How have Data Science Skills Evolved? A case study using embeddings

Maryam Jahanshahi Ph.D.
Research Scientist
TapRecruit.co

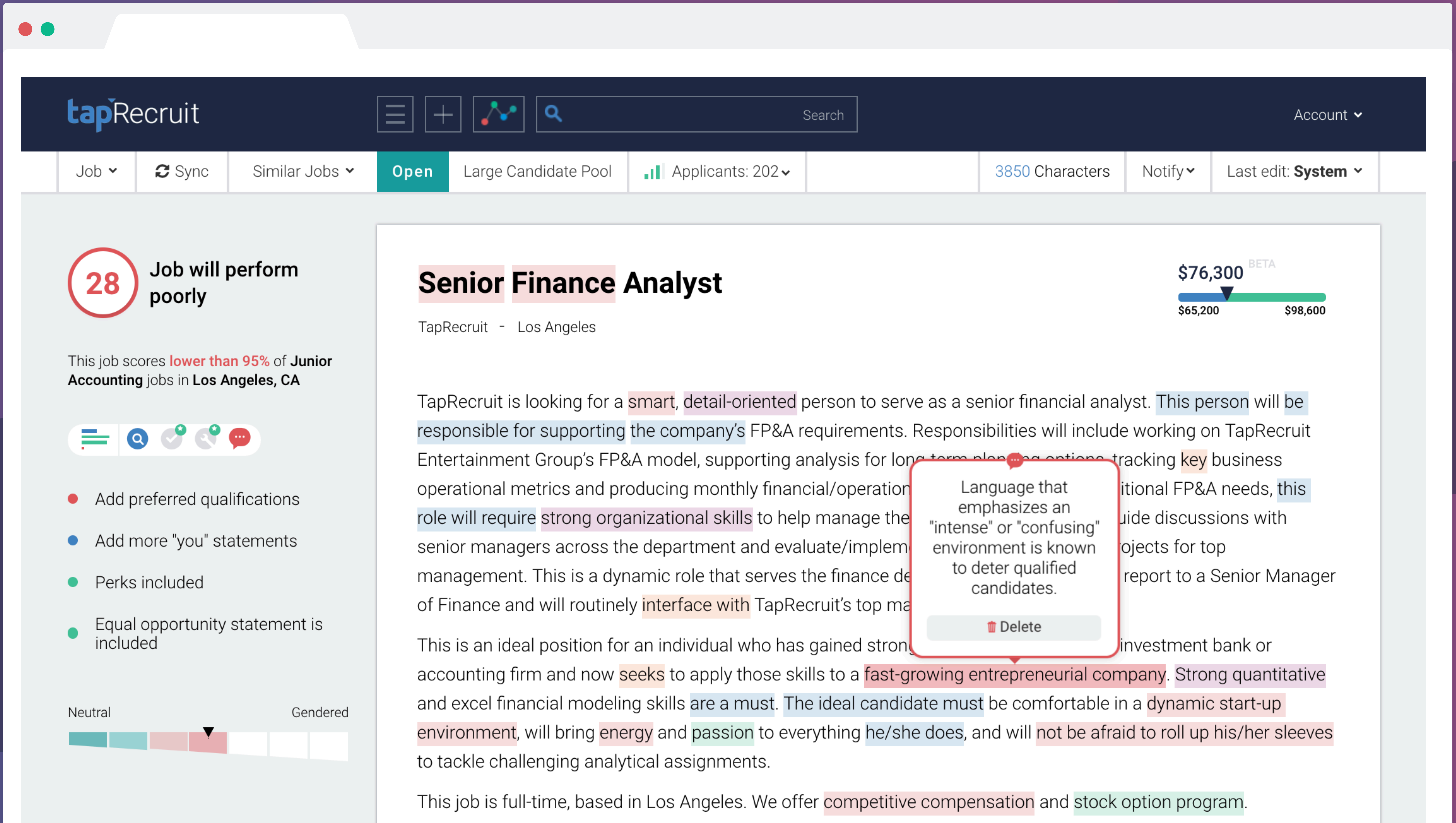
TapRecruit uses NLP to understand career content

Converting unstructured documents into structured data

Smart Editor for JDs
Data-driven suggestions on both the content and language use in job descriptions.

Pipeline Health Monitoring
Analytics dashboards to help diagnose quality and diversity issues in talent pipelines.

Salary Estimation
Data-driven salary estimates based on a job's requirements rather than just title and location.


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
TapRecruit - Los Angeles
$76,300
BETA
$65,200
$98,600
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 lon
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
ide discussions with senior managers across the department and evaluate/impleme
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 stron
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.

Language matters in job descriptions

Same title, Different job	**Finance Manager** **Kraft Foods** Junior (3 Years) No Managerial Experience	**Finance Manager** **Roche** 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** Mid-Level Quantitative Focus iBanking Expertise Data Analysis Tools (SQL) Consulting Experience Preferred MBA Preferred	**Senior Analyst, Customer Strategy** **The Gap** Mid-Level Quantitative Focus Finance Expertise Relational Database Experience External Consulting Experience Preferred BA in Accounting, Finance, MBA Preferred	✓ **Required Experience** ✓ **Required Skills** ✓ **Required Experience** ✓ **Required Skills** ✓ **Preferred Experience** ✓ **Preferred Education**

How have data science skills changed over time?

Strategies to identify changes within datasets

MBA	SQL
PhD	Tableau
Python	PowerBI

1880	1920	1960	2000
force	**atom**	**radiat**	**state**
energy	**theory**	**energy**	**energy**
motion	**electron**	**electron**	**electron**
differ	**energy**	**measure**	**magnet**
light	**measure**	**ray**	**field**

Matter

Quantum

Electron

Manual Feature Extraction:
Require *a priori* selection of key attributes, therefore difficult to discover new attributes

Dynamic Topic Models:
Uses a bag of words approach, and require experimentation with topic number.

Adapted from Blei and Lafferty, ICML 2006.

