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

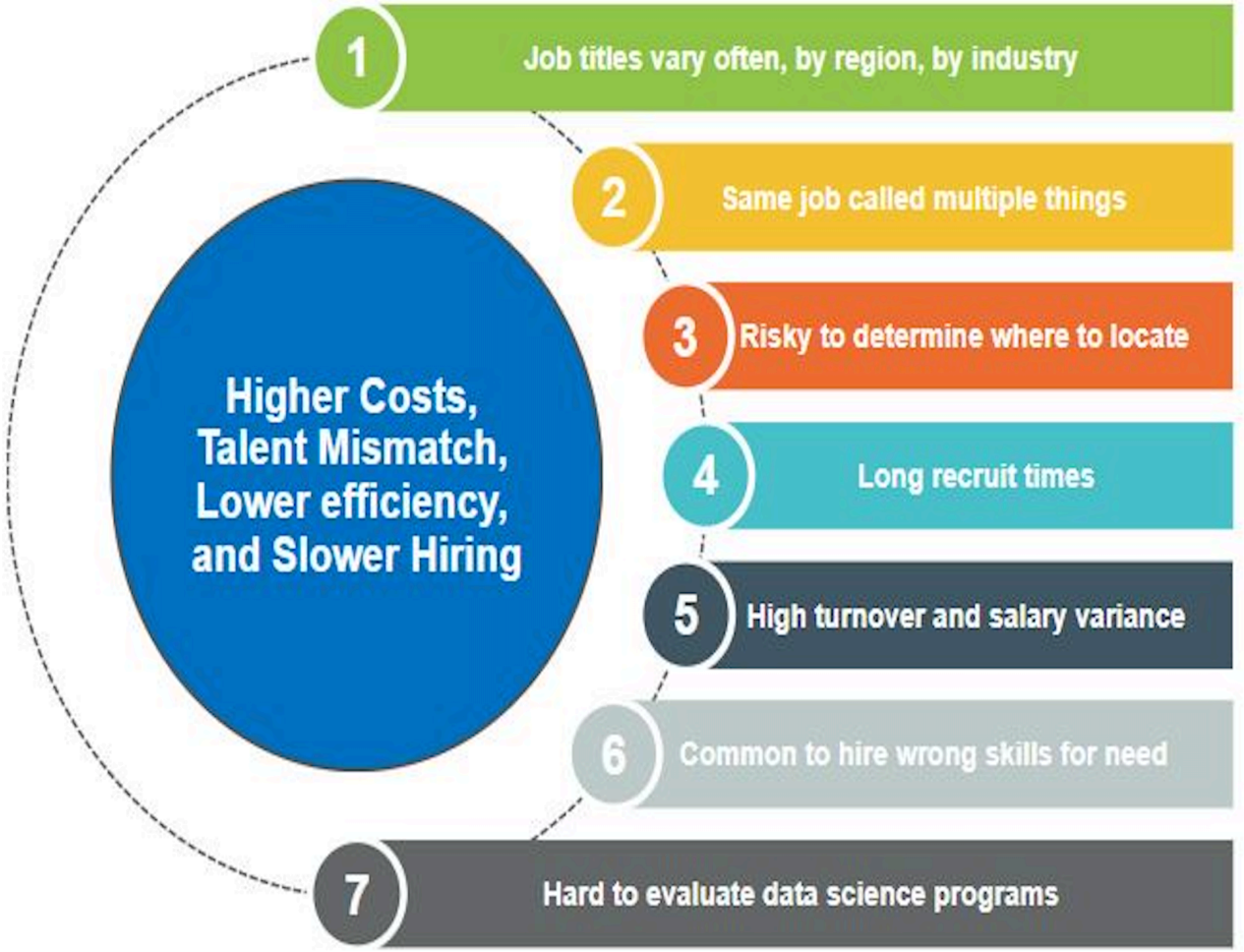
Michael Ryder

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

The Problem

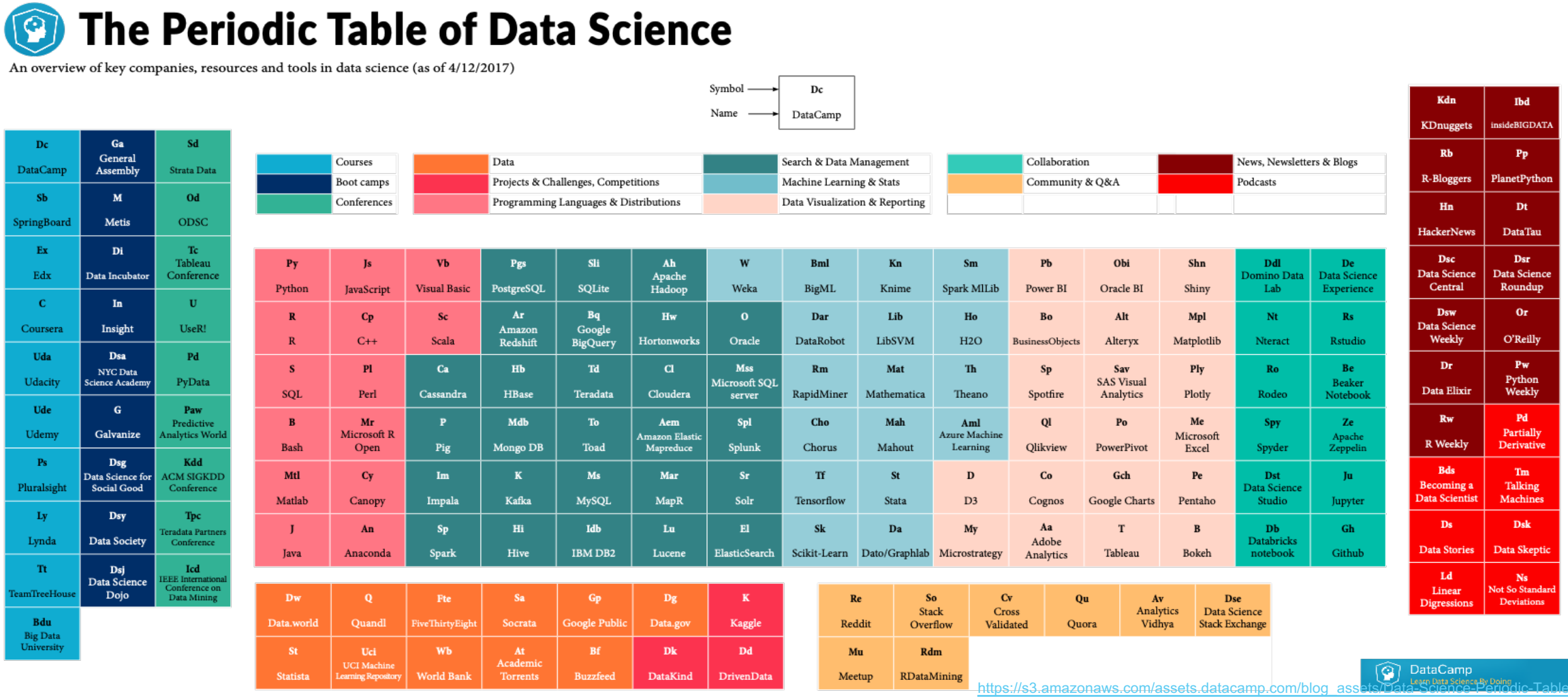
Inefficiencies, shortages, constant change, many gaps

- 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 needs others to source data, do hardware/software engineering, write requirements, deal with growing infrastructure needs, put models in production, and operationalize.



