



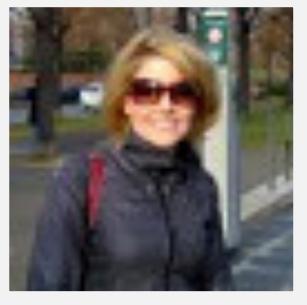
RIGHT HIRES !

Talent Forecasting

The Right People, The Right Skills, The Right Jobs

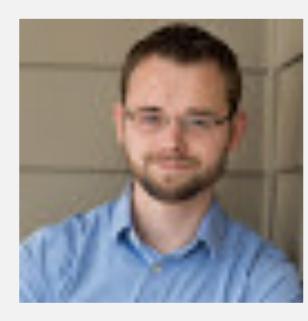
Team

Members have a diverse set of data science, software engineering, consulting, leadership, and creative skills. Team roles filled included:



**Julia
Barnhart**

Business Research, Word2Vec/T-SNE,
Strategy, Test, Presentation



**Brennen
Chadburn**

Modeling/NLP, Clustering, EDA, Data Prep,
Strategy, Test, Presentation



**Jonathan
McKim**

Dashboards, EDA, Data Prep, Scalability,
Strategy, Test, Presentation



**David
Pilkington**

Mobile App, EDA, Marketing, Graphics,
Strategy, Test, Presentation



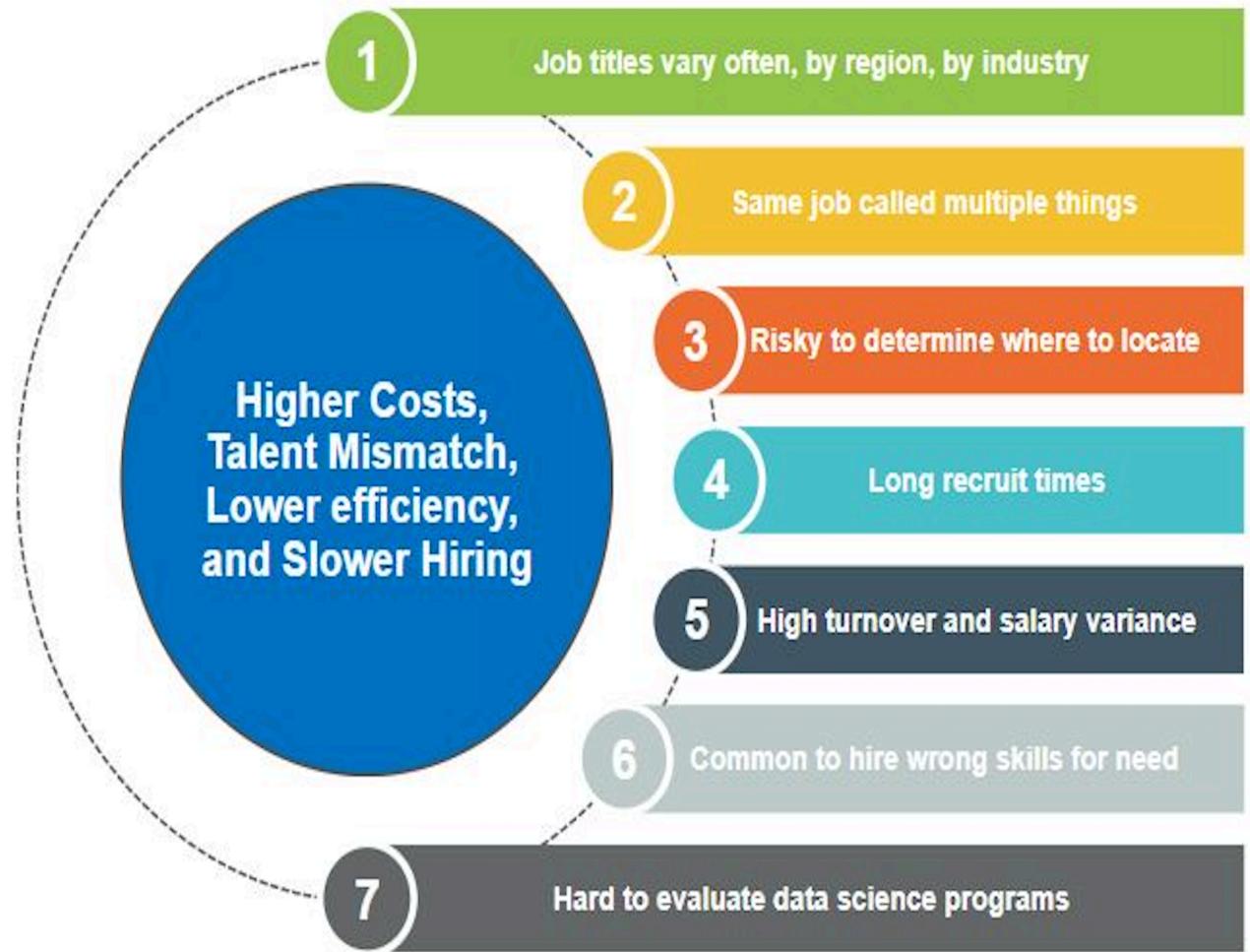
**Michael
Ryder**

Writer, PM, Graphics, Video Edit,
Strategy, Test, Presentation

The Problem

Inefficiencies, shortages, constant change, many gaps

- 30-50% of senior leaders cite finding talent as their most significant managerial challenge.
- 83% of companies report skill gaps in those they hire for Data Science jobs. Gaps affect project success.
- Getting talent doesn't guarantee success. Mismatch of new hires to actual needs is a risk.
- Quickly evolving field requires candidates to drink from "fire hose" to meet new challenges.
- No real consistency across companies, industries, regions, titles, and educational programs.
- Data science alone isn't enough. Teams need others to source data, do hardware/software engineering, write requirements, deal with growing infrastructure needs, put models in production, and operationalize.



It's Complicated!

Many technologies that are always evolving



The Periodic Table of Data Science

An overview of key companies, resources and tools in data science (as of 4/12/2017)

Dc DataCamp	Ga General Assembly	Sd Strata Data
Sb SpringBoard	M Metis	Od ODSC
Ex Edx	Di Data Incubator	Tc Tableau Conference
C Coursera	In Insight	U UseR!
Uda Udacity	Dsa NYC Data Science Academy	Pd PyData
Ude Udemy	G Galvanize	Paw Predictive Analytics World
Ps Pluralsight	Dsg Data Science for Social Good	Kdd ACM SIGKDD Conference
Ly Lynda	Dsy Data Society	Tpc Teradata Partners Conference
Tt TeamTreeHouse	Dsj Data Science Dojo	Icd IEEE International Conference on Data Mining
Bdu Big Data University		

Courses
Boot camps
Conferences

Data		Search & Data Management
Projects & Challenges, Competitions		Machine Learning & Stats
Programming Languages & Distributions		Data Visualization & Reporting

