Using Embeddings to Understand the Variance and Evolution of Data Science Skill Sets

Maryam Jahanshahi Ph.D.
Research Scientist
TapRecruit

TapRecruit uses NLP to understand and organize natural language career content

Smart Editor for Job Descriptions

Active Pipeline Health Monitoring

Multifaceted Salary Estimation

tapRecruit

Search

Account

Job | Sync | Similar Jobs | Open | Large Candidate Pool | Applicants: 202 | 3850 Characters | Notify | Last edit: **System**

28

Job will perform poorly

This job scores lower than 95% of **Junior Accounting** jobs in **Los Angeles, CA**

- Add preferred qualifications
- Add more "you" statements
- Perks included
- Equal opportunity statement is included

Neutral — Gendered

Senior Finance Analyst

$76,300 BETA

$65,200 — $98,600

TapRecruit - Los Angeles

TapRecruit is looking for a smart, detail-oriented person to serve as a senior financial analyst. This person will be responsible for supporting the company's FP&A requirements. Responsibilities will include working on TapRecruit Entertainment Group's FP&A model, supporting analysis for long-term planning options, tracking key business operational metrics and producing monthly financial/operation itional FP&A needs, this role will require strong organizational skills to help manage the uide discussions with senior managers across the department and evaluate/implem ojects for top management. This is a dynamic role that serves the finance de report to a Senior Manager of Finance and will routinely interface with TapRecruit's top ma

Language that emphasizes an "intense" or "confusing" environment is known to deter qualified candidates.

Delete

This is an ideal position for an individual who has gained strong investment bank or accounting firm and now seeks to apply those skills to a fast-growing entrepreneurial company. Strong quantitative and excel financial modeling skills are a must. The ideal candidate must be comfortable in a dynamic start-up environment, will bring energy and passion to everything he/she does, and will not be afraid to roll up his/her sleeves to tackle challenging analytical assignments.

This job is full-time, based in Los Angeles. We offer competitive compensation and stock option program.

Responsibilities:

- Provide financial support to News group and overall FP&A / corporate finance department

Language matters in job descriptions

Same title, Different job

Finance Manager Kraft Foods	**Finance Manager** Kraft Foods	
Junior (3 Years) **No Managerial Experience**	**Senior** (6-8 Years) **Division Level Controller** **Strategic Finance Role** **MBA / CPA**	Same Title Required Experience Required Responsibility Preferred Skill Required Education

Different title, Same job

Performance Marketing Manager PocketGems	**Senior Analyst, Customer Strategy** The Gap	
Mid-Level **Quantitative Focus** **Expertise iBanking** **Data Analysis Tools (SQL)** **Consulting Experience Preferred** **MBA Preferred**	**Mid-Level** **Quantitative Focus** **Expertise iBanking** **Relational Database Experience** **External Consulting Experience Preferred** **BA degree in business, finance, MBA Preferred**	Required Experience Required Skills Required Experience Required Skills Preferred Experience Required and Preferred Education

How have data science skills changed over time?

How have data science skills changed over time?

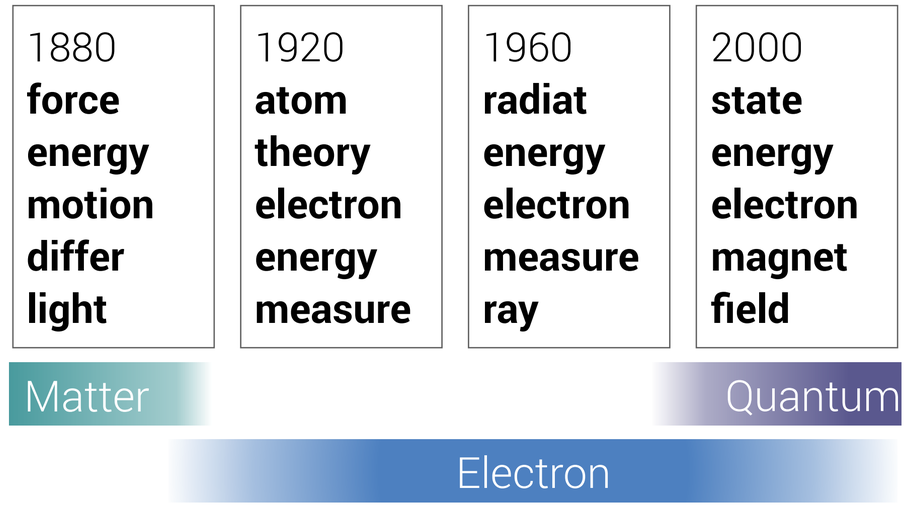
Manual Feature Extraction

MBA | SQL

PhD | Tableau

Python | PowerBI

Dynamic Topic Modeling



Adapted from Blei and Lafferty, ICML 2006.

