Applying Dynamic Embeddings in Natural Language Processing to Analyze Text

Maryam Jahanshahi Ph.D.
Research Scientist

tapRecruit.co

Skills and qualifications matter in job descriptions

Same title, Different job

Finance Manager — Kraft Foods	Finance Manager — Roche
Junior (3 Years)	Senior (6-8 Years)
No Managerial Experience	Division Level Controller
	Strategic Finance Role
	MBA / CPA

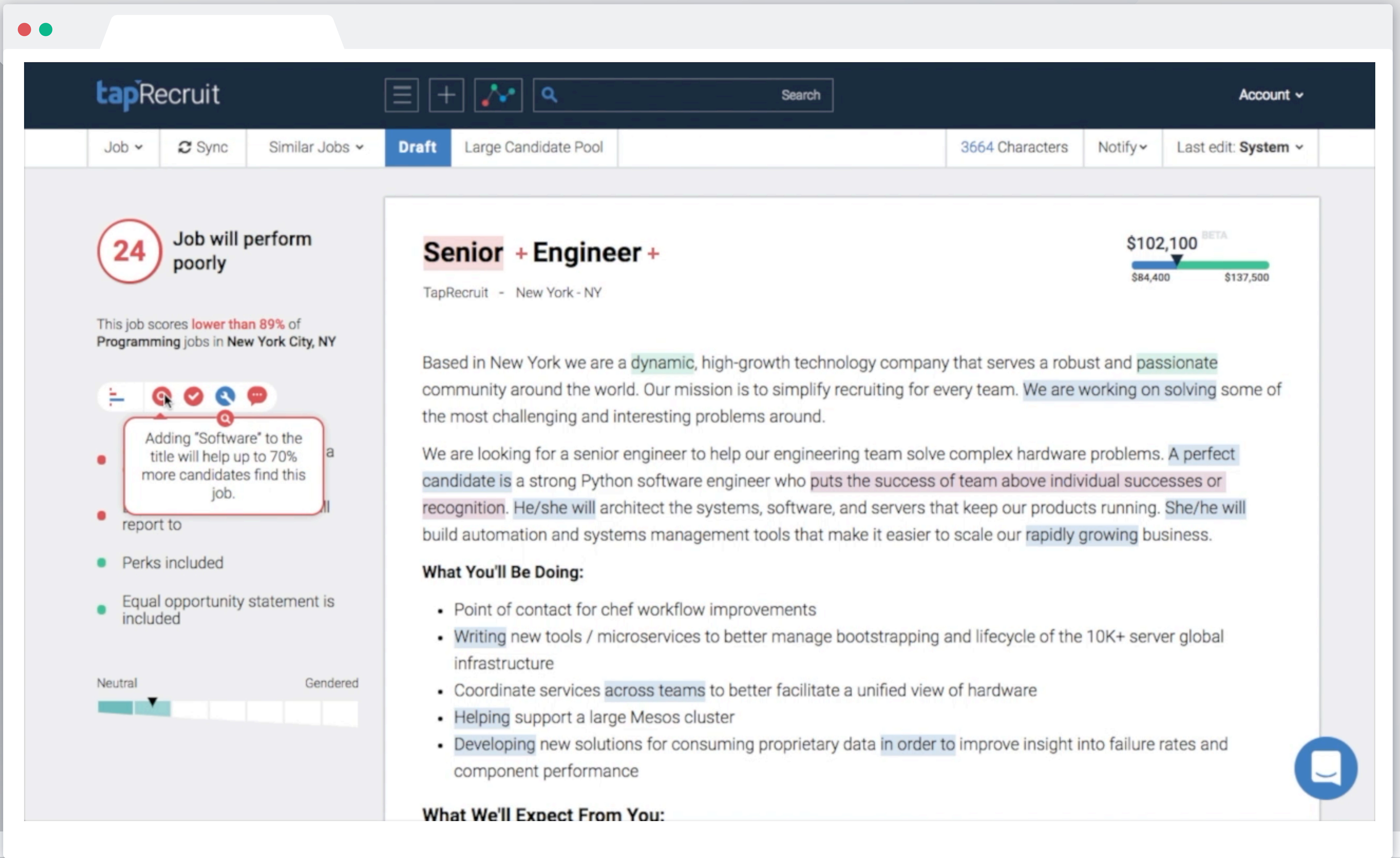
- ✔ **Same Title**
- ✘ Required Experience
- ✘ Required Responsibility
- ✘ Preferred Skill
- ✘ Required Education

Different title, Same job

Performance Marketing Manager — PocketGems	Senior Analyst, Customer Strategy — The Gap
Mid-Level	Mid-Level
Quantitative Focus	Quantitative Focus
iBanking Expertise	Finance Expertise
Data Analysis Tools (SQL)	Relational Database Experience
Consulting Experience Preferred	External Consulting Experience Preferred
MBA Preferred	BA in Accounting, Finance, MBA Preferred