	Collaboration		News, Newsletters & Blogs
	Community & Q&A		Podcasts



Py	Js	Vb	Pgs	Sli	Ah	W	Bml	Kn	Sm	Pb	Obi	Shn	Ddl	De
Python	JavaScript	Visual Basic	PostgreSQL	SQLite	Apache Hadoop	Weka	BigML	Knime	Spark MLLib	Power BI	Oracle BI	Shiny	Domino Data Lab	Data Science Experience
R	Cp	Sc	Ar	Bq	Hw	O	Dar	Lib	Ho	Bo	Alt	Mpl	Nt	Rs
R	C++	Scala	Amazon Redshift	Google BigQuery	Hortonworks	Oracle	DataRobot	LibSVM	H2O	BusinessObjects	Alteryx	Matplotlib	Nteract	Rstudio
S	Pl	Ca	Hb	Td	Cl	Mss	Rm	Mat	Th	Sp	Sav	Ply	Ro	Be
SQL	Perl	Cassandra	HBase	Teradata	Cloudera	Microsoft SQL server	RapidMiner	Mathematica	Theano	Spotfire	SAS Visual Analytics	Plotly	Rodeo	Beaker Notebook
B	Mr	P	Mdb	To	Aem	Spl	Cho	Mah	Aml	Ql	Po	Me	Spy	Ze
Bash	Microsoft R Open	Pig	Mongo DB	Toad	Amazon Elastic Mapreduce	Splunk	Chorus	Mahout	Azure Machine Learning	Qlikview	PowerPivot	Microsoft Excel	Spyder	Apache Zeppelin
Mtl	Cy	Im	K	Ms	Mar	Sr	Tf	St	D	Co	Gch	Pe	Dst	Ju
Matlab	Canopy	Impala	Kafka	MySQL	MapR	Solr	Tensorflow	Stata	D3	Cognos	Google Charts	Pentaho	Data Science Studio	Jupyter
J	An	Sp	Hi	Idb	Lu	El	Sk	Da	My	Aa	B	Db	Databricks notebook	Gh
Java	Anaconda	Spark	Hive	IBM DB2	Lucene	ElasticSearch	Scikit-Learn	Dato/Graphlab	Microstrategy	Adobe Analytics	Tableau	Bokeh		Github

Dw	Q	Fte	Sa	Gp	Dg	K
Data.world	Quandl	FiveThirtyEight	Socrata	Google Public	Data.gov	Kaggle
St	Uci UCI Machine Learning Repository	Wb World Bank	At Academic Torrents	Bf Buzzfeed	Dk DataKind	Dd DrivenData
Statista						

Re	So	Cv	Qu	Av	Dse
Reddit	Stack Overflow	Cross Validated	Quora	Analytics Vidhya	Data Science Stack Exchange
Mu	Rdm				
Meetup	RDatamining				

Kdn	Ibd
KDnuggets	insideBIGDATA
Rb	Pp
R-Bloggers	PlanetPython
Hn	Dt
HackerNews	DataTau
Dsc Data Science Central	Dsr Data Science Roundup
Dsw Data Science Weekly	Or O'Reilly
Dr Data Elixir	Pw Python Weekly
Rw R Weekly	Pd Partially Derivative
Bds Becoming a Data Scientist	Tm Talking Machines
Ds	Dsk
Data Stories	Data Skeptic
Ld Linear Digressions	Ns Not So Standard Deviations



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

What about Real Life?

Fit success dependent on many factors: technical, soft skills, statistics, SME, culture, JIT learning

MODERN DATA SCIENTIST

Data Scientist, the sexiest job of the 21th century, requires a mixture of multidisciplinary skills ranging from an intersection of mathematics, statistics, computer science, communication and business. Finding a data scientist is hard. Finding people who understand who a data scientist is, is equally hard. So here is a little cheat sheet on who the modern data scientist really is.



- MATH & STATISTICS**
 - ★ Machine learning
 - ★ Statistical modeling
 - ★ Experiment design
 - ★ Bayesian inference
 - ★ Supervised learning: decision trees, random forests, logistic regression
 - ★ Unsupervised learning: clustering, dimensionality reduction
 - ★ Optimization: gradient descent and variants
- PROGRAMMING & DATABASE**
 - ★ Computer science fundamentals
 - ★ Scripting language e.g. Python
 - ★ Statistical computing packages, e.g., R
 - ★ Databases: SQL and NoSQL
 - ★ Relational algebra
 - ★ Parallel databases and parallel query processing
 - ★ MapReduce concepts
 - ★ Hadoop and Hive/Pig
 - ★ Custom reducers
 - ★ Experience with xaaS like AWS
- DOMAIN KNOWLEDGE & SOFT SKILLS**
 - ★ Passionate about the business
 - ★ Curious about data
 - ★ Influence without authority
 - ★ Hacker mindset
 - ★ Problem solver
 - ★ Strategic, proactive, creative, innovative and collaborative
- COMMUNICATION & VISUALIZATION**
 - ★ Able to engage with senior management
 - ★ Story telling skills
 - ★ Translate data-driven insights into decisions and actions
 - ★ Visual art design
 - ★ R packages like ggplot or lattice
 - ★ Knowledge of any of visualization tools e.g. Flare, D3.js, Tableau

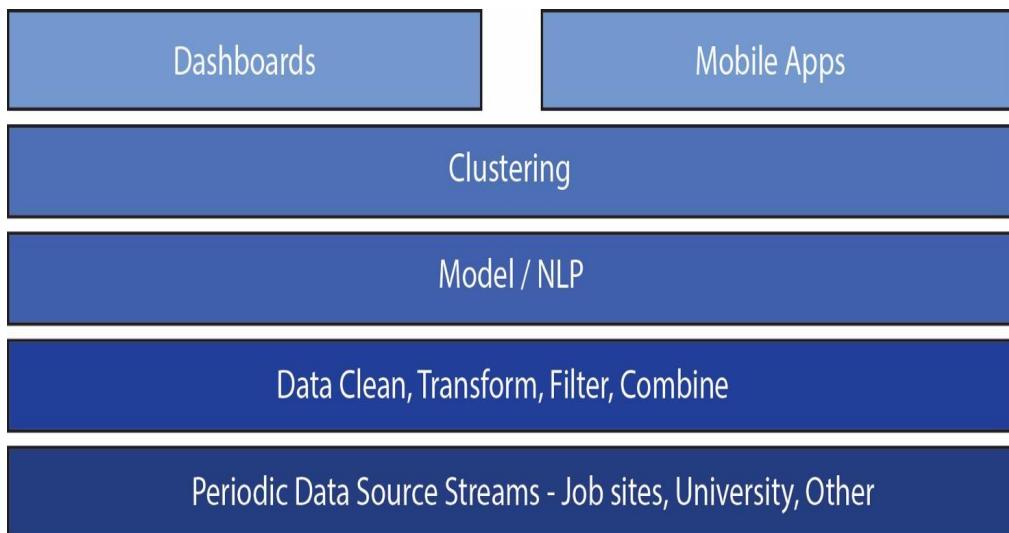
See how many of these challenges you can find on the video !



<https://www.youtube.com/watch?v=UV7f48GYwHY>

The Solution

A multi-layer solution available by subscription to all parties involved in the hire and application process for Data Science jobs.



Key Goals...

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.

Technologies Leveraged

Technologies fell into these categories:

- Data Sourcing – Scrape, API, buy
- Data Preparation – Python, filtering...
- Exploratory Data Analysis (EDA)
- Data Store – Survey, Prepared Data
- Natural Language Processing (TF-IDF, T-SNE, Word2Vec, Doc2Vec ...)
- Clustering (K-Means, LDA, others)
- Mobile (R, Shiny, R Studio)
- Dashboards - Tableau
- Infrastructure – Google Cloud, others
- Experimented with other technologies to supplement data prep, EDA, capacity

UI tools were levered early to aid in EDA.
Multiple tools helped determine “optimal” reached.

RStudio, R Shiny
for the Mobile Apps

Tableau for
dashboards and EDA.

Server-based database,
and the Twitter API for the
mobile application.

Functionality to handle and
specify stop words, and
multiple word to numeric
vector methods including
Doc2Vec and Word2Vec.

Clustering algorithms
including K-means,
Hierarchical, Non-negative
Matrix Factorization (NMF),
and Latent Dirichlet
Allocation (LDA).

Natural Language Processing,
Term Frequency Inverse
Document Frequency (TFIDF)

SAS and Alteryx for
EDA and clustering
experimentation.