Word embeddings capture semantic similarities

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

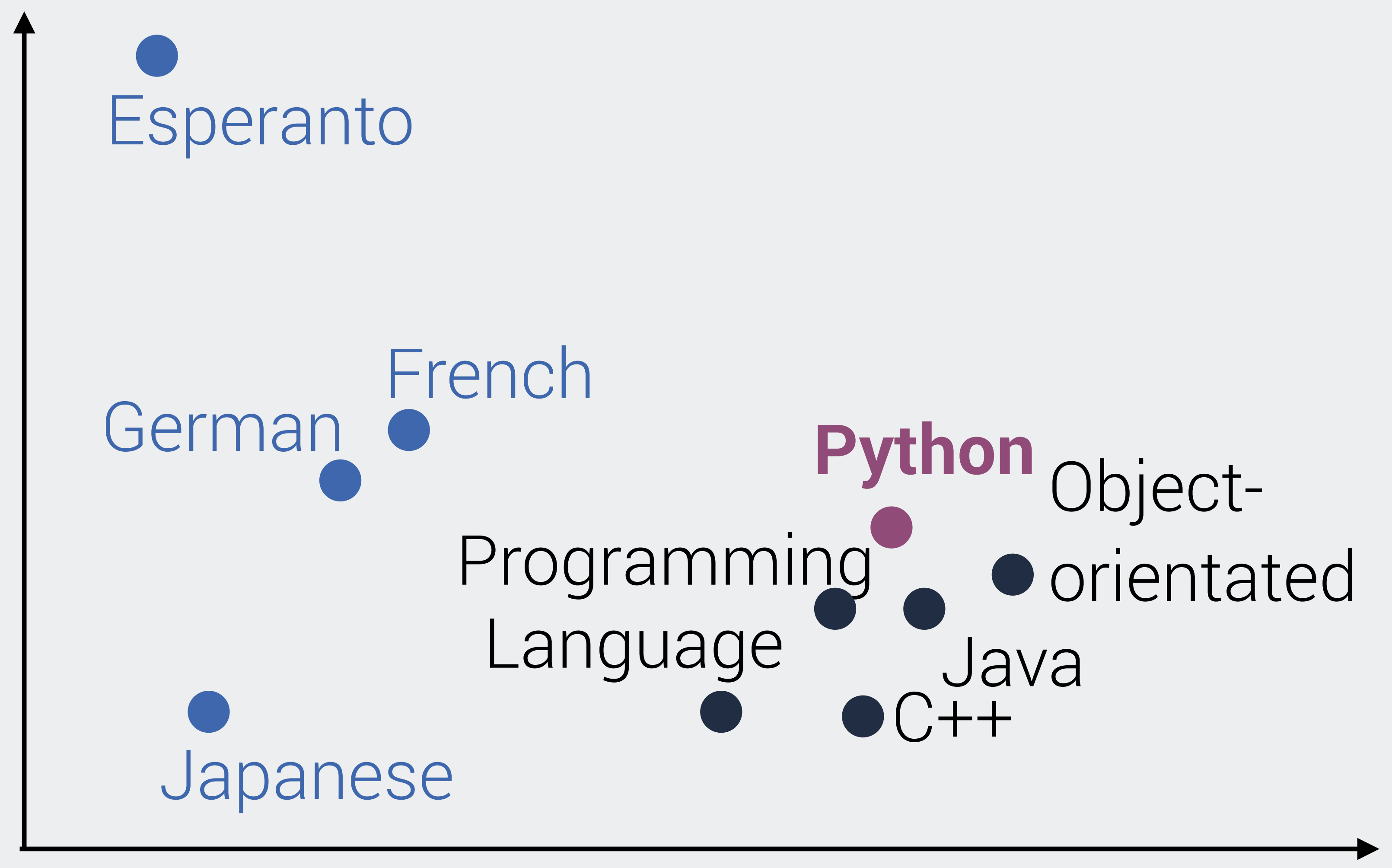
Context **Word**

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

Context **Word** **Context**

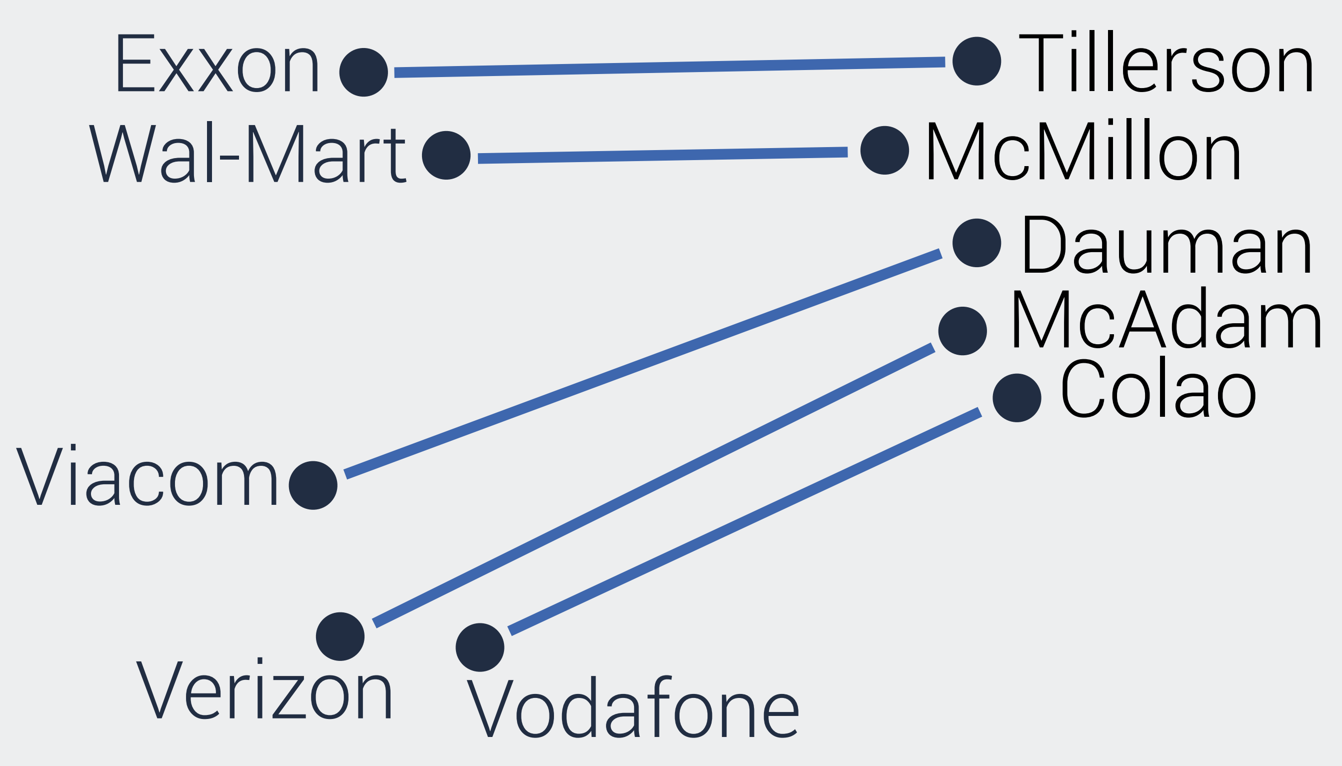
Proficiency programming in **Python**, Java or C++.