- ✔ **Required Experience**
- ✔ **Required Skills**
- ✔ **Required Experience**
- ✔ **Required Skills**
- ✔ **Preferred Experience**
- ✔ **Preferred Education**

tapRecruit.co


tapRecruit
Search
Account
Job
Sync
Similar Jobs
Draft
Large Candidate Pool
3664 Characters
Notify
Last edit: System
24
Job will perform poorly
This job scores lower than 89% of Programming jobs in New York City, NY
Adding "Software" to the title will help up to 70% more candidates find this job.
report to
Perks included
Equal opportunity statement is included
Neutral
Gendered
Senior + Engineer +
$102,100
BETA
$84,400
$137,500
TapRecruit - New York - NY
Based in New York we are a dynamic, high-growth technology company that serves a robust and passionate community around the world. Our mission is to simplify recruiting for every team. We are working on solving some of the most challenging and interesting problems around.
We are looking for a senior engineer to help our engineering team solve complex hardware problems. A perfect candidate is a strong Python software engineer who puts the success of team above individual successes or recognition. He/she will architect the systems, software, and servers that keep our products running. She/he will build automation and systems management tools that make it easier to scale our rapidly growing business.
What You'll Be Doing:
Point of contact for chef workflow improvements
Writing new tools / microservices to better manage bootstrapping and lifecycle of the 10K+ server global infrastructure
Coordinate services across teams to better facilitate a unified view of hardware
Helping support a large Mesos cluster
Developing new solutions for consuming proprietary data in order to improve insight into failure rates and component performance
What We'll Expect From You:

Research at TapRecruit

What are distinguishing characteristics of successful career documents?

NLP and Data Science:

- What are distinguishing characteristics of successful career documents?
- **What skills are increasingly important for different industries?**
  Calibrating labor supply and demand

Decision Science:

- How do candidates make decisions about which jobs to apply to?
- How do hiring teams make decisions about candidate qualifications?

tapRecruit.co

How have data science skills changed over the last few years?

Strategies to identify changes among corpora

Traditional approaches do not capture syntactic and semantic shifts

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 selection of key attributes, therefore difficult to discover new attributes

Dynamic Topic Models

Require experimentation with topic number

Adapted from Blei and Lafferty, ICML 2006.

