# Job Recommendations

Patrick Nieto

# Questions

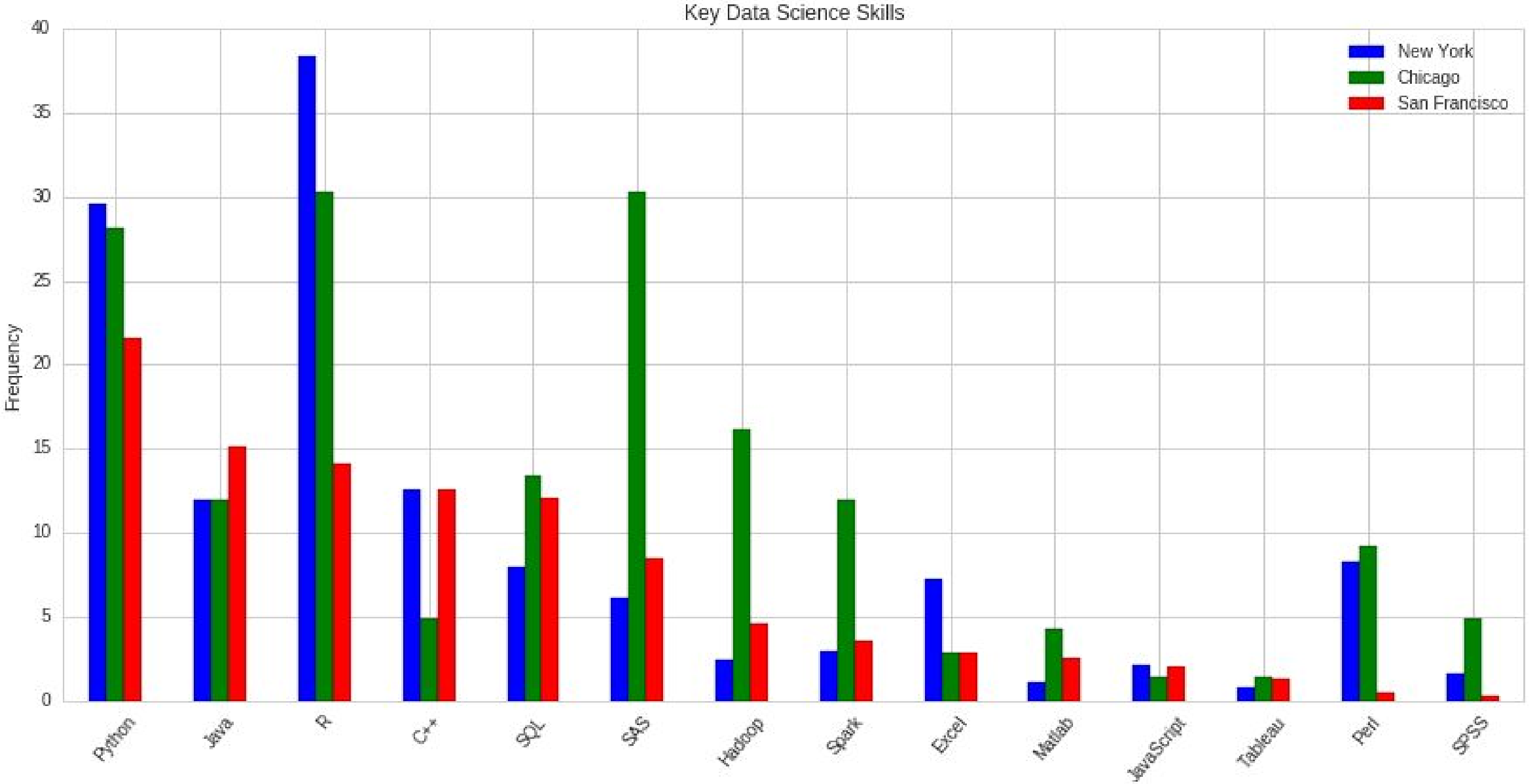
1. Can we determine the most important data science skills?
2. Are there distinguishable language qualities that separate one job description from another?
3. Would it be possible to group job descriptions based on specific tones, sentiment, or topics?

| Aspect | Approach |
|---|---|
| Setup, Design | Spot checking, Individual brainstorming, |
| Data | 10,000 different Data Science job descriptions from Indeed.com |
| Modeling | Unsupervised Learning; Latent Semantic Indexing, Latent Dirichlet Allocation, Approximate Nearest Neighbors, keyword extraction |
| Tools | nltk; gensim; scikit-learn; MongoDB; Flask; AWS |

# Steps:

- Create a generalized web scraping tool.
- Initial understanding and exploratory analysis
- Produce a weighted word vector matrix (TF-IDF)
- Identify patterns and relationships between terms (LSI + SVD)
- Determine observable similarities

# Initial Information Gathering



# TF-IDF → LSI

- Use TF-IDF to include weights in words based on their frequency
- Create a LSI space with 300 dimensions to reduce to after the SVD.
- Comparisons of cosine similarities based on my term vectors