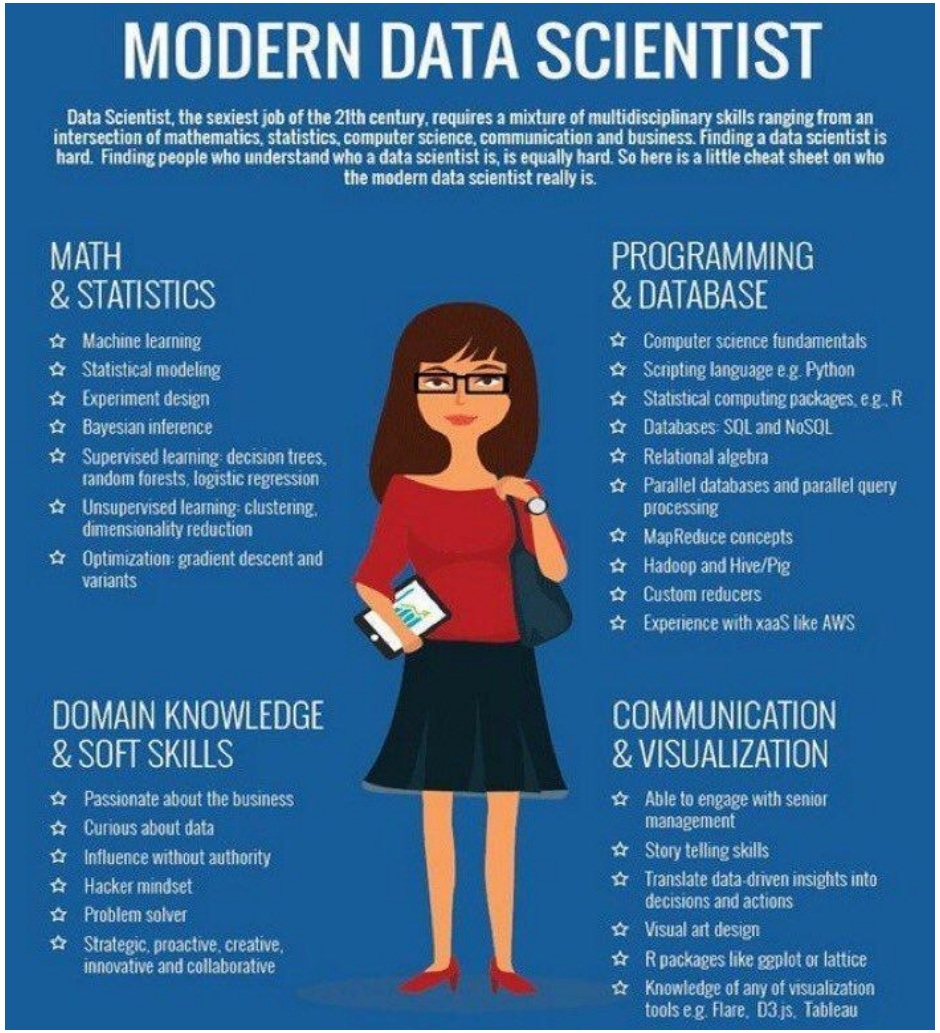
It's Complicated!

Many technologies that are always evolving



What about Real Life?

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

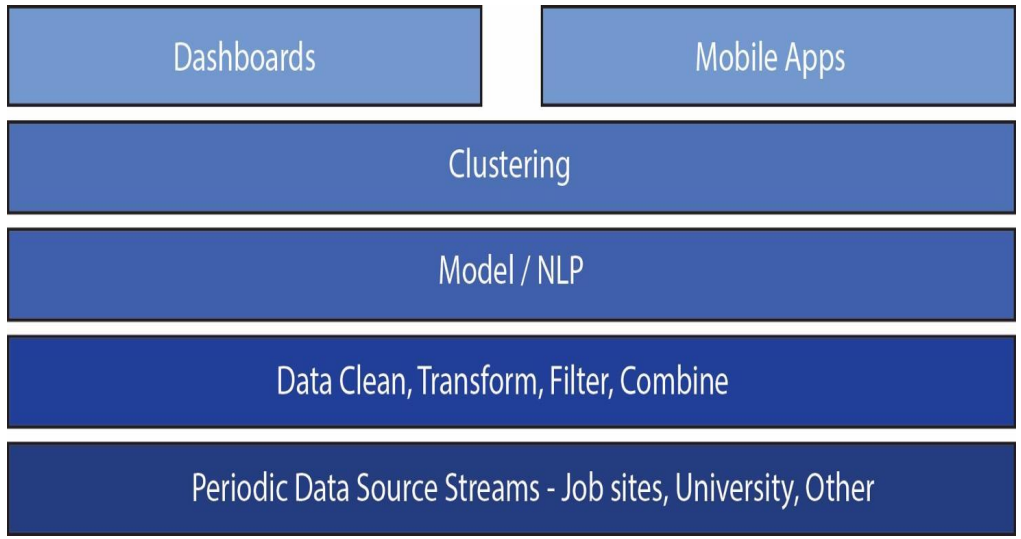


See how many of these challenges you can find on the video !

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.

