineer (25) senior data analyst (19) senior tableau developer (17) business data analyst (16) bi analyst ii (16) data scientist (15)	bi analyst (123) data scientist (121) data analyst (107) senior bi analyst (39) data engineer (38) senior data scientist (37) senior data analyst (29) senior data engineer (20) marketing data analyst (13) bi analyst ii (8)	data analyst (32) bi analyst (22) bioinformatics scientist (21) data scientist (14) director biostatistics (12) research data analyst (12) bioinformatics specialist (9) senior bi analyst (9) healthcare data analyst (8) scientist bioinformatics (6)

Table 2: Top 3 Clusters Covering RightHires Targeted Jobs (continued)

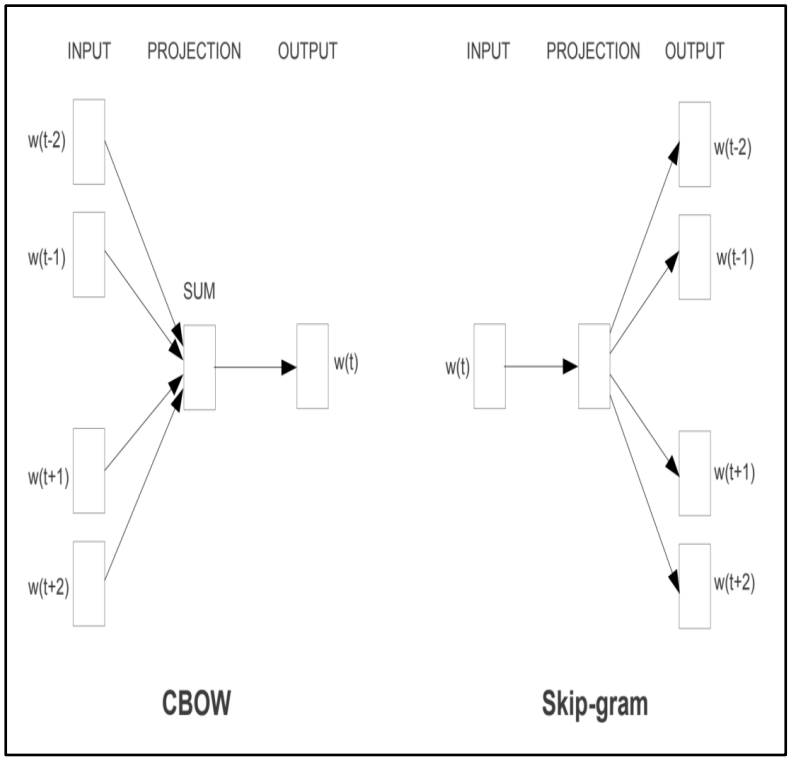
Cluster	1	2	3
Top terms and skills in order of importance that should be in job description (More in plots and data to user RightHires interfaces)	Report, requirements, manage, development, work process, business intelligence, knowledge management, Tableau, SQL, analyst, performance, dashboard, project, tools, design	Machine, learning, design, engine, python, analyze, model, big data, technology, innovate, data scientist, drive, make, build, insight, market, passion	Research, clinic, patient, public, population, care, applications, study, strategist, state, medical, investigate, requirements, staff, federal, scientific, education

Word2Vec

In addition to the methods outlined in the last several sections for modeling and clustering, the team experimented with a method called Word2Vec. Word2Vec can be used to help determine words that are like others. This was deemed useful to the RightHires because shortages of candidates with particular skills (like Python) may be successfully filled with candidates with similar skills (like R or SAS) using different technologies.

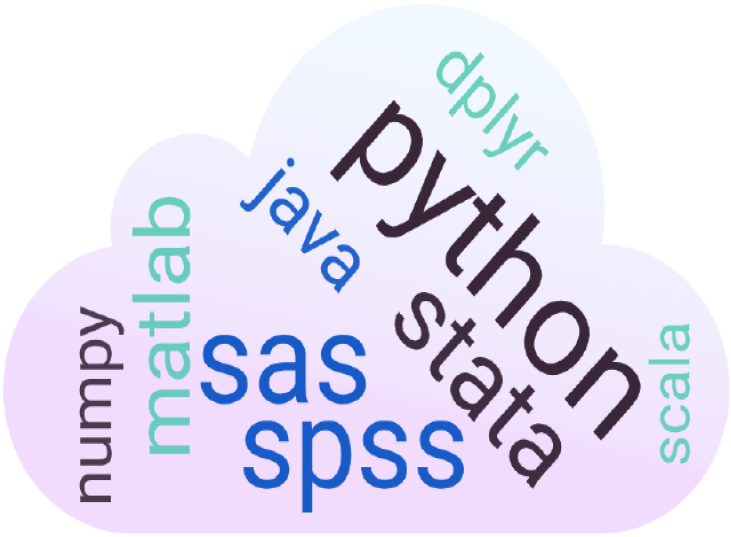
Word2Vec is a popular algorithm used to create word embeddings. "An embedding is a mapping of a discrete, categorical variable to a vector of continuous numbers. In the context of neural networks, embeddings are low-dimensional, learned continuous vector representations of discrete variables. Neural network embeddings are useful because they can reduce the dimensionality of categorical variables and meaningfully represent categories in the transformed space" (Koehrsen, 2018).

The Python package Gensim was used including the default shallow neural-network architecture using Continuous Bag-of-Words (CBOW). The architecture is illustrated to the right. The model focused on learning about words given their local usage context, which was defined by a window of neighboring words. Word order is not considered for CBOW.



In Word2Vec, a configurable window parameter of 5 was tested. Resulting vectors generated were 100-dimensional, which reflected the number of nodes in the hidden layer of the neural network. This is also a configurable parameter in the algorithm tested.

Figure 20: Words Similar to "R"



Word2Vec maps each word on a 100-dimensional space. Based on context, words that are more similar to one another, are placed closer together in this space. This is based on paradigmatic relations, meaning that words with similar contextual words are considered similar. Therefore, there is a function to retrieve the top 10 words most similar to any given word. This is based on cosine similarity. Figure 20 shows the relative similarity metric for those words in the corpus most related to the "R" programming language (the larger the word, the higher the cosine similarity). Languages like Python (with a cosine similarity of 0.92) and SAS (0.90) are similar to R meaning an employer could consider someone with those skills if there was a shortage of R candidates. By comparison, dplyr has a cosine similarity of 0.79.

A cosine similarity score of 1.0 indicates the words are the same. While a score of 0 indicates unrelated words. Words close to a 1 indicate similar skills that may be substituted.

Figure 21: Similar Words to Amazon Web Services

Similarly, organizations may find it difficult to find data engineers that have skills in cloud-based and distributed computing technologies. Figure 21 shows the most similar words to Amazon Web Services, or "AWS". The most similar word is Microsoft Azure (0.92), which is the competing cloud-computing environment to AWS.

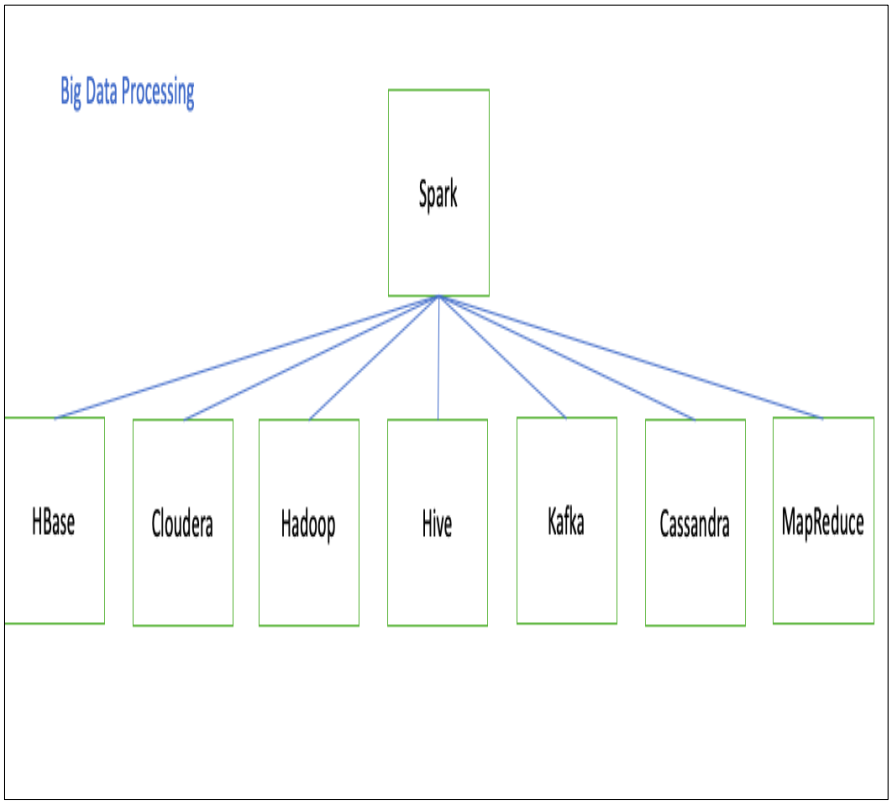
Redshift (0.90), the next word, is a cloud-based data warehouse product that is part of the AWS platform. Snowflake has a cosine similarity of 0.81 and NoSQL 0.87. Therefore, an organization could broaden the candidate pool for data engineers by asking for either Azure or AWS skillsets. The same is true for the distributed processing platforms of Hadoop and Spark, which are also depicted in Figure 21.

This analysis helped in the development of a domain-specific ontology. Ontologies are data structures that consist of hierarchically-organized concepts (entities) and relations that explain objects in the target world of observation (Mizoguchi, 2004). An ontology can be used in feature engineering and to manually disambiguate clusters produced through the K-means clustering algorithm.

Feature engineering for NLP tasks include the task of vectorization. This is based on developing a vocabulary (V) for the (training) corpus. The more fine-tuned the vocabulary, the better the output from the various vectorization techniques.

One method of refining a vocabulary is defining equivalence classes with the guidance of a domain ontology. This was done by mapping semantically-similar words onto another. For example, there are several tools and platforms available in the industry for big-data processing. This includes Spark, Hadoop, DataBricks, Cloudera, etc.

Figure 22: Sample Ontology for Big Data



In order to disambiguate these terms, one can form equivalence classes.

The sample ontology to the left shows available tools and platforms related to big data processing. For example, this would include Spark, Hadoop, DataBricks, Cloudera.

The structure in Figure 22 guide mapping any word in the leaves of the ontology onto its parent. In the above example, all similar words in the leaves are mapped onto Spark. This was done as it can help evolve clusters further, so they become even more distinct.

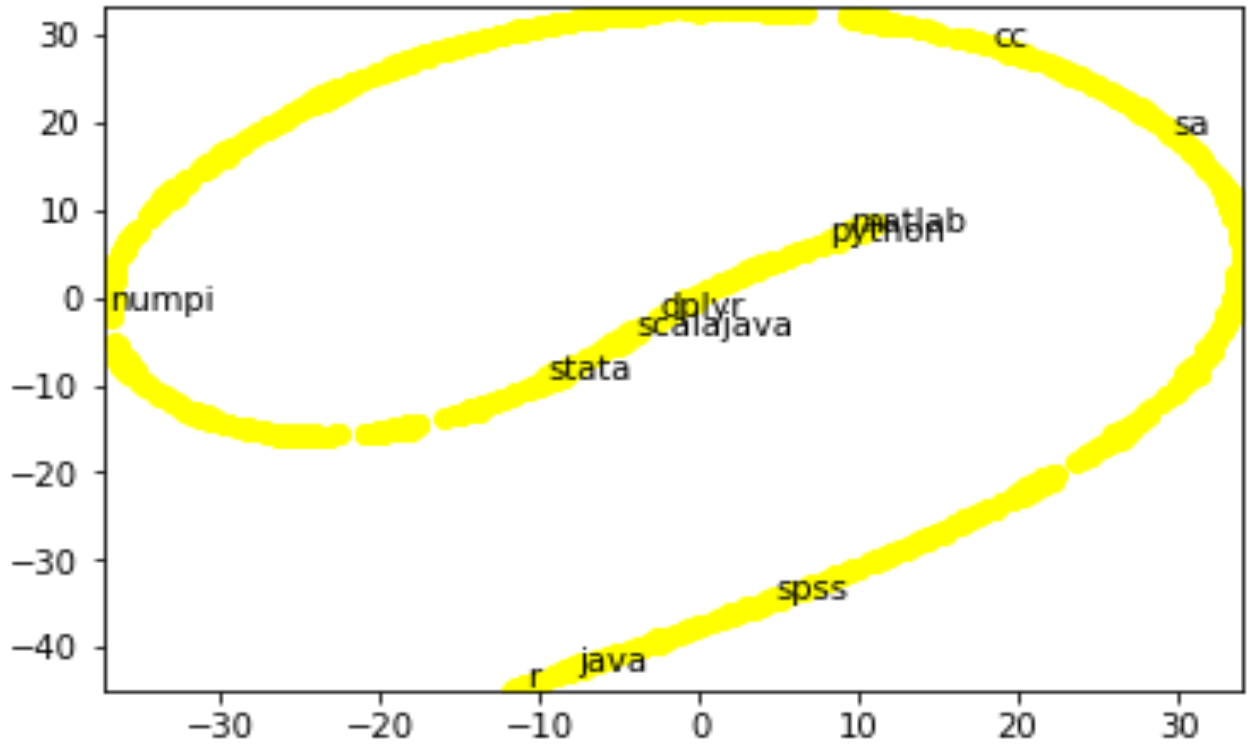
Figure 23: Similar Words to Dashboarding, Visualization, Tableau

This method was also used to map lesser known words found in the corpus to more known ones. For example, Figure 23 shows the top 10 words most similar to the dashboarding and visualization software "Tableau". These words all refer to other BI tools commonly used for similar purposes. Therefore, if an organization is looking for a BI Analyst or a Data Analyst with "Looker" (0.84) visualization skills (which is a less commonly-known BI tool) then they could change the position description to also include Tableau or QlikView (0.81) experience. Cognos leads the group with a cosine similarity of 0.89.