Context **Word** **Context**

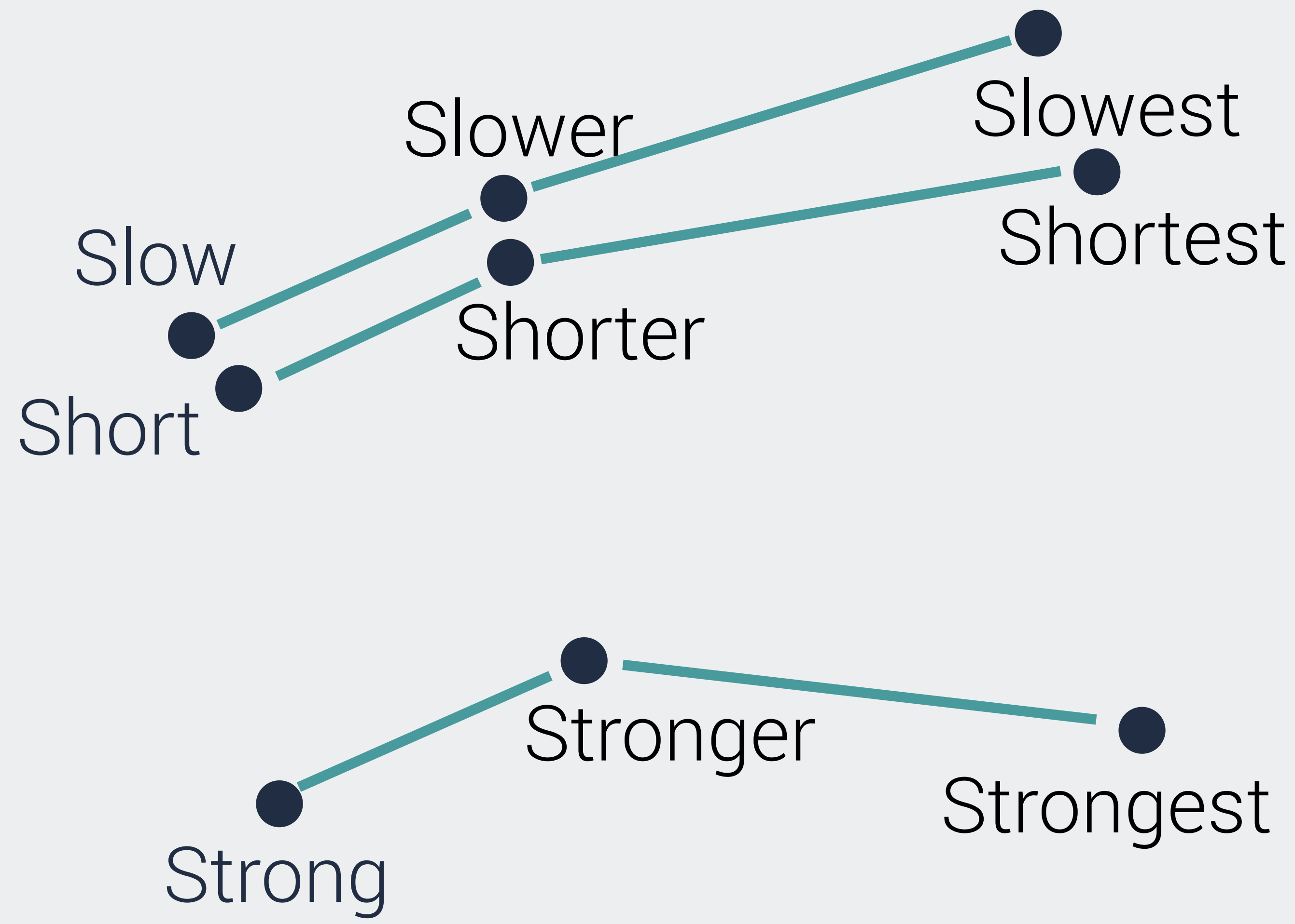


Embeddings capture entity relationships

Dimensionality enables comparison between word pairs along many axes



Hierarchies



Comparatives and Superlatives

Woman :: Queen as Man :: ?

Adapted from Stanford NLP GLoVE Project

Pretrained embeddings facilitate fast prototyping

- Corpus Generation
- Corpus Processing
- Language Model Generation
- Language Model Tuning
- Final Application

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

Modularized for different data and modeling requirements

spaCy OpenNLP

CoreNLP SyntaxNet

Corpus Processing

Tokenization, POS tagging, Sentence Segmentation, Dependency Parsing

gensim PYTORCH

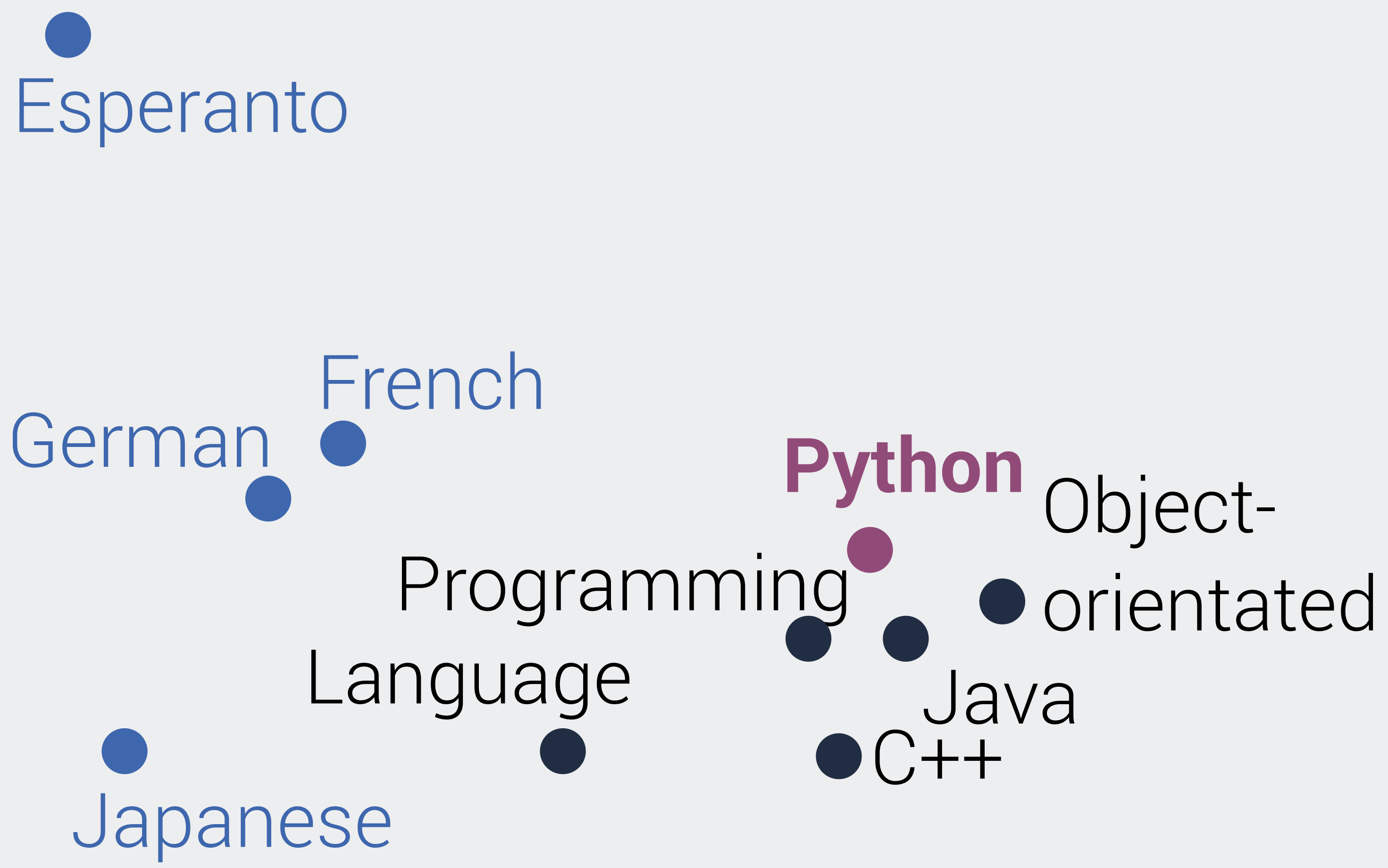
TensorFlow Amazon SageMaker

Language Modeling

Different word embedding models (GLoVE, word2vec, fastText)

