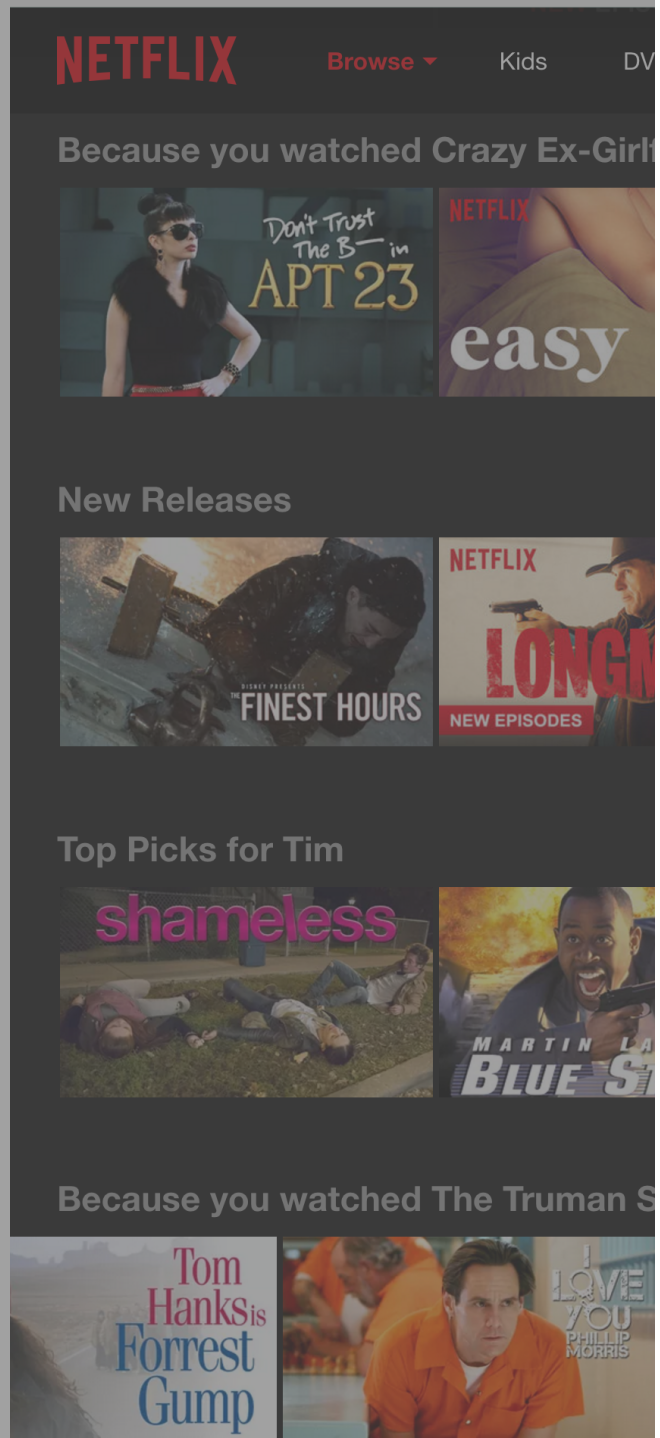
NLP for Classification:

The Language of Job Postings

Using data science to find a data science future



Imagine…

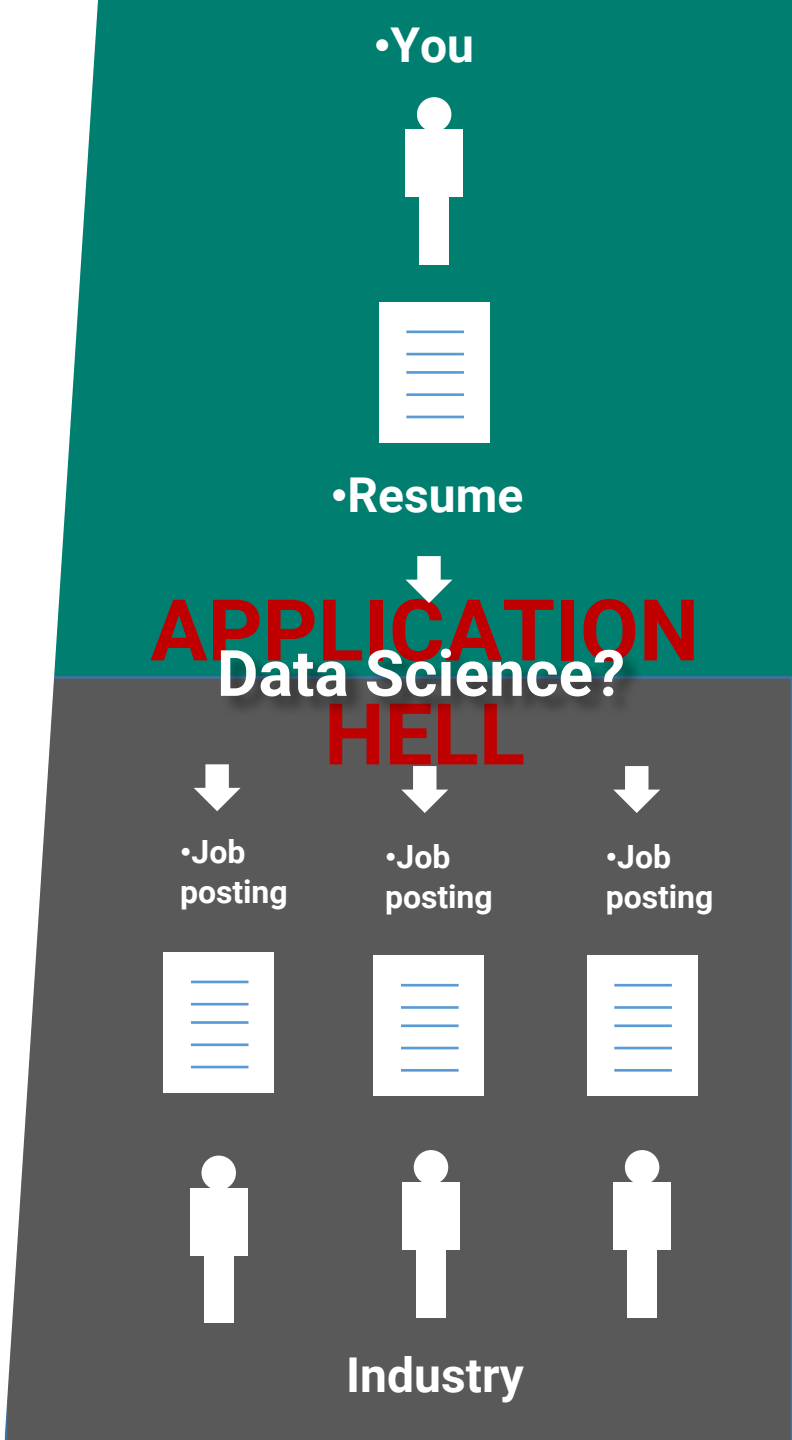
- Imagine having to search for new movies with a search bar:
  - Terminator with time travel
  - Star Wars in medieval times
  - Sleepless in Seattle without Seattle, or being sleepless
- Frustrating, inefficient, and slow
- We don't know the title of what we haven't seen