T-SNE

Word embeddings created using Word2Vec can be visualized using t-distributed stochastic neighbor (T-SNE) embeddings. This is a dimension-reduction technique used to map n-dimensional vectors into 2-dimensional vector spaces. The team used this tool to further visualize words that have similar meanings (to help customers know “like skills” to consider substituting if there are resource shortages).

Figure 24: Words Most Similar to “R” Programming Language



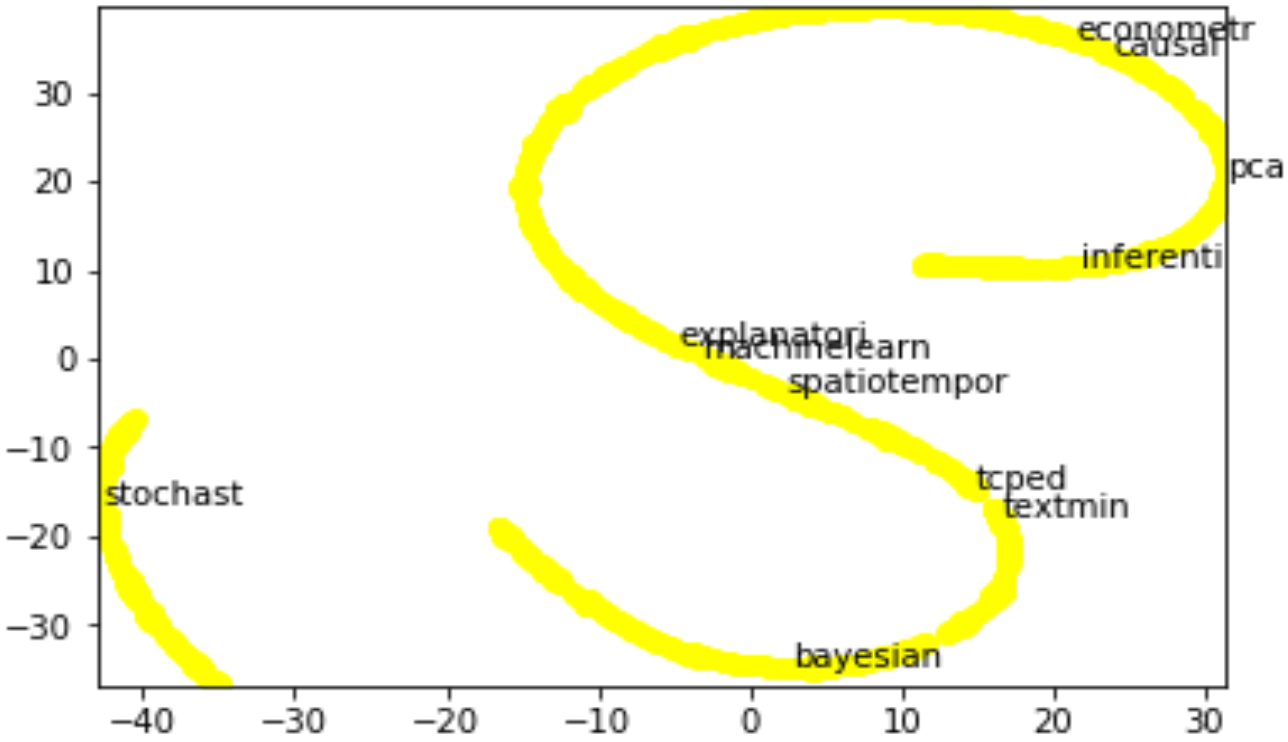
This allows one to reduce the number of dimensions in order to be able to explore relationships in the data visually (Pathak, 2018). Figure 24 is a visualization of the 10 most similar words to the “R” programming language.

One can see several programming languages or Python packages (i.e. dplyr) clumped together in the middle, while other languages such as SAS and Java are further away. This indicates that “stata”, “python”, “Scala”, and “MATLAB”

are found in a more similar context than “SPSS”, “SAS”, or the “NumPy” Python package. However, all of these words are similar to one another in comparison to other words in the corpus.

Similarly, Figure 25 is a visualization showing the 10 most similar words to “machine learning”. This included words such as stochastic, “Bayesian”, “econometrics”, “PCA”, “spatio-temporal”, and “text mining”.

Figure 25: Words Most Similar to Machine Learning



Knowing what words can be substituted for others with similar meaning can help RightHires suggest alternative job description text. This can help customers unable to find Python developers and are willing to be flexible and accept “R” developers with the ability to learn (as model building theory is similar, so the adoption of a new language is a smaller step).

Combining Mappings with Clustering

Previously we reviewed how we identified three clusters, including Cluster 2 - Data Science/Data Engineering, which is of particular interest. Our goal in this section was to describe how we further broke this cluster of job titles. Although conducting further clustering using the same methods didn't result in clear sub-clusters, instead our Doc2Vec ontology comes in handy.

As was already shown, we can use Doc2Vec to create mappings and map similar skills to a more generalized skills set (e.g. Hadoop, Spark, and DataBricks were mapped to 'BigData'). Since the topic modeling has limited data, it can be difficult for the model to understand the relationships between words and to learn to treat synonymous words equally during topic modeling. With enough data, the model can learn these vector representations, however with mapping related terms we hope to assist the process.

Once mappings were completed, the data science cluster was re-evaluated to determine if any improvements were made. Some parameter tuning was required given that we have some words with larger counts. These same parameters were also adjusted in the initial data science cluster model without discernible improvements. For three clusters, there did seem to be a marked improvement with clusters being much more compact as shown in the clustering again using T-SNE.

Looking at the job posting tables created by using K-means, we can even see the clusters are pretty separated by their job titles, with data engineer, data scientist, and data analyst/BI making up separate clusters. From the T-SNE plot we can see the most distinct clusters seem to be clusters 0 and 1 which is really an improvement from our previous clustering. Cluster 0 includes as some of the most common terms the words 'big data', 'engine', 'structured', and 'pipeline'. While cluster 1 includes words related to machine learning, statistics and python. Lastly, cluster 2 includes analyst, report, analysis, and skill.

Right Hires!

Figure 26: Cluster Map

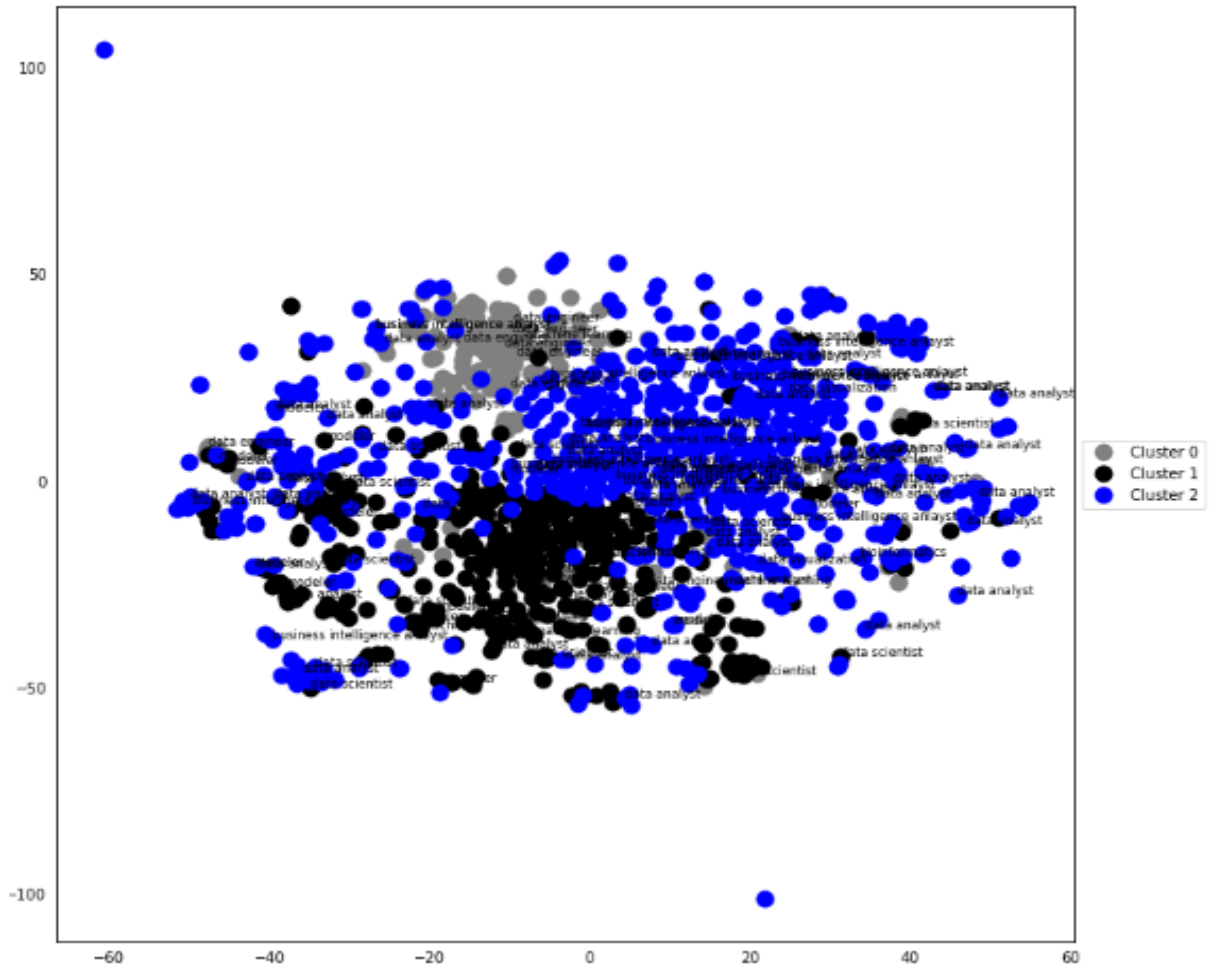


Figure 27: Job Title Split Across 3 Cluster Model

Cluster	0	1	2
job_title			
bioinformatics	NaN	3.0	4.0
business intelligence anlayst	15.0	9.0	256.0
data analyst	23.0	44.0	324.0
data engineer	94.0	6.0	17.0
data scientist	23.0	226.0	39.0
data visualization	4.0	NaN	32.0
machine learning	3.0	38.0	NaN
modeler	4.0	88.0	17.0

NMF also found similar clusters with one cluster including the words ‘machine’, ‘model’, ‘scientist’, ‘python’ etc. The second cluster included analyst, market, report, analysis and the final ‘bigdata’, ‘engine’, ‘structure’, ‘pipeline’, ‘cloud’, and ‘unstructured’.

Interestingly, LDA seemed to only find a single cluster with data science and data engineer words such as ‘model’, ‘bigdata’, ‘python’ etc., and another cluster had more general analyst words such as ‘analyst’, ‘market’, ‘report’, ‘structured’, and ‘analysis’.

We also used Silhouette Analysis to evaluate the optimal number of clusters. Silhouette analysis allows us to see how well the points are separated by cluster by visualizing plotting how a metric from -1 to 1 for how close each point in a cluster is to other clusters.

A score of 1 indicates the point is far from other clusters, 0 indicates the point is very close to the decision boundary, and negative numbers indicate the point may have been put in the wrong cluster. The width of the silhouette also portrays the size of the clusters. We also computed the average silhouette scores based on the number of clusters and we can see the highest score (0.6074) is for three clusters.