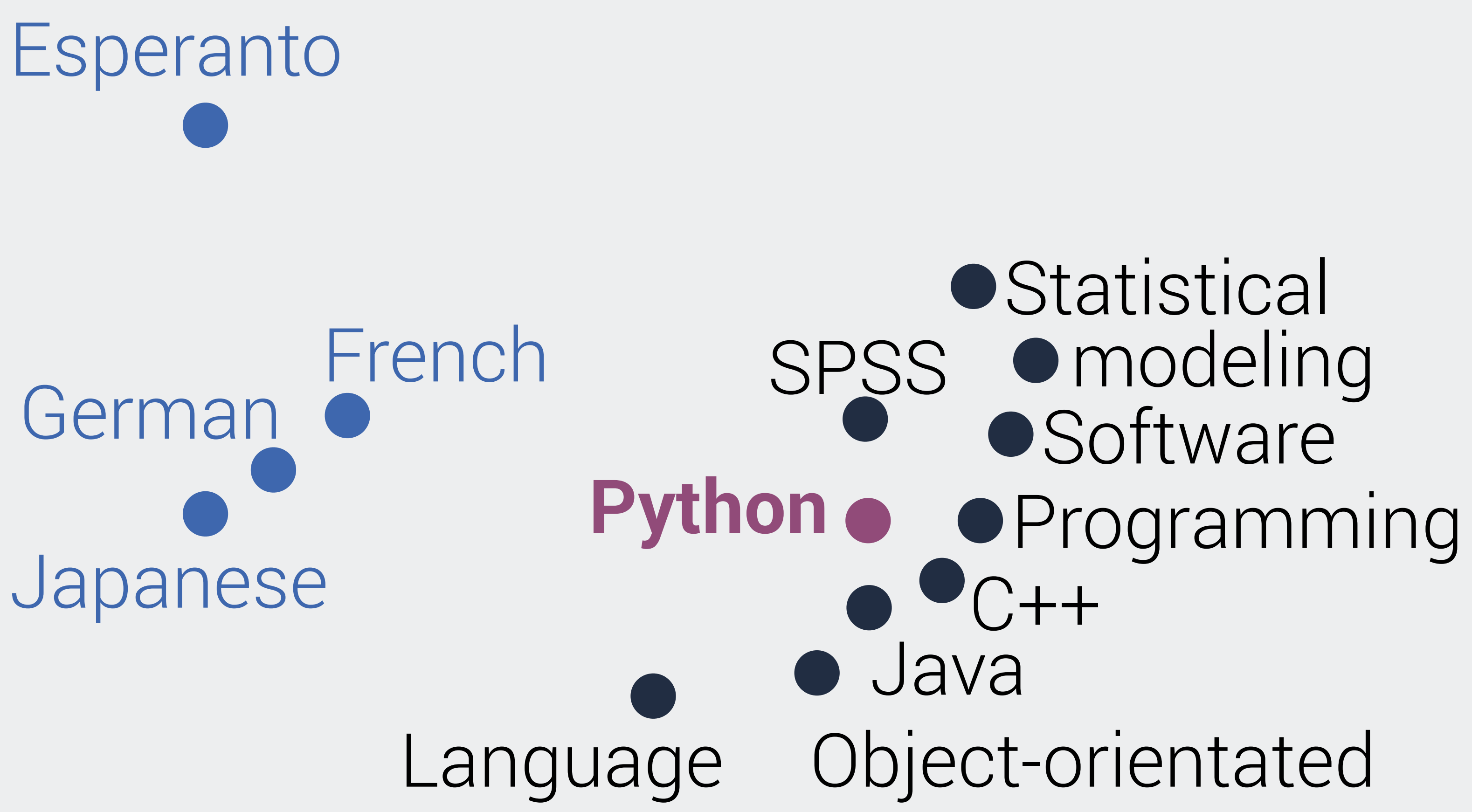
Hyperparameter tuning on final model outputs