tapRecruit.co

Word embeddings use context to extract meaning

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

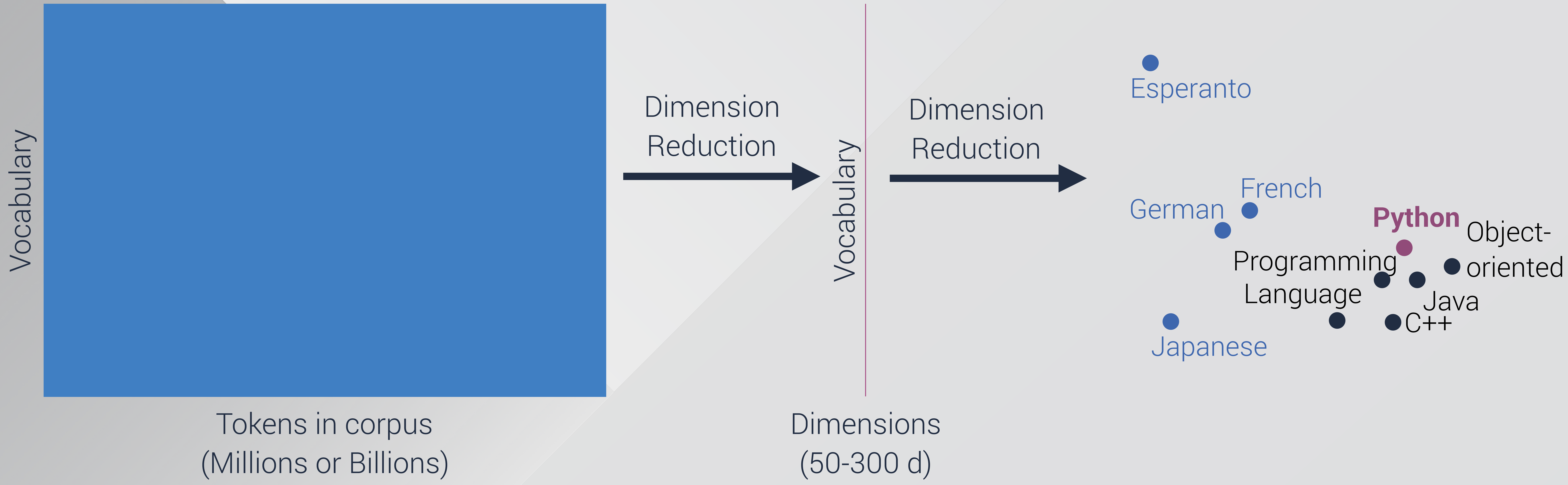
Context Word Context

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

Context Word Context

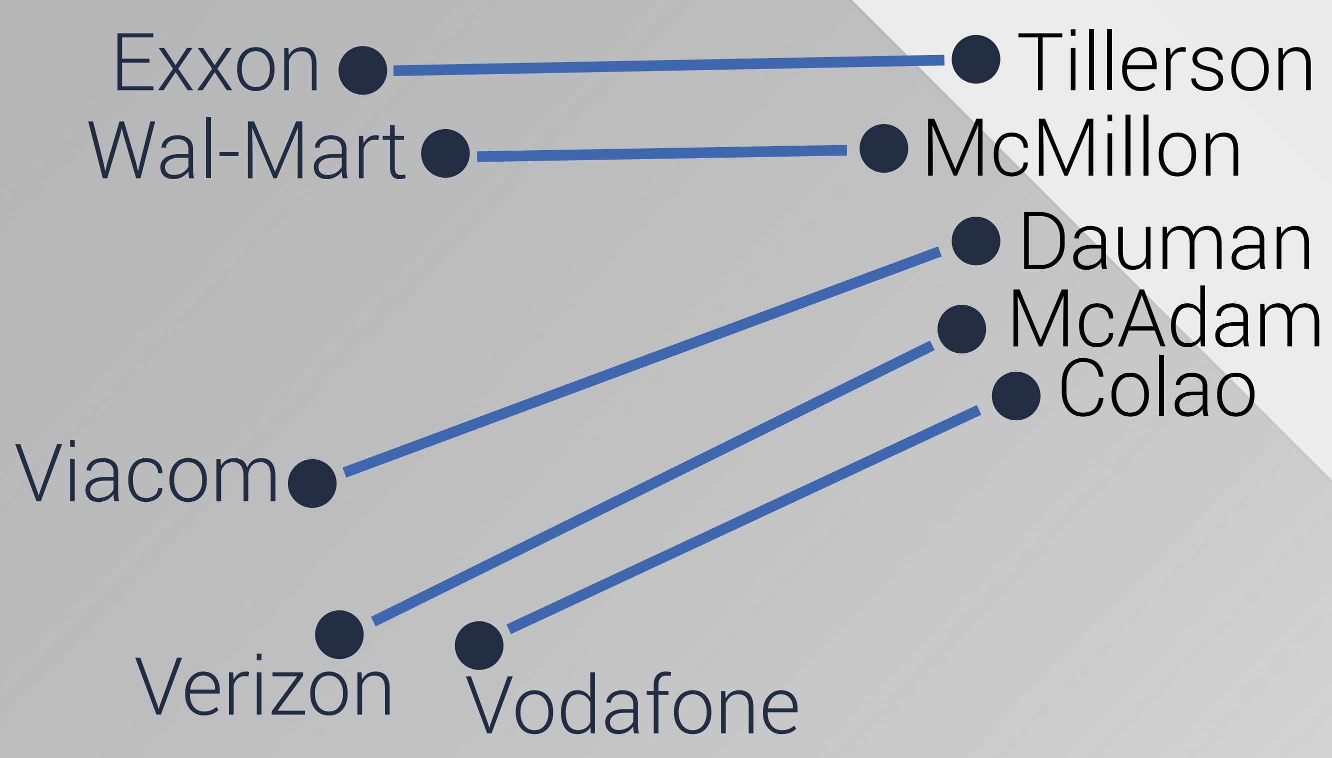
A simplified representation of word embeddings