Python 3.x , Anaconda and
Jupyter Notebook.
for general, EDA, data prep

Google Cloud environment
to provide processing power
to handle larger clusters



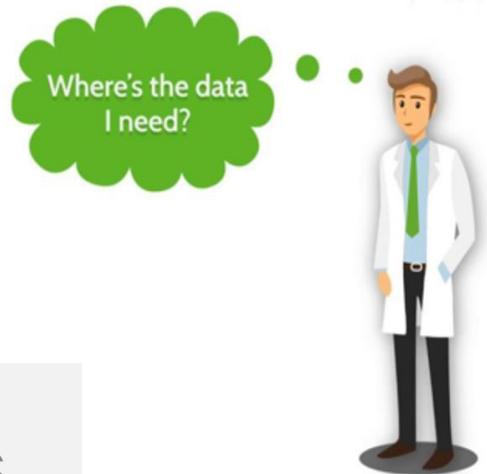
- **Right Hires** – Outward facing. Looks at job and education market to help formulate job titles, descriptions, locate clusters of jobs, talent, critical skills, and education for Data science positions. Subscription based covering 2 audiences (HR/Corp/VC/recruiters, candidates/students).
- **Aspiring Minds** – Supplements the interview and evaluation process with AI/online tools to help HR/Managers
- **People Insights** – Inward facing. Focused on looking for trends with employees, why they leave, who might leave, identifies factors to tweak to retain best talent.

Competition

Each competitor has a unique niche.

Partnership or future growth opportunities identified.

Data Sources



Jobs

2018 Indeed job Postings

Data bought from Data Stock

Other data sources via API or screen scrape



Educational

Masters of Data Science Programs in the US



Social

Twitter



Futures



Other education, conference, partner, community sources



EDA Summary

Traditional and UI tools used for EDA



3.125M 2018 postings acquired with 3% related to Data Science jobs
Top Industries: Finance, Healthcare, Retail, Manufacturing, Consulting, Gov.

Key Fields

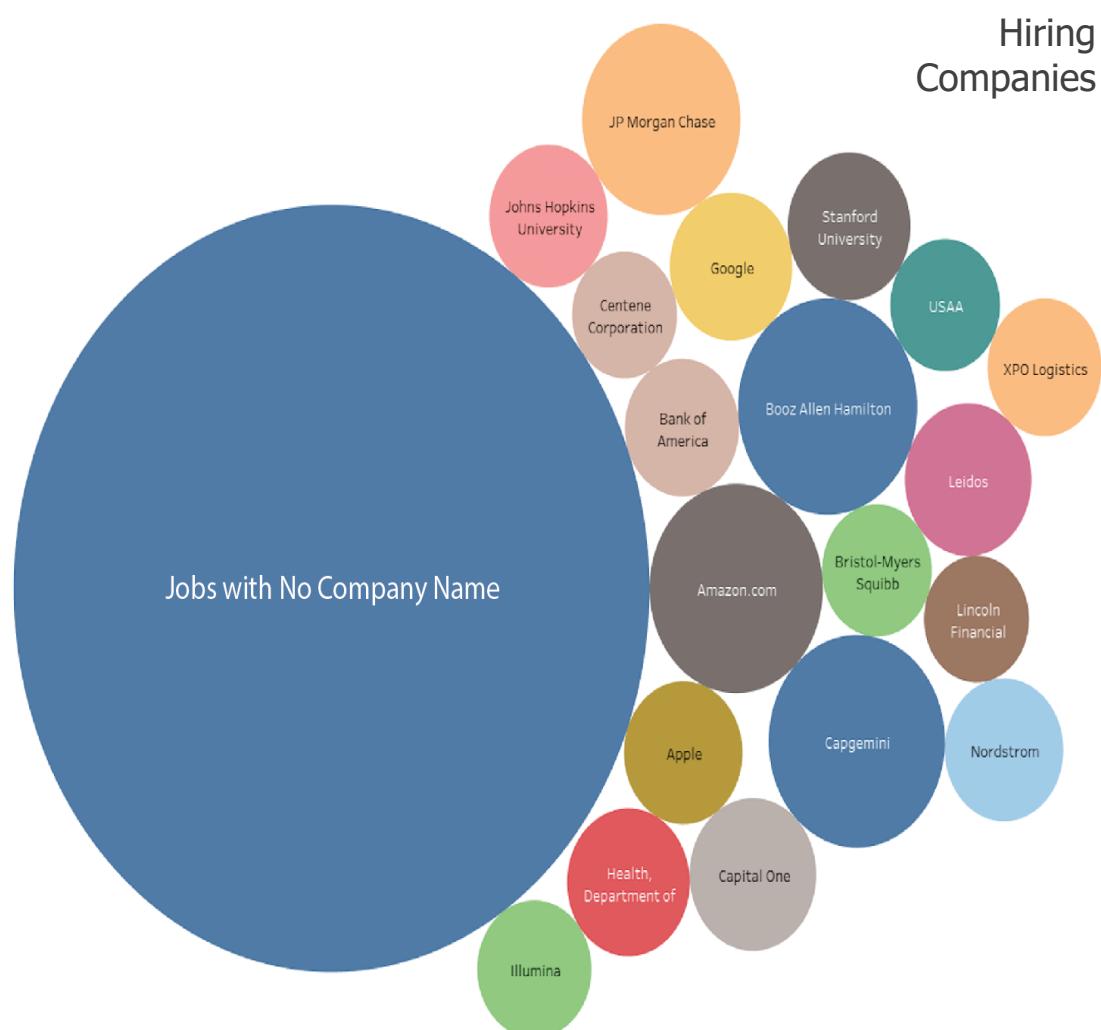
Job Title
Job Description
Company
Location
Posting Date
Salary (few)
Confidential,
Promotional, Fishing
Searches

Key Jobs

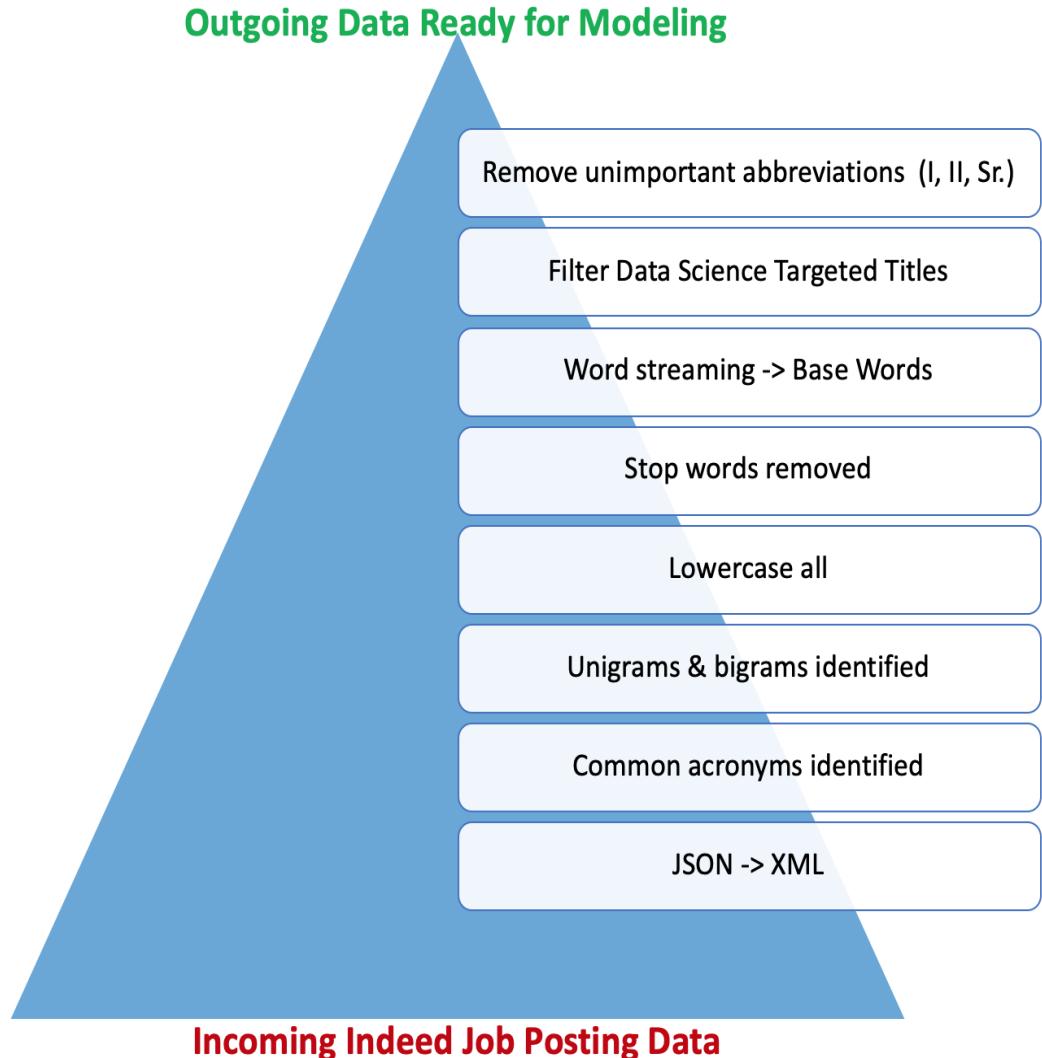
Data Scientist
BI Analyst,
Data Engineer,
Tableau Developer,
Software Engineer,
Statistician,
Bioinformatics,
Scientist,
SME Analyst