Window sizes capture semantic similarity vs semantic relatedness



Small Window Size

Capture Semantic similarity,
Substitutes and Word-level differences

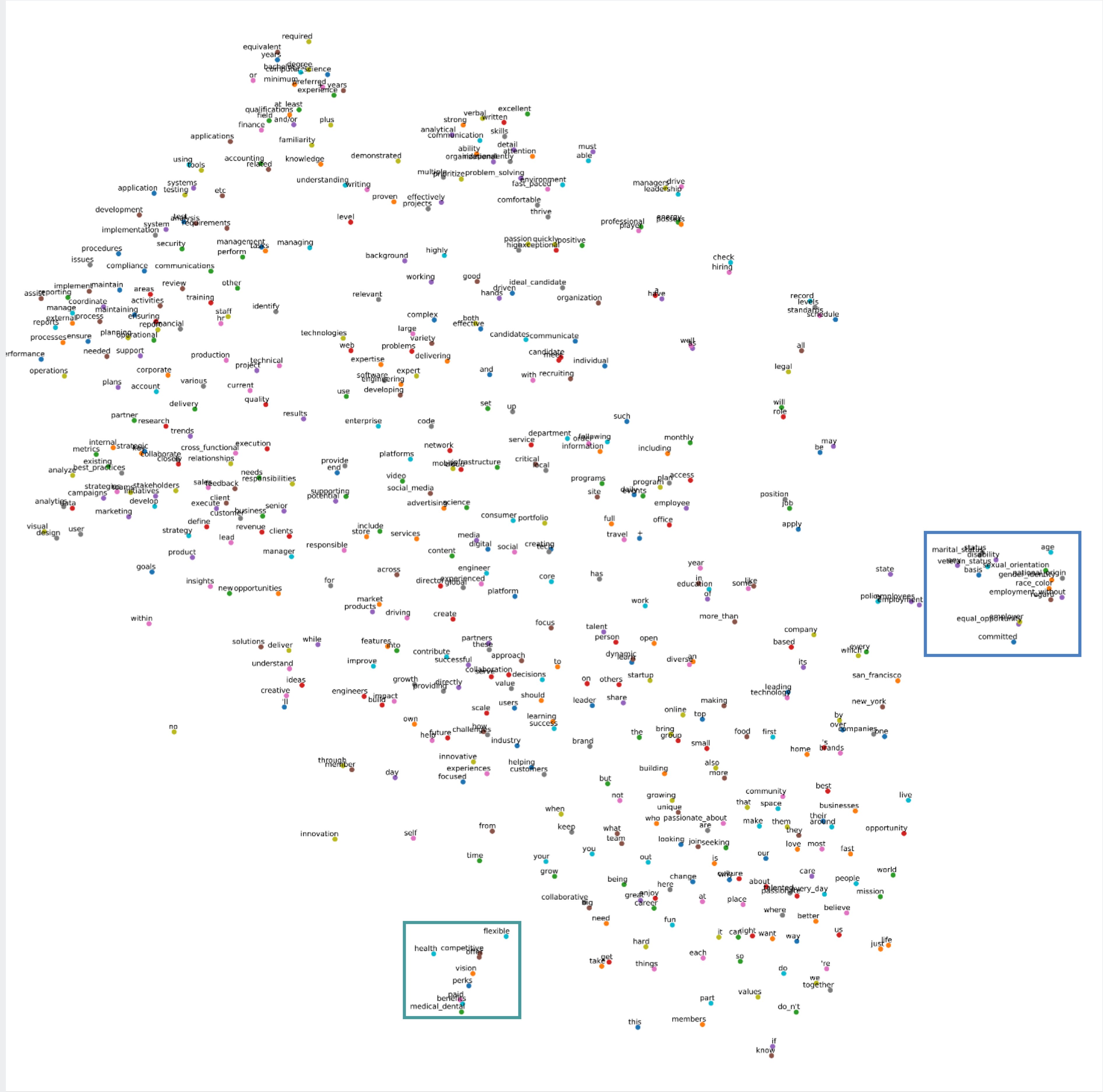


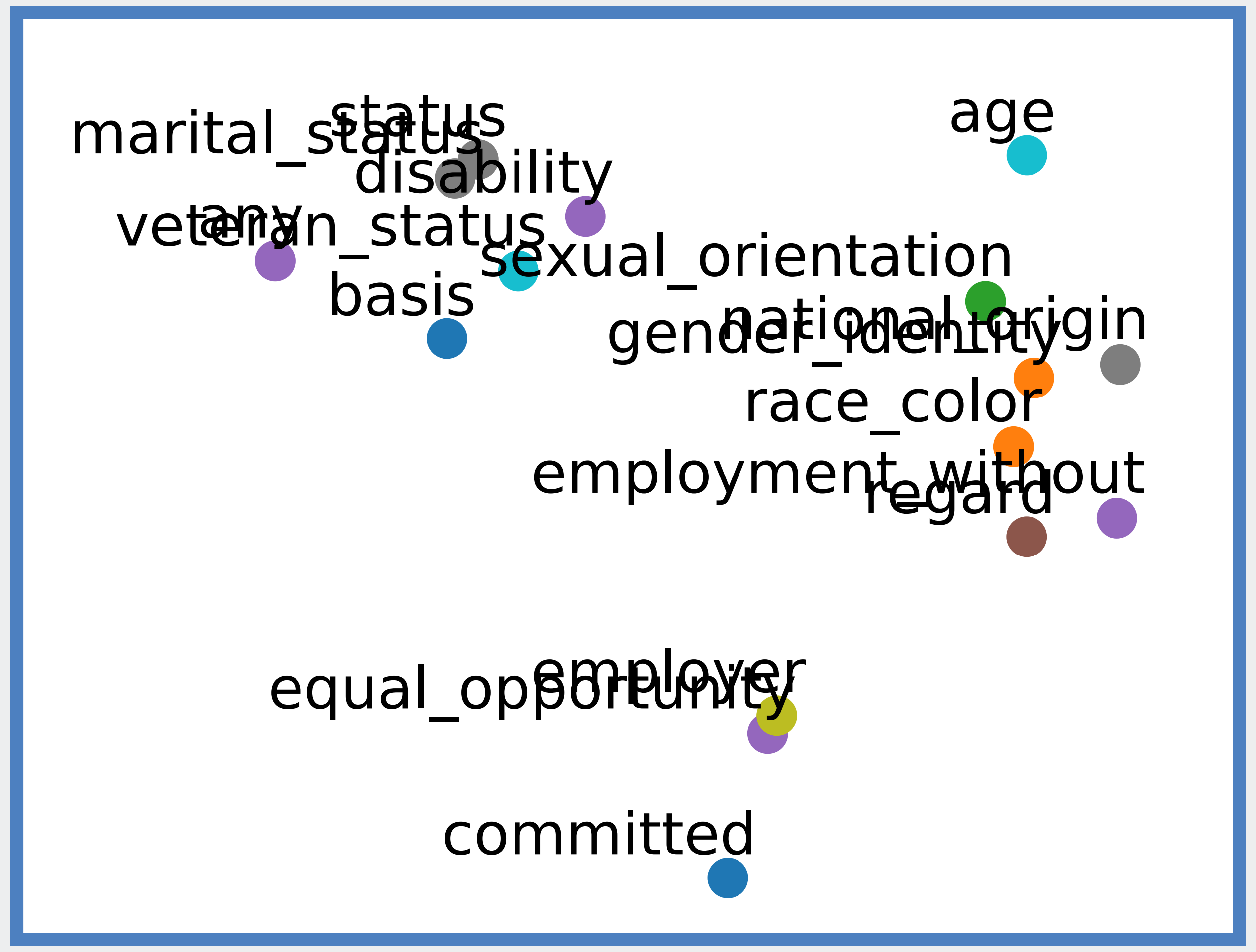
Large Window Size

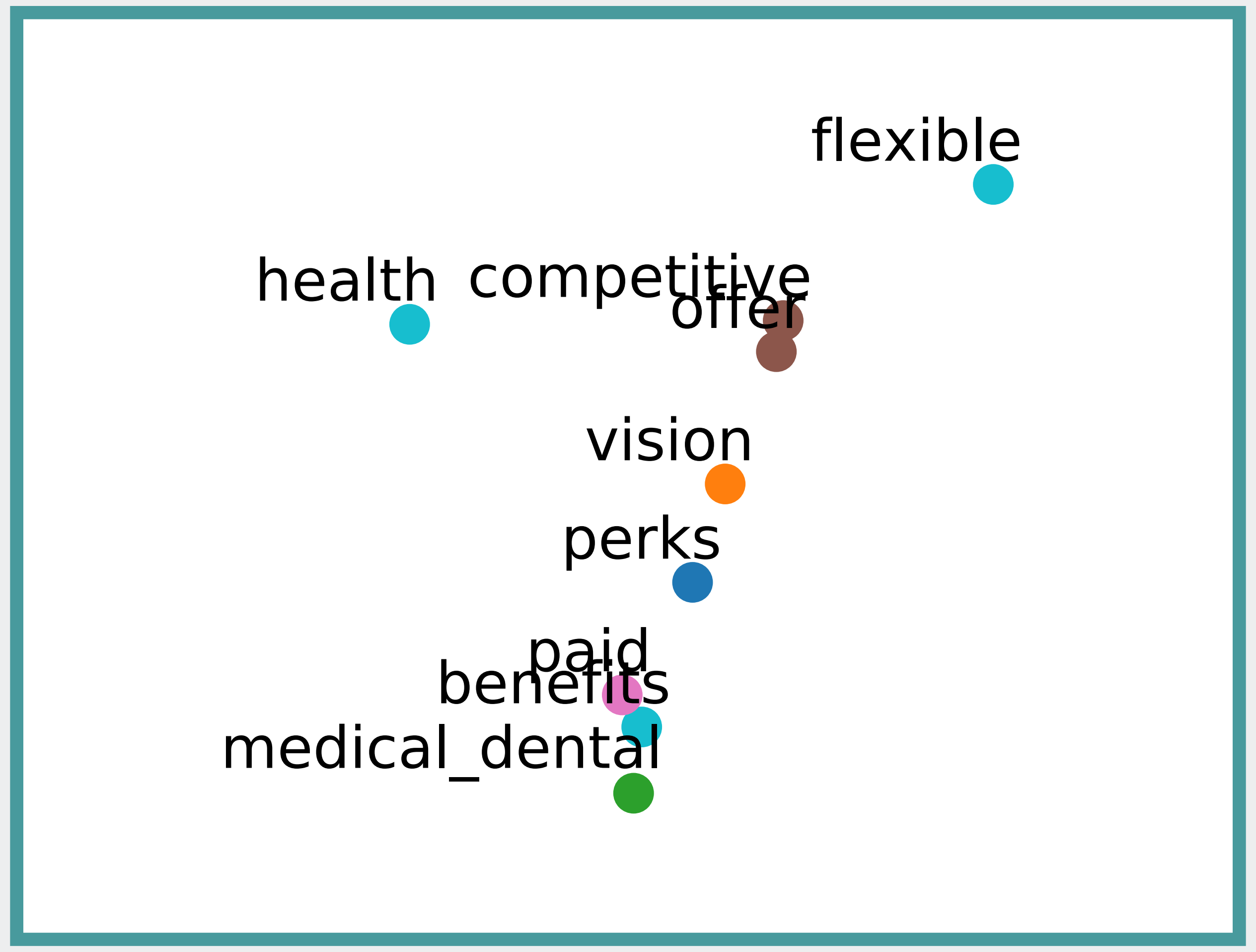
Capture Semantic relatedness,
Alternatives and Domain-level differences