Dimension reduction is key to all types of embeddings models

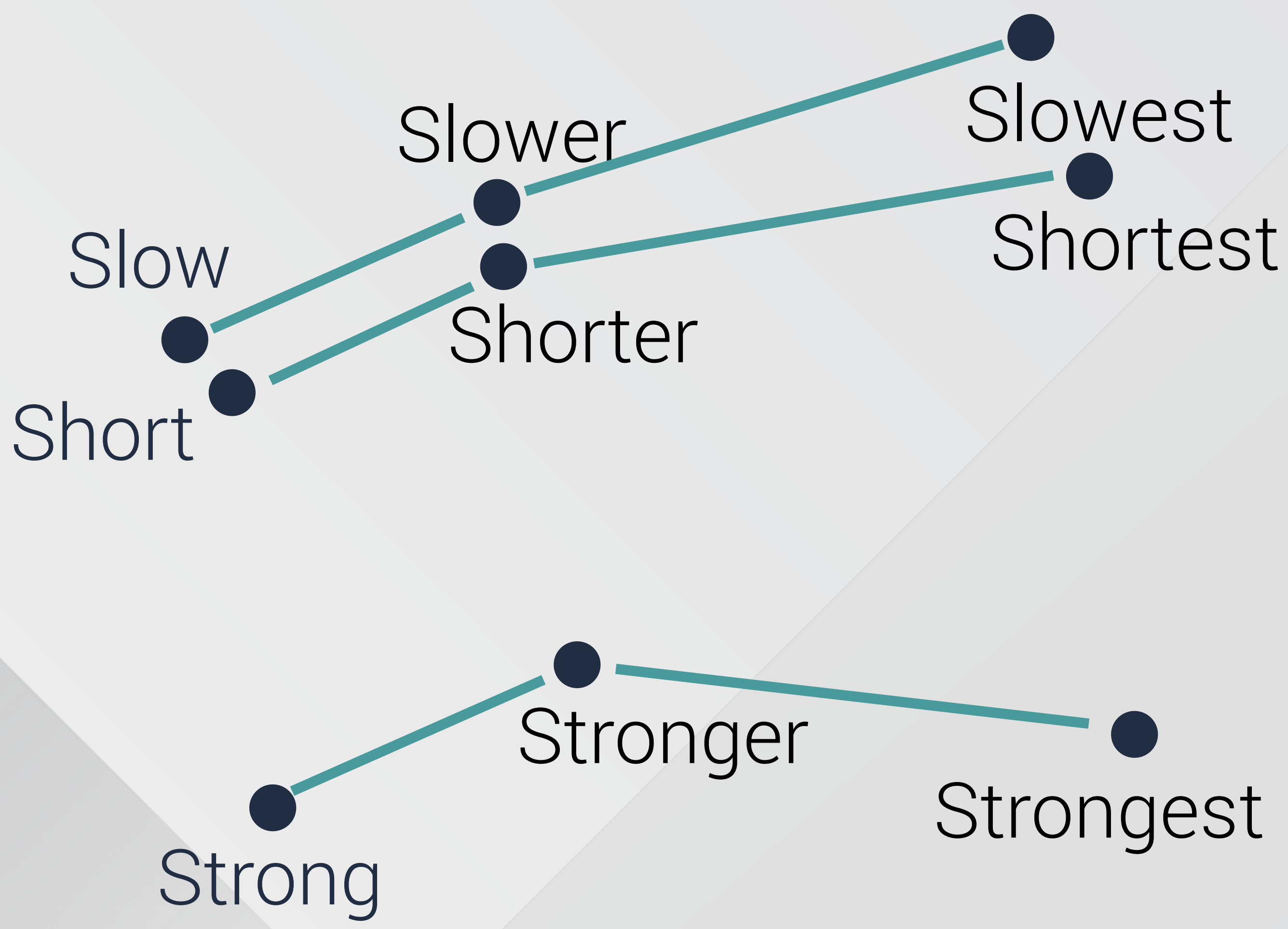


Word embeddings capture entity relationships

Dimensionality enables comparison between word pairs along many axes



Hierarchies



Comparatives and Superlatives

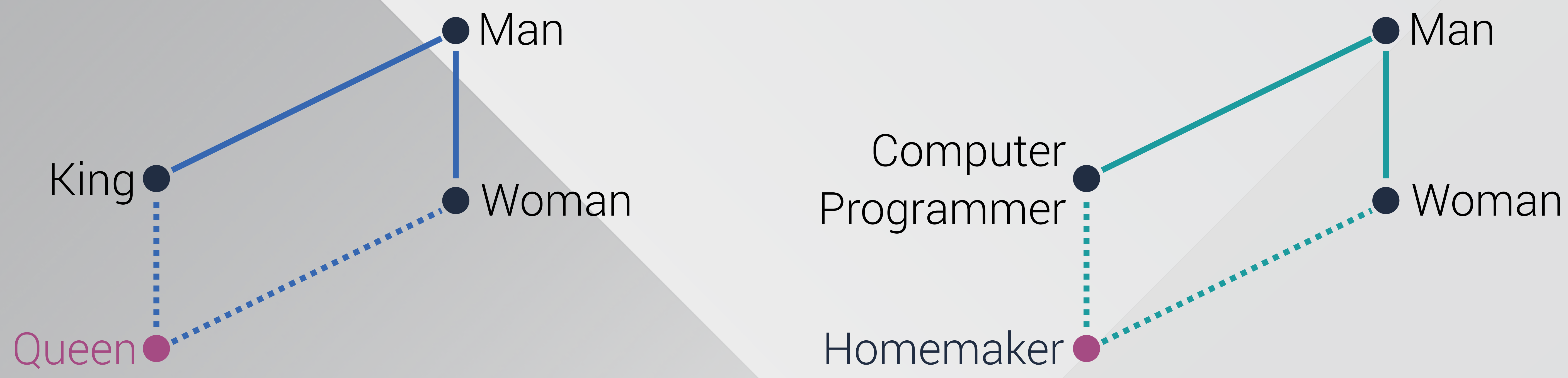
Man :: King as Woman :: ?

Adapted from Stanford NLP GLoVE Project

tapRecruit.co

Word embeddings reflect cultural bias in corpora

High dimensionality enables some bias reduction