- **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

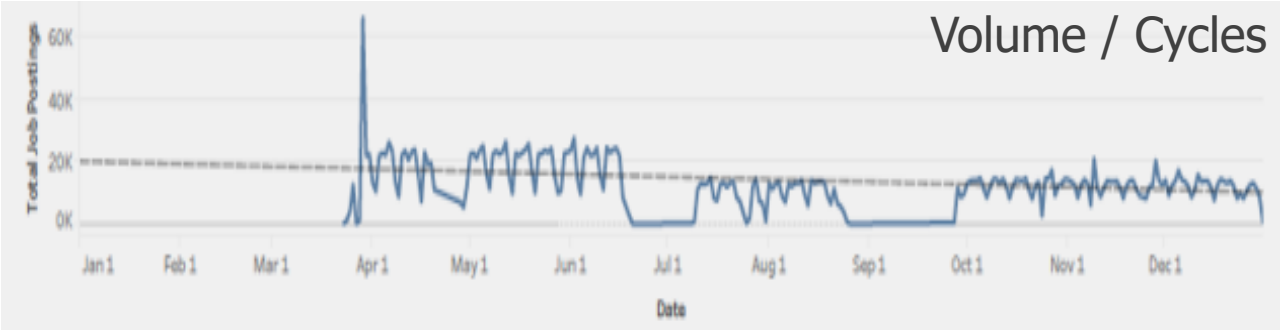
Twitter

Futures

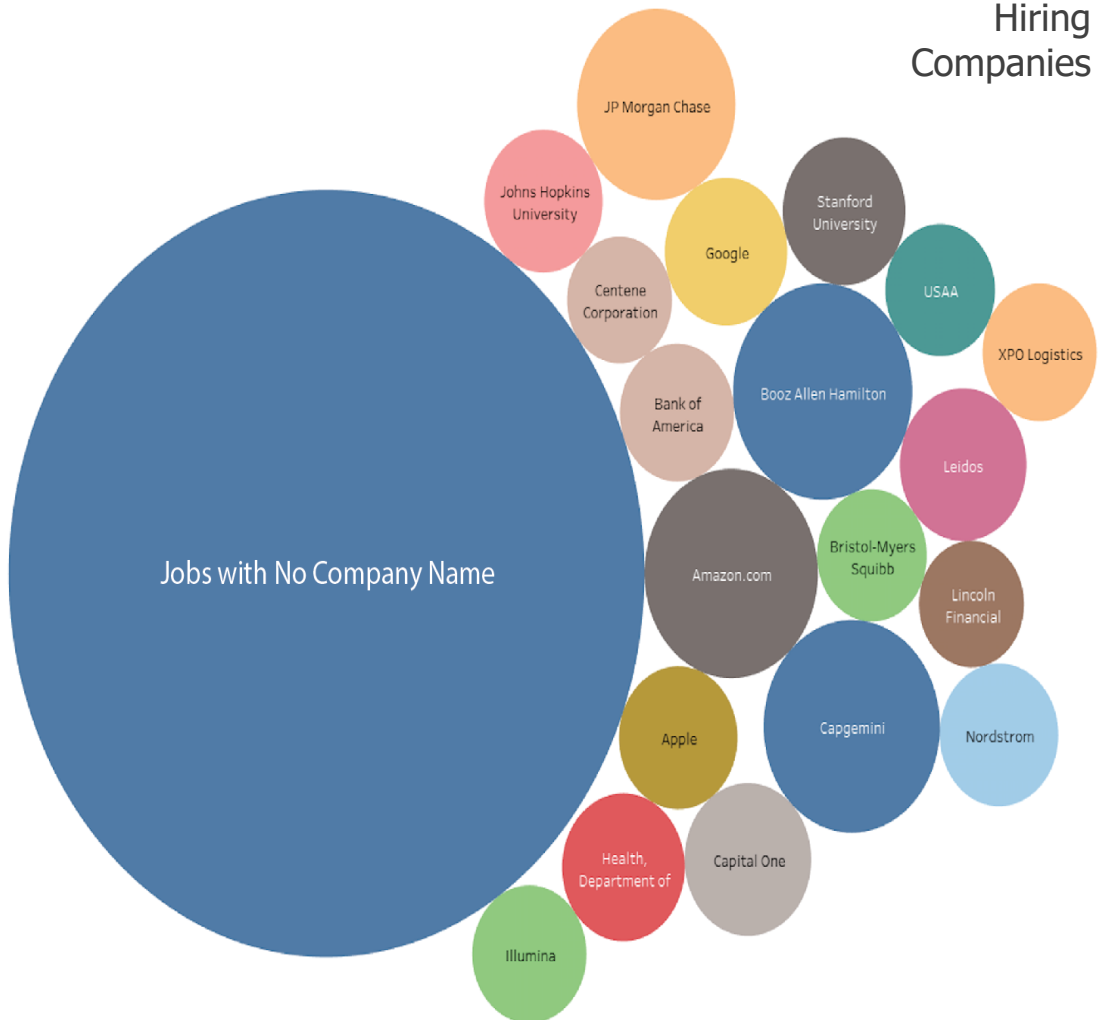
Other education, conference, partner, community sources

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

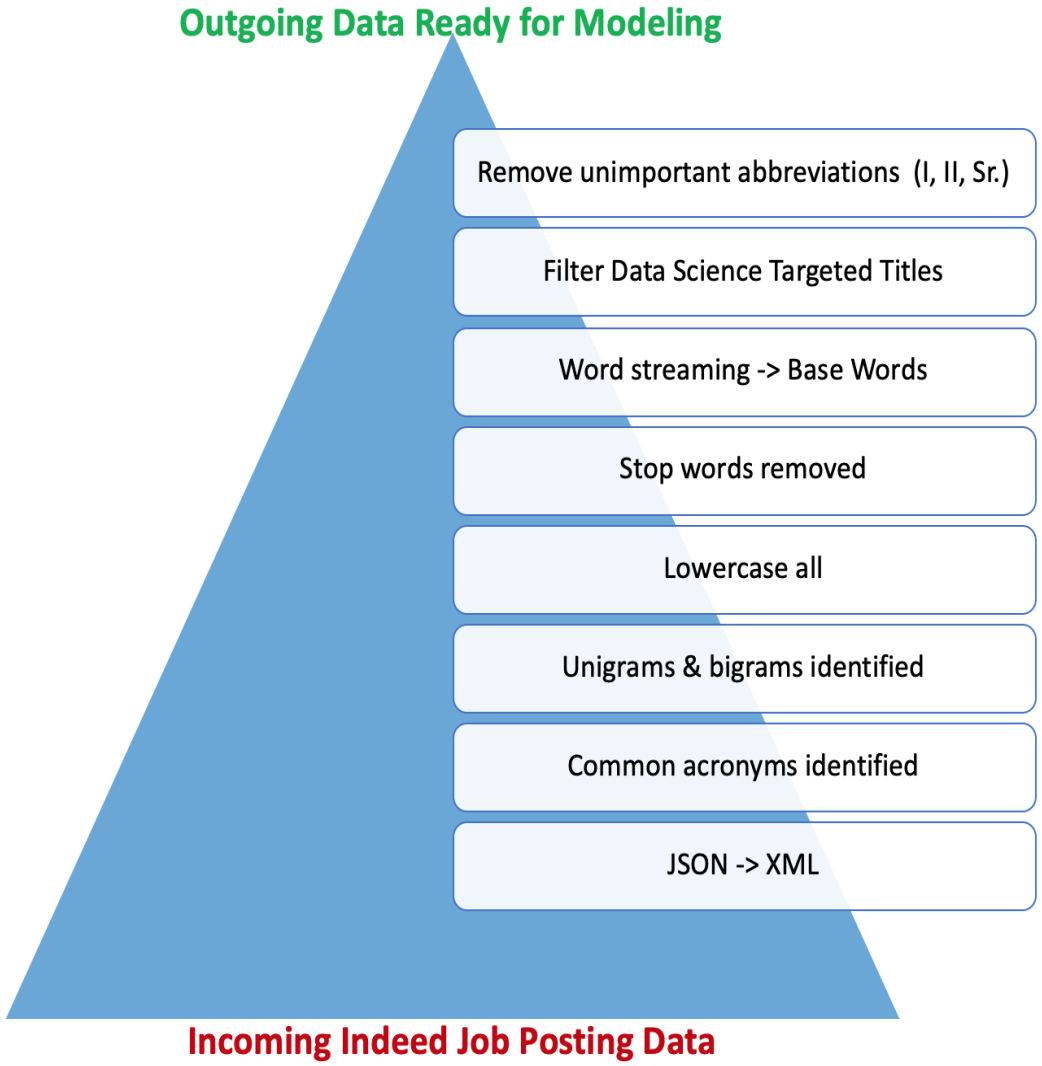
Key Jobs

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.