- This is **exactly** the current process for **Job Searching**

Introduction: Job Searching

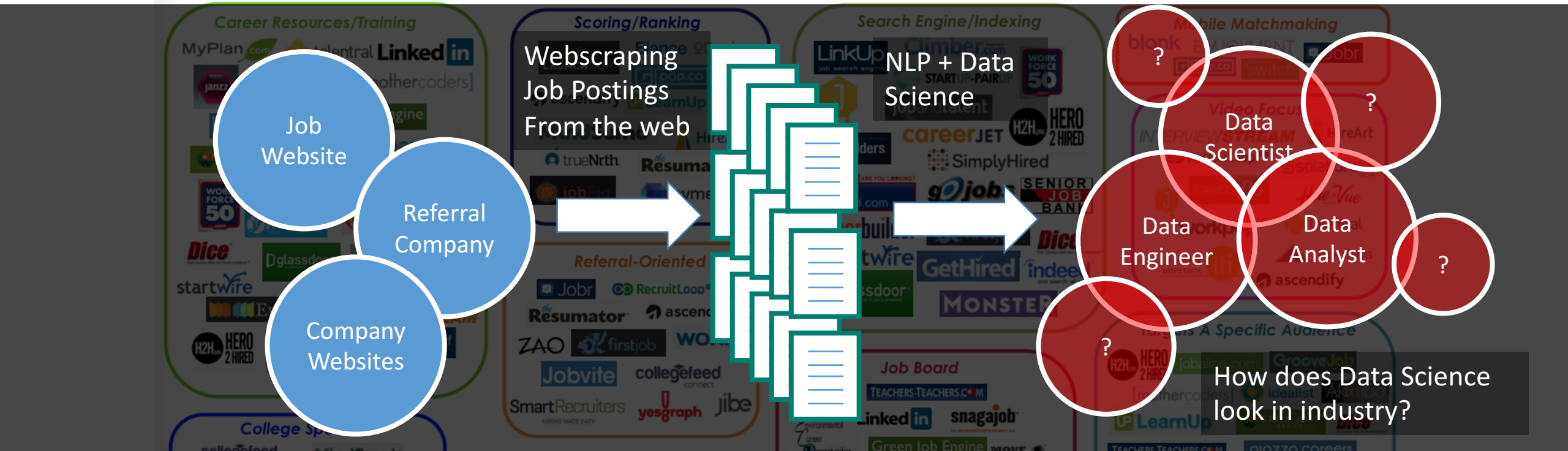
- Job Searching is difficult
  1. Guess some search terms
     1. Either skills or titles
     2. Unfortunately “python”, “data”, and “science” are very generic. Compared to “dentist” or “CPA accountant”
  2. Searching on the web: Spray and Pray

- **Can’t Data science help us find/understand the Data science industry?**



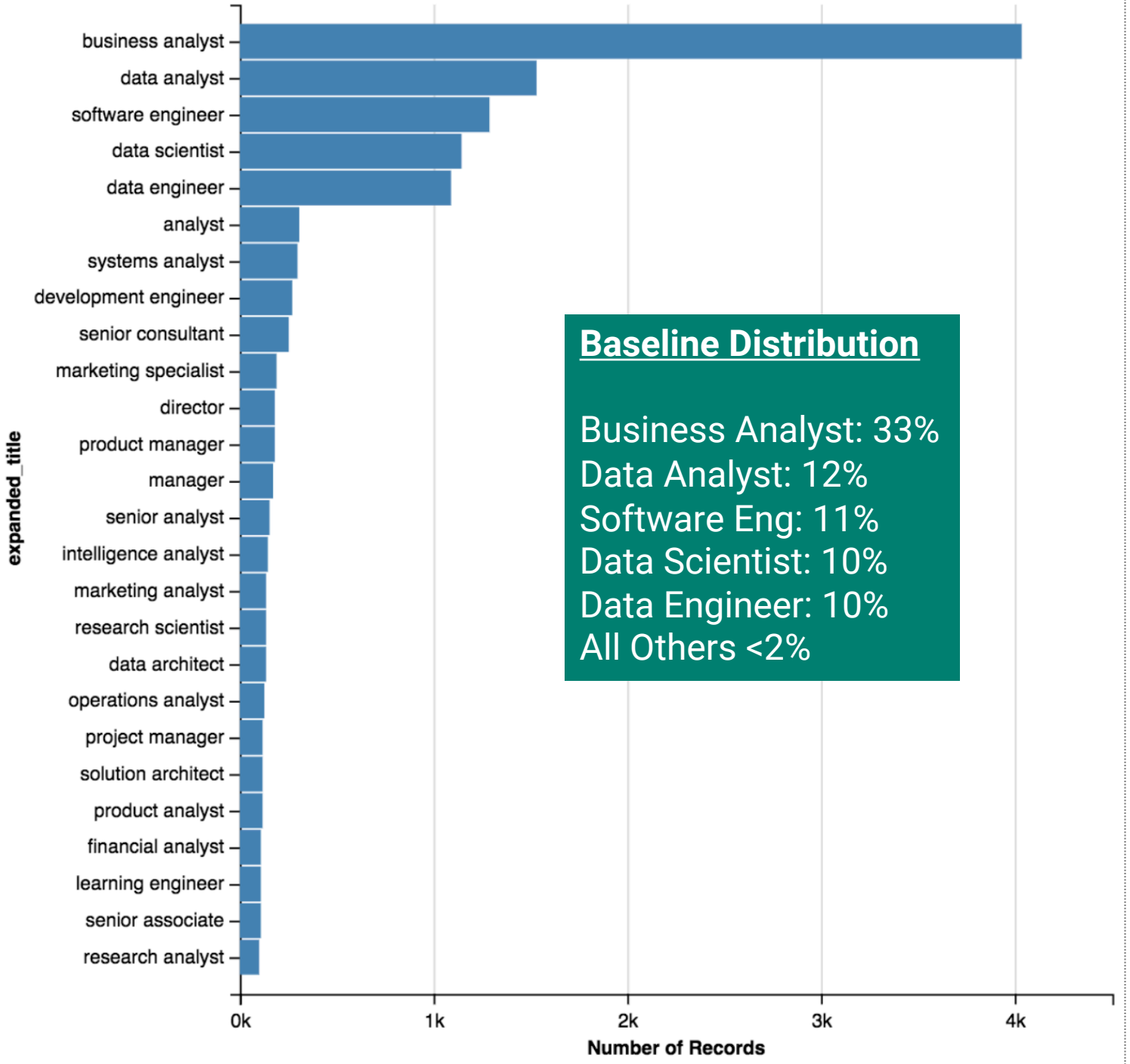
Problem Statement

- Can job postings be used to **predict job titles**?
- Can data science help us apply to jobs?
- What can job postings tell us about the Data Science Industry?