tapRecruit.co

Word embeddings capture semantic similarities

Statistical modeling through software (e.g. SPSS) or programming language (e.g. **Python**)

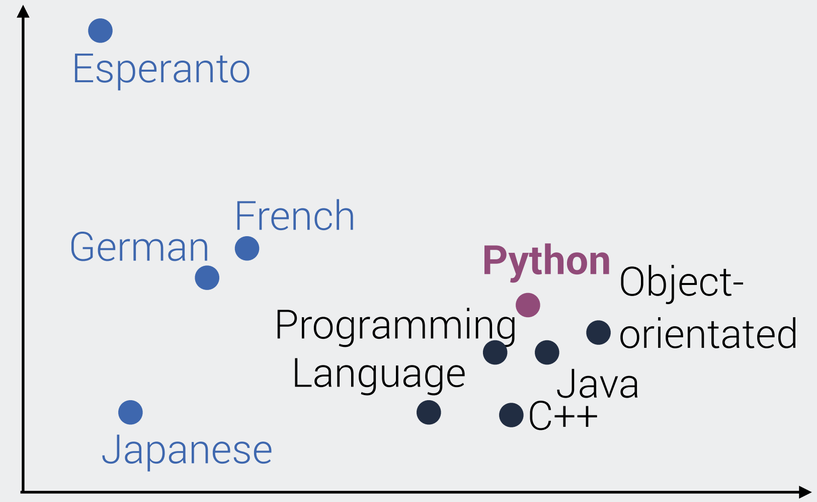
Context **Word**

Experience in **Python**, Java or other object-oriented programming languages

Context **Word** **Context**

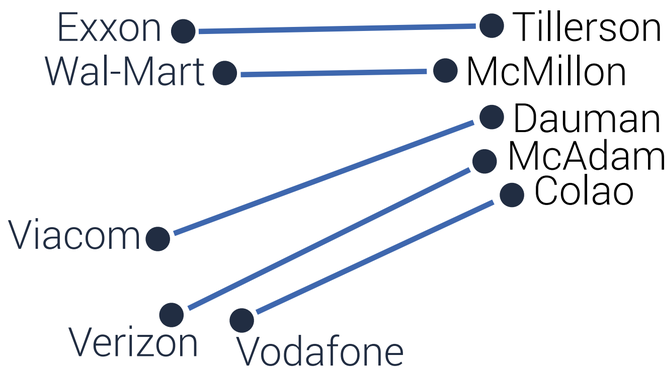
Proficiency programming in **Python**, Java or C++.