Brennen

Modeling Methods

Method for conducting analysis of job functions and related skills

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

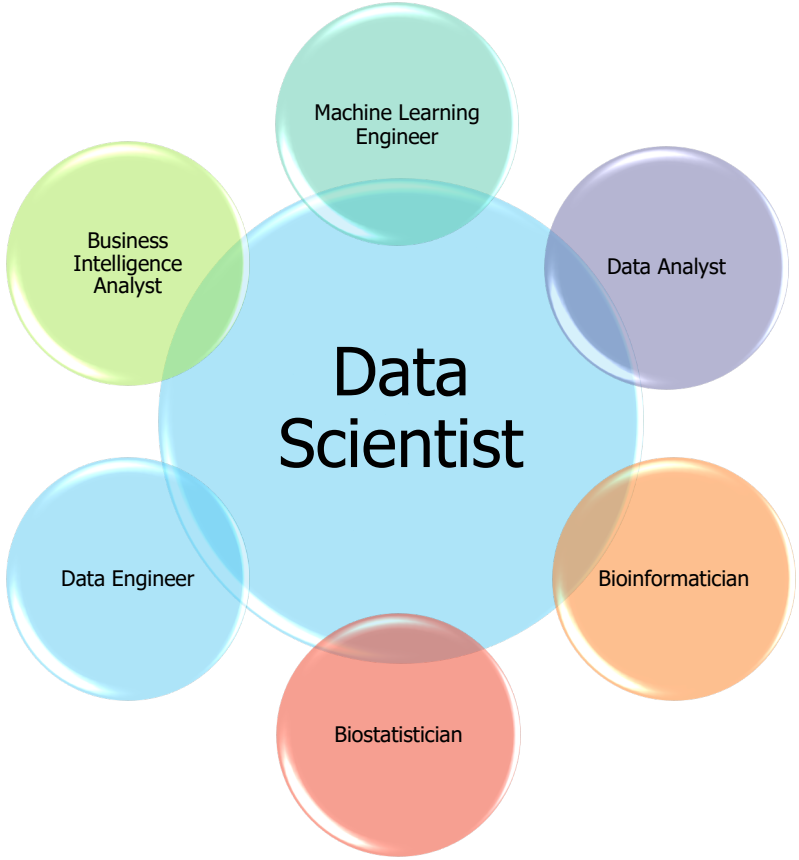
Identify Relevant Job Postings – Data Science

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



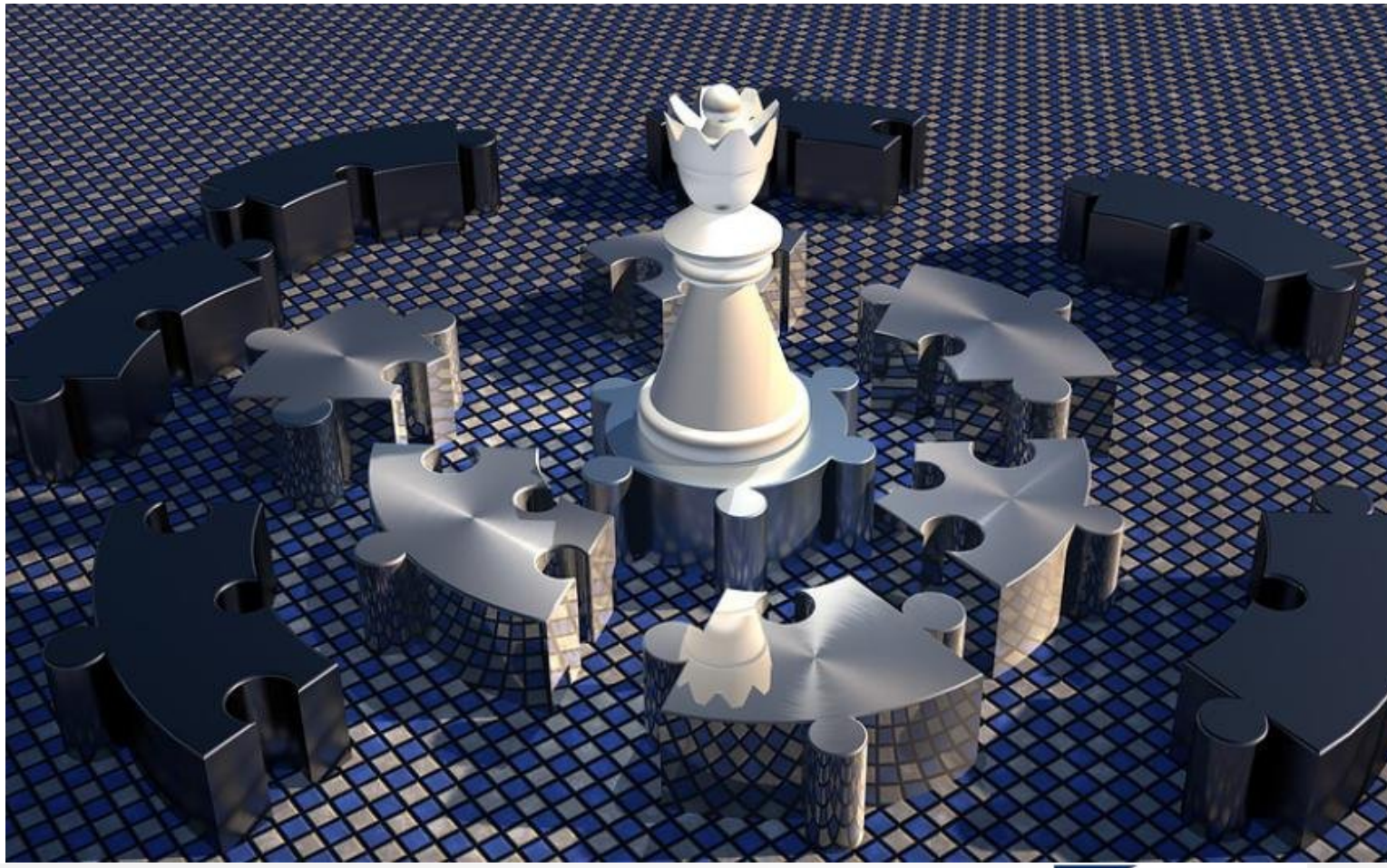
Topic Modeling: Triangulate with Multiple Models

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:

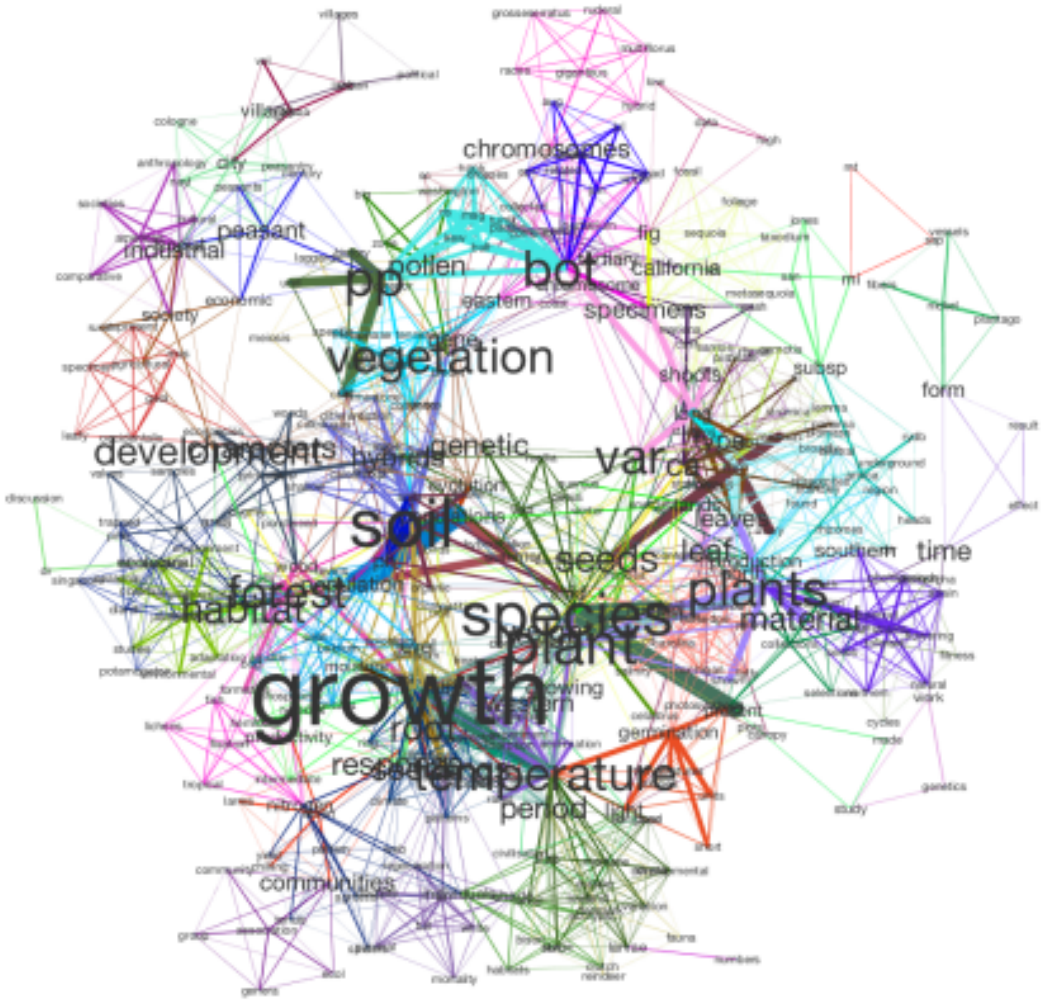
Data Science & Data Engineer

Cluster 1:

Healthcare (data science jobs in healthcare)

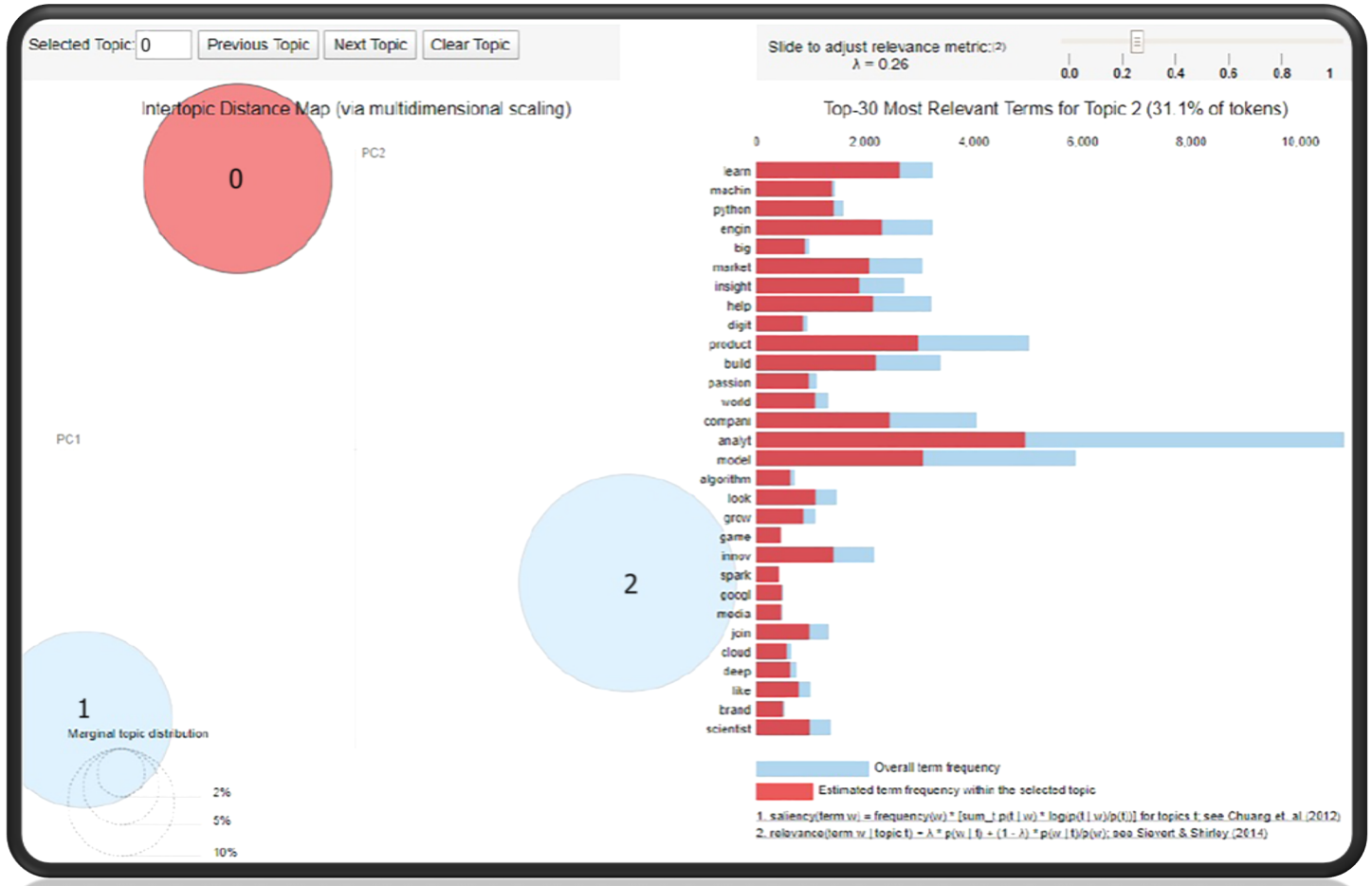
Cluster 2:

Business Intelligence (BI)