Man :: King as Woman :: ?

Man :: Programmer as Woman :: ?

Adapted from Bolukbasi et al., arXiv: 1607:06520.

Pretrained word embeddings enable fast prototyping

Corpus Generation

Corpus Processing

Language Model Generation

Language Model Tuning

Final Application

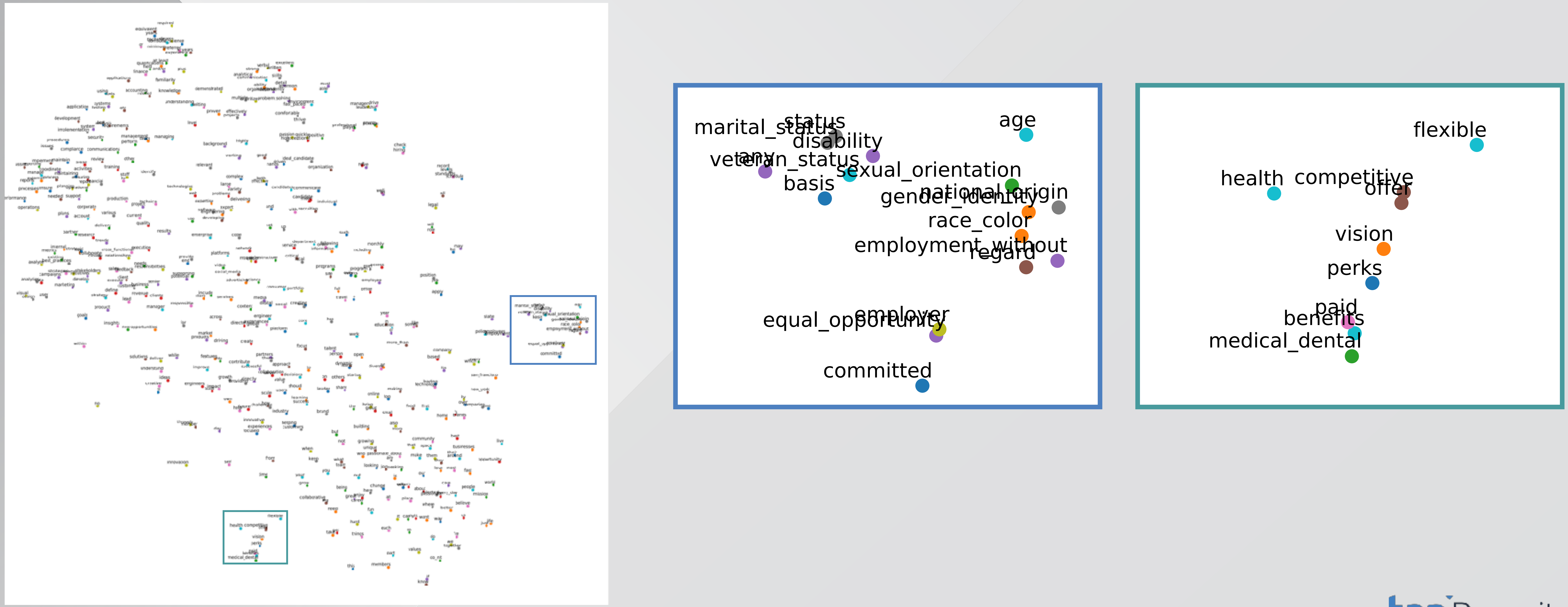
Corpus	Twitter	Common Crawl	GoogleNews	Wikipedia
Tokens	27 B	42-840 B	100 B	6 B
Vocabulary Size	1.2 M	1.9-2.2 M	3 M	400 k
Algorithm	GLoVE	GLoVE	word2vec	GLoVE
Vector Length	25 - 200 d	300 d	300 d	50 - 300 d