Context **Word** **Context**

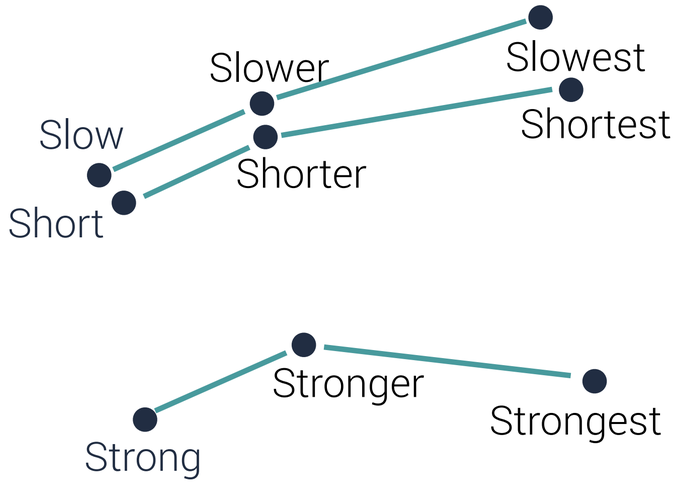


Embeddings capture entity relationships

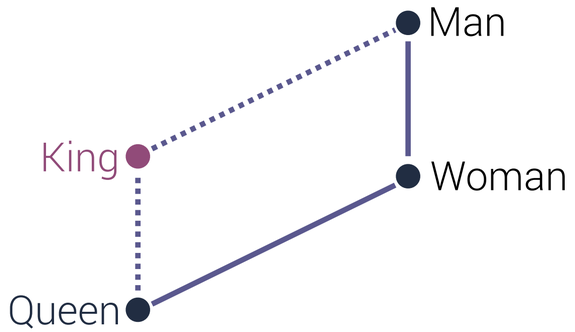
Hierarchies



Comparatives and Superlatives



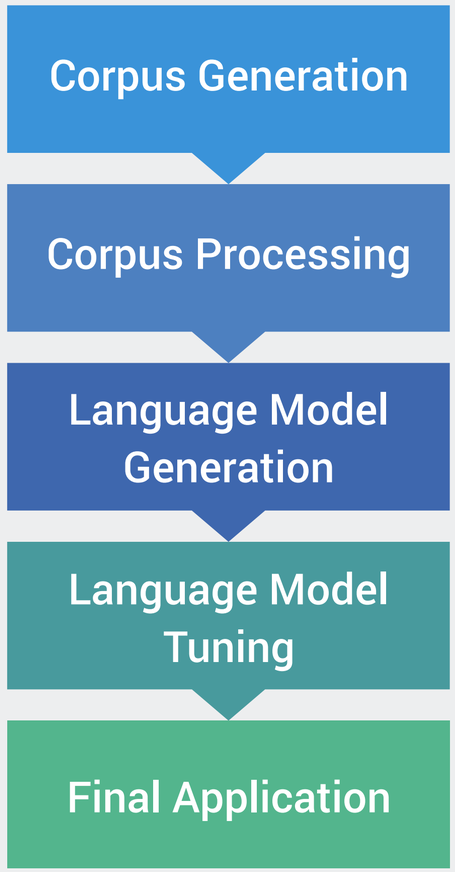
Woman :: Queen as Man :: ?



Adapted from Stanford NLP GLoVE Project

tapRecruit.co

Pretrained embeddings facilitate fast prototyping



Corpus	Twitter	Common Crawl	GoogleNews	Wikipedia
Tokens	27 B	42-840 B	100 B	6 B
Vocabulary Size	1.2 M	1.9-2.2 M	3 M	400 k
Algorithm	GLoVE	GLoVE	word2vec	GLoVE
Vector Length	25 - 200 d	300 d	300 d	50 - 300 d

Problems with pretrained embedding models

Casing	Abbreviations vs Words e.g. IT vs it
Out of Vocabulary Words	Domain Specific Words & Acronyms
Polysemy	Words with multiple meanings e.g. drive (a car) vs drive (results) e.g. Chef (the job) vs Chef (the language)
Multi-word Expressions	Phrases that have new meanings e.g. Front-end vs front + end

Tools for developing custom language models

Corpus Processing

Tokenization, POS tagging, Sentence Segmentation, Dependency Parsing

spaCy

OpenNLP

CoreNLP

SyntaxNet

Language Modeling

Different word embedding models (GLoVE, word2vec, fastText)