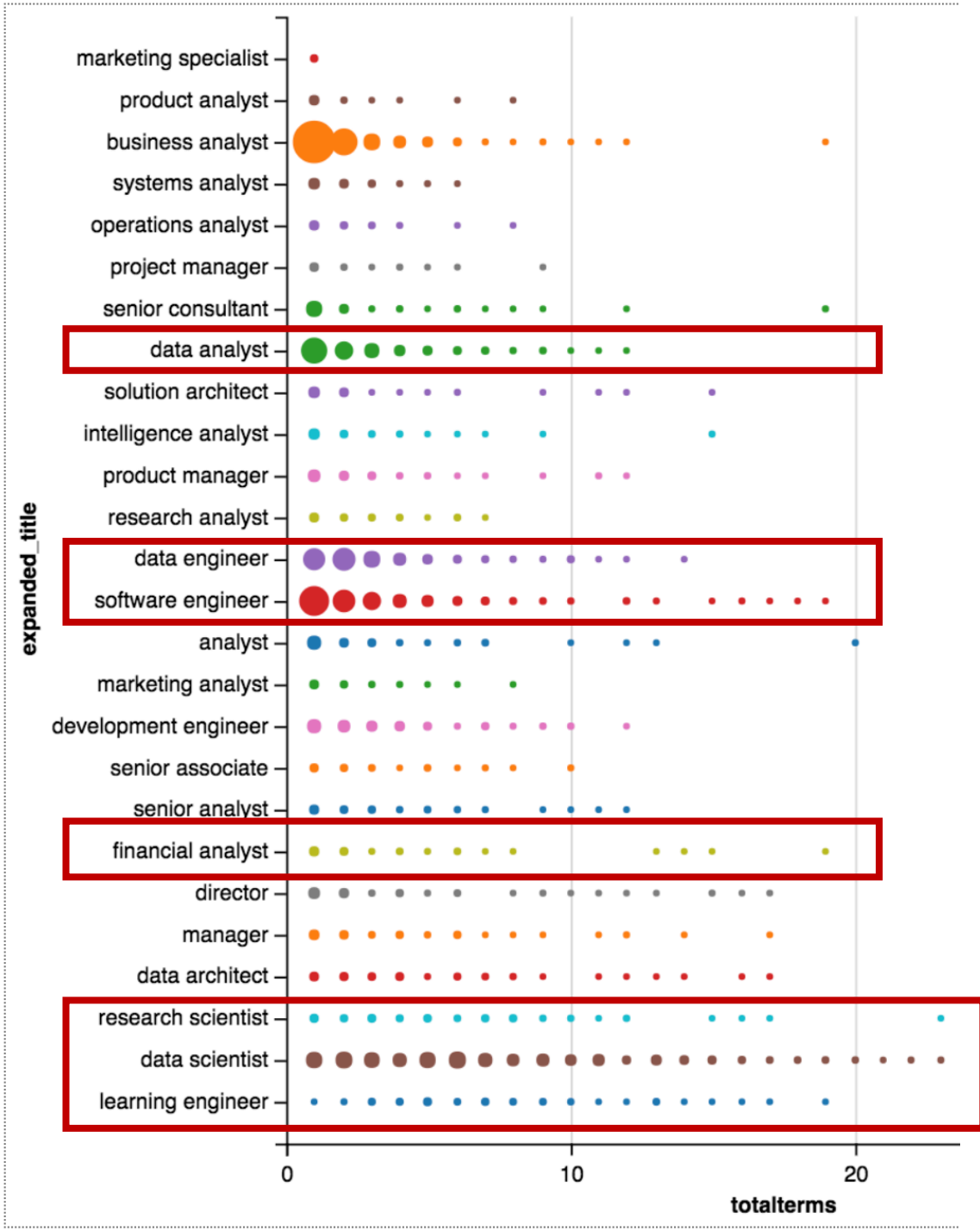
Baseline Titles

- Out of **22,000** job postings
- High frequency titles totaling **12,000** will be considered
- Long titles are reduced. E.g.“Senior Marketing Data Analyst”, and “Data Analyst of Cancer Research” are reduced down to “Data Analyst”



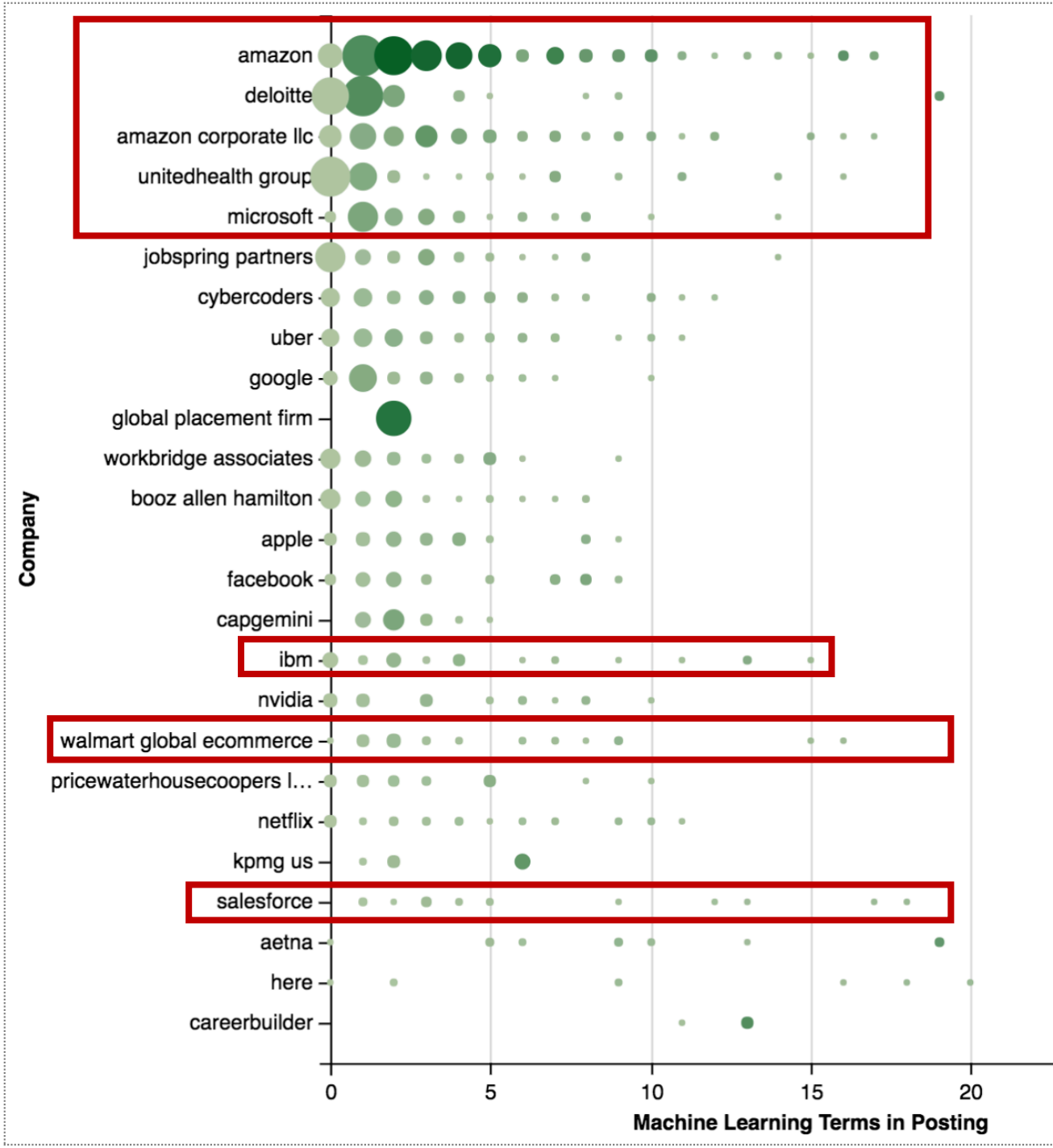
Postings with Data Science Terms

- ***'data science', 'deep learning', 'machine learning', 'prediction'…*** were counted in all job postings collected
- Top job titles with these terms
  - Data scientist
  - Research Scientist
  - Learning Engineer
  - Software engineer
  - Data Engineer
  - Financial Analyst
  - Data Analyst



Companies with Data Science Terms

- Top companies with concentration of machine-learning related job postings companies from extraction (Sep 2016)
  - Amazon
  - UnitedHealth
  - Microsoft
  - Google
  - IBM
  - SalesForce



3 Levels of Text Detail Examined

Dataset #1: Full Text Posting

Scaped 22,000 full HTML job postings from the internet

Predict Job Title

Dataset #2 Line Items

Basic Qualifi
Bachelors degree in a quantitative field such as Computer Science, Engineering, Operational Research, Statistics, Economics, etc.
3+ years' experience working in an Analytics, Data Science, or Engineering function
Experience using SQL
Experience coding in Python, Java, or equivalent
Experience communicating with non-technical colleagues