Drawbacks of pretrained word embeddings

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

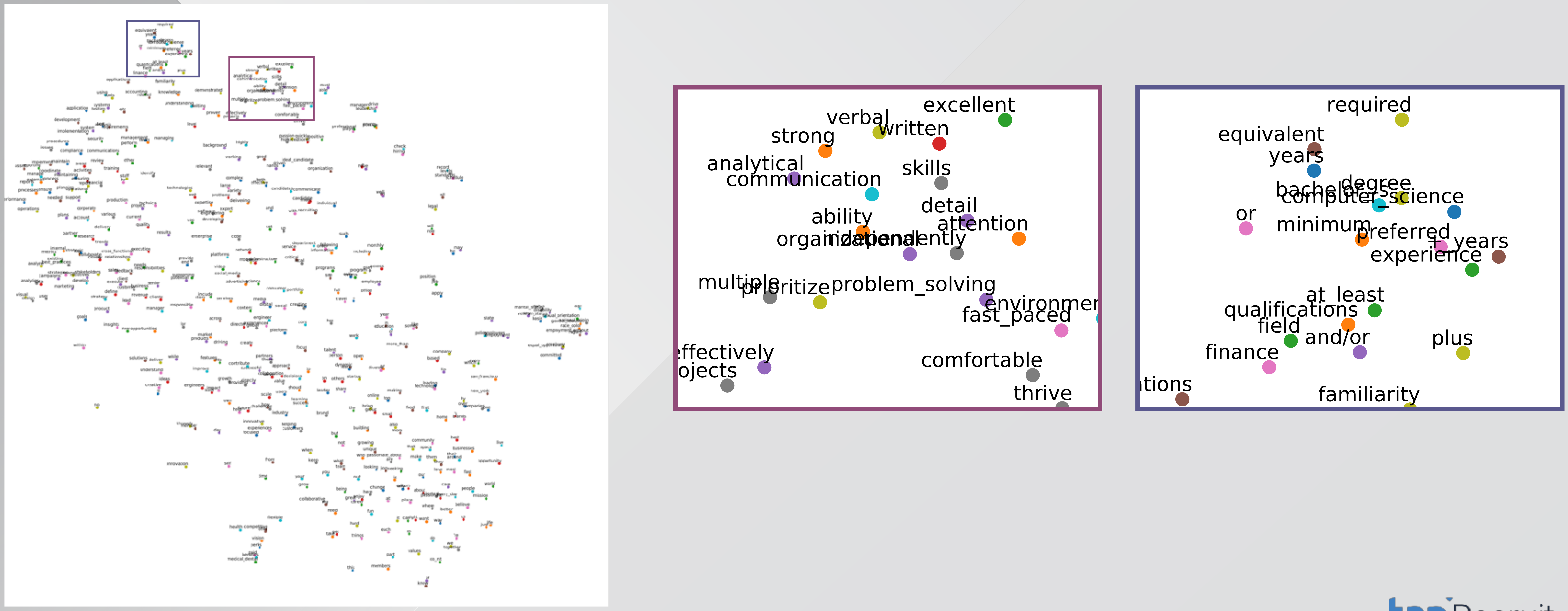
Career language embedding model

Identified equal opportunity and perks language



Career language embedding model

Identified 'soft' skills and language around experience

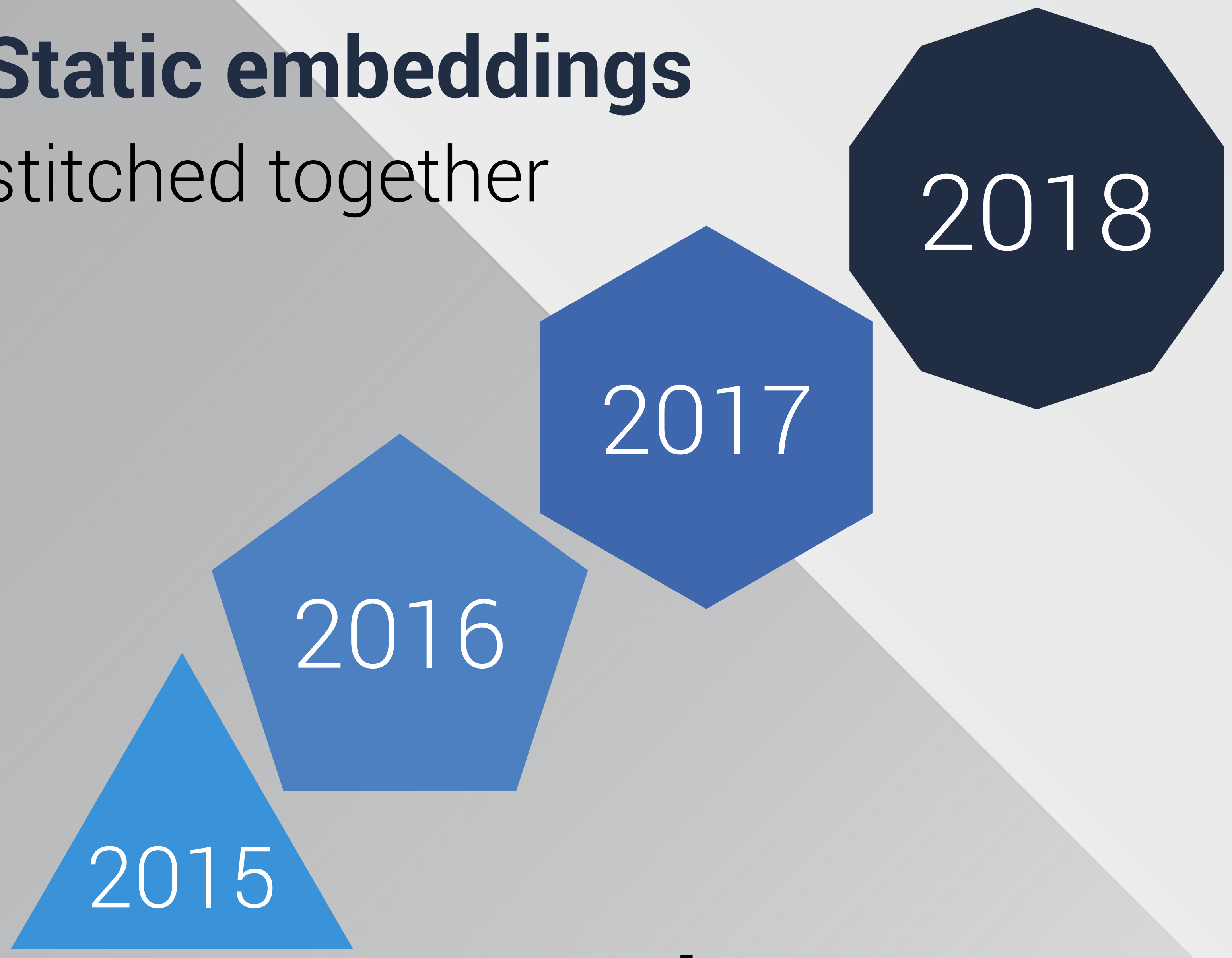


tapRecruit.co

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

Two approaches to connect embeddings

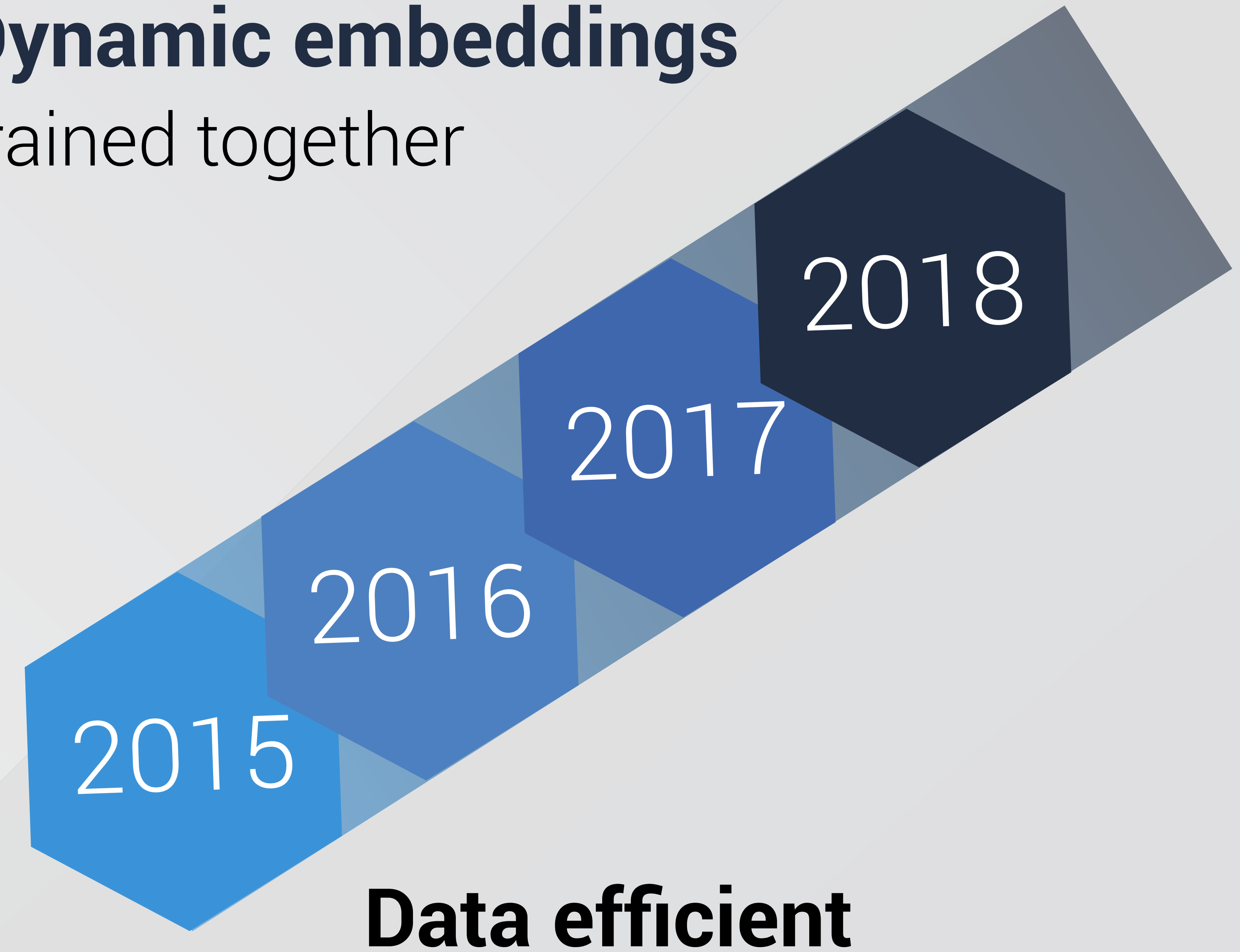
Static embeddings
stitched together



Data hungry
Requires alignment

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

Dynamic embeddings
trained together



Data efficient
Does not require alignment

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

tapRecruit.co

Dynamic Bernoulli embeddings

Outputs facilitate quick analysis of trends

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 embeddings

	Small Corpus
Job Types	All US Jobs
Time Slices	3 (2016-2018)
Number of Documents	**50 k**
Vocabulary Size	10 k
Data Preprocessing	Basic
Embedding Dimensions	100 d

tapRecruit.co

Small corpus identified MBAs and PhDs

Reduced requirement for advanced degrees in many jobs

Demand for MBAs is Falling in US Roles | **and in Roles based in the UK**