gensim

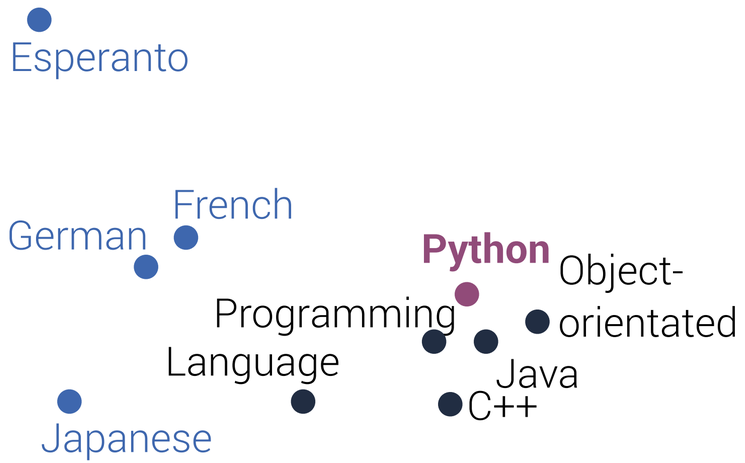
Amazon SageMaker

TensorFlow

Windows capture semantic similarity vs relatedness

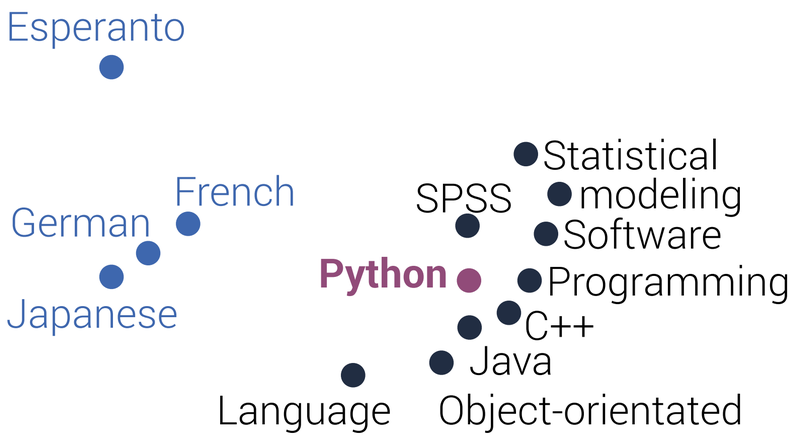
Small Window Size

Captures Semantic similarity, Substitutes and Word-level differences

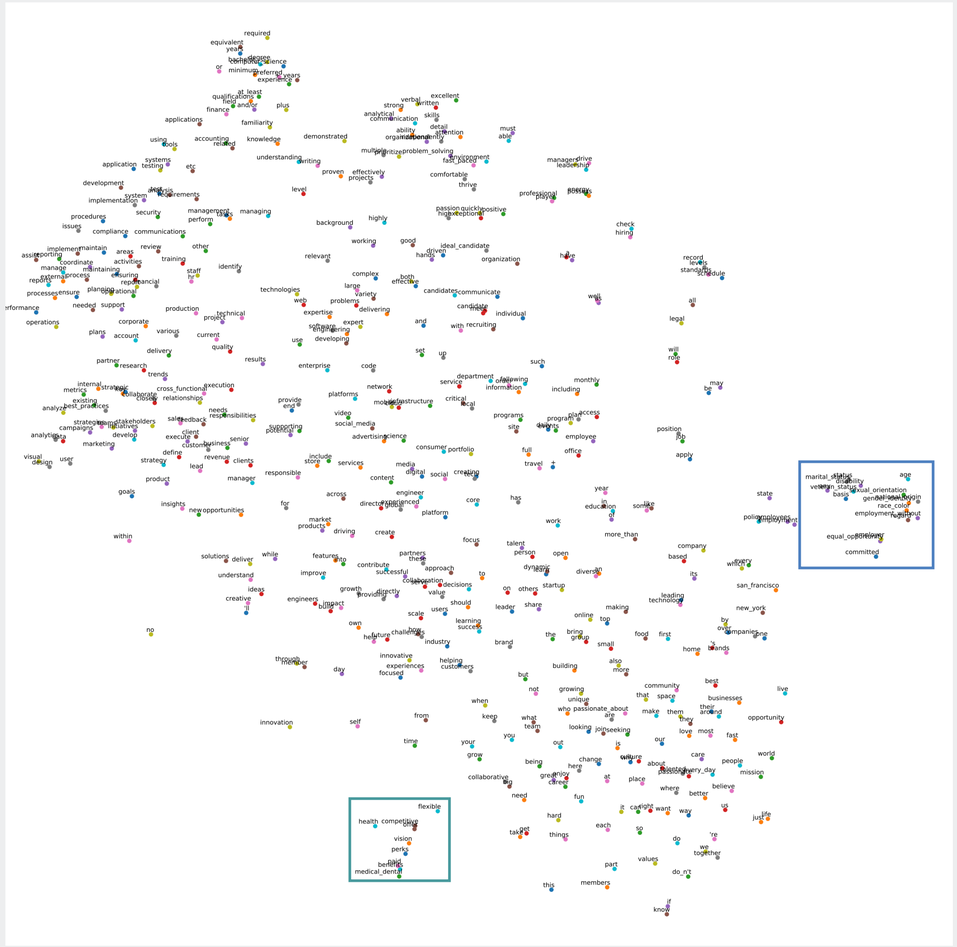


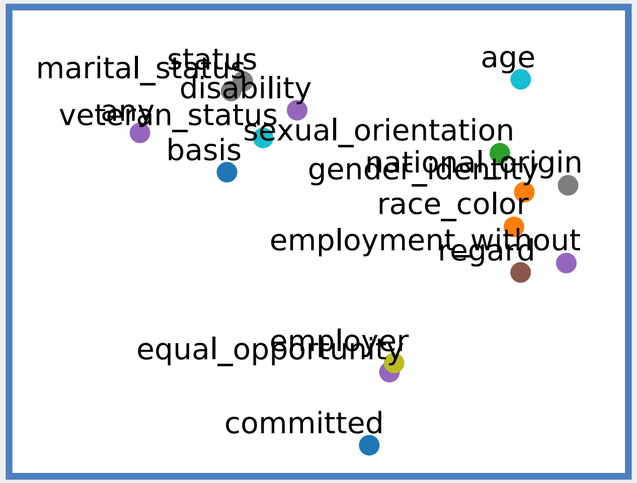
Large Window Size

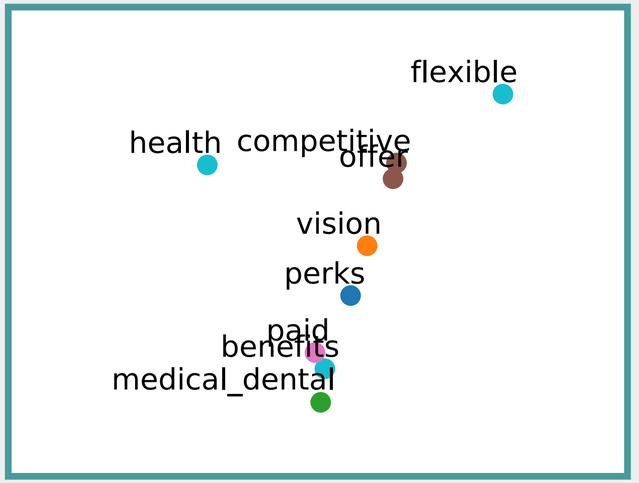
Captures Semantic relatedness, Alternatives and Domain-level differences