Highly Preferred Qualifications:
Masters degree in a quantitative field such as Computer Science, Engineering, Operational Research, Statistics, Economics, etc.
5+ years' experience working in an Analytics, Data Science or Engineering function
Experience working in the consumer web/internet domain
Experience with large-scale data on Hadoop
Web development experience (Javascript/CSS/HTML/D3/Highcharts)
Experience with statistic ogramming environments like R or equivalent
Experience with data min modeling, or machine learning a pl

Predict Job Title

Dataset # 3 Hard Skills

Basic Qualifications:
Bachelors degree **in a quantitative field such as Computer Science, Engineering, Operational Research, Statistics, Economics, etc.**
3+ years' experience working **in an Analytics, Data Science, or Engineering function**
Experience **using SQL**
Experience coding **in Python, Java, or equivalent**
Experience communicating with non-technical col gues

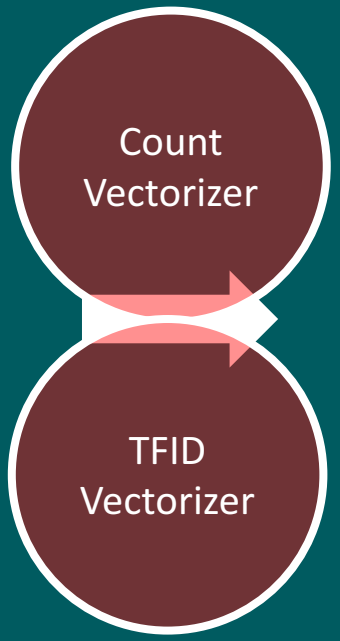
Predict Job Title

Features: Word Vectorizing

Raw Text

Experience working in the consumer web/internet domain Experience with large-scale data on Hadoop Web development experience (Javascript/CSS/HTML/D3/Highcharts) Experience with statistical programming environments like R or equivalent. Experience with data mining, modeling, or machine learning a plus

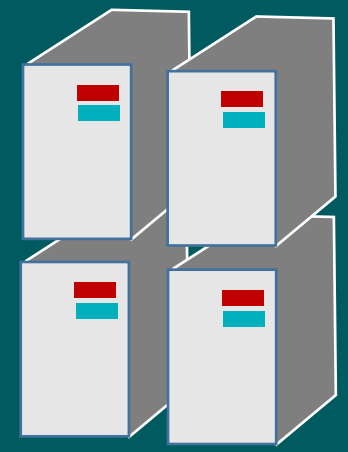
Python Tools



Abstract Text into Math Matrices

['Experience', 'working', 'in', 'the', 'consumer', 'web/internet', 'domain', 'Experience', 'with', 'large-scale', 'data', 'on', 'Hadoop', 'Web', 'development', 'Javascript', 'CSS', 'HTML', 'D3', 'Highcharts', 'Experience', 'with', 'statistical', 'programming', 'environments', 'like', 'R', 'or', 'equivalent', 'Experience', 'with', 'data', 'mining,', 'modeling,', 'or', 'machine', 'learning', 'a', 'plus']

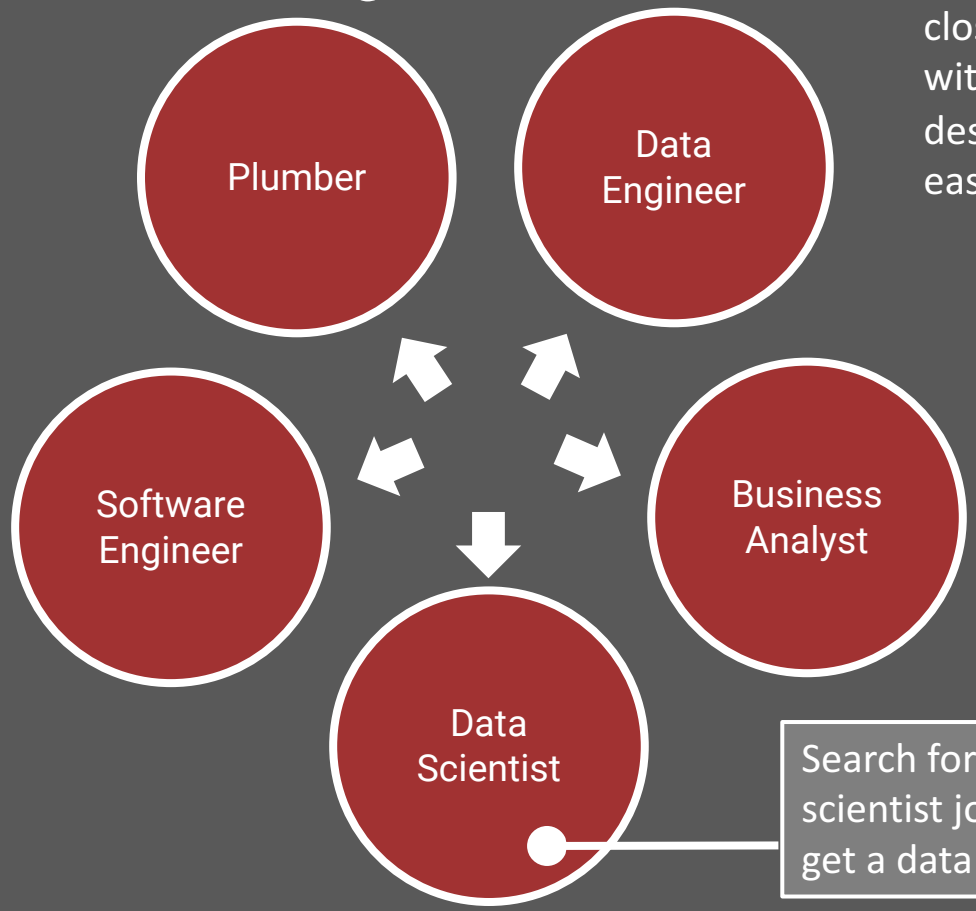
Run Data science Statistical Models for predictions



Distributed Computing

Job Posting to Predicted Titles Scoring

High Score ~ 1.00

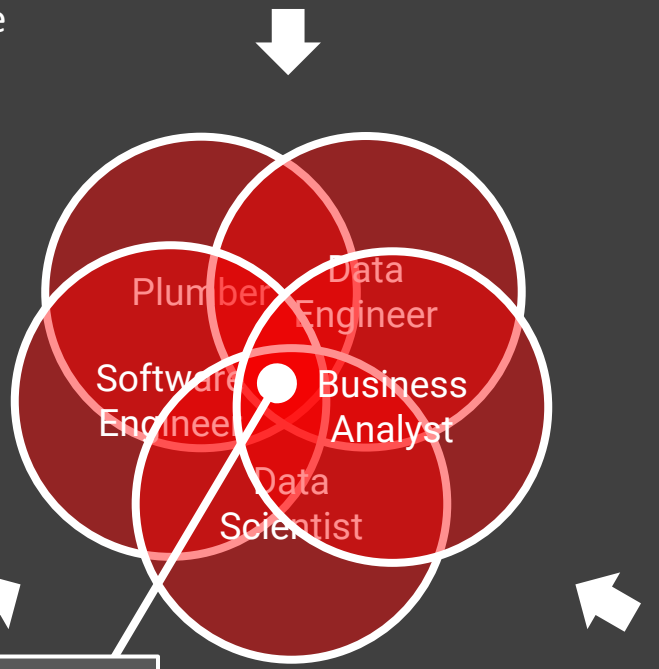


A high score, titles are closely associated with their job descriptions and can easily be separated.