Cluster 0: Data Science Cluster

LDA clustering

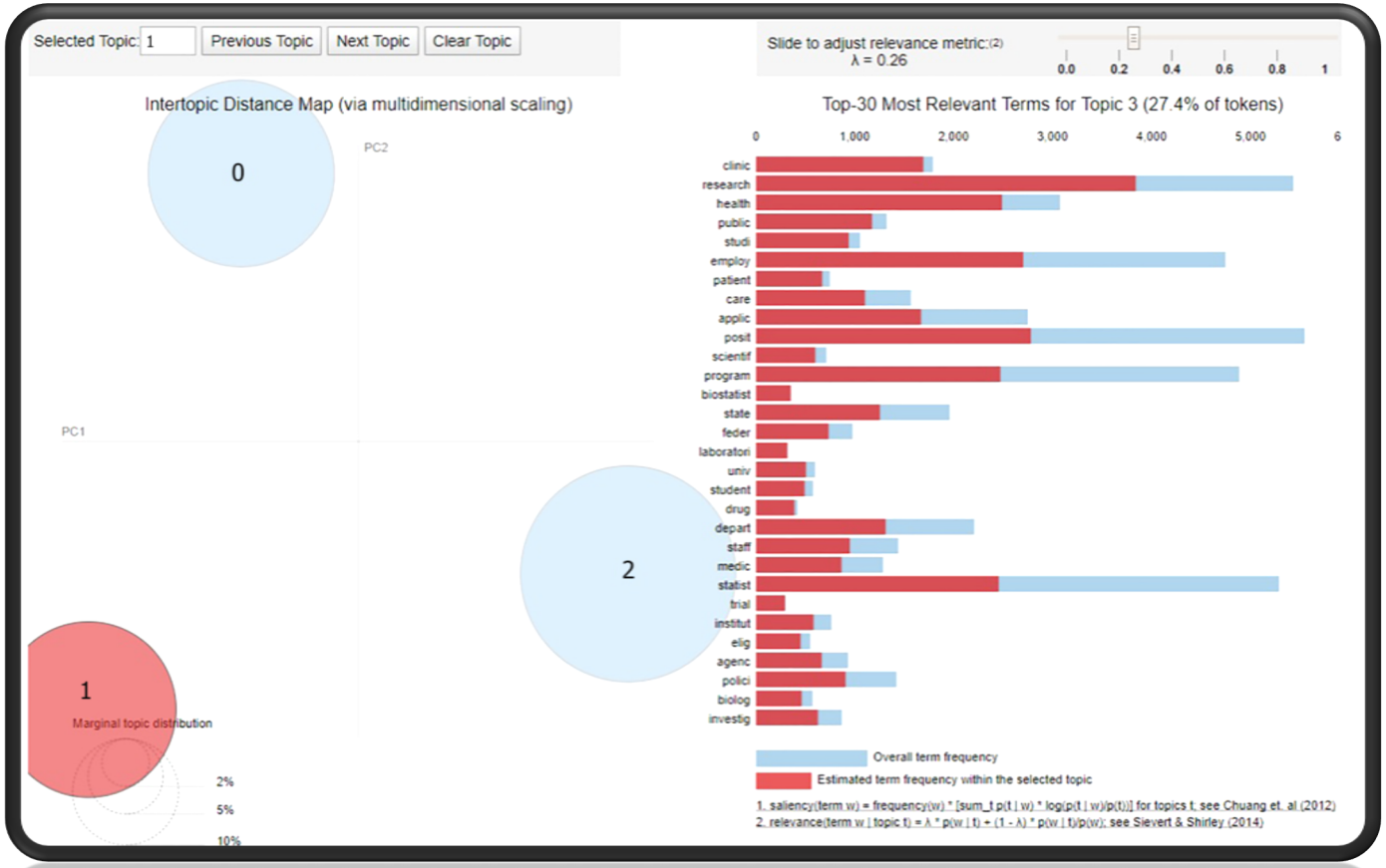


K-means clustering

Cluster	0	1	2
job_title			
bioinformatics	7.0	153.0	2.0
biostatistics	NaN	94.0	1.0
business intelligence	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

Cluster 1: Healthcare

LDA clustering

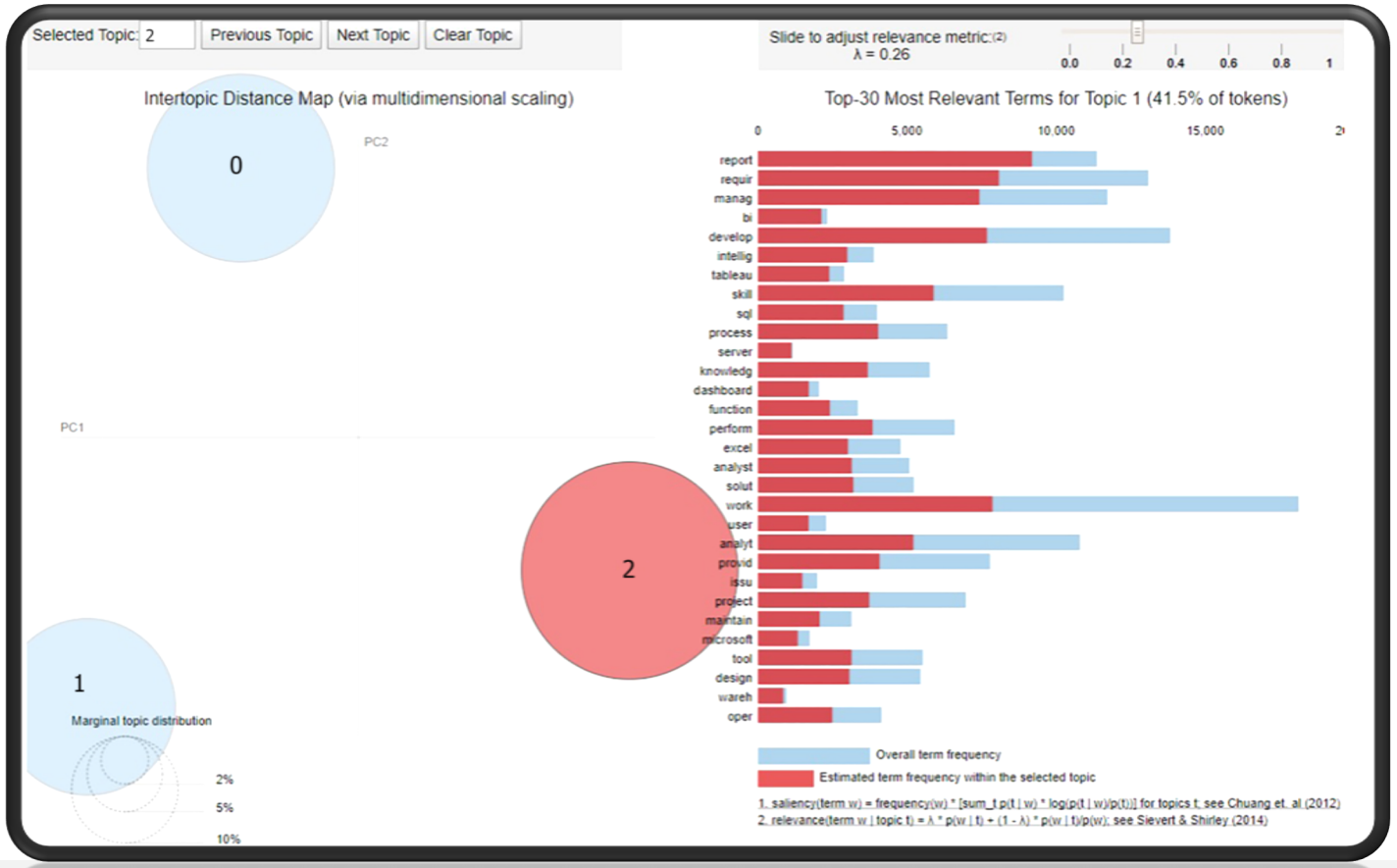


K-means clustering

Cluster	0	1	2
job_title			
bioinformatics	7.0	153.0	2.0
biostatistics	NaN	94.0	1.0
business intelligence	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

Cluster 2: Business Intelligence

LDA clustering



K-means clustering

Cluster	0	1	2
job_title			
bioinformatics	7.0	153.0	2.0
biostatistics	NaN	94.0	1.0
business intelligence	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

Word2Vec: Skill taxonomy

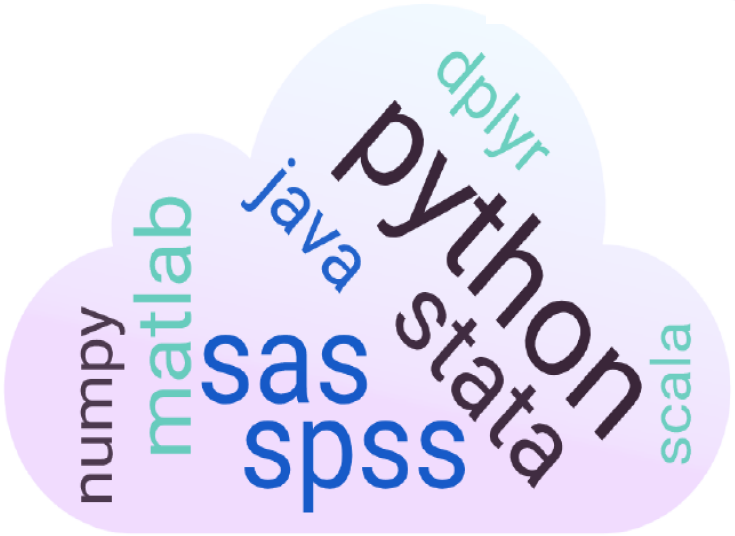
Goal

- 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.