# Similarities using LSI

Percent similarity = 28

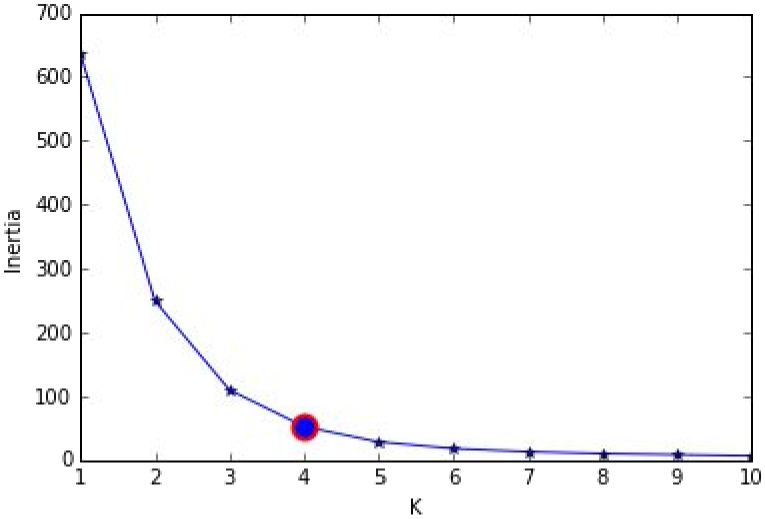
Great location in the heart of Manhattan, Career growth opportunities, competitive salary, plus full benefits including medical, dental, vision, stock, and vacation days (strong work/life balance). You will work on a new innovative team with great engineers that are created a new product with a start-up culture. Company is well-known and is financially stable and profitable. Company has a fully stocked fridge, espresso machine, company sporting activities, weekly happy hour events etc. You would be offered a matching 401K, equity, full benefits for you and your dependents.

We are a well-funded start-up founded in 2011 that has built the leading Machine Learning platform that is used by some of the biggest companies around the world. We are headquartered in California, and we are expanding into the New York City market. We are growing our team right now, but we expect to have a large presence within the year.

Our product is an open source platform that works seamlessly with Hadoop environments to made better predictions with minimal effort. We expect to go public, and are growing to meet that goal

# Dimensionality Reduction

1. Reducing the feature space to 2 dimensions for visualization
2. Then clustering the points for different values of K (number of clusters) to find an optimum value.
3. Running a full LDA transform against the BoW corpus, with the number of topics set to 4



# Word Clouds

Mars, password,
Chocolate', cacao,
Accounts, Enter, functionality,
Dod, start, reset

Loading, ihs, glassdoor, linguist
Splunk, earnest, password, language
Wait, linguistics

Roche, self, sequencing,
Peoplesoft, Caesars
Hart, santa, Oracle
Alto, Palo

Memorial, michigan, cardiovascular,
Arbor, ann, investigator,
Charts, temporary, university,
atherosclerosis

Flask app: http://54.89.23.164/

# Conclusion

- There was as significant tradeoff when choosing the number of dimensions.

- Startups and corporations have unique characteristics in their job descriptions

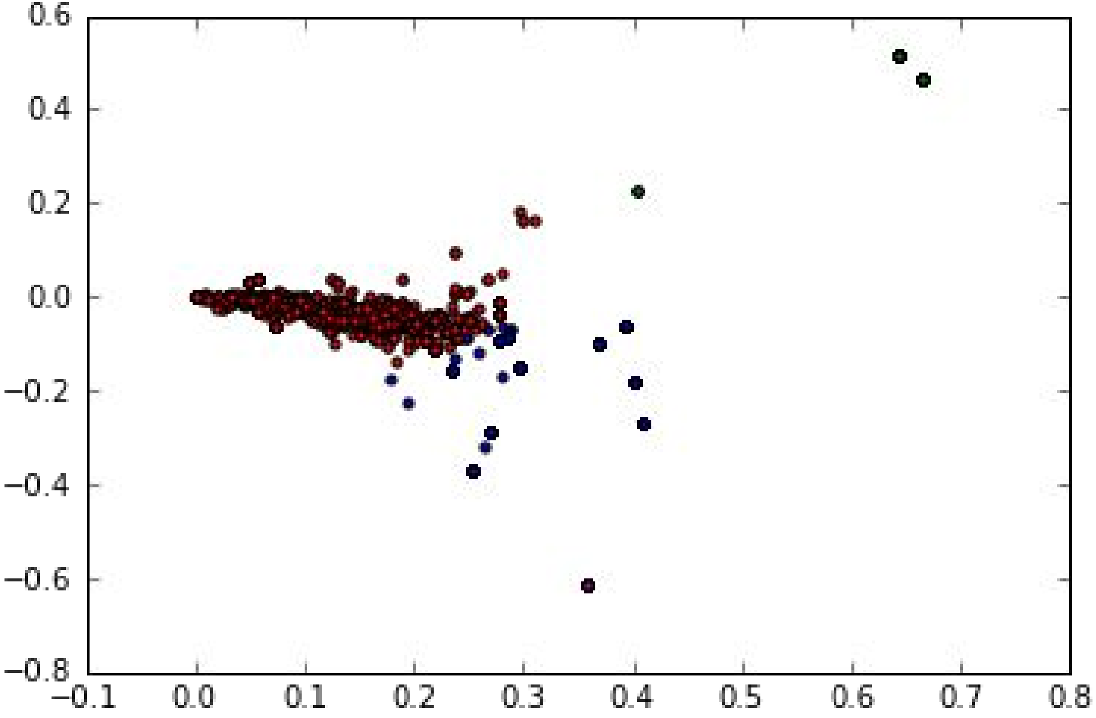
- Python, R and Java are the most frequently mentioned data science skills

# Next Steps

- Create a more refined web scraping tool

- Create a job recommendation tool using a combination of my LSI model and Flask

# Thank you!


0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
−0.8
−0.1
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8