Search for a data scientist job, you will get a data scientist job

Low Score ~ 0.00

At a low score, the titles are not correlated to the job descriptions, a.k.a there is large overlap between different job postings

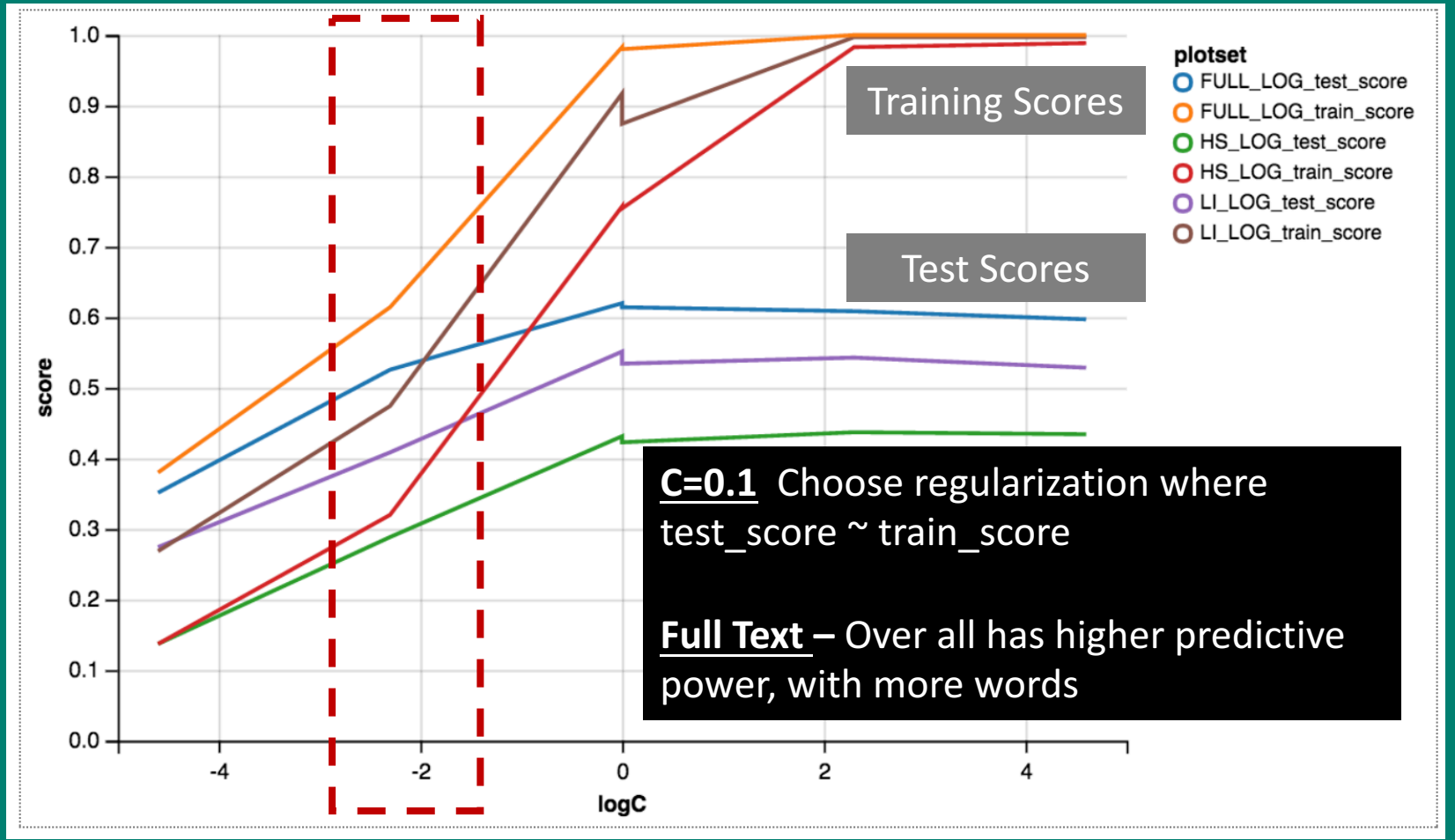


Search for a data scientist job, and no clue what type of job you will get?

Logistic Regression Score ~0.60

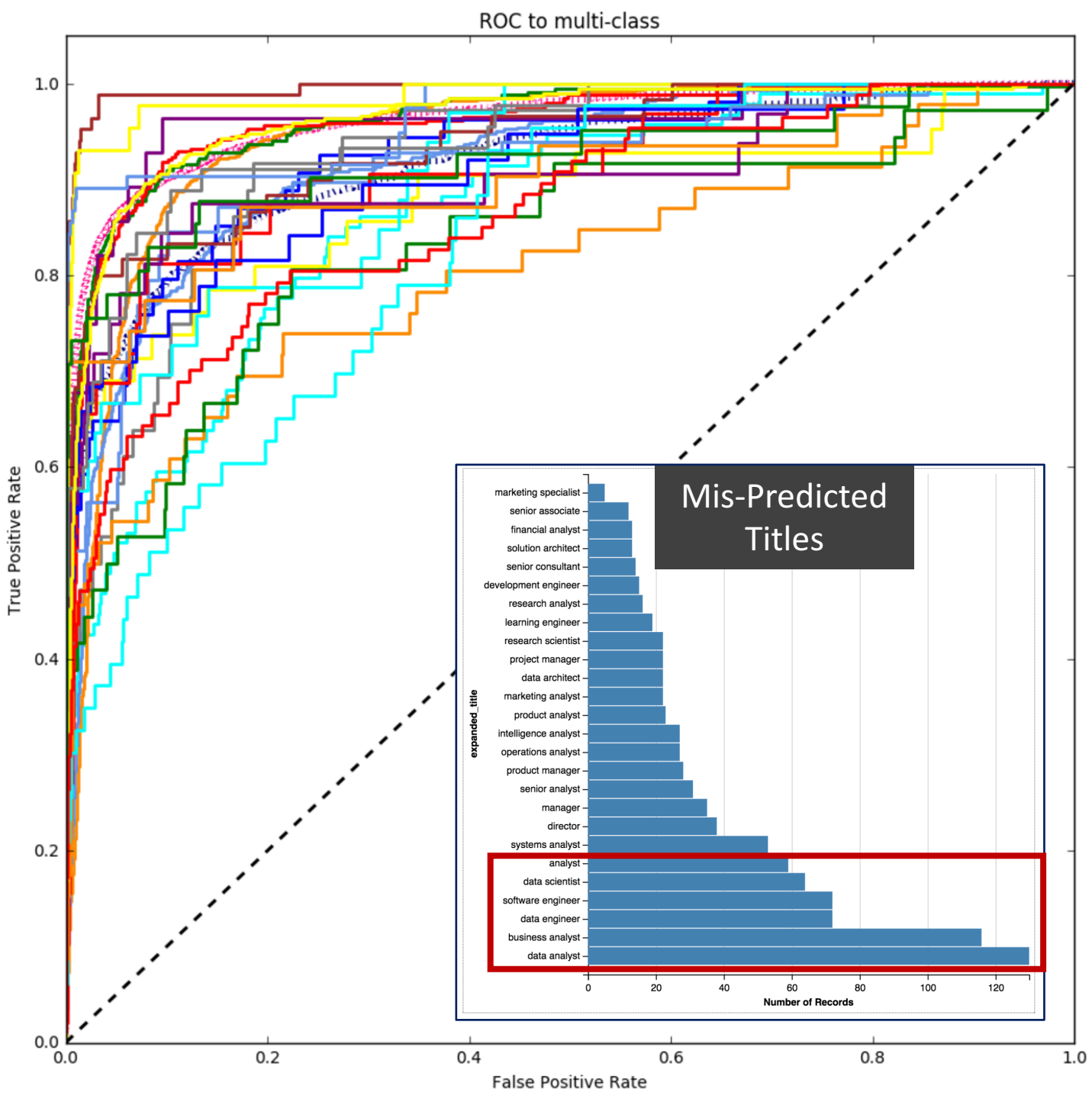
- Multi-class logistic regression
- Classification scores by Log C
- C = regularization constant