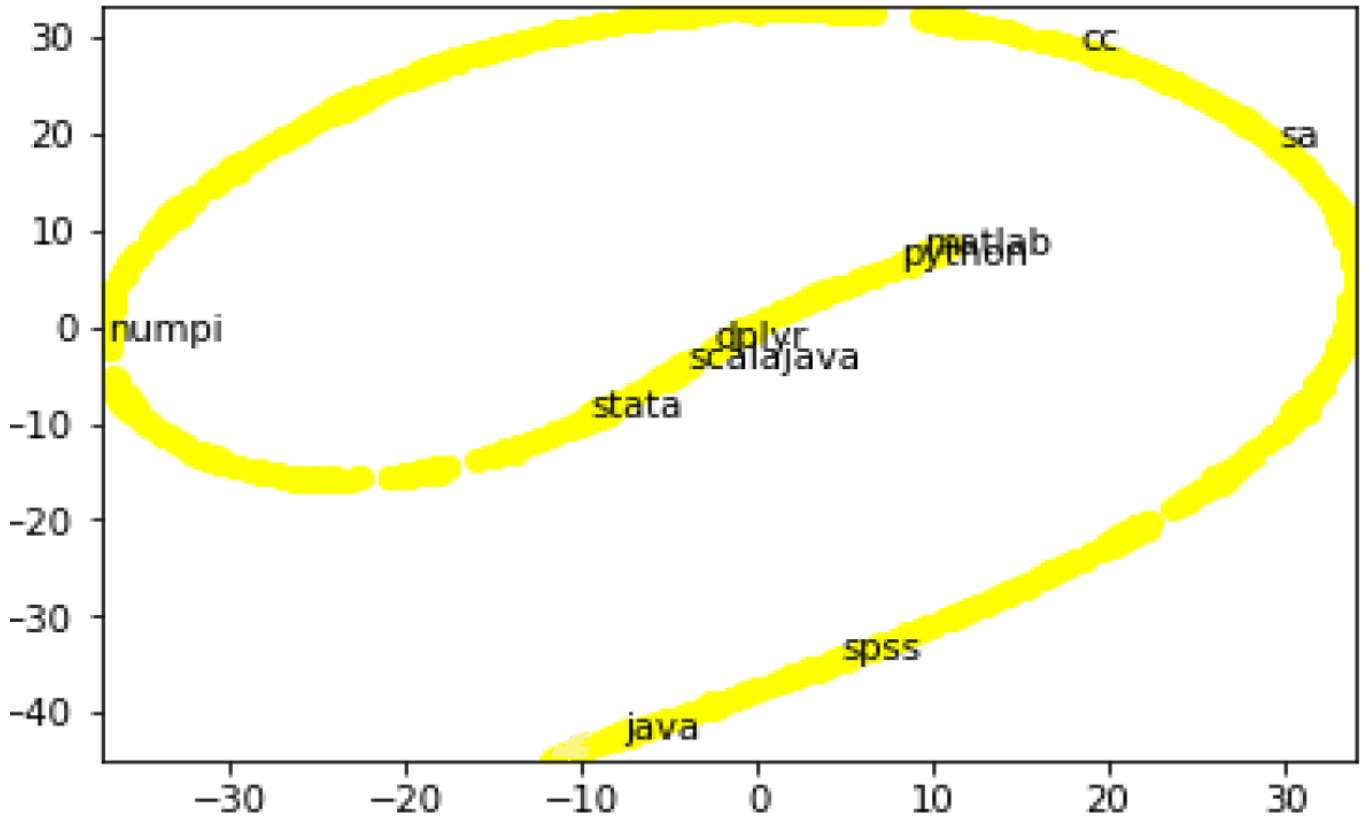


Amazon Web Services (AWS)

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’, ‘dpylr’, 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



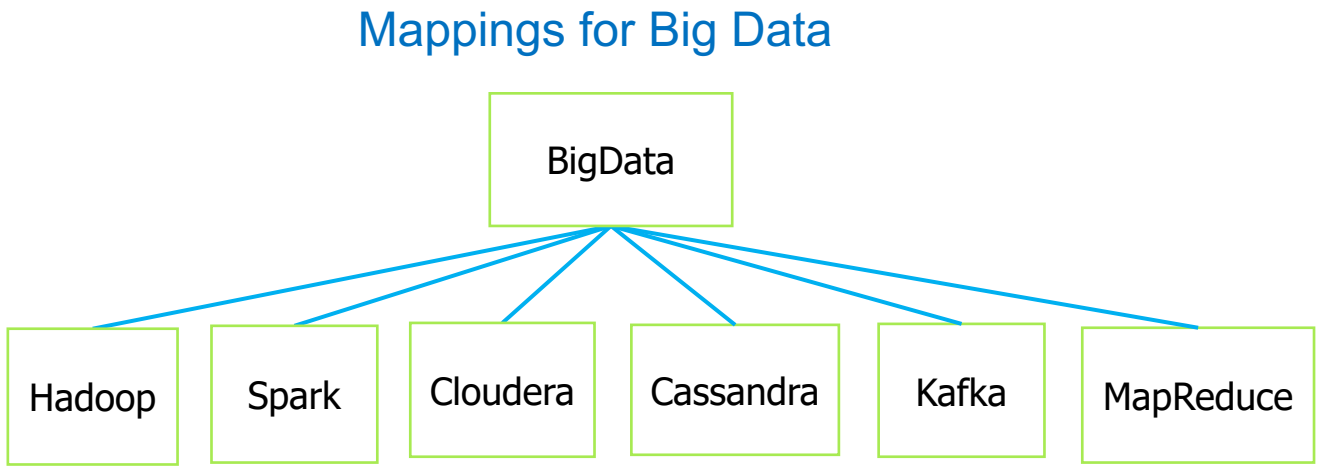
Stage 2: Combining mappings with Clustering