Career language embedding model

Identified equal opportunity and perks language

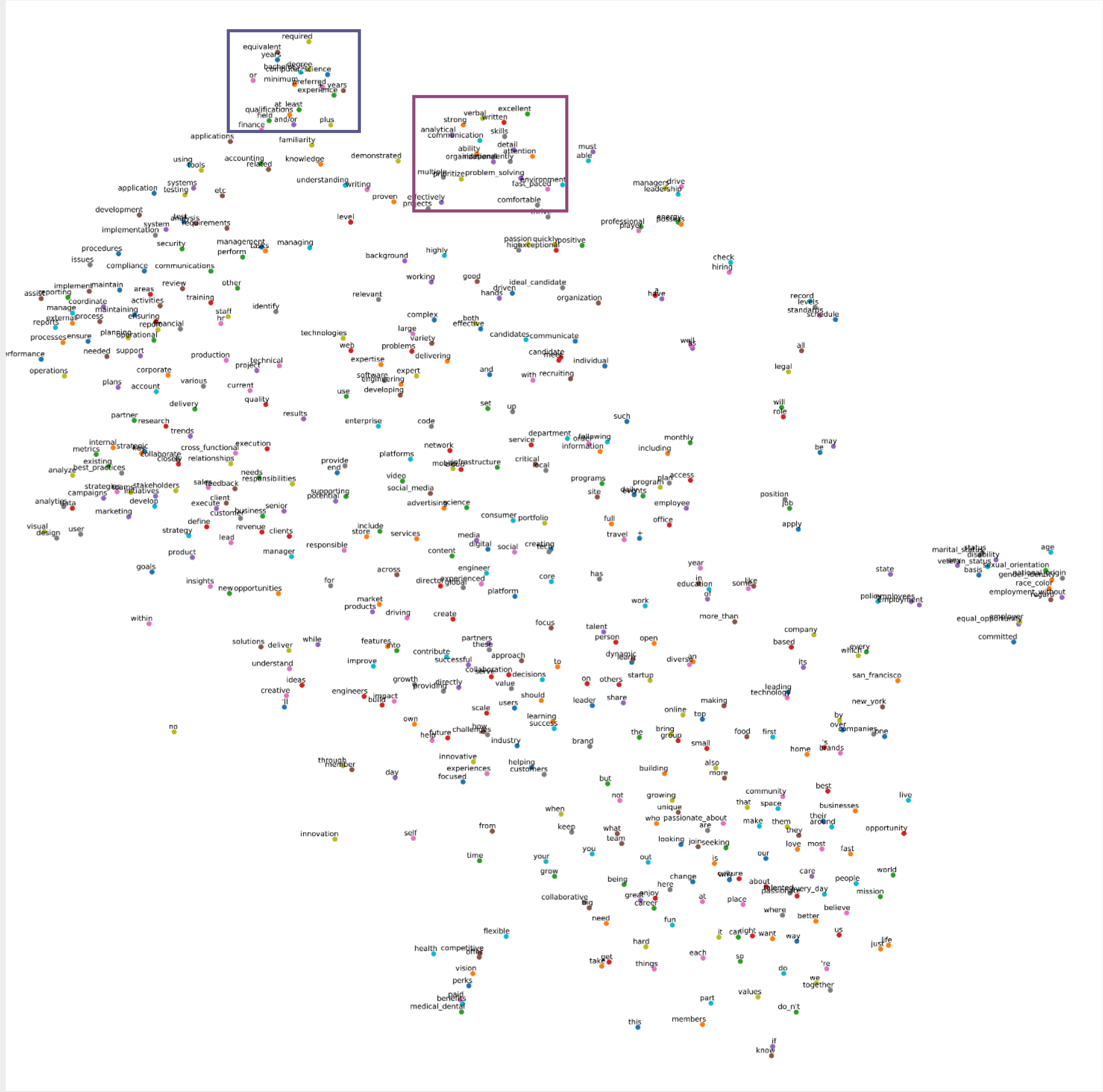


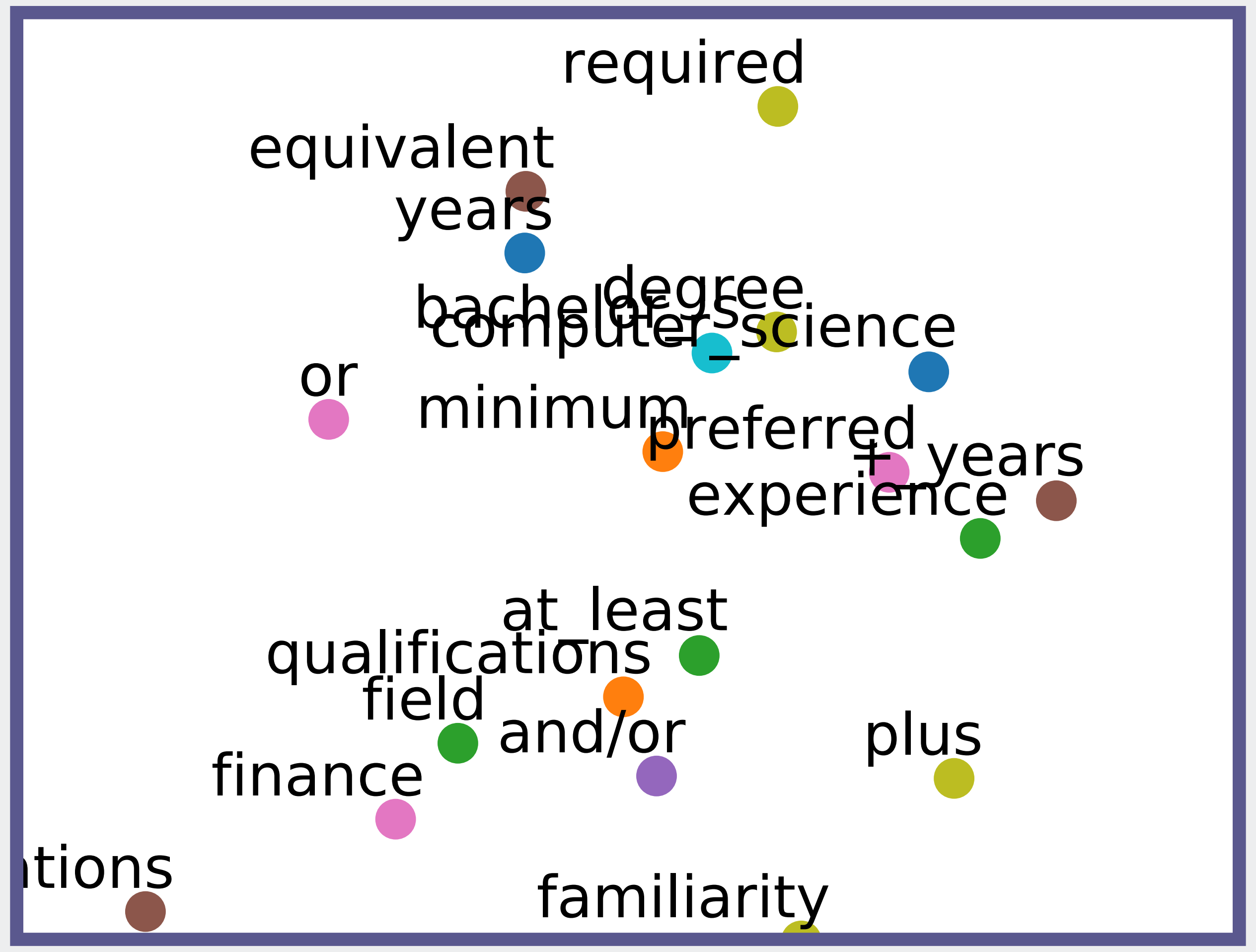




Career language embedding model

Identified 'soft' skills and language around experience



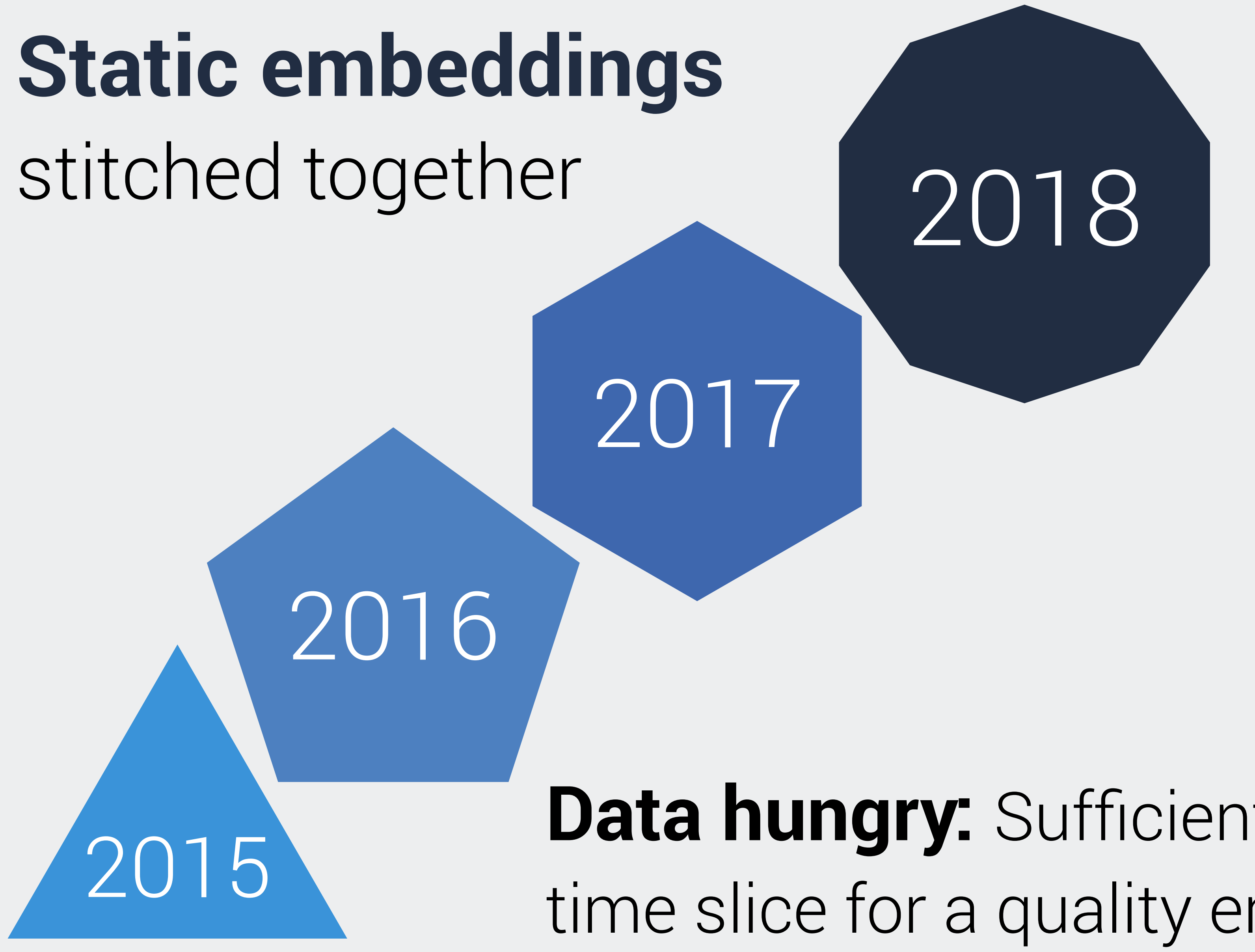


tapRecruit.co

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

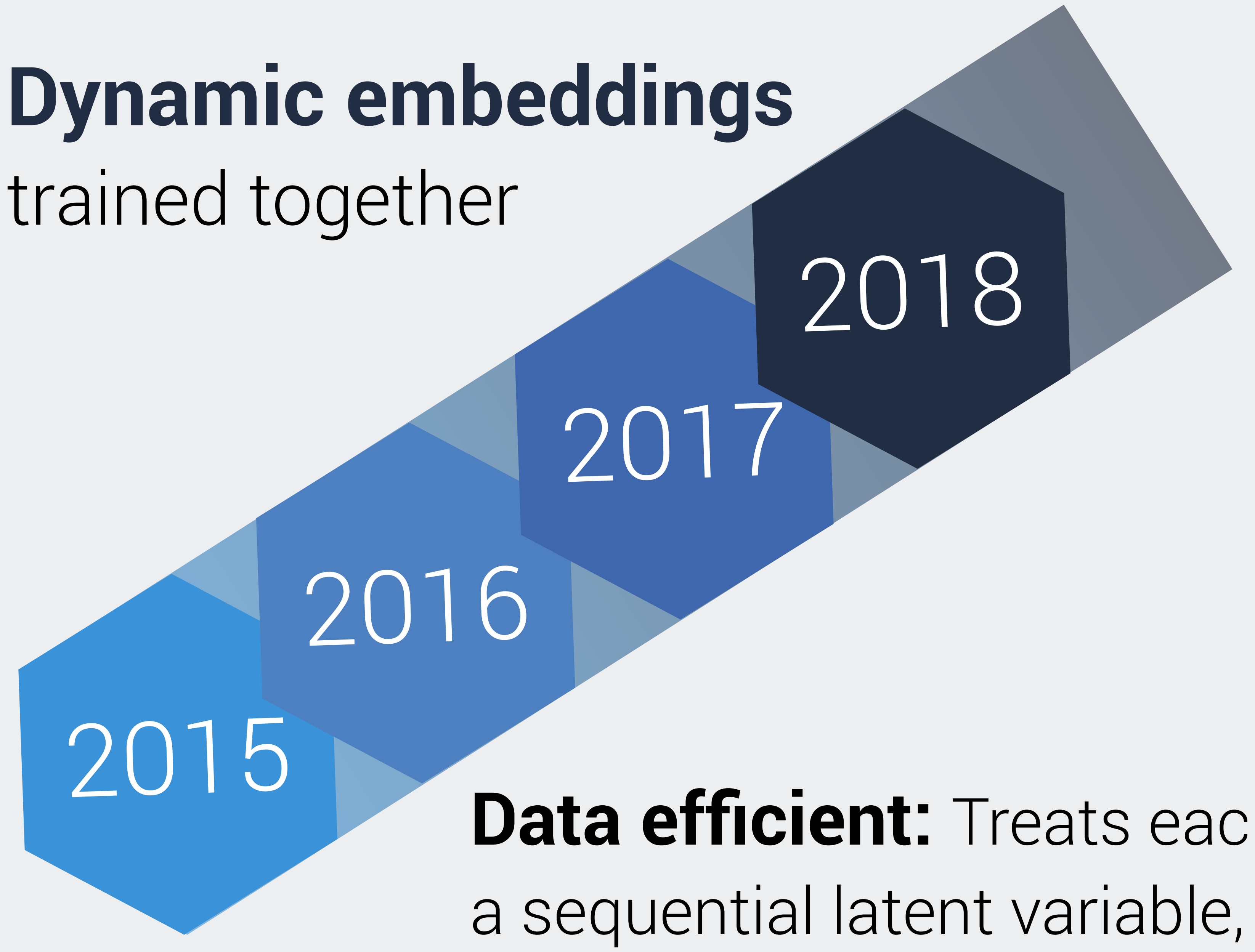
Two approaches to connect embeddings

Static embeddings
stitched together



Data hungry: Sufficient data for each time slice for a quality embedding.
Requires alignment: Each time slice is trained independently, therefore dimensions are not comparable across slices.

Dynamic embeddings
trained together



Data efficient: Treats each time slice as a sequential latent variable, enabling time slices with sparse data.
Does not require 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.