Key Skills/Words

Machine Learning,
Python, R,
Application Dev,
Statistics, SQL,
Analysis, BI, Cloud,
Tableau, Big Data,
Innovate, report,
Requirements, Make,
Build, dashboard,
Performance, Tools



Data Prep



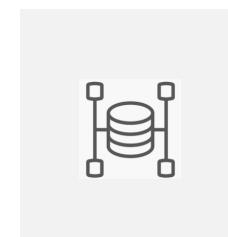
- NLP/Modeling/Clustering requires words to be represented as numeric vectors.
- Initial raw data provided as JSON formatted files containing job postings in English.
- Initial 3.12M jobs had ~35M words in their descriptions. Not all words are created equal!
- Most data work came in the form of filtering, preparation, transformation, removal of unimportant ("stop words") or repetitive or low value words.
- Did not need to clean, deal with NAs or bad values
- Combined similar/alternate job titles or description skills words to formulate clusters of the most important job titles and skills words for targeted Data Science jobs.



Modeling Methods

*Method for conducting analysis of job functions
and related skills*

Brennen



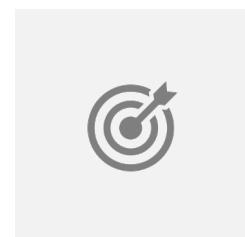
Pull data for
specific field

Extract well-known as
well as newly trending
job titles



Outline Job
Functions

Conduct Topic
Modeling



Pinpoint Skills &
Experience

Provide targeted skills
& experience by job
function



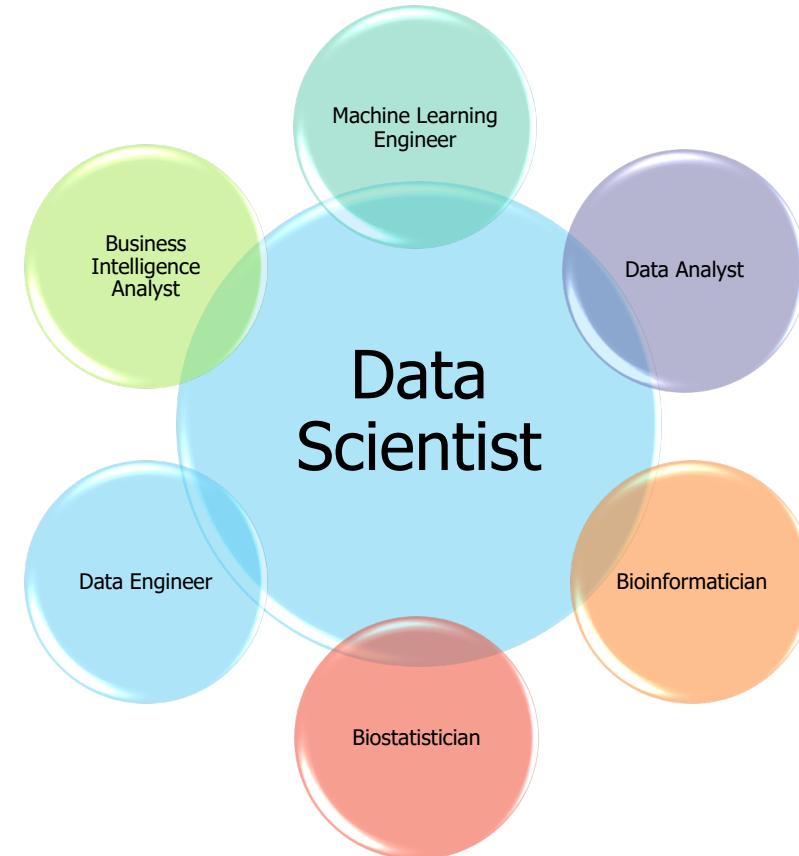
Identify Relevant Job Postings – Data Science

The Purpose

- RightHires keeps a pulse on the industry.
Understanding current trends and identifying where the market is going.

Method

- Evaluated the rapidly developing field of *data science* for demonstration purposes
- Start with popular titles (e.g. data scientist, data analyst)
- Used TF-IDF to identify relevant words to find related job postings
- Grouped related job titles together
- Kept most popular titles





Topic Modeling: Triangulate with Multiple Models

Strategy

- No one size fits all - Aggregate models
- Use multiple metrics to compare performance
- Evaluate results using subject matter experts
- in the field to provide further guidance

Methods Used

- TF-IDF to convert terms to word embeddings
- K-means clustering
- T-distributed stochastic neighbor (T-SNE)
- Non-negative Matrix Factorization (NMF)
- Latent Dirichlet Allocation (LDA)
- Silhouette Analysis





Stage 1: Topic Modeling

Cluster 0:

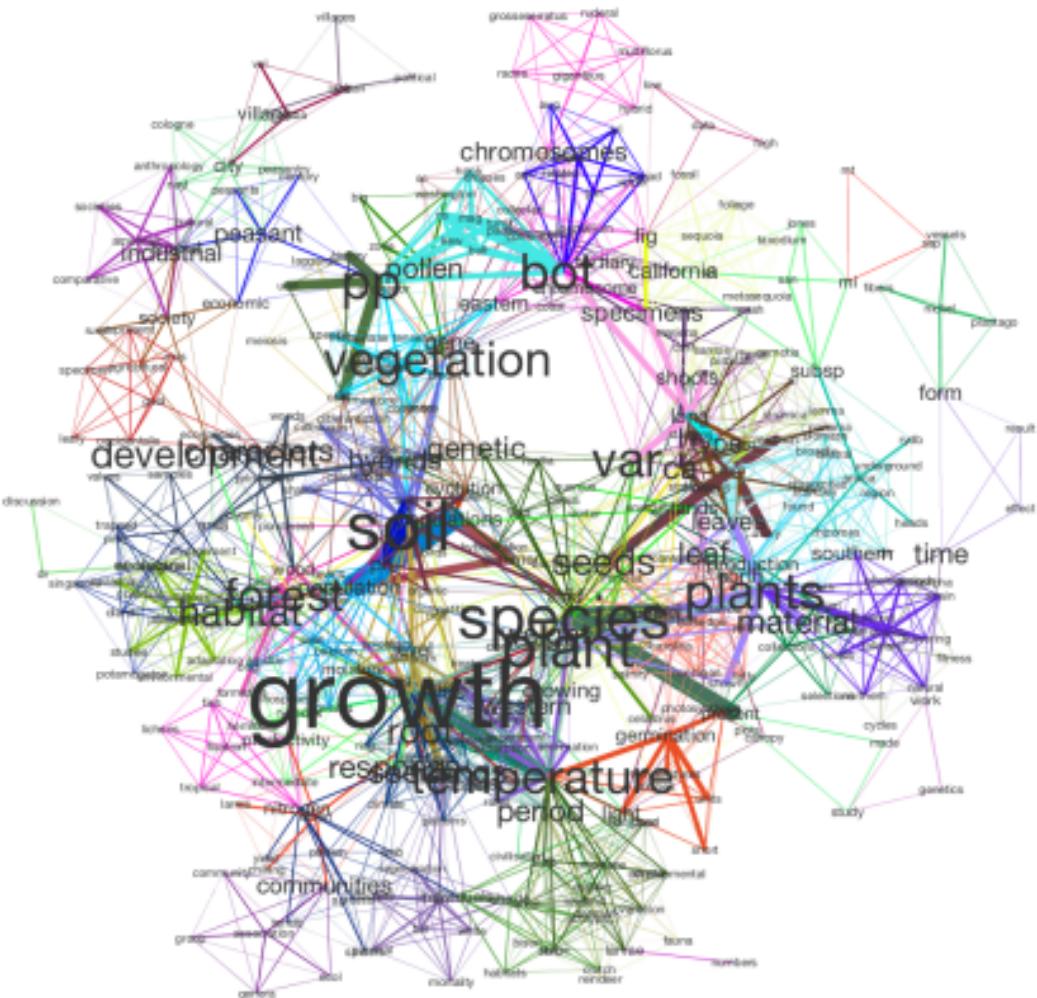
Data Science & Data Engineer

Cluster 1:

Healthcare (data science jobs in healthcare)

Cluster 2:

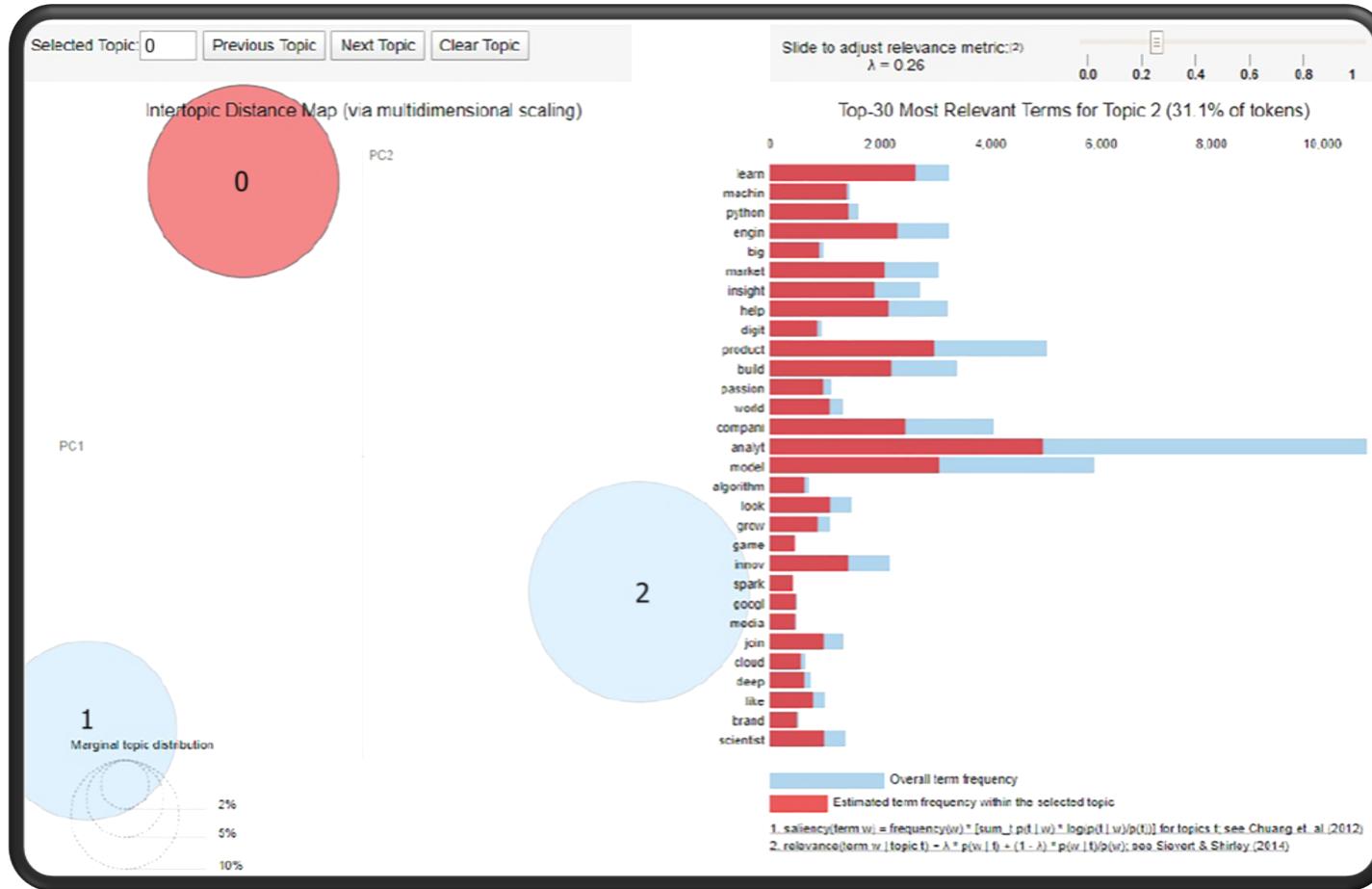
Business Intelligence (BI)