Problem

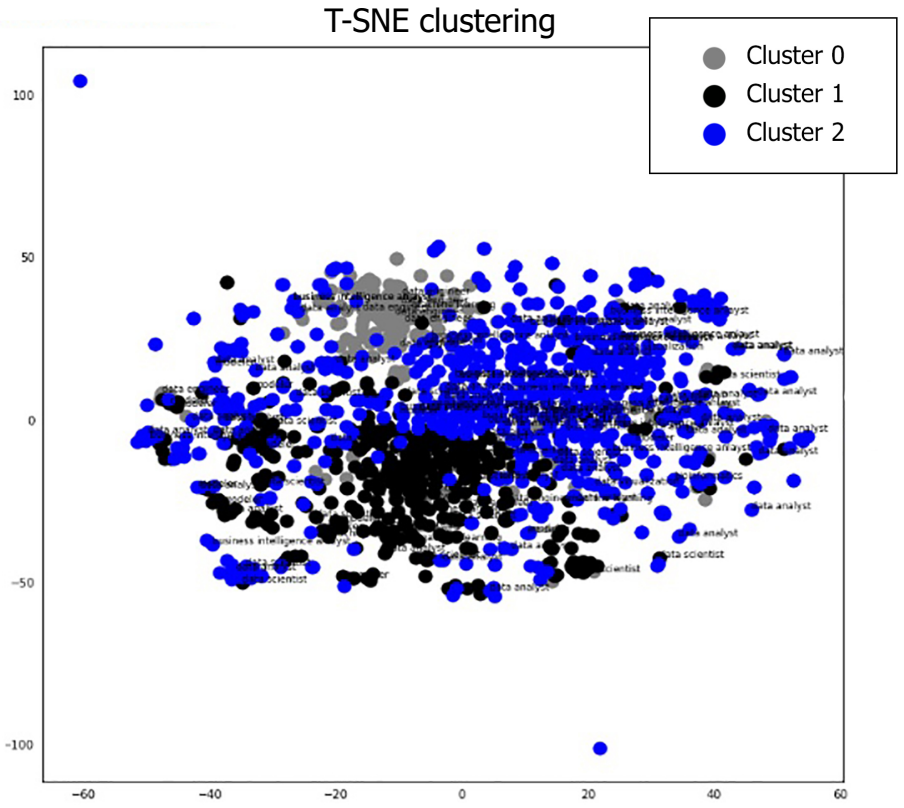
- Initial clusters (healthcare, Business Intelligence, DS/Data Engineer) are still quite broad
- Distinct job functions (e.g. Data Science and Data Engineer) need to be broken up
- Companies need focused skills for job functions

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’



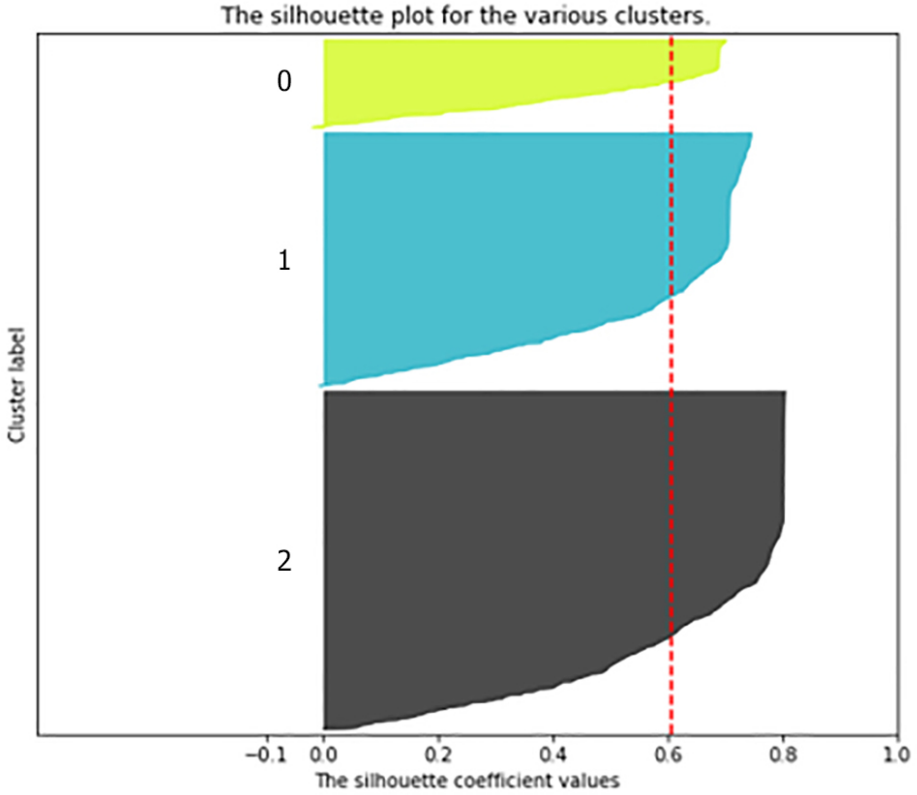
Topic Modeling Findings



K-means clustering

Cluster	0	1	2
job_title			
bioinformatics	NaN	3.0	4.0
business intelligence	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

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):

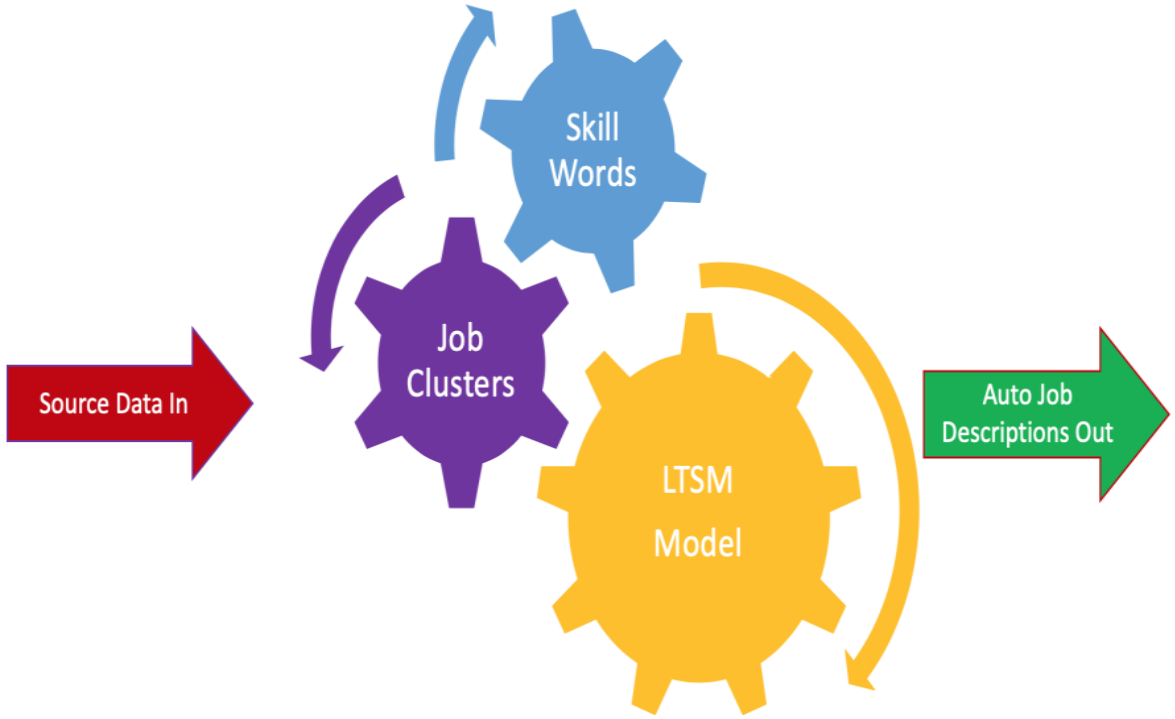
- BI and Data Analyst: Included 'analyst', 'report', 'analysis', and 'skill', '**structured***'.

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