Custom embeddings identified equal opportunity and perks language

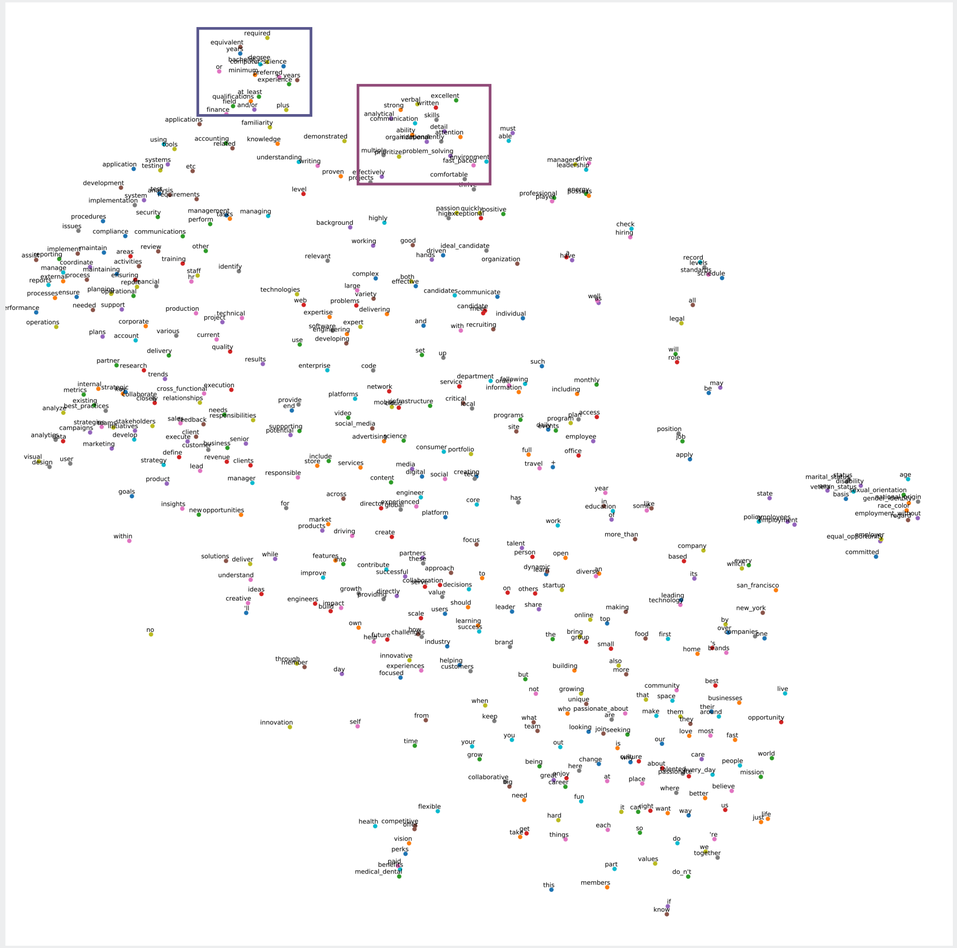


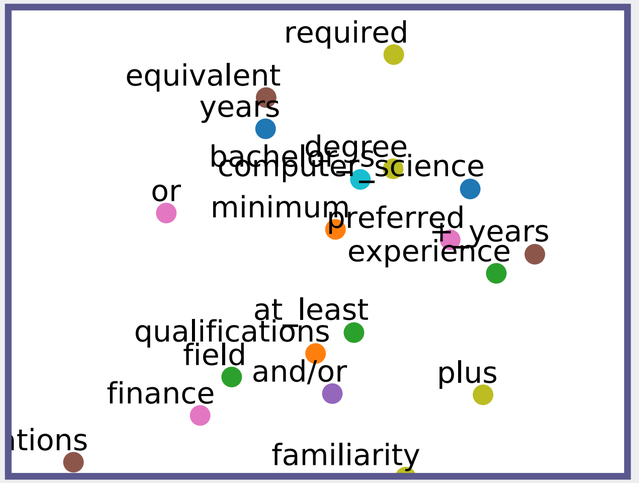




tapRecruit.co

Custom embeddings identified ‘soft’ skills and language around experience



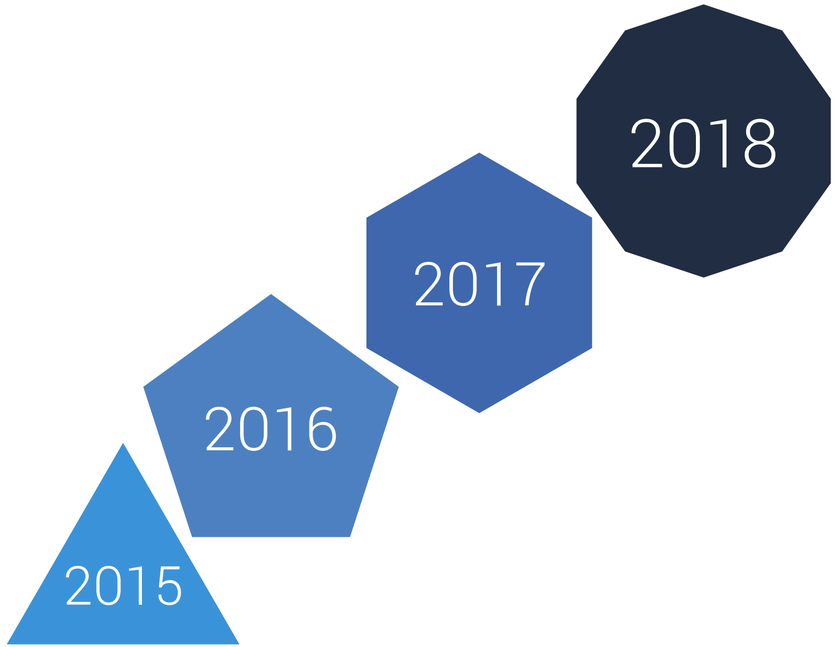


tapRecruit.co