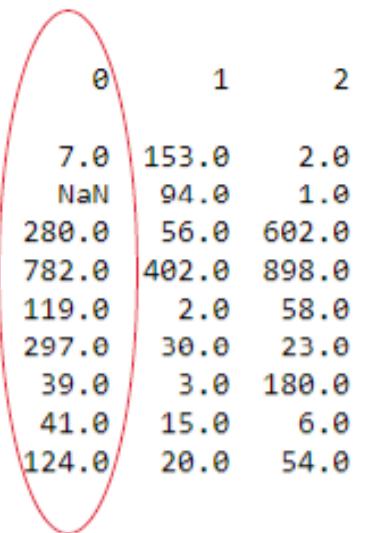
Cluster 0: Data Science Cluster

LDA clustering



K-means clustering

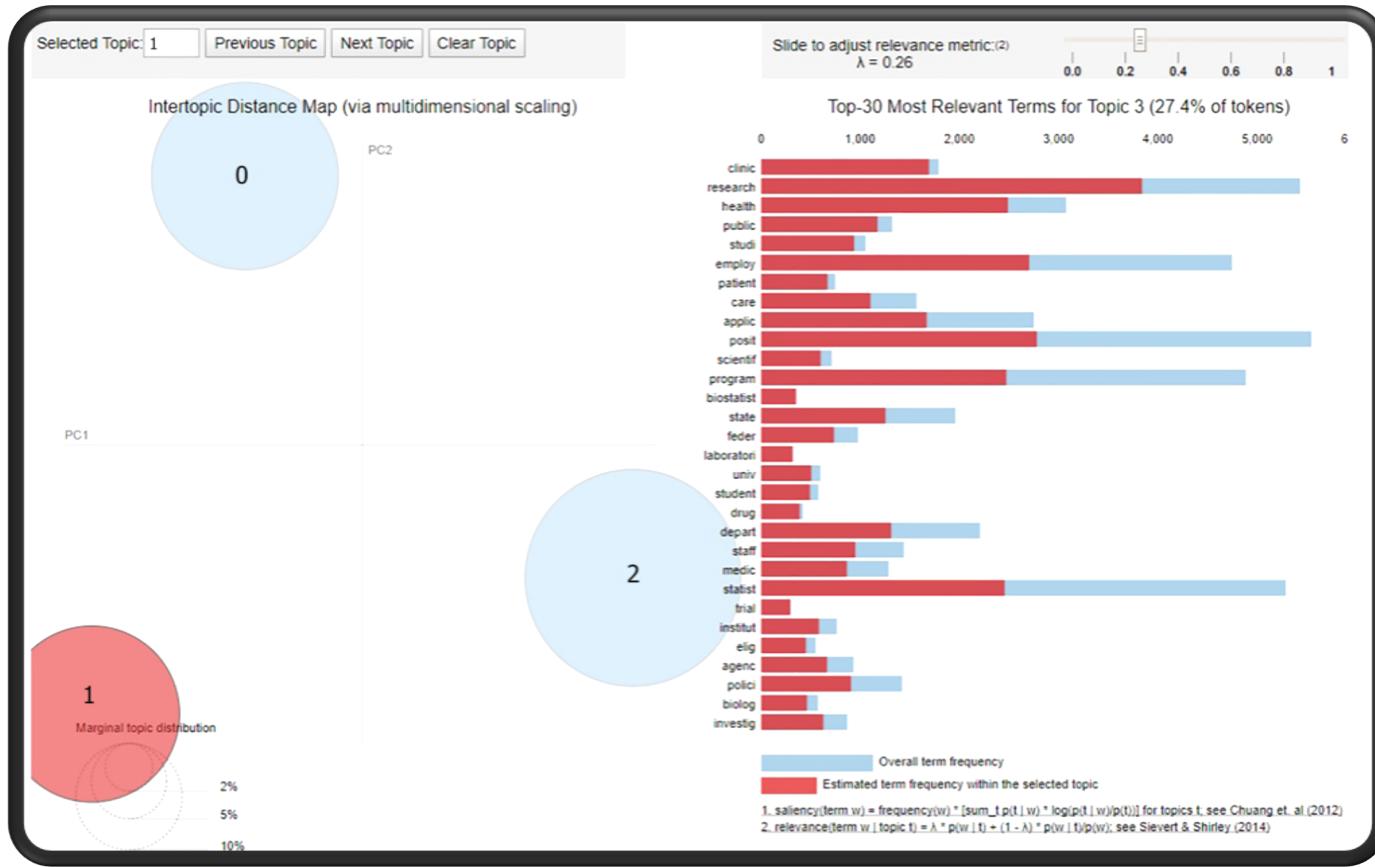
Cluster	job_title	bioinformatics	biostatistics
0	7.0	153.0	2.0
1	NaN	94.0	1.0
2	280.0	56.0	602.0
0	782.0	402.0	898.0
1	119.0	2.0	58.0
2	297.0	30.0	23.0
0	39.0	3.0	180.0
1	41.0	15.0	6.0
2	124.0	20.0	54.0





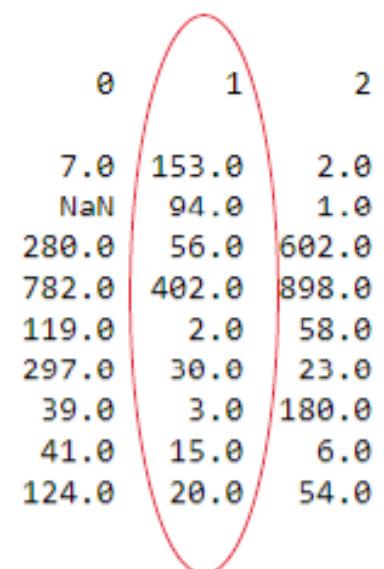
Cluster 1: Healthcare

LDA clustering



K-means clustering

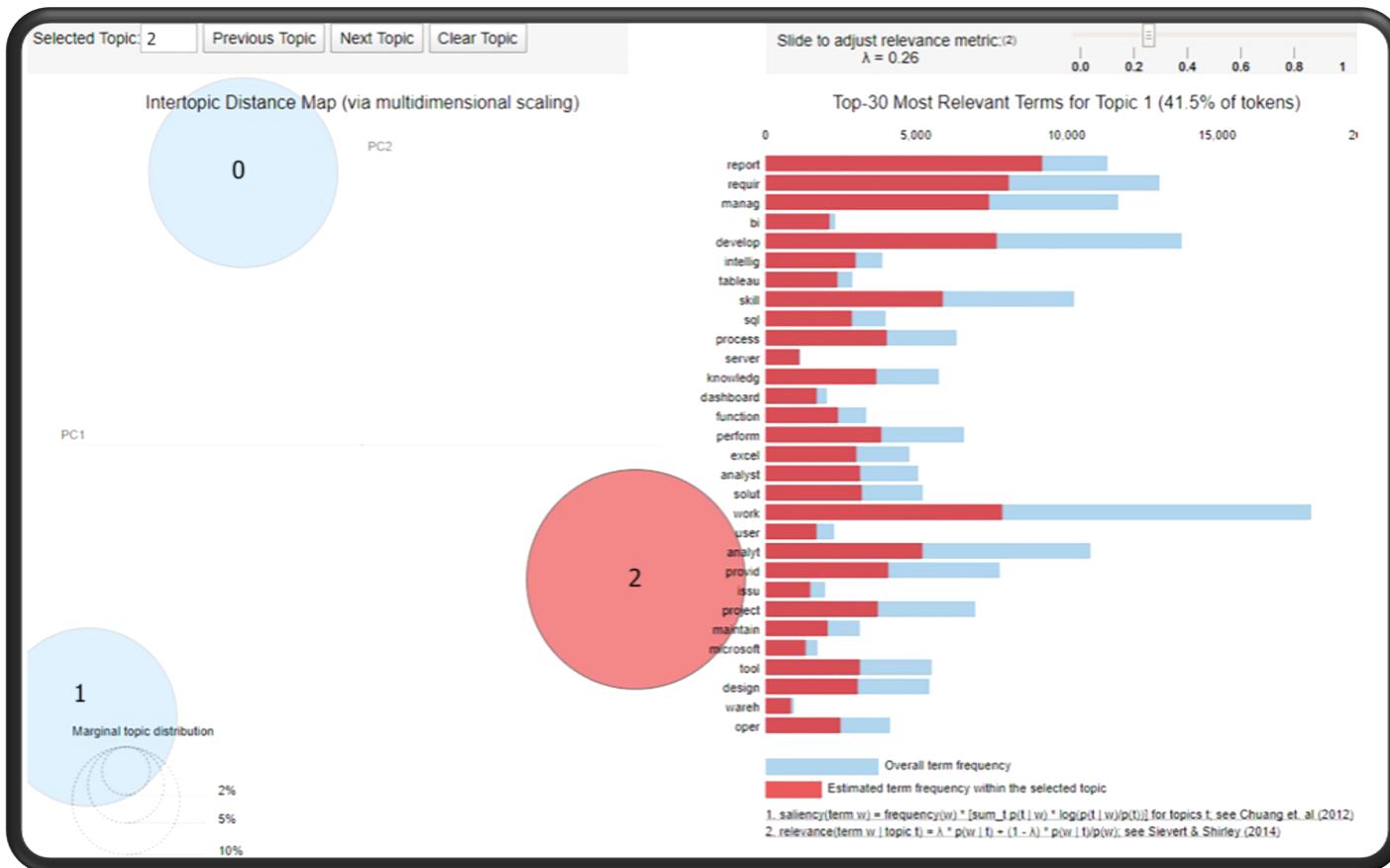
Cluster
job_title
bioinformatics
biostatistics
business intelligence
data analyst
data engineer
data scientist
data visualization
machine learning
modeler





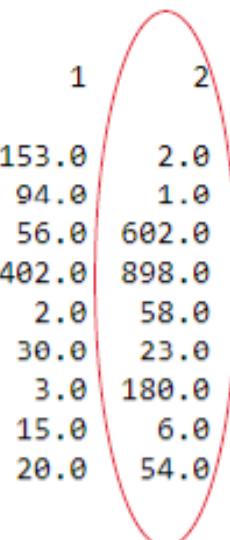
Cluster 2: Business Intelligence

LDA clustering



K-means clustering

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





Word2Vec: Skill taxonomy

Goal

- Provide companies with insight into related skills currently used in the market
- Identify skill proximity

Methods Used

- Word2Vec with Cosine Similarity:

A cosine similarity score of 1.0 indicates the words are the same. While a score of 0 indicates unrelated words.