MBAs in All Jobs	MBAs in DS Jobs	MBAs in Tech Jobs	MBAs in All Jobs	MBAs in Tech Jobs
-35%	-15%	+30%	-40%	-40%

Demand for PhDs is Falling in US Roles | **and in Roles based in the UK**

PhDs in All Jobs	PhDs in DS Jobs	PhDs in ML Jobs	PhDs in DS Jobs	PhDs in ML Jobs
-35%	-20%	-30%	-40%	-50%

tapRecruit.co

Small corpus identified skill demands

Data Viz is up in lots of different roles

Demand for Data Visualization tools is up

Tableau	PowerBI
+20%	+100%

Data Viz growth in US Non-DS Roles

Data Viz in DS Jobs	Data Viz in Other Jobs
+30%	+90%

and UK Non-DS Roles

Data Viz in DS Jobs	Data Viz in Other Jobs
+30%	+200%

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

tapRecruit.co

Small corpus identified skill demands

Demand for Hadoop (but not Spark) is down in Data Science jobs

Data Science Jobs

Hadoop	Spark	AWS
-30%	Steady	+100%

Tech Jobs (non-DS)

Hadoop	Spark	AWS
Steady	+60%	+100%

Blue boxes indicate phrases identified from top drifting words analysis.
Grey and pink boxes indicate ‘control’ skills.

tapRecruit.co

Battle of the languages: Supply vs Demand

Demand for Perl is down in Data Science Roles

Perl	Python	Java
-40%	Steady	-60%

Blue boxes indicate phrases identified from top drifting words analysis.
Grey and pink boxes indicate ‘control’ skills.