C	Log C
0.01	-2
0.1	-1
1.0	0
10.0	1
100.0	2



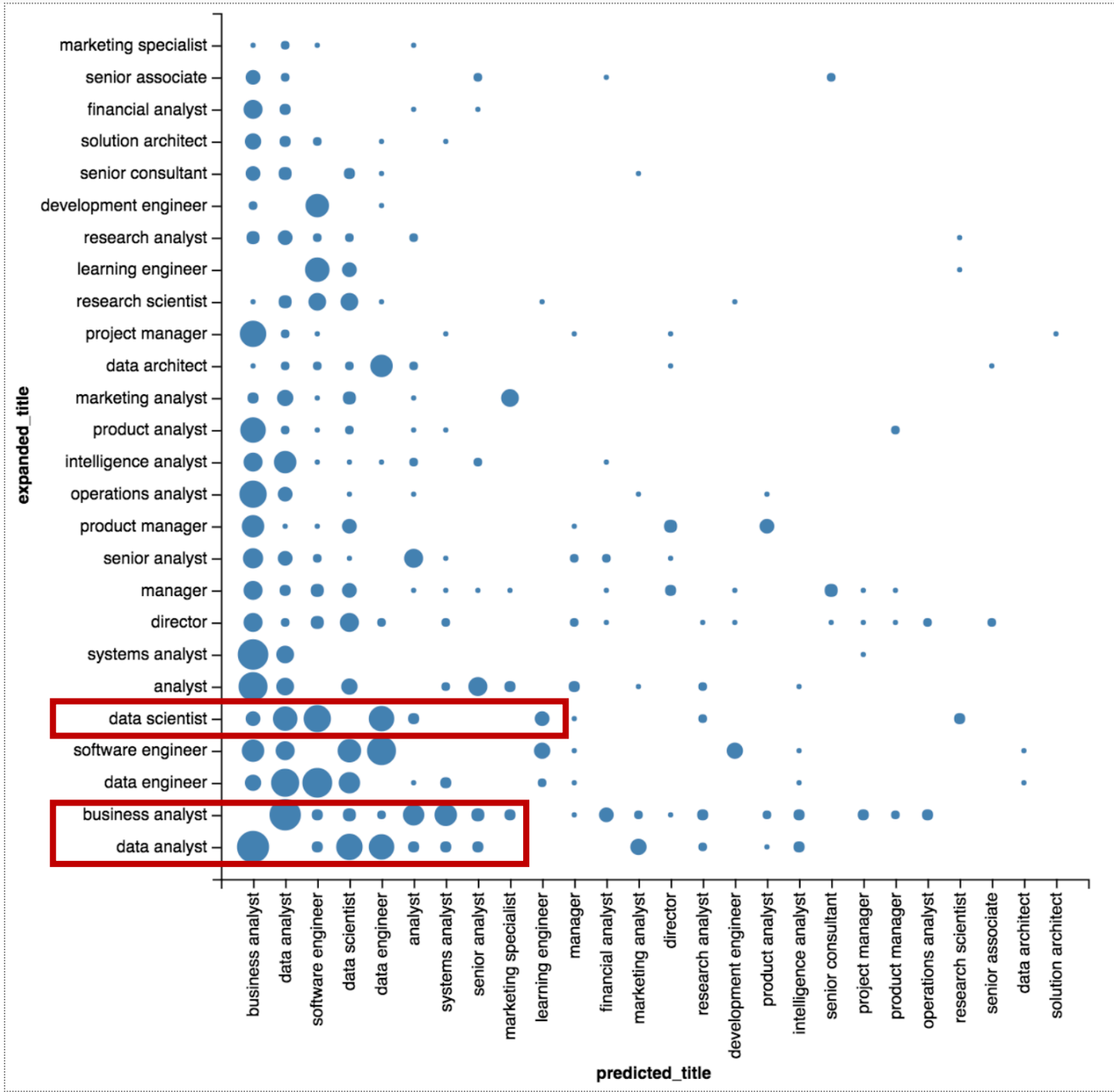
ROC Curve

- To the right is the ROC curves for all 26 different predicted titles.
- The poorest performing titles:
  - business analysts
  - data analysts
  - data engineer
  - Software engineer
  - data scientist



What wrong titles were assigned?

- All the ”analyst” positions are very similar
- Data scientist mis-predicted as these:
  - data analyst,
  - software engineer,
  - data engineer,
  - learning engineer,
  - business analyst,
  - research scientist



Takeaways

- **Data Scientist industry is partially defined, but not completely:** job postings have many same characteristics with data engineering, learning engineer, or data analyst
- **Unfortunately data analyst titles are general :** and are intermingled with all other analysts, system analyst, business analyst, senior analyst
- **Data science is out there, though small:** Even though CA is the biggest data science hub, there are smaller pockets of data science in the US even throughout the US. An additional time analysis can measure the growth of data science

Questions?

Thank you for your time!

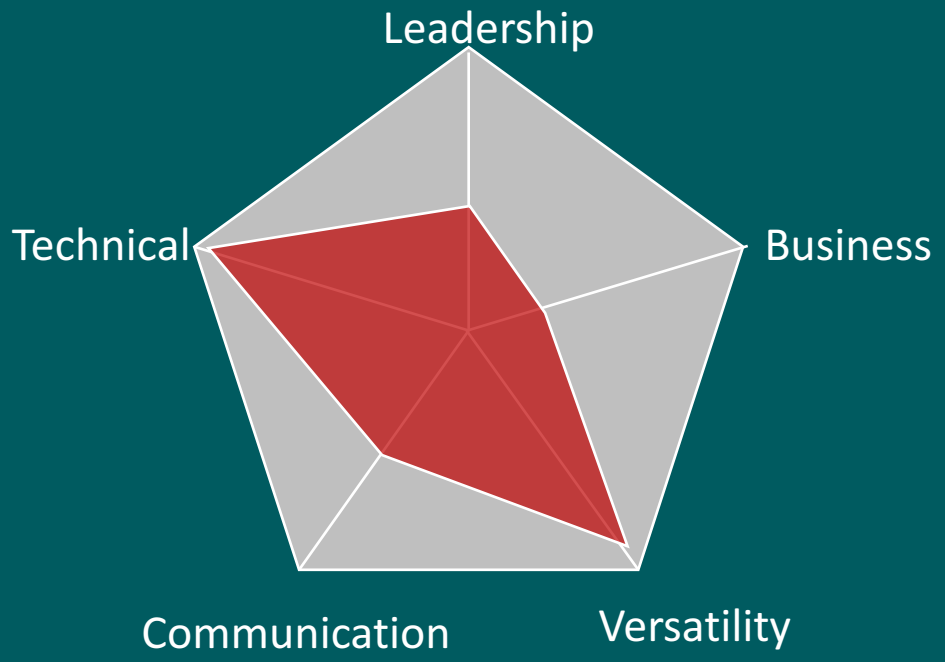
Otherwise … on to bonus materials

Next Steps:

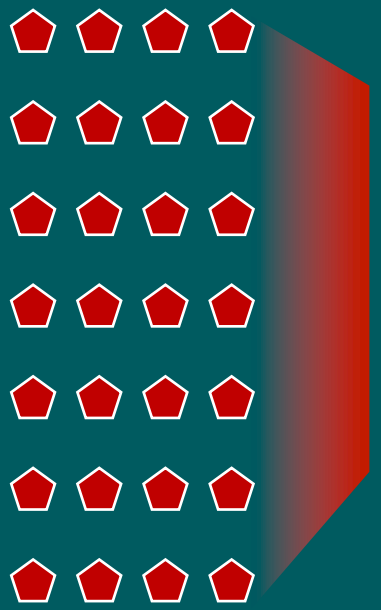
Step 1: Similar to sentiment analysis…

Admires	+1
Disappointed	– 2.5
Encouraged	+1
Hopeful	+ 0.6
Great	+1.2
Lackluster	-0.6
Letdown	-1
Poor	-5
…	

Step 2: Make New Metrics Per job Posting



Step 3: Look at all postings



For a Job Title:

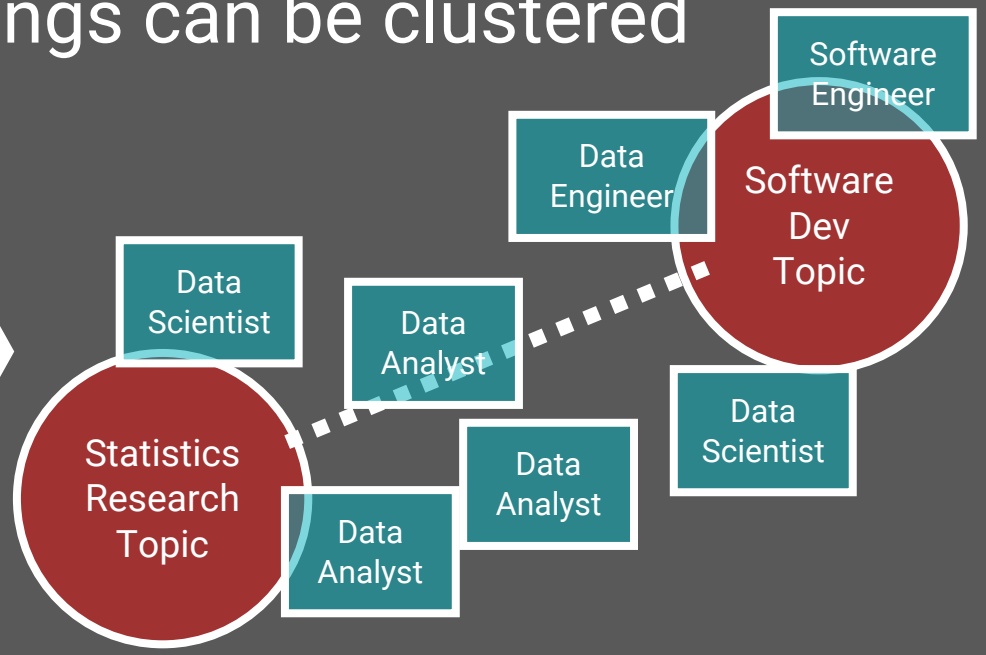
- Avg Stats
- Dev Stats
- Outliers
- Trends
- Not limited by titles
- Easier comparison to other postings

Next: NLP Topic Modeling for Better Recommendations

- LDA Topic modeling was performed with full text, line item, and hard skill excerpts.
- The topic probabilities are then used to build an recommendation engine, and job postings can be clustered around topics

['Experience',
'working',
'in',
'the',
'consumer',
'web/internet',
'domain',
'Experience',
'with',
'Javascript',
'CSS',
'HTML',
'D3',
'Highcharts',
'Experience',
'with',
'statistical',

LDA
Latent Dirichlet Allocation



Sample Topics Extracted from Different Data Sources

	Full Job Posting Source Materials				Hard Skill Posting Source Materials			
General Topic	Generic Business	Machine Learning Software	Data / Stats Analyst	Web Developer	Generic Business	Big Data / Tech Heavy	Consulting /	Reporting / Databases
Top Words, ordered by topic importance	business +requirements+ experience + project + management + skills + ability + work')	experience + learning + data + team + machine + work + software + science')	data + analytics + business + experience + analysis + statistical + insights + skills')	experience + software + development + design + technical + systems + applications + web')	business + requirements + project + process + development + management + technical + functional')	data + java + hadoop + big + experience + systems + python + technologies')	environment + fast + paced + work + business + team + data + effectively')	data + business + sql + reporting + tools + analysis + analytics + reports')

Data Scientist N-grams – Hard Skills

Words	
data	business
experience	large
learning	computer
scientist	strong
work	**python**
statistics	ability
analysis	**algorithms**
environment	processing
machine	years

2 N-gram	
Machine Learning	Understand algorithms
Computer science	Amounts data
Big data	Mathematical viewpoint
Data mining	Methods mathematical
Data science	Linux environment
Operations Research	Viewpoint strong
Algorithm Methods	Processing large
Intuitive viewpoint	Statistics mathematics
Large amounts	Environment processing

3 N-gram	
Large amounts data	Computer science statistics
Methods mathematical viewpoint	Large machine learning data
Algorithms methods mathematical	Data cloud environment
Viewpoint intuitive viewpoint	Work linux environment
Understand algorithms strong	Operations research related
Processing large amounts	Communicate results educate
Environment processing large	Engineering bioinformatics physics

Data Scientist Words (1,200 postings)

Full Posting	
data	**statistical**
experience	**models**
business	**scientist**
Work	ability
team	years
learning	strong
science	working
machine	product
analytics	**statistics**

Line Items	
data	science
experience	skills
learning	years
ability	**sql**
machine	**statistical**
work	**statistics**
strong	**python**
business	knowledge
analysis	large

Hard Skill	
data	business
experience	large
learning	computer
scientist	strong
work	**python**
statistics	ability
analysis	**algorithms**
environment	processing
machine	years