Balmer and Mandt, arXiv: 1702:08359
Yao, Sun, Ding, Rao and Xiong, arXiv: 1703:00607
Rudolph and Blei, arXiv: 1703:08052

Dynamic embeddings models

Rudolph and Blei, arXiv: 1703:08052

Absolute drift

Identifies top words whose usage changes over time course

words with largest drift (Senate)			
IRAQ	3.09	coin	2.39
tax cuts	2.84	social security	2.38
health care	2.62	FINE	2.38
energy	2.55	signal	2.38
medicare	2.55	program	2.36
DISCIPLINE	2.44	moves	2.35
text	2.41	credit	2.34
VALUES	2.40	UNEMPLOYMENT	2.34

Embedding neighborhoods

Extract semantic changes by nearest neighbors of drifting words

UNEMPLOYMENT		
1858	**1940**	**2000**
unemployment	unemployment	unemployment
unemployed	unemployed	jobless
depression	depression	rate
acute	alleviating	depression
deplorable	destitution	forecasts
alleviating	acute	crate
destitution	reemployment	upward
urban	deplorable	lag
employment	employment	economists
distressing	distress	predict

Repository Link: http://bit.ly/dyn_bern_emb

Experiments with Dynamic Bernoulli Embeddings

	Small Corpus	Large Corpus
Job Types	All	All
Time Slices	3 (2016-2018)	3 (2016-2018)
Number of Documents	**50 k**	**500 k**
Vocabulary Size	10 k	10 k
Data Preprocessing	Basic	Basic
Embedding Training	100 dimensions, 10 epochs	100 dimensions, 10 epochs

Repository Link: http://bit.ly/dyn_bern_emb

Dynamic Bernoulli embeddings

Small corpus identified gains and losses

Demand for PhDs and MBAs is Falling

PhDs in All Jobs	PhDs in DS Jobs
-23%	**-30%**

MBAs in All Jobs
-35%

Data Science skills showing significant shifts

Tableau	Spark	Python
+20%	**Steady**	**Steady**

PowerBI	Hadoop	Perl
+100%	**-30%**	**-40%**

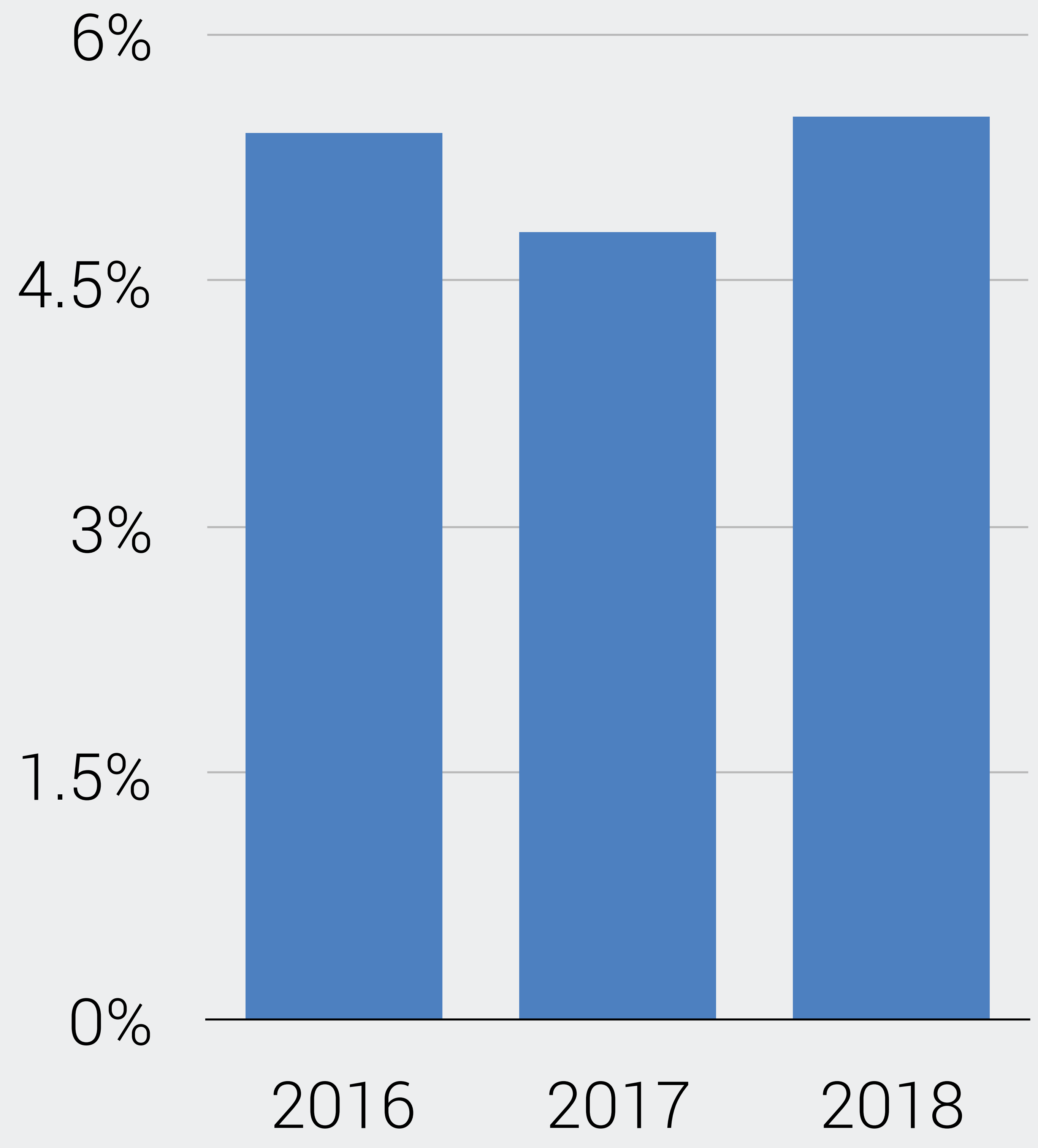
Blue boxes indicate phrases identified from top drifting words analysis.
Grey boxes indicate 'control' skills.

tapRecruit.co

Dynamic Bernoulli embeddings

Large corpus identified role-type dependent shifts in requirements

No change to SQL demand



SQL requirement increases in specific functions

FP&A Roles **+70%**	Sales Roles **Steady**	Marketing Roles **Steady**
FinTech Roles **Steady**	BizDev Roles **+50%**	HR Roles **+25%**

tapRecruit.co

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

Exponential Family Embeddings

Conditional probabilistic models generalize the spirit of embeddings to other data types

	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

Bernoulli Embeddings

Binary Data

Presence of word, given surrounding words

Poisson Embeddings

Count or Ordinal Data

Number of item purchased, given number of other items purchased in the same cart.

Gaussian Embeddings

Continuous Data

Weight of an edge, given other edges on the same node.

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

Exponential Family Embeddings

Poisson embeddings capture item similarities from shopper behavior

	Mini Bagels
Context	Cream cheese
	Milk
Datapoint	**Coffee**
Context	Orange Juice

Poisson Embeddings
Count or Ordinal Data

262	**Maruchan chicken ramen**
223	Maruchan creamy chicken ramen
162	Maruchan oriental flavor ramen
137	Maruchan roast chicken ramen

293	**Yoplait strawberry yogurt**
69	Yoplait apricot mango yogurt
176	Yoplait strawberry orange smoothie
241	Yoplait strawberry banana yogurt

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

Exponential Family Embeddings

Inner product of vectors identify substitutes and alternatives

High Inner Product Combinations:
Yield products that are frequently bought together

Old Dutch potato chips & Budweiser Lager beer

Lays potato chips & DiGiorno frozen pizza

Low Inner Product Combinations:
Yield products that are rarely bought together

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 DataEngConf!

Maryam Jahanshahi Ph.D.
Research Scientist
@mjahanshahi
maryam-j