Amazon Web Services (AWS)

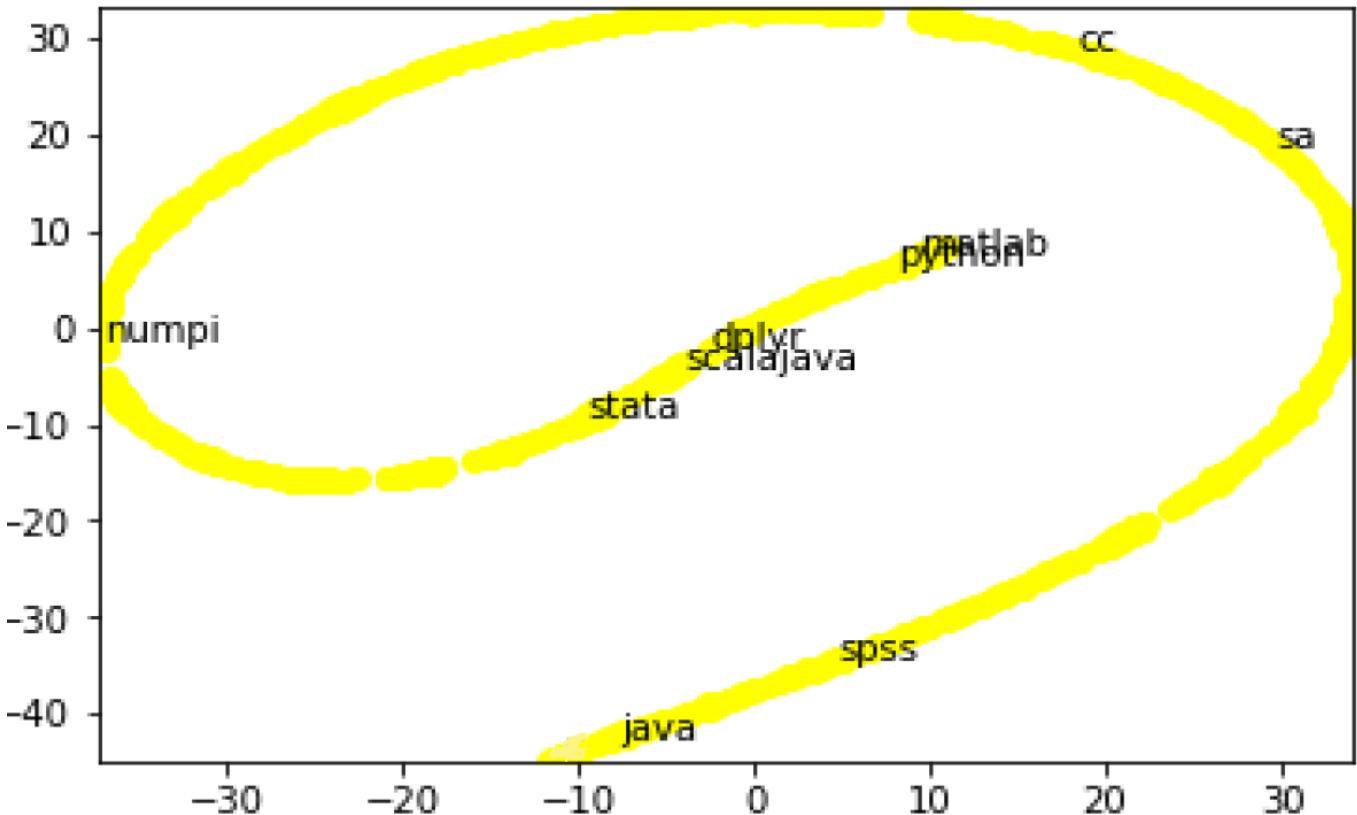
Azure (0.92), Redshift (0.9), NoSQL (0.87) and
Snowflake (0.81)





Word2Vec: T-SNE

- T-SNE can be used to visualize word similarities
- Chart to our right displays the 10 most similar words to R Programming
- The words clumped in the middle such as 'matlab', 'dplyr', and 'scala' indicate words are used in more similar context than 'spss', 'java' or 'numpy'
- Keep in mind, all of these words are still very similar



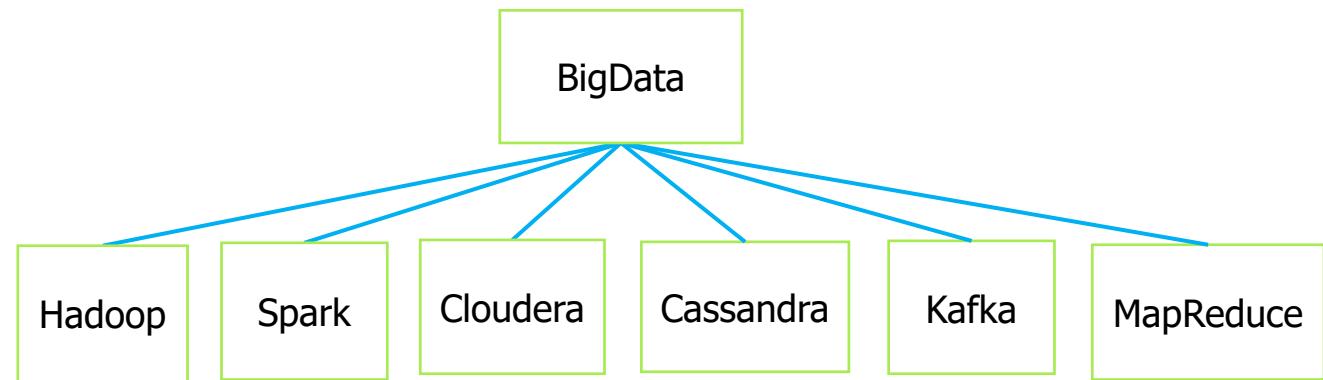


Stage 2: Combining mappings with Clustering

Problem

- Initial clusters (healthcare, Business Intelligence, DS/Data Engineer) are still quite broad
- Distinct job functions (e.g. Data Science and Data Engineer) need to be broken up
- Companies need focused skills for job functions