Figure 28: Silhouette Score for Different Number of Clusters (higher is better)

```
For n_clusters = 2 The average silhouette_score is : 0.5271527052257654
For n_clusters = 3 The average silhouette_score is : 0.6073948808235134
For n_clusters = 4 The average silhouette_score is : 0.5558710636930293
For n_clusters = 5 The average silhouette_score is : 0.522453271368437
For n_clusters = 6 The average silhouette_score is : 0.5490437756441872
```

Figure 29: Three Clusters – Silhouette Analysis

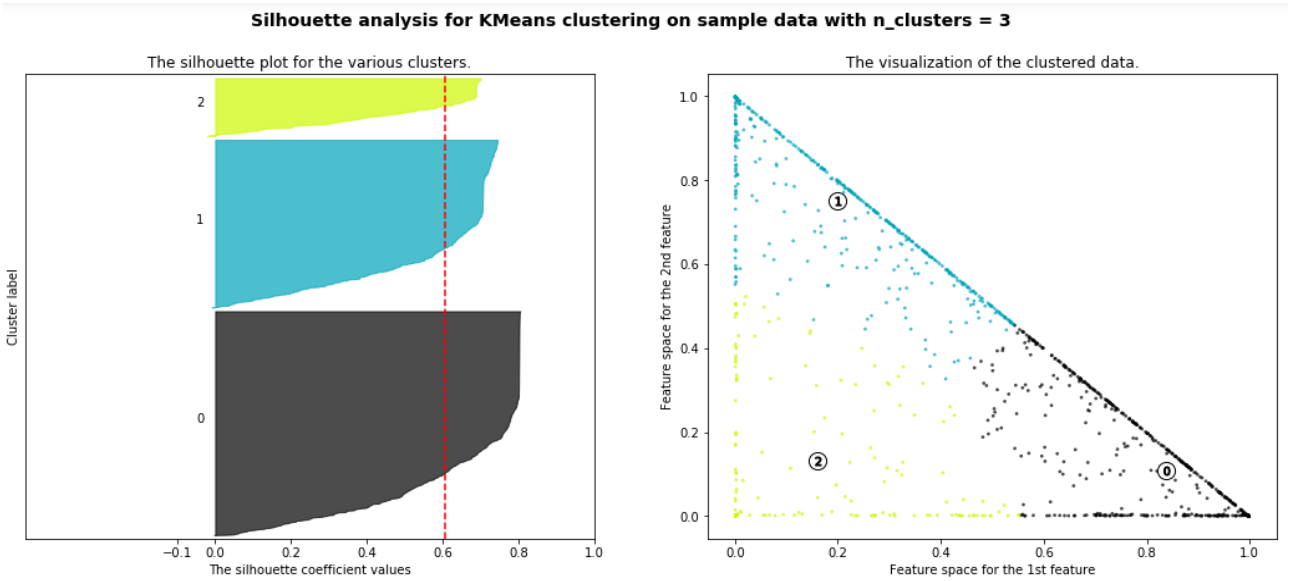


Figure 30: Two Clusters – Silhouette Analysis

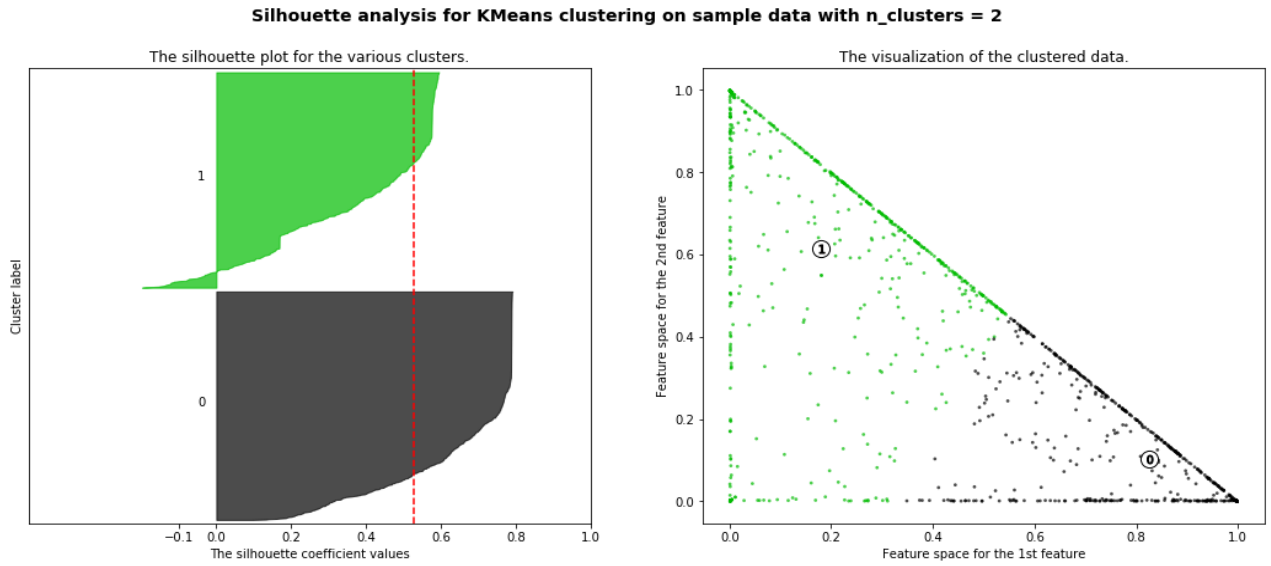
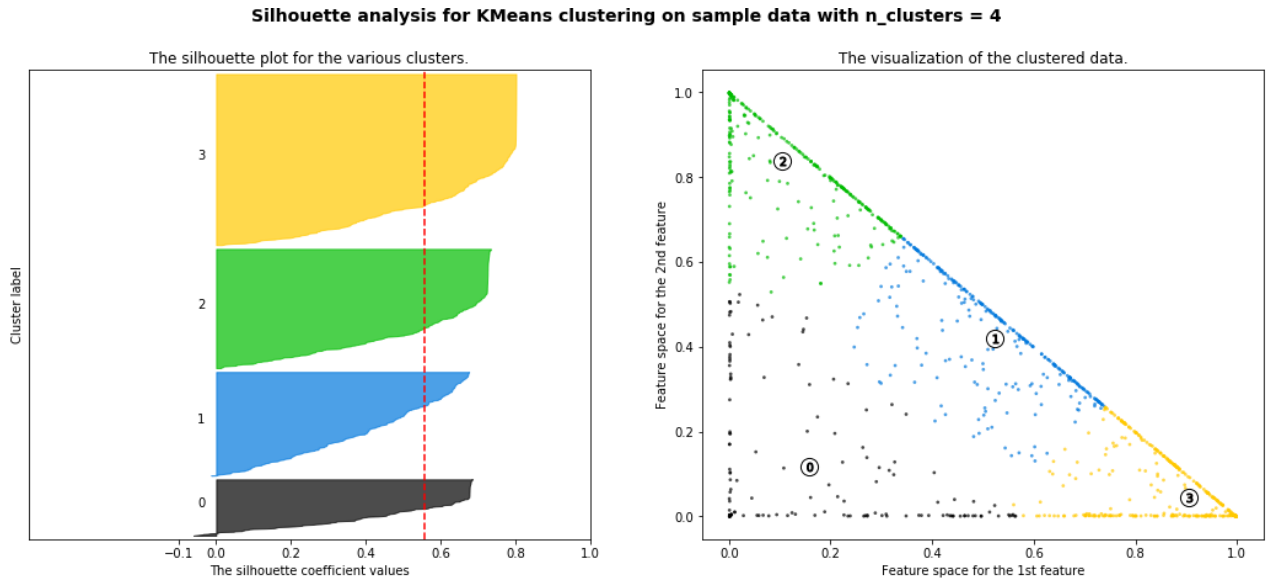


Figure 31: Four Clusters – Silhouette Analysis



Looking at the visualization for three clusters, we can see coefficients close to 1. We can see that the clusters are different sizes which seems to coincide with our job title counts (e.g. analyst/bi analyst have much more samples). Unlike two and four clusters, we also don't see negative numbers. We do see some low values, but for the most part the data points have higher scores (with an average of .6 delineated by the red line.)

These findings indicate that our data science cluster can be broken up into three distinct categories. This indicates there is less fragmentation or confusion with what these roles entail than previously expected. While the roles are very closely related the word prevalence and skills indicate the job postings are distinct. BI analyst and Data Analyst roles, on the other hand, couldn't be separated from these two roles but the two titles seem to have a lot of overlap.

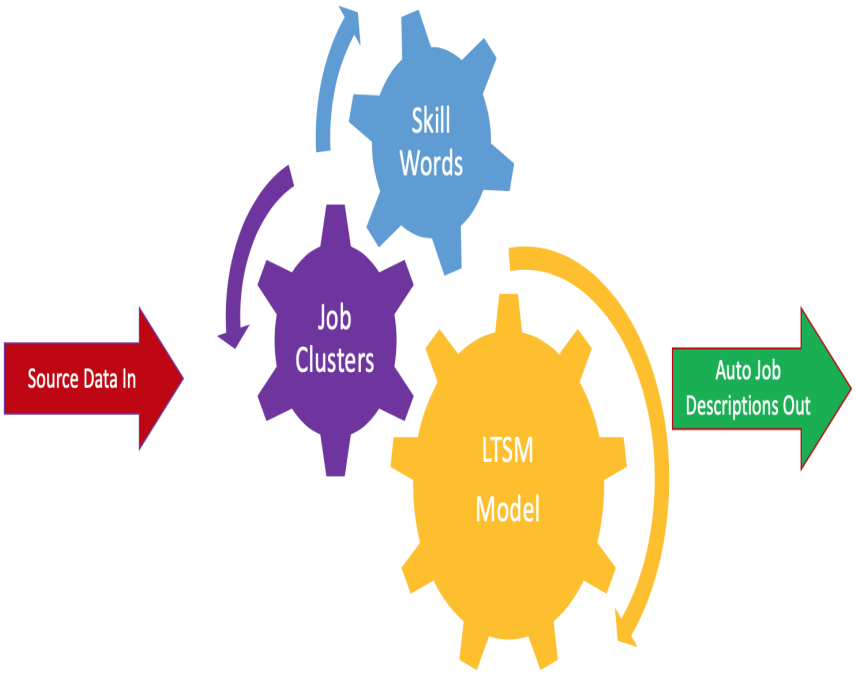
Lastly, due to the close proximity of these roles, we believe RightHires can provide the unique advantage of guiding companies to better tailor job postings to attract the right employees and avoid mis-labeled, outdated, and ill targeted hiring efforts.

Right Hires!

Research – Automatically Generating Job Postings

During this project, initial explorations (Proof of Concept, POC) were conducted for creating a solution to help companies in generating job postings. While still in its infancy, we investigated the use of creating a text generator to create job postings that could be used to assist companies with drafting job postings.

A large amount of jobs requiring different skills are lumped under the same titles and it can be difficult for companies to set themselves apart. However, our proposal is (as a future RightHires enhancement) to use a text generator to artificially create job postings related to the skills required for a job.



We set out to design a text generator to learn job requirements, experience, technical and non-technical skills that are currently required for the role from other companies. To further set the company apart, we would further train the model on job descriptions and refine the model to provide more targeted insights based on current industry trends.

We initially created a Long Short-Term Memory (LSTM) model and trained it on our job postings using the LSTM structure below. We leveraged base code provided by Jason Brownlee, a data scientist practitioner. We wanted the model to learn the relationships between these embeddings and to predict output words. We used an input sequence of fifty words to make this prediction, but found the model performed rather poorly. To work better the model would require an immense amount of training, network depth and design as well as additional computing resources.