I’ve got 300 dimensions…
but time ain’t one

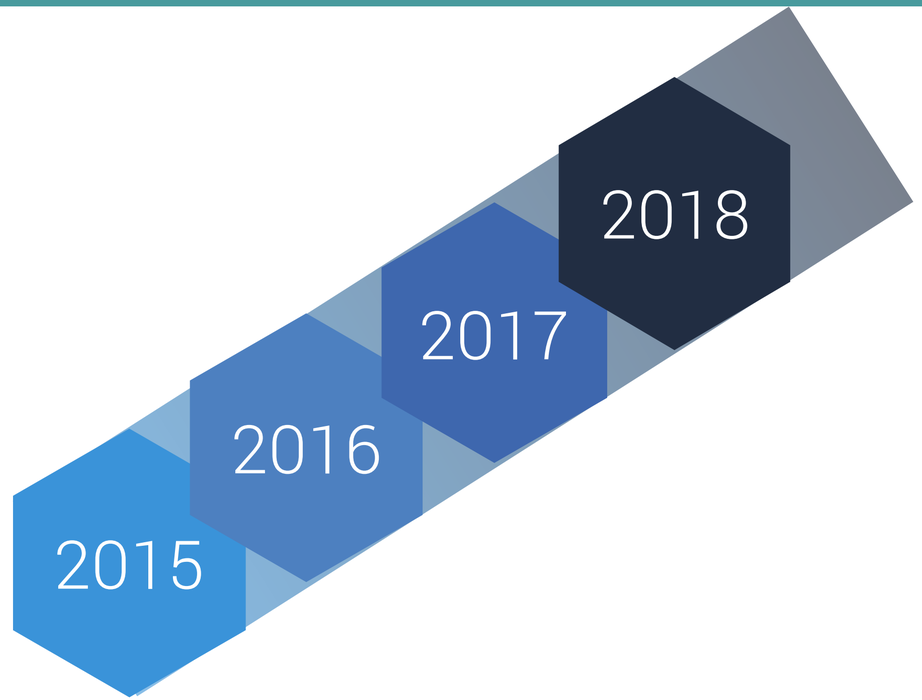
Two flavors of dynamic embeddings

Independently Trained



Kim, Chiu, Kaneki, Hedge and Petrov, arXiv: 1405:3515.
Kulkarni, Al-Rfou, Perozzi and Skiena, arXiv: 1411:3315.
Hamilton, Leskovec and Jurafsky, arXiv: 1605:09096.