Mappings for Big Data

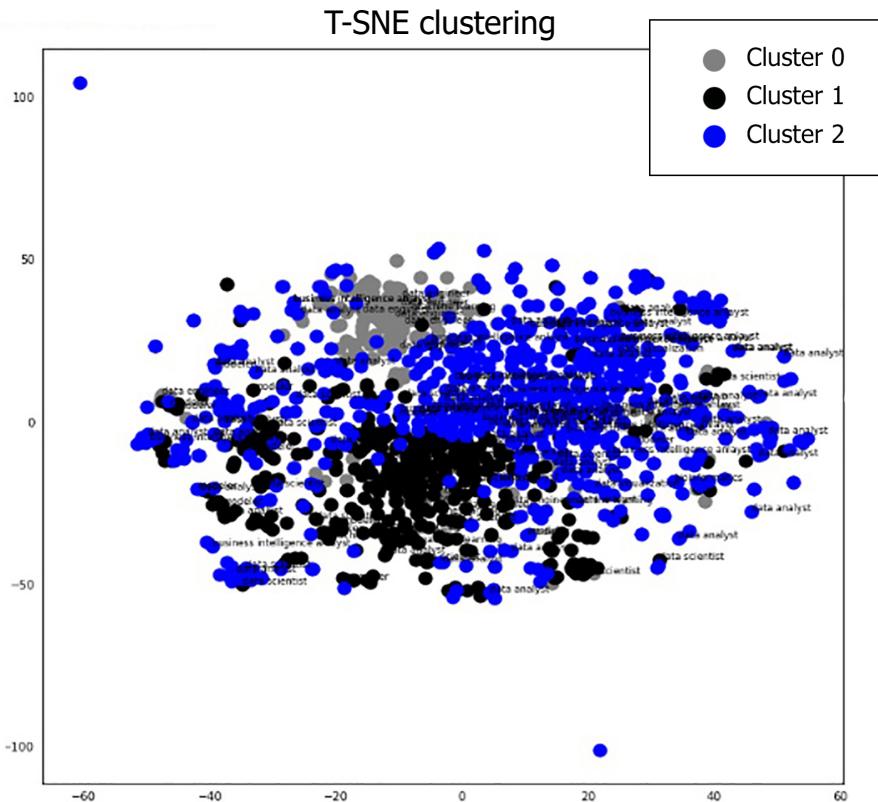


Solution

- Use Doc2Vec to return similar words
- Subject Matter Experts can assist with creating ontology
- Simplify by conflating skills
 - MapReduce, 'Hadoop', 'Hive' changed to 'BigData'



Topic Modeling Findings



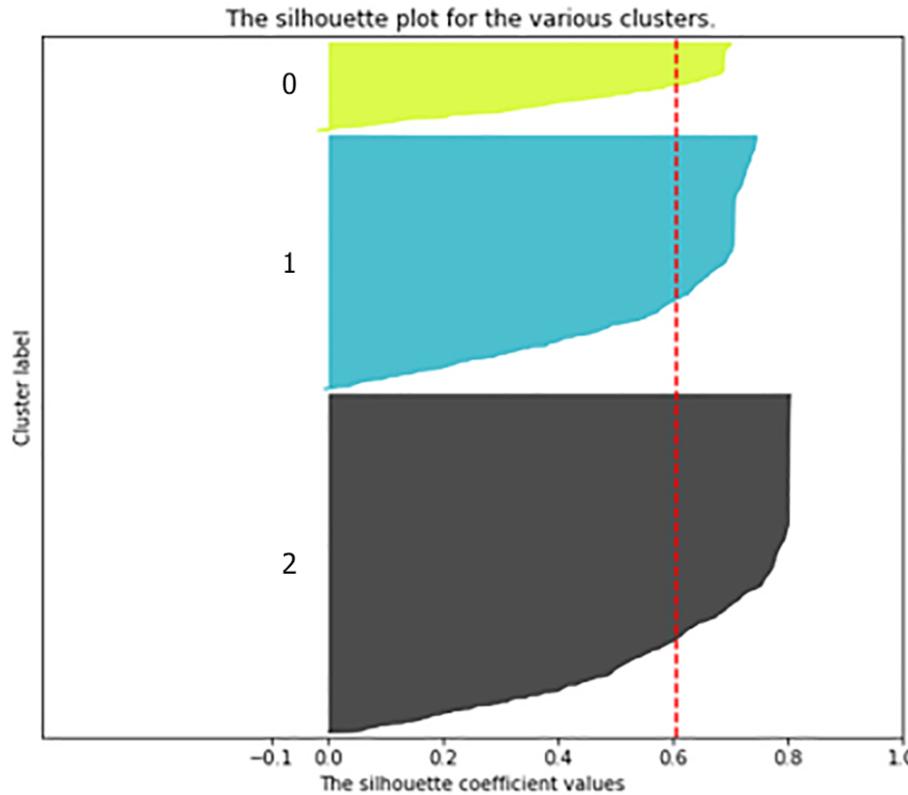
K-means clustering

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

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



Topic Modeling continued...



Width: Indicates size of cluster

Coefficient Values: +1 indicates point is far from other clusters and -1 it may have been put in wrong cluster

Red Line: Average Score (highest for 3 clusters)

Common skills/words

Cluster 0 (green)

- Data Engineering: Included as some of the most common terms the words '**bigdata***', 'engine', 'structured', cloud, '**pipeline***', '**unstructured***'.

Cluster 1 (blue)

- Data Scientist: Included 'machine learning', 'statistics' and '**python***', algorithm and similar words.

Cluster 2 (black):

- BI and Data Analyst: Included 'analyst', 'report', 'analysis', and 'skill', '**structured***'.

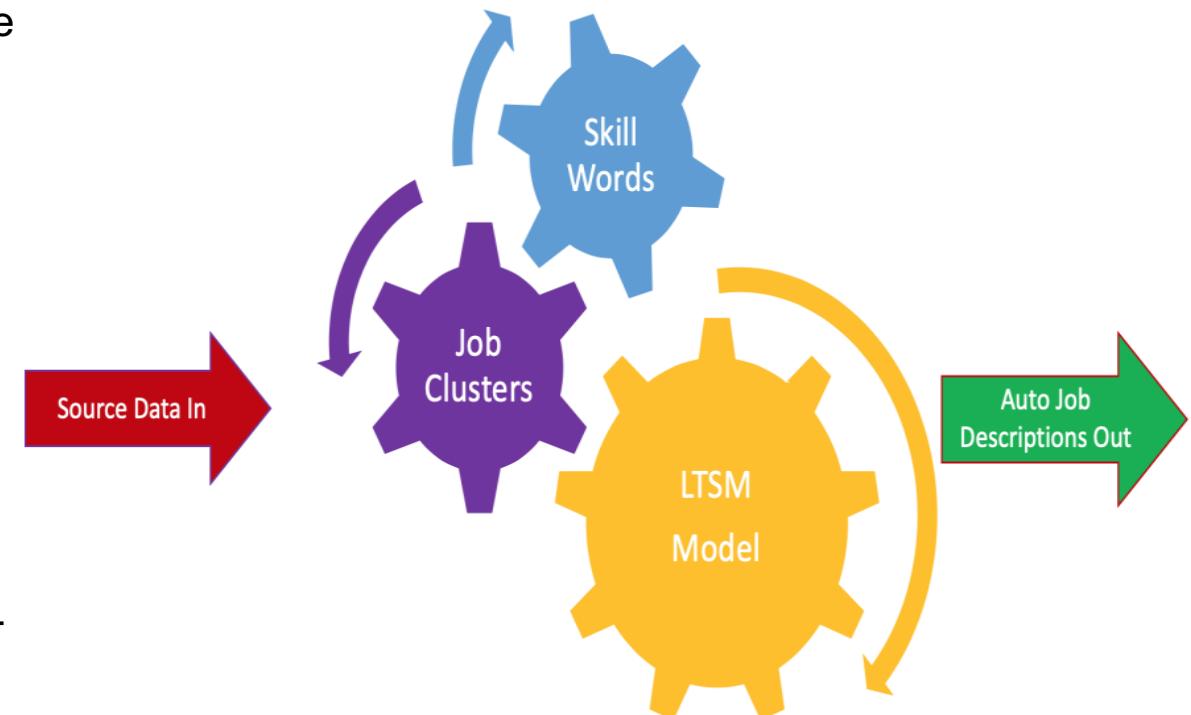
Mapped words*



Proof of Concept (POC): Automatically Generating Job Postings

Goal

- Use insights from Topic Modeling and Doc2Vec to generate relevant job postings
- Companies can identify target applicants and we generate industry standard skills and education requirements
- Job Postings are meant to inform, not replace the work needed to create readable and relevant postings



Method

- Start with state of the art NLP models such as XLNet, GPT and GPT-2 models
- Use Doc2Vec, WordNet and Subject matter experts to add 'sense level' tags for each word (potential categories)



POC: Text Generation Results

Job Posting input into GPT-2 model

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

Text Generated

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