Figure 32: LSTM structure

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
embedding_1 (Embedding)      (None, 50, 50)            681000
_________________________________________________________________
lstm_1 (LSTM)                (None, 50, 100)           60400
_________________________________________________________________
lstm_2 (LSTM)                (None, 100)               80400
_________________________________________________________________
dense_1 (Dense)              (None, 100)               10100
_________________________________________________________________
dense_2 (Dense)              (None, 13620)             1375620
=================================================================
Total params: 2,207,520
Trainable params: 2,207,520
Non-trainable params: 0
```

We then decided to use pre-trained models that could later be modified to fit our purposes. These models represent some of the most state-of-the-art NLP algorithms and initial findings were promising. While we also experimented with XLNet and GPT, GPT-2 seemed to generate the most relevant information. All of these models are unsupervised models called text generators that have been trained on Gigabytes of curated online data. GPT-2 performed so well, the full model wasn't released due to concerns about abuse and "deep fakes". For demonstration purposes, we pulled one of our data science job postings and extracted different sections for the model to generate output as shown below.

Figure 33: Input Data Example 1

Summary

Data analysis, modeling, and forecasting is at the heart of our business and service. This senior position requires a strong applied statistics and mathematics background, with experience in managing a small team. Knowledge and experience in math, statistics, neural networks, and machine learning are needed in order to analyze company metrics, in order to develop high-quality models.

Figure 34: Generated Text Example 1

We are also applying some statistical techniques to our application, analyzing data from international broadcast media and our internal network or geographic networks. All of our data and data analysis will be analytical and preferably mathematical and data integration will not take over all the time we manage our company.

Figure 35: Input Data Example 2

The Data Scientist will be responsible for end-to-end analytic projects including:

- **The understanding of business and data needs**
- **Discovering, cleaning, and transforming data as needed**
- **Designing and building analytical models**
- **Prototyping**
- **Performing statistical analyses**
- **Providing diagnostic, descriptive, prescriptive, and predictive analytics**
- **Determine opportunities and needs around the use of machine and deep learning for help in prescriptive and predictive analytics, automation, and model training**

Figure 36: Generated Text Example 2

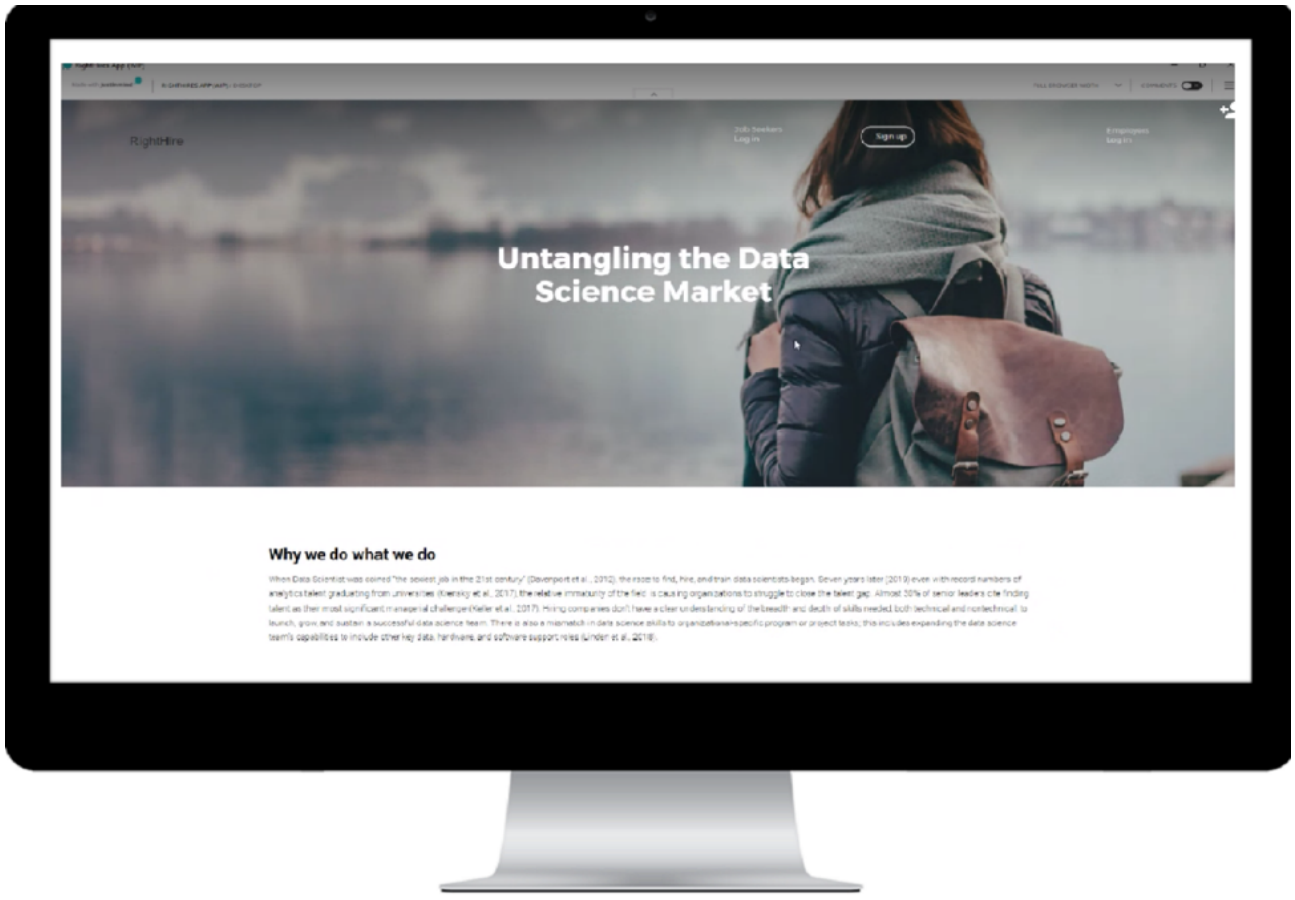
- **Understanding the role of cross-services and data science in business**
- **Operating logistics, software and services areas**
- **Evaluating scalability in business**
- **Learning from operating systems**
- **Development of effective product and service controls**
- **Automation of data under normal and special circumstances**
- **Enterprise environments**
- **Extending and utilizing the term "stability," the study enables developers and new data scientists to focus on scalable engineering solutions and capture technological growth**

In theory, it will be possible to create useful synthetic data that can help our customers develop focused job descriptions that reflect current industry trends and attract the right talent. This potential feature can be developed further with additional investment in our proposed Phase 2.

The User Experience - Dashboards

The insights driven through our modeling will be packaged into an interactive web and mobile application experience for two sets of audiences, the people that are looking to gain employment into the Data Science field and HR, hiring managers, executives, or VCs tasked to build or grow a data science team or practice. RightHires plans initially to focus the web experience towards the corporate audience and a mobile application for people looking for a job in the Data Science field but will expand the offerings of the web and mobile application in future releases to allow overlap across those audiences.

Figure 37: Web Experience Splash Screen (user logon will determine what they can see)

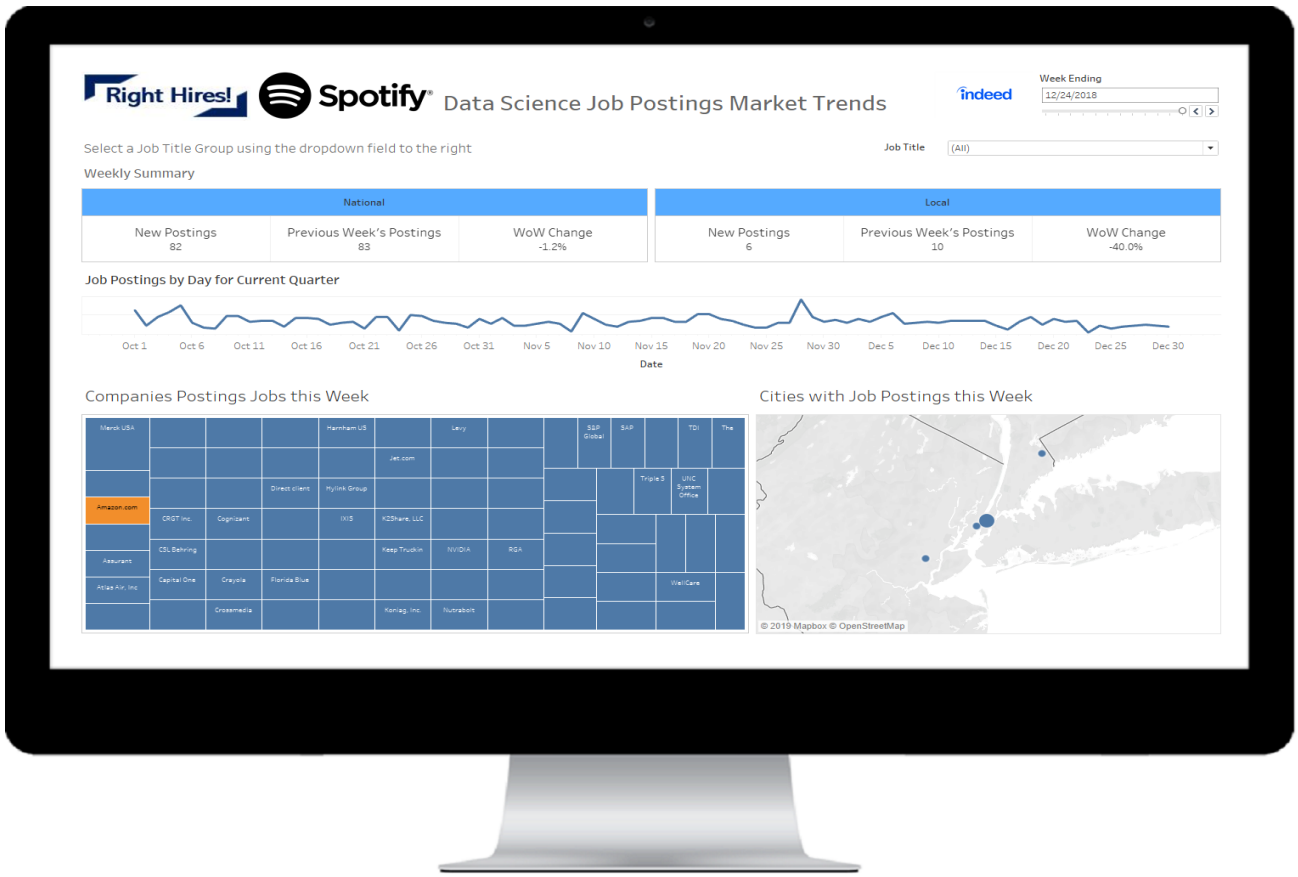


The web experience consists of dashboards that with help assist Human Resource departments, contract placement companies and headhunters to

use the correct descriptions, requirements and titles within their position postings so they attract the right type of candidates and advertise within the right media platforms. These dashboards will also help this audience understand trends in the market, opportunity size, geographic characteristics of postings, identification of competition for candidates, and time-based change in the market over the last year.

The sections below show a sample of the features within the web experience tailored for a particular company; in this case we show how an HR resource at Spotify may experience the website.

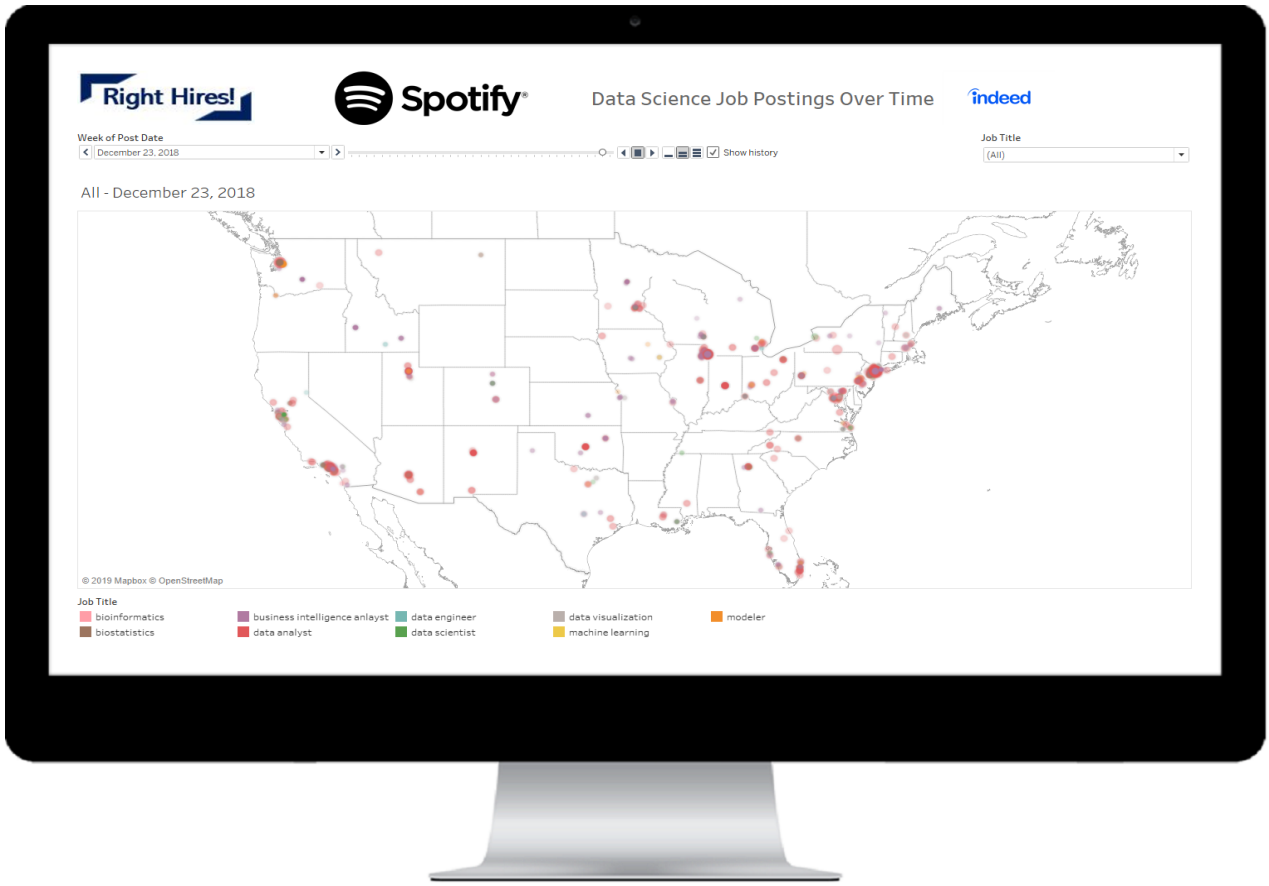
Figure 38: Data Science Job Posting Trends Dashboard



Like the mobile application, the RightHires dashboard will feature a branded splash (Figure 37) screen and provide the user with a place to logon as

shown below. Because RightHires will sell multiple levels of subscriptions, the role associated with each user will determine what they can see. The user will then have access to a menu of actions that reflect different levels of insights – a section for descriptive data science market trends, a section where the user can upload a job description and understand what type of candidates it is likely to attract, and finally a section with resources and examples of how to build out or grow a data science team within their particular industry and location.

Figure 39: Location of Data Science Jobs Across the U.S

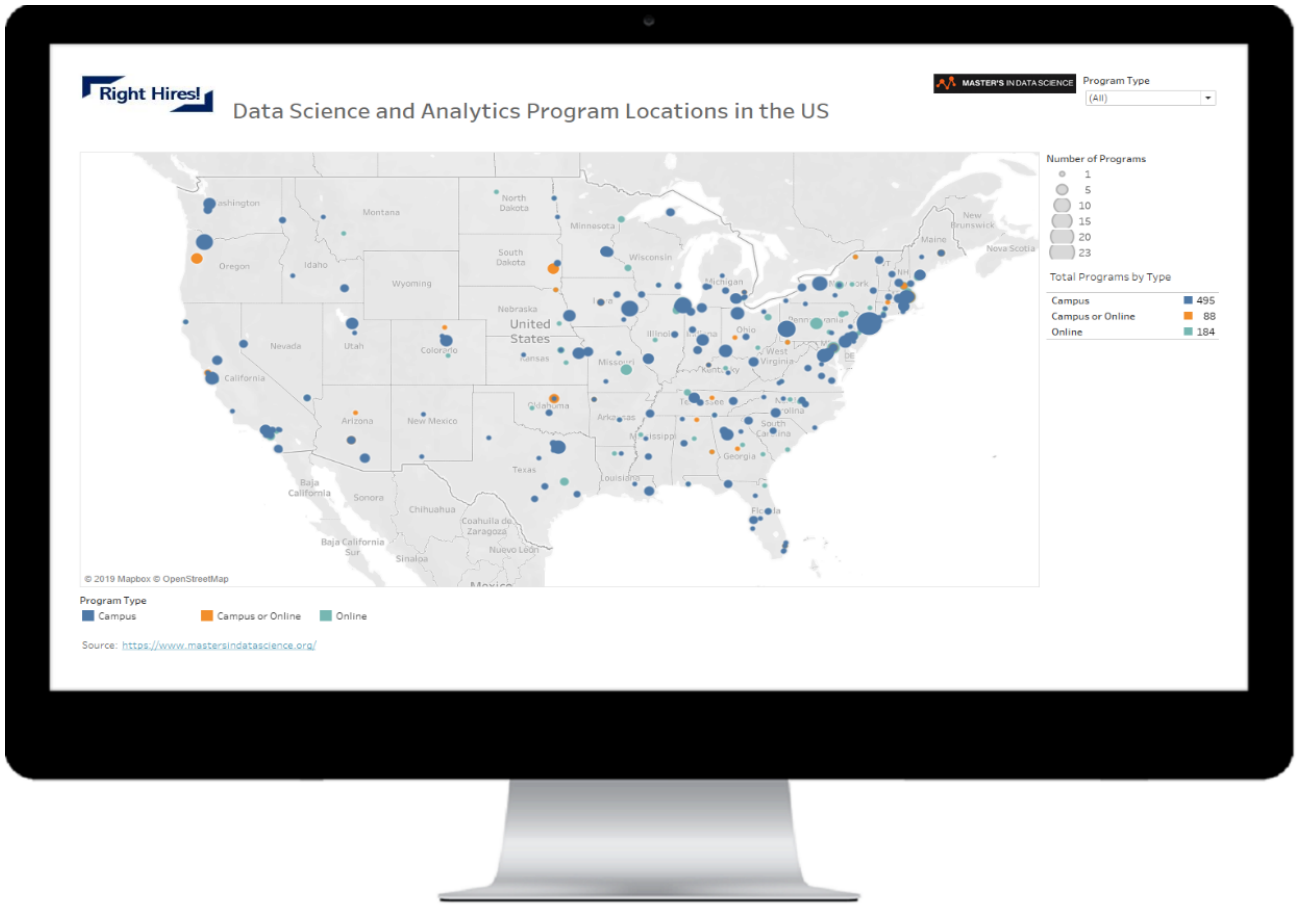


In the market trends section (Figure 38), the HR resource has access to an interactive dashboard where they can see the number of new job postings in the data science field nationally and locally, understand how the market is

increasing or decreasing over time, understand which companies (and which of their competitors) are ramping up their data science teams, and finally see the jobs that are being posted closest to their headquarters. There is a date toggle field at the top that allows the user to go back to any week in the past year for further investigation.

Additionally, the user will have access to a dashboard that shows the number of job postings of a job title by geography over time. As the job titles change over time, the user can understand what new job titles are being posted in Data Science market hot beds such as NYC and Seattle and get ahead of the curve to attract the best talent in their location.

Figure 40: Location of Data Science Programs Across the U.S