Trained Together



Rudolph and Blei, arXiv: 1703.08052.

The benefits of dynamic Bernoulli embeddings

Dynamic embeddings

Data hungry: Sufficient data for each time slice for a quality embedding.

Requires stitching: Each time slice is trained independently, therefore dimensions are not comparable across slices.

Dynamic Bernoulli embeddings

Data efficient: Treats each time slice as a sequential latent variable, enabling time slices with sparse data.

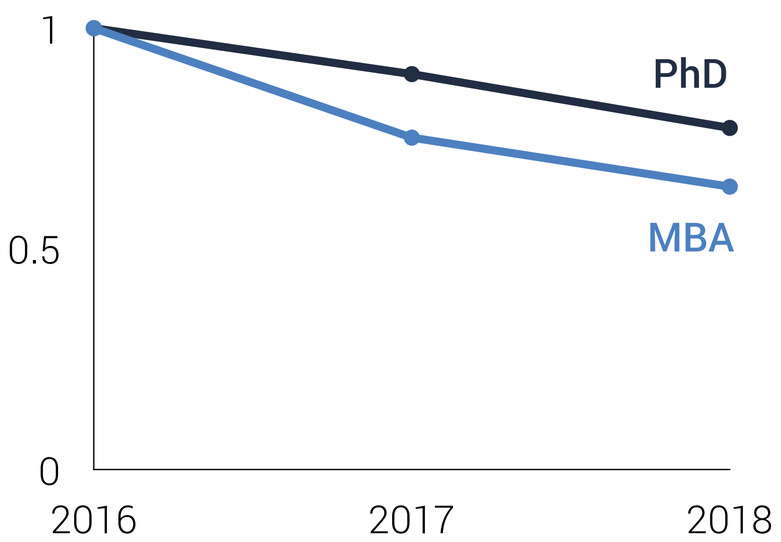
Does not require stitching/alignment: Treating time slice as a variable ensures embeddings are connected across slices.

Kim, Chiu, Kaneki, Hedge and Petrov, arXiv: 1405:3515.
Kulkarni, Al-Rfou, Perozzi and Skiena, arXiv: 1411:3315.
Hamilton, Leskovec and Jurafsky, arXiv: 1605:09096.

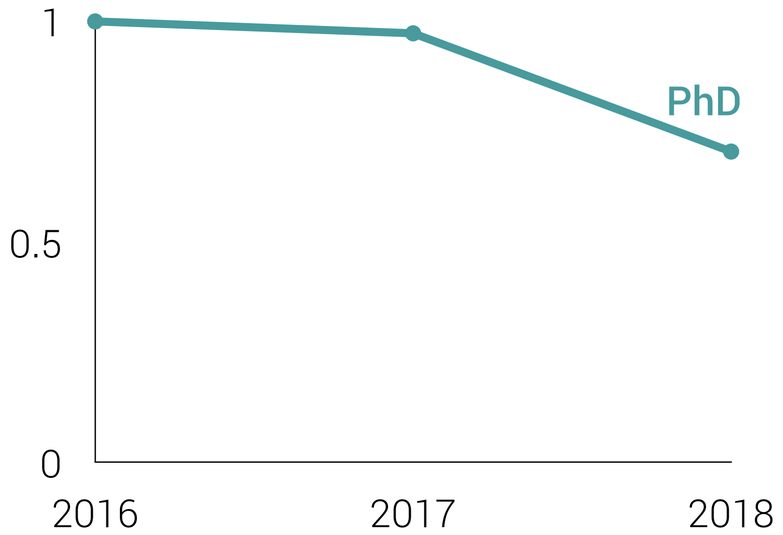
Balmer and Mandt, arXiv: 1702:08359.
Rudolph and Blei, arXiv: 1703:08052.