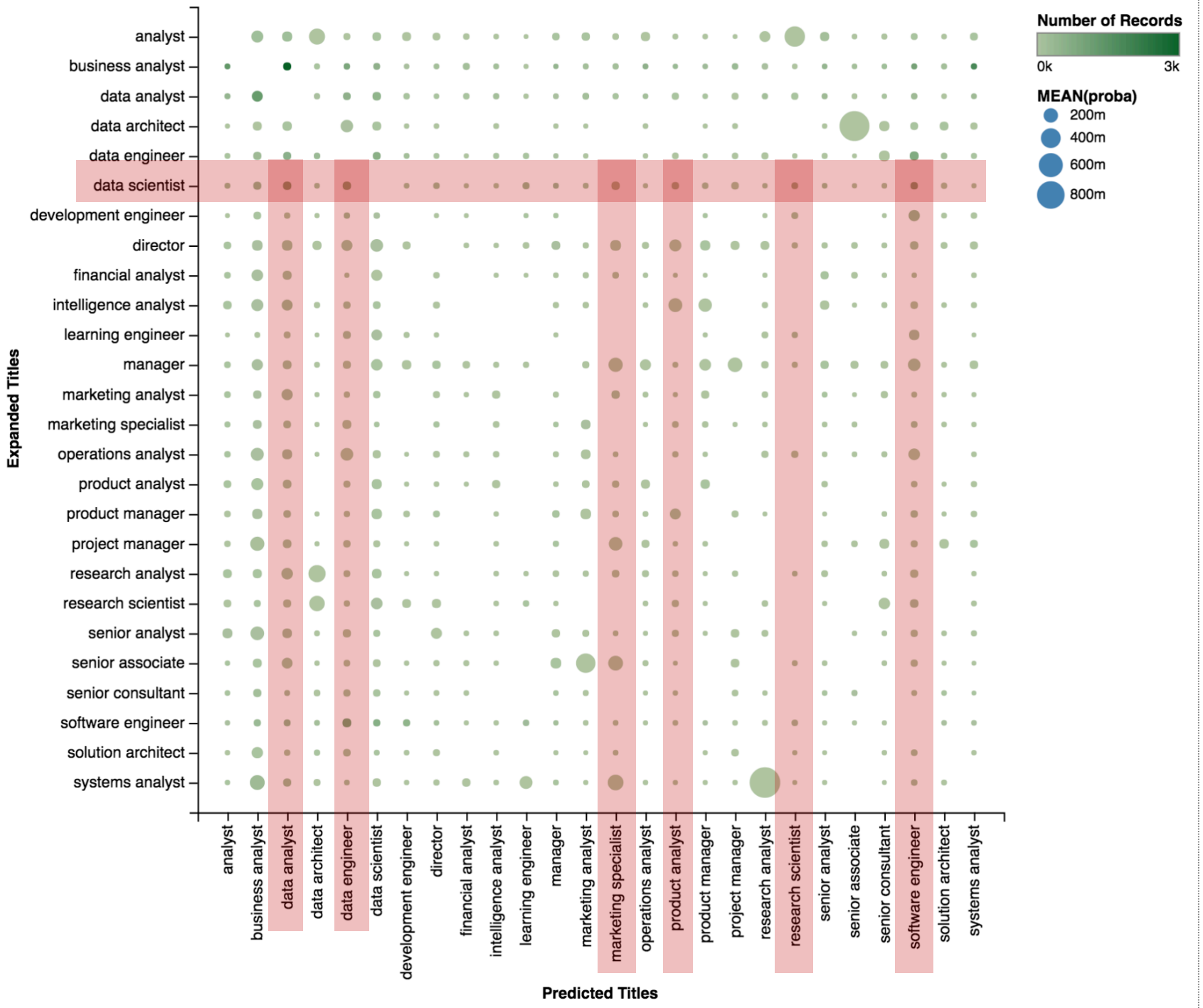
Data Scientist CountVec vs. TFIDF

Count Vect	Hard Skill
data	business
experience	large
learning	computer
scientist	strong
work	**python**
statistics	ability
analysis	**algorithms**
environment	processing
machine	years

TFIDF	Hard Skill
data	computer
considered	**python**
experience	analysis
work	years
learning	business
statistics	quantitative
science	strong
machine	environment
subject	staistical

Title Similarity

- After using logistic regression, we have a probability table.
- Data scientist has relations to:
  - Data analyst
  - Data Engineer
  - Marketing Spec.
  - Product Analyst
  - Research Sci.
  - Software Eng.



About Me

- **Other Projects**
  - Data Science
    - Kaggle : HS Grad Rates
    - Kaggle : West Nile Virus
    - Movie Poster Recognition*
  - Reporting/Analytics
    - Enterprise Contra Discount ReEngineering
    - M&A SAP Financials slicer
    - MS SQL to Oracle ETL library
    - JQuery P&L dashboard
    - Balance Sheet dummy (accounting) data generator

- **Interests**
  - Making things faster/more efficent:
  - Image Recognition
  - Neural Networks
  - Audio Recognition
  - Natural Language Processing
  - Big Data / Spark
  - Python
  - Scala

Tim Lee
MS + BS. Mech Eng.
@ UCLA
G.A. Data Science Alum

Previous Hats:
Data Analytics Manager
@ PWC SJ

Stress Engineer
@ Rolls Royce

ERP Sales Engineer
@ GoEngineer

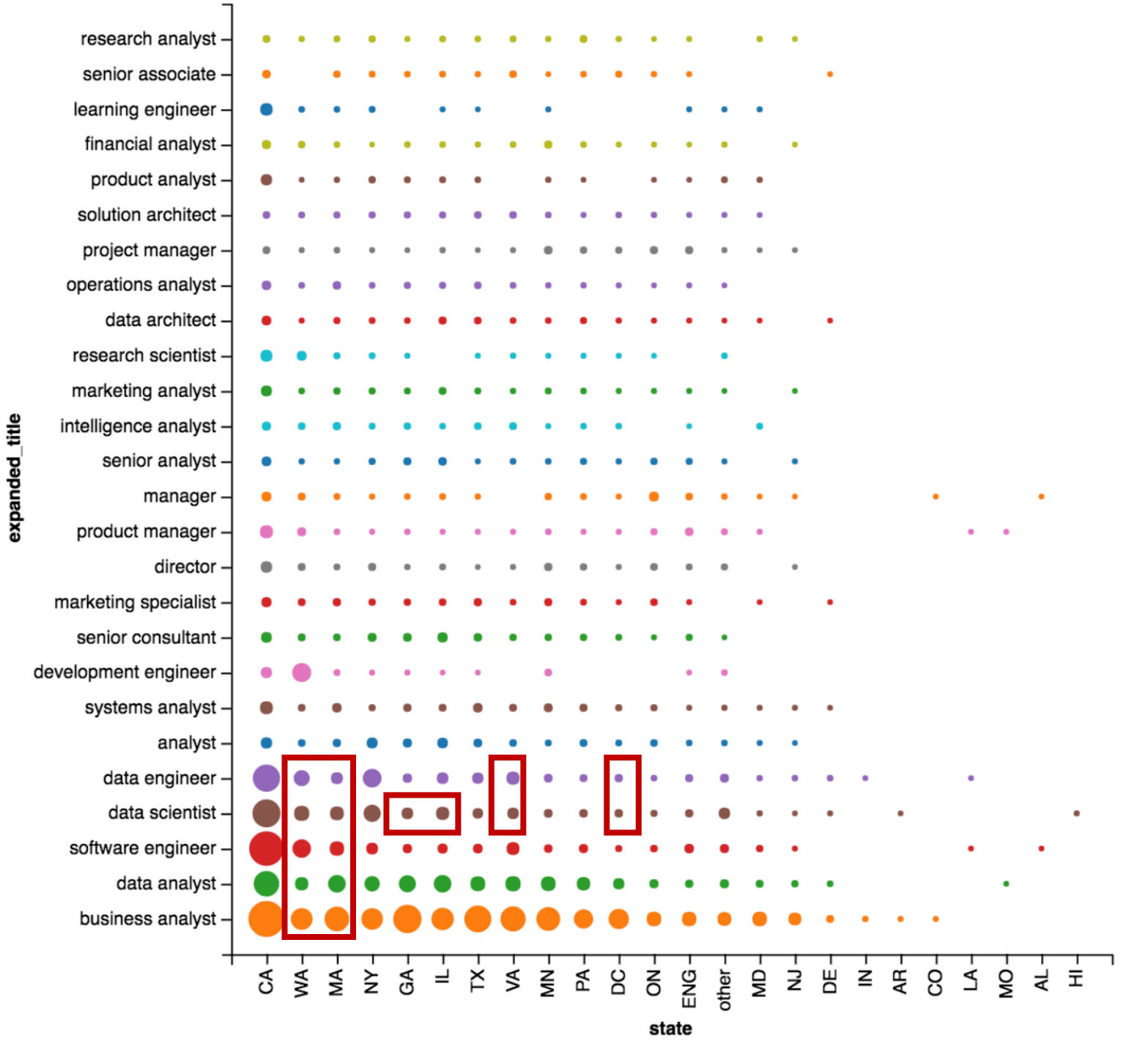
Free Lance Webdesigner

Appendix

Additional Charts

Job Titles by State

- Aside from CA, NYC, there are machine learning opportunities in
  - WA
  - DC
  - MA
  - GA
  - VA
  - TX



Confusion Matrix

- Overall high accuracy. Lowest recall is clustered around the “Analyst” job titles