The next section of dashboards leverages the model output and help HR resources put together the best job description to attract talent. The 'Analyze My Posting' section allows the user to upload a job description and analyzes the terms used to determine what type of candidate within the modeling clusters the job posting may attract and provide suggestions on what to change if desired.

The 'Help Me Find Similar Skills' allows the user to input the skills or technologies used by a particular job posting to see what other terms they could be adding to attract more talent. HR representatives do not always know the technologies they are being asked to recruit for but having a tool that can expand the number of job candidates with similar skills can help fill that gap. For example, if a Data Science executive let an HR representative know they use R within their department, the HR rep can use the tool to understand that candidates skilled in SAS are highly linked to candidates skilled in R and may help find bridge a potential talent gap.

Finally, the 'Help Me Build My Team' provides resources for the HR representative to take a proactive approach to building a larger pipeline to attract future talent by connecting with universities in the area that offer data science programs and understand the skills and technologies being taught in the programs. Figure 40 shows an interactive dashboard map view of the Data Science programs currently available in the US broken out by traditional on campus and online methods. The HR rep has the ability to drill down into their location and click into the colleges and the names of the program, along with skills and technologies used for each class, a contact person for that college, job fair and other recruiting event information, and a best practice document with a list items to help establish a long term relationship with that college.

During the CEO presentation, the team will take the audience through this workflow to share what the user interaction looks like.

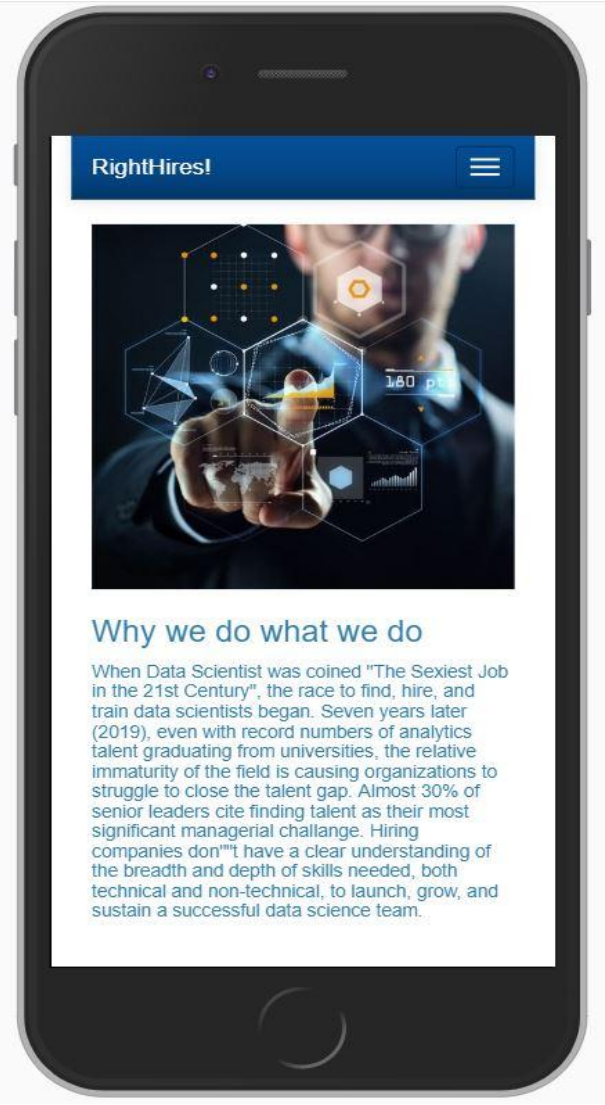
The User Experience - Mobile Application

RightHires understands that mobile users are individuals looking for jobs and not corporations working on talent placements. Thus, our mobile experience will be focused on individuals looking to perform tasks and queries quickly using a familiar smartphone mobile interface.

Our mobile strategy is to help individuals looking for a job to use the correct terms on their resumes, find out where the market and data science leaders are headed, finding the correct titles for jobs that match their skills, and submit a resume.

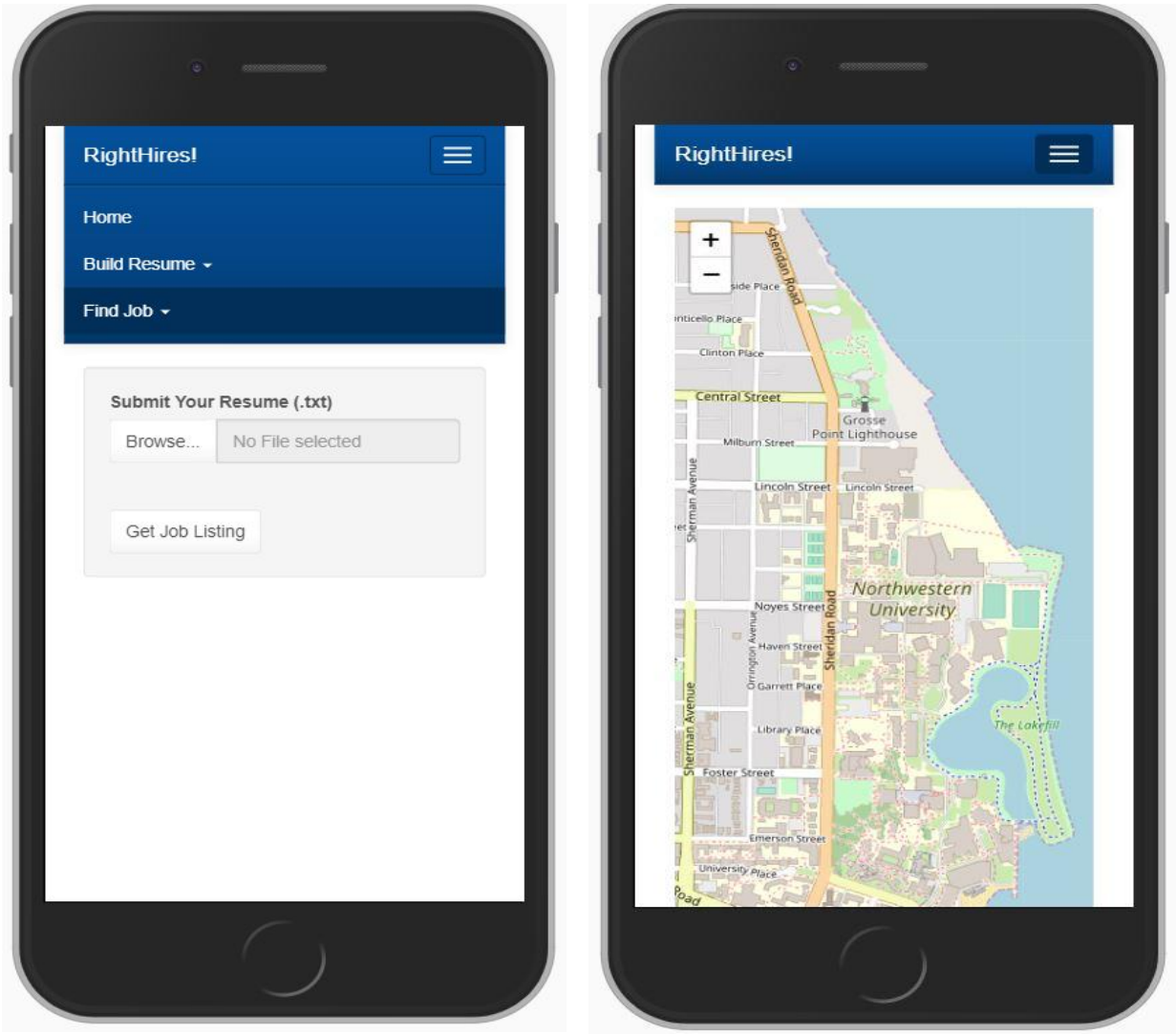
The application will also help candidates find the regions that have the highest number of roles that appropriately match the skillsets they have or are building (in the case of students).

The screen shot to the right shows the RightHires mobile home screen. Like the dashboards, it will be branded. Functions within the mobile application are available by clicking on the “hamburger” (three horizontal line) icon at the upper right-hand side of the screen.



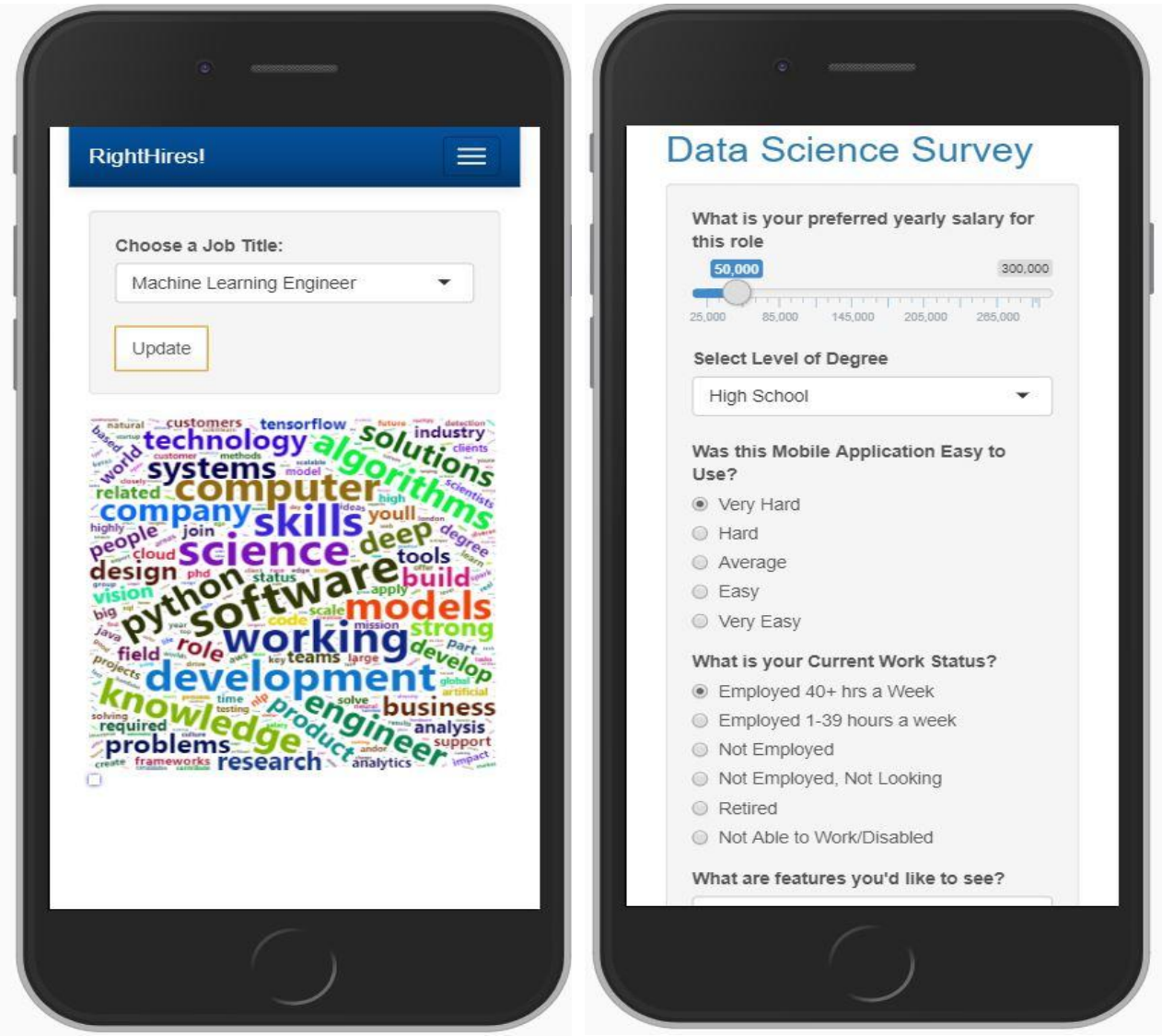
The RightHires mobile application will not only leverage data from the model/clustering components of the platform but also use other external sources for added value for our customers. For example, the Twitter API is being leveraged to gather and display trending words up to the minute on Data Science. This feature will help candidates understand what's most talked about in Data Science, by industry leaders, give them insight into skills trending up they should add to their experience (like AI), and help prepare candidates for an upcoming interview by knowing what foundation and leading-edge skills they may be asked about.

Figure 41: Submit Resume and Find Data Science Programs (Northwestern)



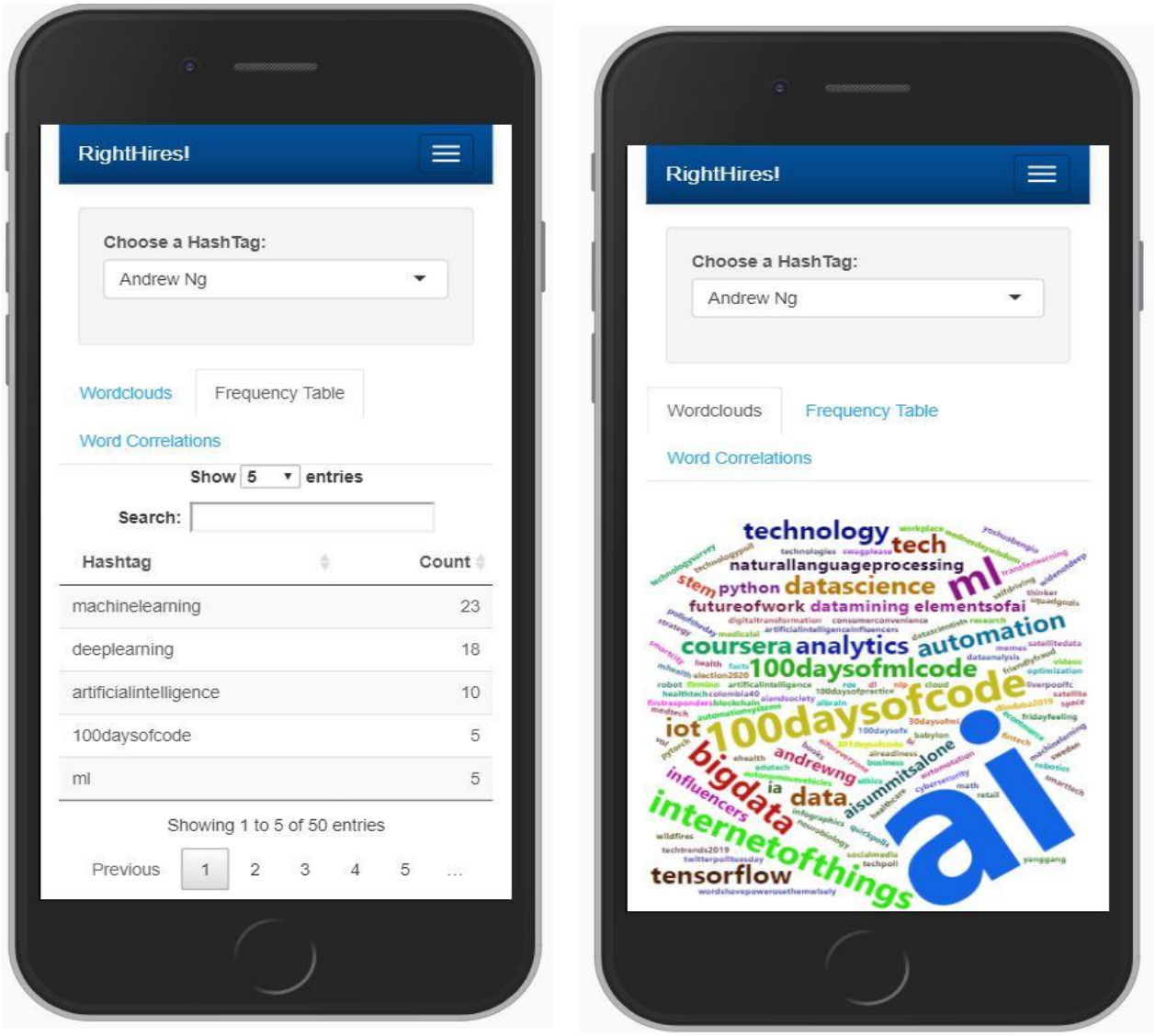
The mobile application will also be a source channel for the “voice of the customer”. This channel will be leveraged by the RightHires Marketing and Product Management Teams. Because the RightHires business model is based on monthly and yearly subscriptions, it’s important to offer more value over time to keep renewal revenue coming in.

Figure 42: Word Map by Job Title (left), Surveys (right)



Several ways the mobile application will help facilitate the customer feedback channel is to periodically collect customer opinions about ideas for new features RightHires is considering, preview those features on a free trial basis, and collect satisfaction surveys. The data for these will be stored in a central location at RightHires and mined by various groups in the company to help with planning future releases and monitor trends in customer satisfaction.

Figure 43: Leading Hash Tags by Data Science Leaders (left), Word Cloud by Data Science Leaders (right)



The mobile application will leverage many technologies including R, Shiny, cloud database, filtering including stop words that will help provide the most important material on a small smartphone screen, and interfaces to Twitter and Indeed (for trending terms and up to date job postings).

During the CEO presentation the mobile application will be demonstrated to share its workflow and functionality.

Scalability

The Data Science job market is not the only field where job titles and descriptions cause confusion for candidates and employers, and the proprietary modeling methodology and expertise gained by the team in this project allow for vast expansion into other markets. In order to achieve this however, RightHires! plans to move its data processing and analytics operations to the cloud.

The team has done extensive research into the cloud platforms available in the market that includes Amazon Web Services, Microsoft Azure, and Google Cloud Platform (GCP), and has decided to use Google Cloud Platform as the future state solution.

Google Cloud Platform is a suite of cloud computing services and management tools offered by Google. This includes compute services, networking services, cloud AI services, big data services, and storage services, all with a highly active and knowledgeable community and the reliability and security one associates with the Google brand.

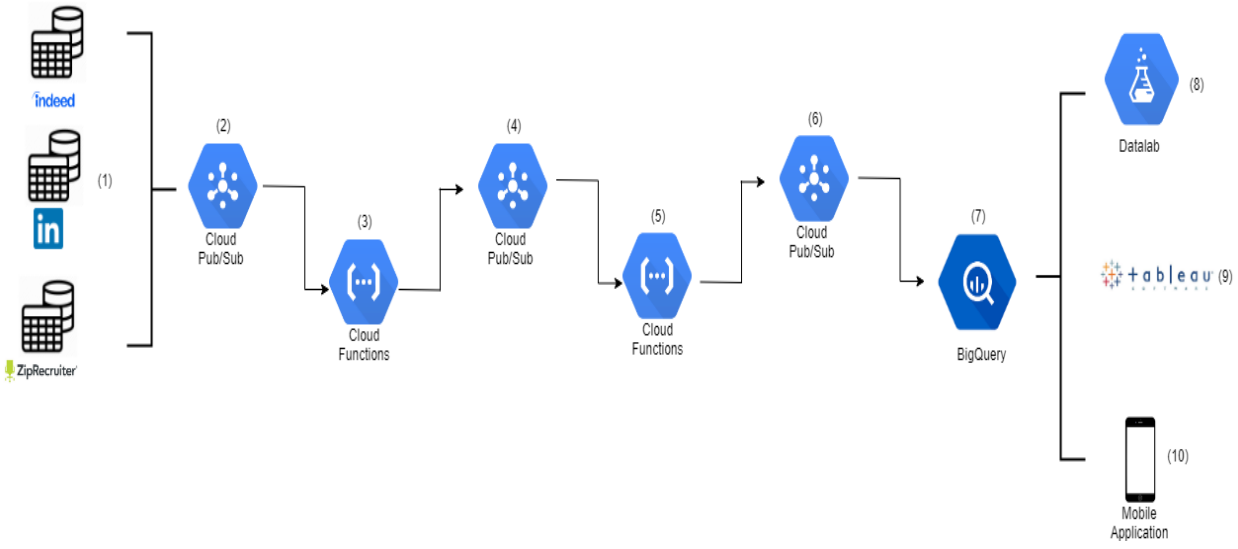
The team expects substantial benefits of going to the cloud in areas of data processing, data storage, and computational power for analytics, especially for high CPU requirements models in natural language processing. One of the main attributes of GCP is that there is zero upfront cost to get started, and the platform will scale in sync with our data and analytics needs so we only pay for what we need and when we need it.

The remainder of this section outlines the specific modules within Google Cloud Platform that the team plans to implement in the future to meet the scalability needs of the company.

The solution to be built needs to be an automated process that (1) ingests and scrapes job posting data, (2) cleans, aggregates, and stores the data in a central location, (3) has the computational power to run sophisticated models, (4) has the ability to connect model output to reporting tools and a mobile application.