Demand for MBAs and PhDs falling

All Jobs

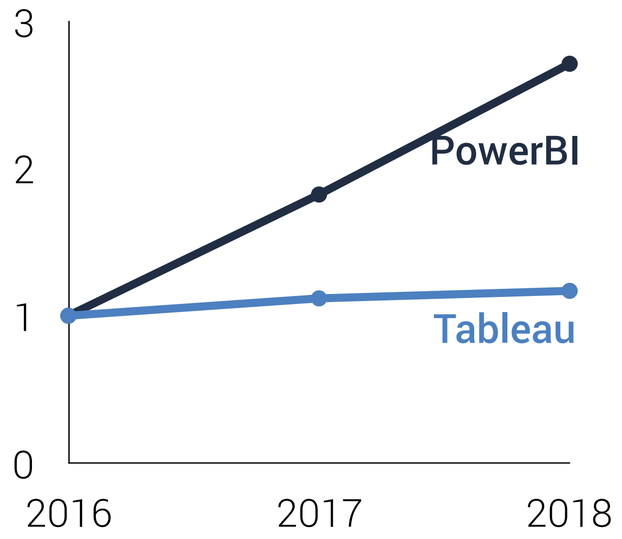


Data Science Jobs

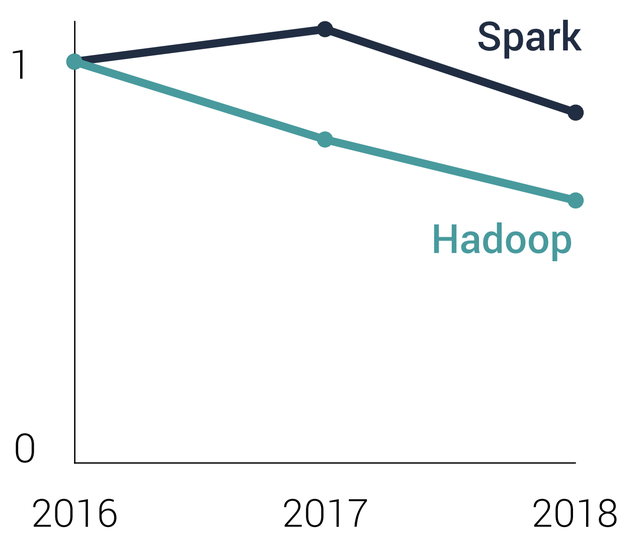


Data Science skills showing significant shifts

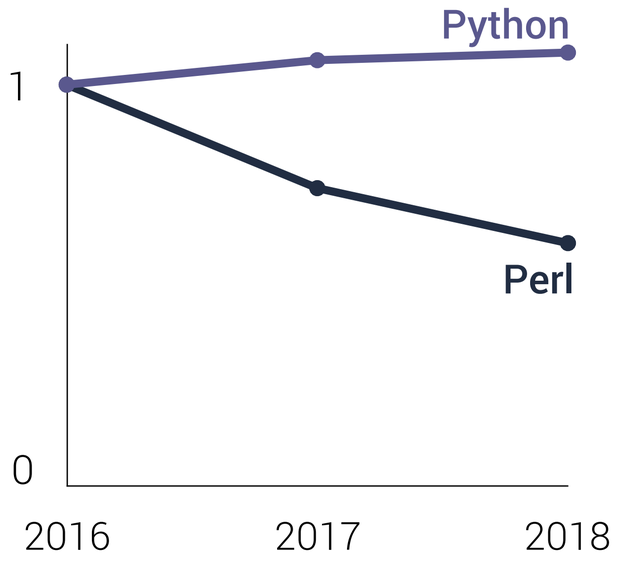
Tableau and PowerBI



Hadoop vs Spark



Python vs Perl



regression :: Generalized Linear Models as word2vec :: Exponential Family Embeddings

Members of the Exponential Family of Embeddings

Bernoulli Embedding	Poisson Embedding	Gaussian Embedding
Binary Data	Count or Ordinal Data	Continuous Data
Proficiency	Mini Bagels	JFK-CDG
Context programming	Context Cream cheese	Context LGA-DCA
Python	Milk	JFK-DFW
Datapoint Java	**Datapoint Coffee**	**Datapoint LAX-JFK**
Context C++	Context Orange Juice	Context LAX-LGA

Poisson embeddings capture item similarities from shopper behavior

Maruchan chicken ramen

Maruchan creamy chicken ramen
Maruchan oriental flavor ramen
Maruchan roast chicken ramen

Yoplait strawberry yogurt

Yoplait apricot mango yogurt
Yoplait strawberry orange smoothie
Yoplait strawberry banana yogurt

High Inner Product Combinations

Old Dutch potato chips & Budweiser Lager beer

Lays potato chips & DiGiorno frozen pizza

Low Inner Product Combinations

General Mills cinnamon toast & Tide Plus detergent

Beef Swanson Broth soup & Campbell Soup cans

Adapted from Rudolph, Ruiz, Mandt and Blei, arXiv: 1608.00778.

How have data science skills changed over time?

- Flavors of static word embeddings: The Corpus Issue
- Considerations for developing custom embedding models
- Flavors of dynamic models: Dynamic Bernoulli embeddings
- Other members of the Exponential Family of Embeddings

Thank you PyData NYC!

Maryam Jahanshahi Ph.D.
Research Scientist
TapRecruit