With GCP, RightHires can achieve this using the module services of BigQuery (main analytics engine), DataLab (interactive notebook style analytics), Cloud Functions (event-driven serverless compute platform), and Cloud Pub/Sub (event management service). The figure below shows how this works.

Figure 44: RightHires GCP Architecture



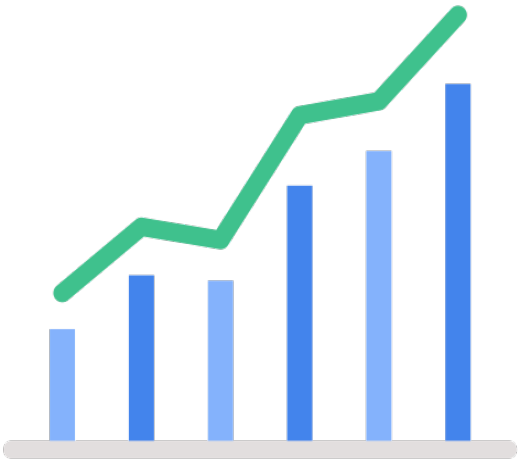
In the above example, we've partnered with major job posting sites Indeed, LinkedIn, and ZipRecruiter (and can add others) to allow us to connect to their data via API, however we have a mechanism to scrape data off their sites as well.

The planned process to support this is outlined in the table below.

Table 3: Using GCP Process Overview

Step	Description
1	Indeed, LinkedIn, and ZipRecruiter send job posting data in JSON format directly to a Google Cloud Pub/Sub queue through a client embedded in the application.
2	Pub/Sub opens an incoming data connection that occurs at the same time or times each day via the Cloud Scheduler module.
3	The Pub/Sub also kicks off the Cloud Function module which contains the Python code that scrapes data from job posting sites.
4 & 5	Again using Pub/Sub and Cloud Scheduler, it kicks off another Cloud Function module which contains the Python code that cleans, transforms, and aggregates the job posting JSON data.
6	Processed data is written back to another Pub/Sub queue that will be loaded to the BigQuery module.
7	BigQuery is the analytics engine that stores the processed data and productionalized NLP models.
8	DataLab is an interactive notebook environment, similar to Jupyter Notebooks, where the analytics team can explore the data and collaborate before finalizing the model.
9	Tableau and other reporting tools can access the data directly through BigQuery for near real-time dashboards.
10	The mobile application can also tap into the model data and model output directly through BigQuery for automatic updates.

The above process highlights the main benefits of going to the cloud such as an automated end to end data pipeline, multiple analytics engines (one for data/model exploration, and the other for models in production), as well as the ability to connect various reporting and dashboard tools in near real-time (the only barrier to being real time is waiting for partners to send us data, or the time it takes to scrape data from the sites).



The virtual machine that processes each of the modules can easily be scaled up and down in terms of memory and CPU required. As the data keeps growing in size and the number of models grows with opening new lines of business, we do not have to worry about our technology stack keeping up with the demand.

Right Hires!

Conclusions and Recommendations

Conclusions

The following conclusions are the result of completing the design-build portion of the project. This 10-week project has put in place a platform that can be expanded further to obtain more value and customers in the future.

Table 4: Project Findings Observations and Learnings

Track	Observations and Learnings
Data Sourcing (Bought, screen scraping, other) And… EDA and Data Preparation (Transformation, cleaning, filtering, etc.)	Data is well understood, and the cleaning strategy developed worked as needed. However, there were some issues with the bought job posting data from the Indeed site. For the project, the team worked around it to complete the schedule. Going forward the team will either negotiate a data quality SLA with the current vendor, find a new vendor, or build a data gathering mechanism on its own or by leveraging tools. Team has identified other sources of data that will be used in addition to our modeling efforts. These extra sources will add more value to the dashboard and mobile application(e.g. salary, other job and educational sources).
Modeling	For topic modeling data cleaning has a large impact on locating relevant job postings. There are also a lot of common industry acronyms that can be transformed to help the analysis. We found that the choice of applying lemmatization and/or stemming depends on the application. In our case we found stemming to work best. Additionally, we learned that TF-IDF worked very well for transforming the text and seemed to work best for topic modeling. However, Word2Vec was able to be used in conjunction with topic modeling to give further insights in the relatedness of important words (where Doc2Vec was not usable). Finally, we found that using several model and clustering methods was an effective way to verify results and get additional gains.

Table 4: Project Findings Observations and Learnings (continued)

Track	Observations and Learnings
Clustering	Evaluating the most common job titles including terms like data science type jobs allowed us to locate industry standards for labeling such jobs. We were able to limit the scope to the most common titles which could be compared against the clusters found by our algorithm. We also expanded the number of titles after our initial explorations to prove the base strategy is sound and can be expanded in the future to cover more types and jobs and industries.
Word2Vec	Experimentation for this unplanned bonus feature uncovered potential value by using this method to find alternative job description key words. In comparison of the POC results, combining Word2Vec with other topic modeling models was found to provide better results and was leveraged in the final portion of this project. It also worked as a stand-alone model for returning related skills.
Dashboard and Mobile Application	Verified the architectures and technologies or both user interfaces are solid, performance of early implementations is within reasonable bounds, and we have an alternative for when they are not. We leave this phase knowing the high level hierarchical functional structure of both interfaces and have built them out to deliver the first phase. The team did observe benefits of offering some similar functionality across both interfaces (even though we initially thought of keeping them separate). Therefore the team will experiment with methods to do that where feasible in a Phase 1 pilot.
Infrastructure	During the project the team scoped the number of targeted jobs or broke the whole problem into parts to run most of the time on high end laptops with occasional runs on Google's cloud environment. As RightHires expands, it will be necessary to run on a cloud environment with flexible capacity to handle the load.

Right Hires!

Table 4: Project Findings Observations and Learnings (continued)

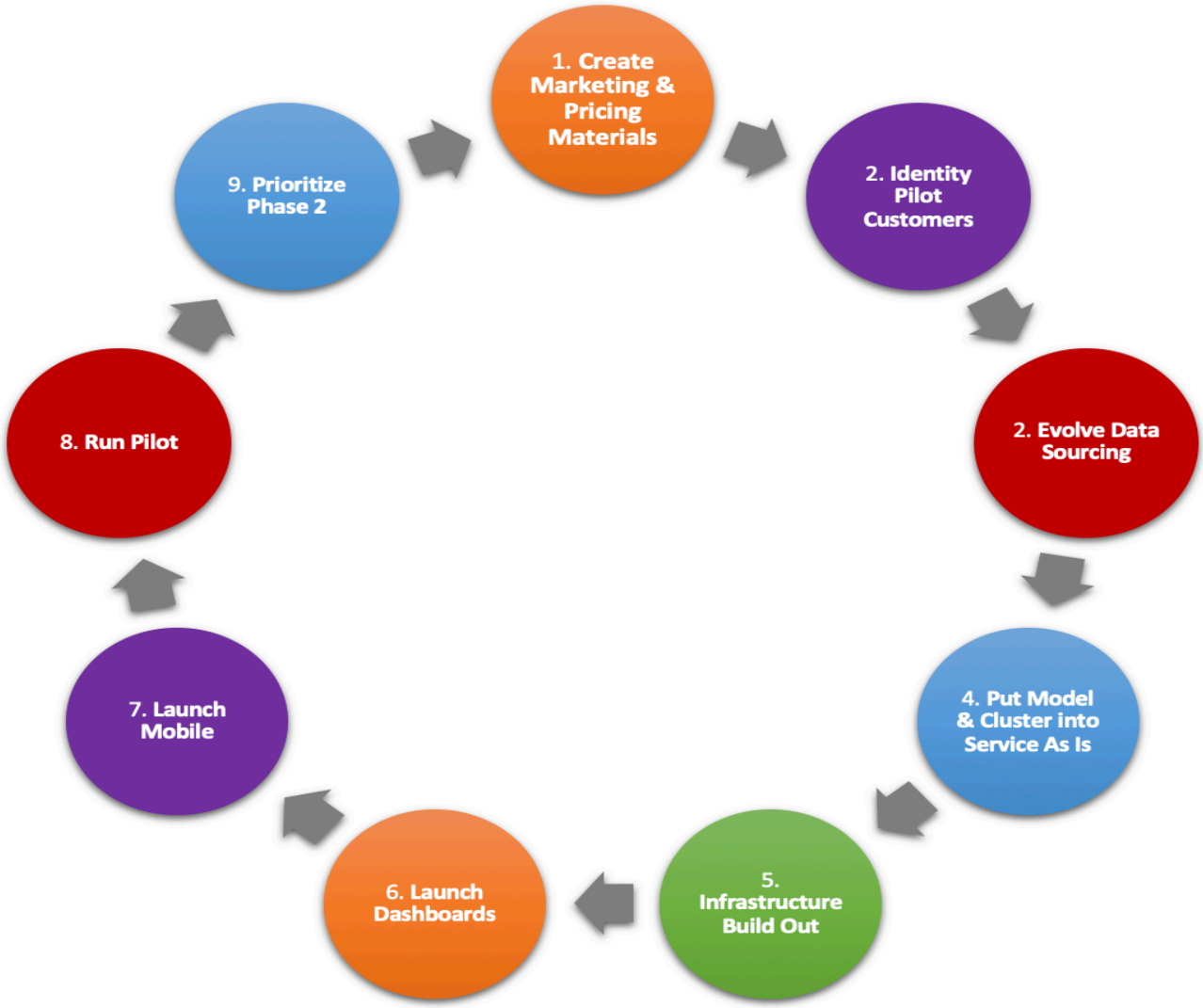
Track	Observations and Learnings
Expansion	During the 10 week project the team POC'ed several future feature ideas and identified other extensions that would add value (see "Future Investments" section) to our offering. The platformed as designed will provide a solid base to build these additional features on in the future.
Partnerships	We've confirmed RightHires has a unique offering. But because we're in startup mode, funding to cover the whole lifetime span of the hire and retain process today is impossible. Two competitors (also in startup) may be good partnerships to allow all 3 to serve a wider audience. Fallback can be built into the partnerships by interfacing in ways that allows any partner to leave.

Right Hires!

Recommendations – Phase 1

During this project the team has proven out strategies around business need, technology selection for modeling and clustering, targeted jobs to launch with, data sourcing options, multi-user interfaces and initial functionality through those interfaces, and infrastructure required to support demand. As a result, the team makes the following recommendations to the CEO for taking RightHires into the market (Phase 1) and then extending our offering (Phase 2). Figure 45 illustrates the sequence of steps for Phase 1. The next section provides a list of potential Phase 2 recommendations.

Figure 45: Phase 1 Recommendation Sequence



Right Hires!

Description of Phase 1 Recommendation Steps

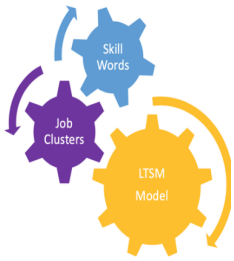
1. **Create Marketing/Pricing Materials** – Leveraging findings from the problem statement, environment stats, goals, findings work done to date.
2. **Identity Pilot Customers** – for the targeted spaces, namely HR, Headhunters, VC firms, job candidates, and students.
3. **Evolve Data Sourcing** – Move data acquisition in house by partnering directly with job posting websites to obtain API access or invest more in our data scraping capabilities for use across additional job posting websites.
4. **Put Model/Cluster into Service As Is** – The tested design that leveraged data cleaning steps, TF-IDF, Word2Vec, T-SNE and K-Means has worked well for the jobs targeted. Put those in production for the pilot.
5. **Infrastructure Build Out** – Implement the cloud Infrastructure as specified to handle the workload of running the model and clustering on the latest data every week.
6. **Launch Dashboards** – Target dashboards built to date for our business customers while collecting feedback over the pilot to provide insight into required enhancements before General Availability (GA).
7. **Launch Mobile** – Target mobile applications built to date for all pilot customers. This is being done to test the original theory that mobile is preferred by non-business users. Feedback collected will be used to tweak the offering (and user population targeted) for GA.
8. **Run Pilot** – Run pilot, collect feedback through UIs and user groups, use the feedback to determine next steps before GA.
9. **Prioritize Phase 2** – Over all the above steps, continually gather data to help prioritize and enhance the list of Phase 2 recommendations. Use that insight to determine the sequence and content of Phase 2 potential solutions.

Recommendations – Phase 2

During the 10-week project the team has identified, and in some cases experimented, with proofs of concepts (POC) for ways to expand the value of the RightHires platform and business.

Each potential solution below will add value and require additional investment. As a next step the executive team should prioritize these ideas based on investment needed, value to our customers, and enhancement the uniqueness of our offering.

Table 5: Platform Expansion Opportunities

Potential Solutions	Target Segment	Benefits	Methods
Automatically generate job postings using XLNet 	Talent Management	Assist companies with building job postings based on industry standards. With Doc2Vec, generate similar skills & experience to attract more talent.	Develop GPT-2 and XLNet. Add ontological insight to models using WordNet, Word2Vec and topic modeling to provide relevant word 'categories'.

Right Hires!

Table 5: Platform Expansion Opportunities (continued)

Potential Solutions	Target Segment	Benefits	Methods
Voice Mining and Analysis 	Talent Management and Students	Expand platform functionality and make it more natural to interact with using voice. Use for customer support, submit voice resume, collect ideas and customer feedback, automated pre-screen interviews, interview practice, grade pre-interview tests, support audio FAQs.	Leverage a collection of automated techniques to analyze audio in real time to add features to RightHires
Add Salary 	Talent Management and Students	Would help provide guidance to all in the latest salary ranges by job and region.	Identify sources for salary ranges, bring that data in, and combine with the results of our modeling, Make available through our two UIs. Offer a way for users to submit their current ranges, jobs, and regions to add a crowdsource capability.
Expand educational data, functionality, and partnerships 	Talent Management and Students	Provide students with more info through our UIs about a wider range of on-line, under grad, contest, and grad opportunities to add to skill set. Provide classes and program by level, topic, region, cost, goals.	Scrape, API to more education sources. Partner with course sources, do revenue sharing to sign up new students.

Right Hires!

Table 5: Platform Expansion Opportunities (continued)

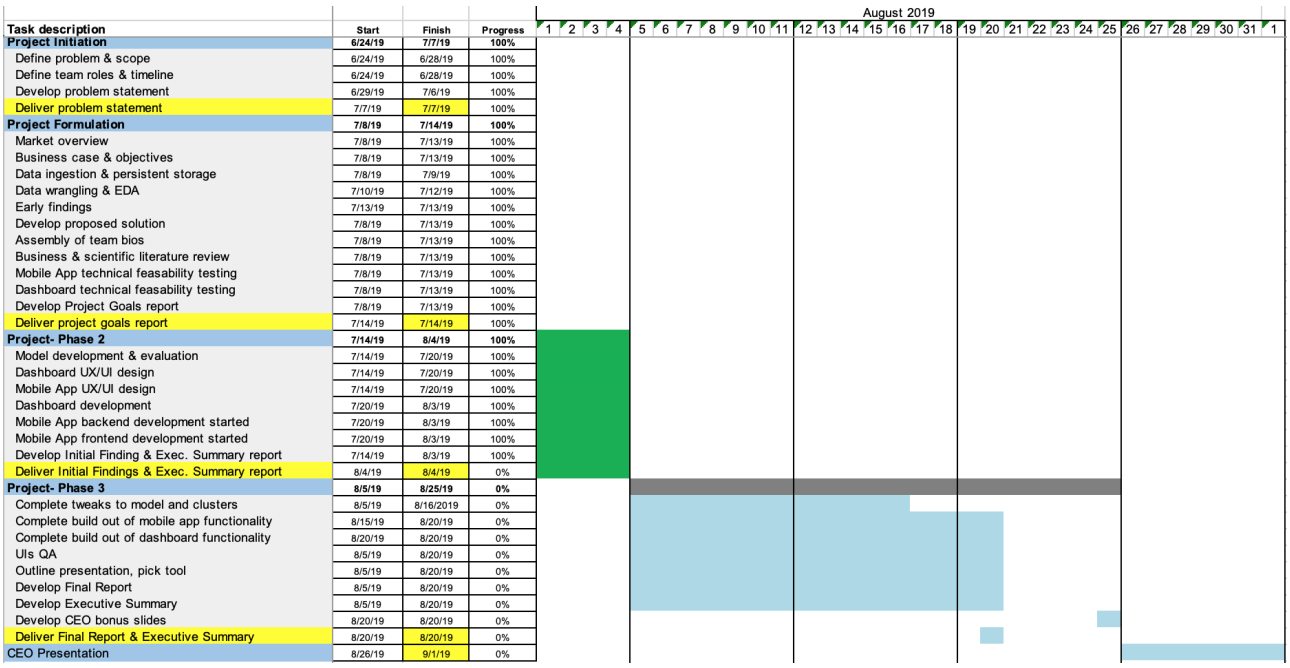
Potential Solutions	Target Segment	Benefits	Methods
Expand functionality through Partnerships 	Talent Management	Provide more value to our customers by covering more of the lifetime the recruit, hire, retain process	As outlined in the Competitors section, RightHires identified 2 other startups that complement our offering. Forming a partnership can provide each with the ability to offer more without all the upfront investment.
Expand Job Listing sources in US and other countries 	Talent Management and Students	More coverage of jobs, industries, job titles, a countries leveraging the base platform built.	Scrape, acquire data, interface to APIs from Glassdoor, LinkedIn, ZipRecruiter, Monster, Kaggle, GitHub. Then connect to non-US job sites in a 2nd Phase,
Cross Web and Mobile Functionality 	All	Provide user of both our user interfaces with access to common functionality	Use componentize architectures to make it easy to connect common functions to both UIs while centralizing the code.

Appendix

Project Plan and Next Steps

A high-level project plan is shown in Figure 46 for the Team 52 Capstone project. Project tasks are grouped in support of major deliverables (milestones) due on July 7, July 14, August 4, August 20, and September 1, 2019. Currently the project is on time, and the team will begin Phase 4 (CEO presentation) of the project on August 21 as planned. All Phase 1 and 2 deliverables (June and July) have been completed on time and not shown in the timeline below.

Figure 46: RightHires Project Plan and Current Status



Task description	Start	Finish	Progress
Project Initiation	**6/24/19**	**7/7/19**	**100%**
Define problem & scope	6/24/19	6/28/19	100%
Define team roles & timeline	6/24/19	6/28/19	100%
Develop problem statement	6/29/19	7/6/19	100%
Deliver problem statement	7/7/19	7/7/19	100%
Project Formulation	**7/8/19**	**7/14/19**	**100%**
Market overview	7/8/19	7/13/19	100%
Business case & objectives	7/8/19	7/13/19	100%
Data ingestion & persistent storage	7/8/19	7/9/19	100%
Data wrangling & EDA	7/10/19	7/12/19	100%
Early findings	7/13/19	7/13/19	100%
Develop proposed solution	7/8/19	7/13/19	100%
Assembly of team bios	7/8/19	7/13/19	100%
Business & scientific literature review	7/8/19	7/13/19	100%
Mobile App technical feasability testing	7/8/19	7/13/19	100%
Dashboard technical feasability testing	7/8/19	7/13/19	100%
Develop Project Goals report	7/8/19	7/13/19	100%
Deliver project goals report	7/14/19	7/14/19	100%
Project- Phase 2	**7/14/19**	**8/4/19**	**100%**
Model development & evaluation	7/14/19	7/20/19	100%
Dashboard UX/UI design	7/14/19	7/20/19	100%
Mobile App UX/UI design	7/14/19	7/20/19	100%
Dashboard development	7/20/19	8/3/19	100%
Mobile App backend development started	7/20/19	8/3/19	100%
Mobile App frontend development started	7/20/19	8/3/19	100%
Develop Initial Finding & Exec. Summary report	7/14/19	8/3/19	100%
Deliver Initial Findings & Exec. Summary report	8/4/19	8/4/19	0%
Project- Phase 3	**8/5/19**	**8/25/19**	**0%**
Complete tweaks to model and clusters	8/5/19	8/16/2019	0%
Complete build out of mobile app functionality	8/15/19	8/20/19	0%
Complete build out of dashboard functionality	8/20/19	8/20/19	0%
UIs QA	8/5/19	8/20/19	0%
Outline presentation, pick tool	8/5/19	8/20/19	0%
Develop Final Report	8/5/19	8/20/19	0%
Develop Executive Summary	8/5/19	8/20/19	0%
Develop CEO bonus slides	8/20/19	8/20/19	0%
Deliver Final Report & Executive Summary	8/20/19	8/20/19	0%
CEO Presentation	8/26/19	9/1/19	0%

Operationally, the project team formally meets three times a week using Google Hangouts in order to provide task updates to other team members, synchronize upcoming tasks, and to keep the project within the defined scope. This more closely aligns with industry-standard agile development methodology. The forum is also used to solve more complex problems

together and review major deliverables like the software or reports. In between, the team uses the chat function to provide quick status updates and ask questions. Project deliverables are stored daily on a Google team drive which facilitates easy sharing, collaboration, and safely stores deliverables on the cloud.

Project Next Steps

The last phase of the project involves building slides and a demo and presenting that to the CEO. The team will also store all project collateral in a place where it can be obtained by others in the future.

Table 6: Remaining Work

Track	Activities to Complete by Week 10
File Share	Set up and populate a public file share so others can obtain the code, reports, and user interfaces in the future.
Oral Presentation to the CEO	Layout presentation outline. Identify collateral to leverage or create. Pick technology to record presentation and publish recorded presentation.

Right Hires!

Project Team

A highly-qualified team with interdisciplinary skills will deliver the project using an interactive, collaborative, Agile approach. Primary/secondary area leads are indicated below. In true Agile form, team members share work beyond primary assignments to keep the team on schedule.

Table 7: RightHires Project Team and Roles

Roles P: Primary S: Secondary C: Co Contribute T: Tertiary * all participating	Julia Barnhart 	Brennen Chadburn 	Jonathan McKim 	David Pilkington 	Mike Ryder 
Project Manager	P				S
Developer/Models/ EDA		P	S	S	
Validation/Testing		S	P	S	T
Dashboards			P	S	
Mobile App		T	S	P	
Writer/Research	S				P
Marketing/Graphics				C	C
Oral Presentation/Prep	*	*	*	*	*

Julia Barnhart has several years' experience leading technical and analytics projects, including guiding technical architecture and data pipelines. She brings with her the ability to translate business problems into technical solutions. Her interests include NLP and Deep Learning

Brennen Chadburn works at Willis Towers Watson and was recently moved to the Global Data Services and the Innovation team developing AI applications including conducting job matching through Natural Language processing and automation through machine learning and expert systems. He'll be providing analytics support and building topic modeling and classification models to assist with improving company surveys and developing new products.

Jon McKim is a solution architect within the Data & Analytics practice at Slalom Consulting. He has 10+ years of experience in business intelligence and data strategy. He is skilled in aligning analytics with business strategy and believes in measurable, actionable and repeatable solutions to deliver value to a customer. He has experience implementing solutions in data warehousing, data visualization, advanced analytics, data engineering and strategy.

David Pilkington acts as the projects Mobile Application/Dash-board Developer and Programming validation and testing duties. In his free time, he will assist in the presentation and media department. He works in leadership in his day job as a General Manager of a software division focused on Cloud (SaaS) Products.

Mike Ryder has an MS in Medical informatics, MBA, and MS in Computer Engineering. He works in healthcare leading a group that does technology research, starting an innovation center, and move technology into early adoption. Prior background includes software engineer, consultant, program/product manager, and strategist in a wide variety of markets. Data Science interests are to leverage AI, IoT, and machine learning for better patient care.

References

ATD - Association for Talent Development (2019). Bridging the Talent Gap. Retrieved from https://d22bbllmj4tvv8.cloudfront.net/83/74/450e8cb644188b984d6528d43d58/2018-skills-gap-whitepaper-final-web.pdf

Brownlee, J. (2019). Develop Deep Learning for Natural Language Processing in Python. Machine Learning Mastering Series.

Dataquest (2019). Working with Large Data Sets using Pandas and JSON in Python. Retrieved from https://www.dataquest.io/blog/python-json-tutorial/

Davenport, T., Patil, D. (Oct 2012). Data Scientist: The Sexiest Job of the 21st Century. *Harvard* Business Review. Retrieved from https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century

Keller, S., Meaney, M. (Nov 2017). Attracting and Retaining the Right Talent. Mckinsey & Company. Retrieved from https://www.mckinsey.com/business-functions/organization/our-insights/attracting-and-retaining-the-right-talent

Koehrsen, W. (2018). Neural Network Embeddings Explained. Retrieved from https://towardsdatascience.com/neural-network-embeddings-explained-4d028e6f0526

Krensky, P., Linden, A. (Oct 2018). Data Science and Machine Learning Solutions: Buy, Build or Outsource? Gartner Research. Retrieved from https://www.gartner.com/en/documents/3892470/data-science-and-machine-learning-solutions-buy-build-or

Krensky, P., Vashisth, S., and Laney, D. (Oct 2017). Leading Upskilling Initiatives in Data Science and Machine Learning. Gartner Research. Retrieved from https://www.gartner.com/en/documents/3816863/leading-upskilling-initiatives-in-data-science-and-machi

Linden, A., Idoine, C., Hare, J., Brethenoux, E. (Aug 2018). Staffing Data Science Teams: Map Capabilities to Key Roles. Gartner Research. Retrieved from https://www.gartner.com/en/documents/3888468/staffing-data-science-teams-map-capabilities-to-key-role

Mizoguchi, R. (2004). Part 3: Advanced course of ontological engineering. New Generation Computing, 22(2), 193-220. Retrieved from https://link-springer-com

Pathak, M. (Sep 2018). Introduction to T-SNE. Retrieved from https://www.datacamp.com/community/tutorials/introduction-t-sne