tapRecruit.co

Battle of the languages: Supply vs Demand

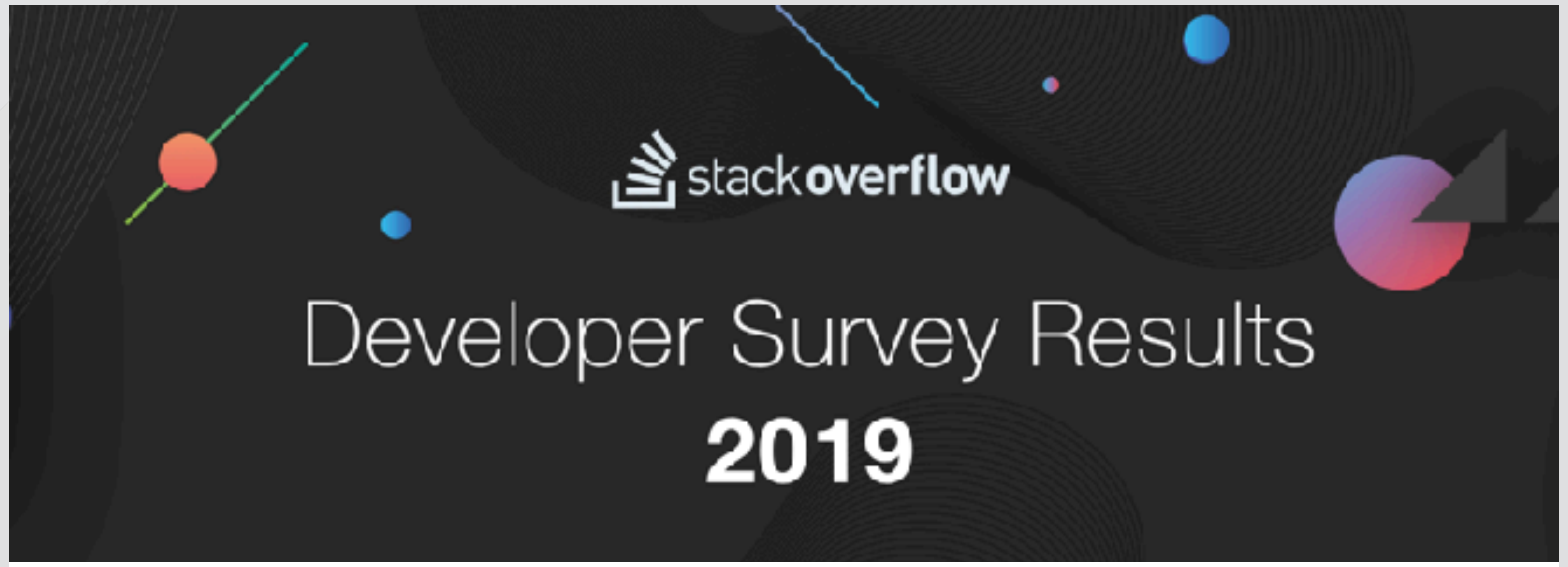
Scripting Languages absorbing changes

Data Science Jobs

Java	Python	Scripting Languages
-60%	Steady	+140%

Tech Jobs (non-DS)

Java	Python	Scripting Languages
Steady	+30%	+200%



- Python, the fastest-growing major programming language, has risen in the ranks of programming languages in our survey yet again, edging out Java this year and standing as the second most loved language (behind Rust).

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

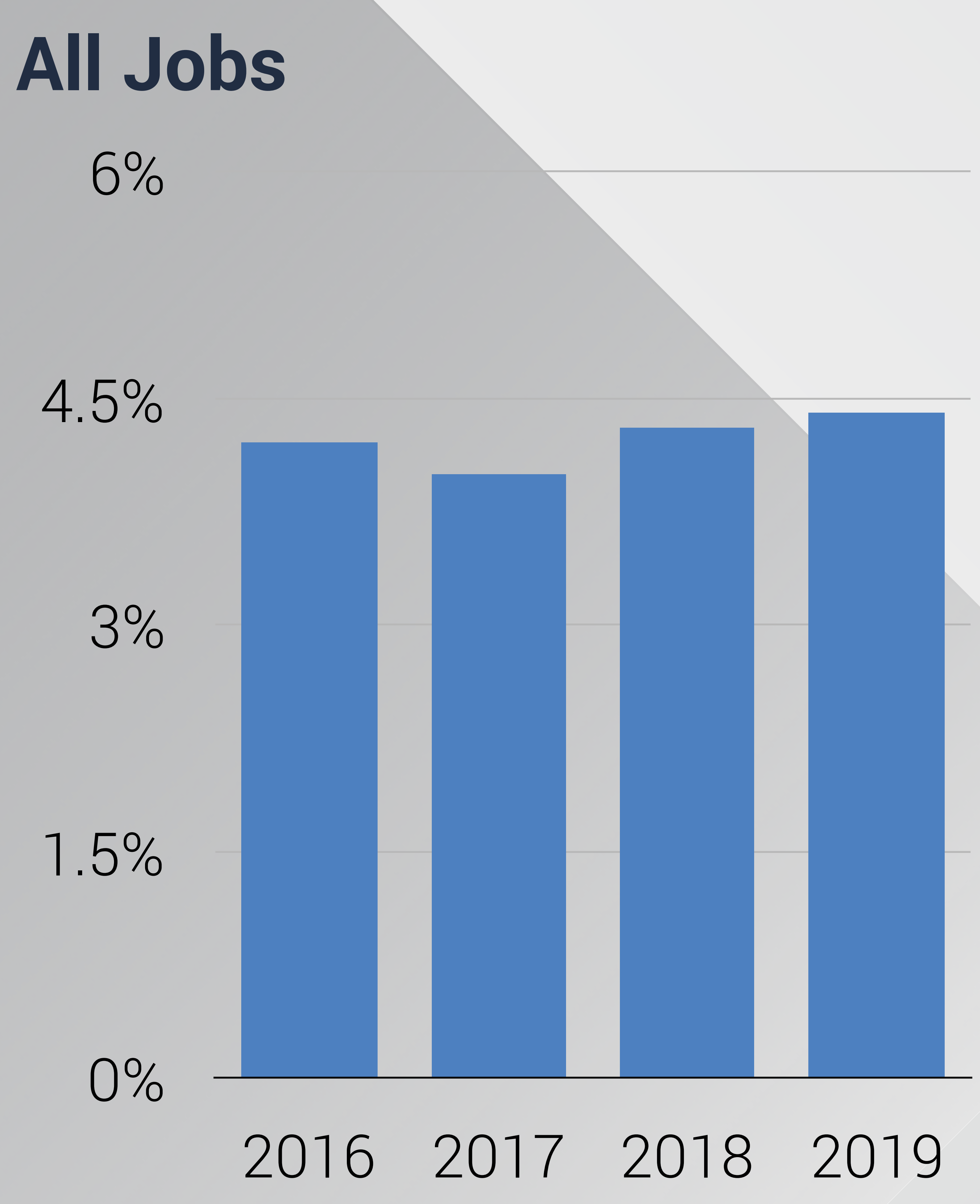
tapRecruit.co

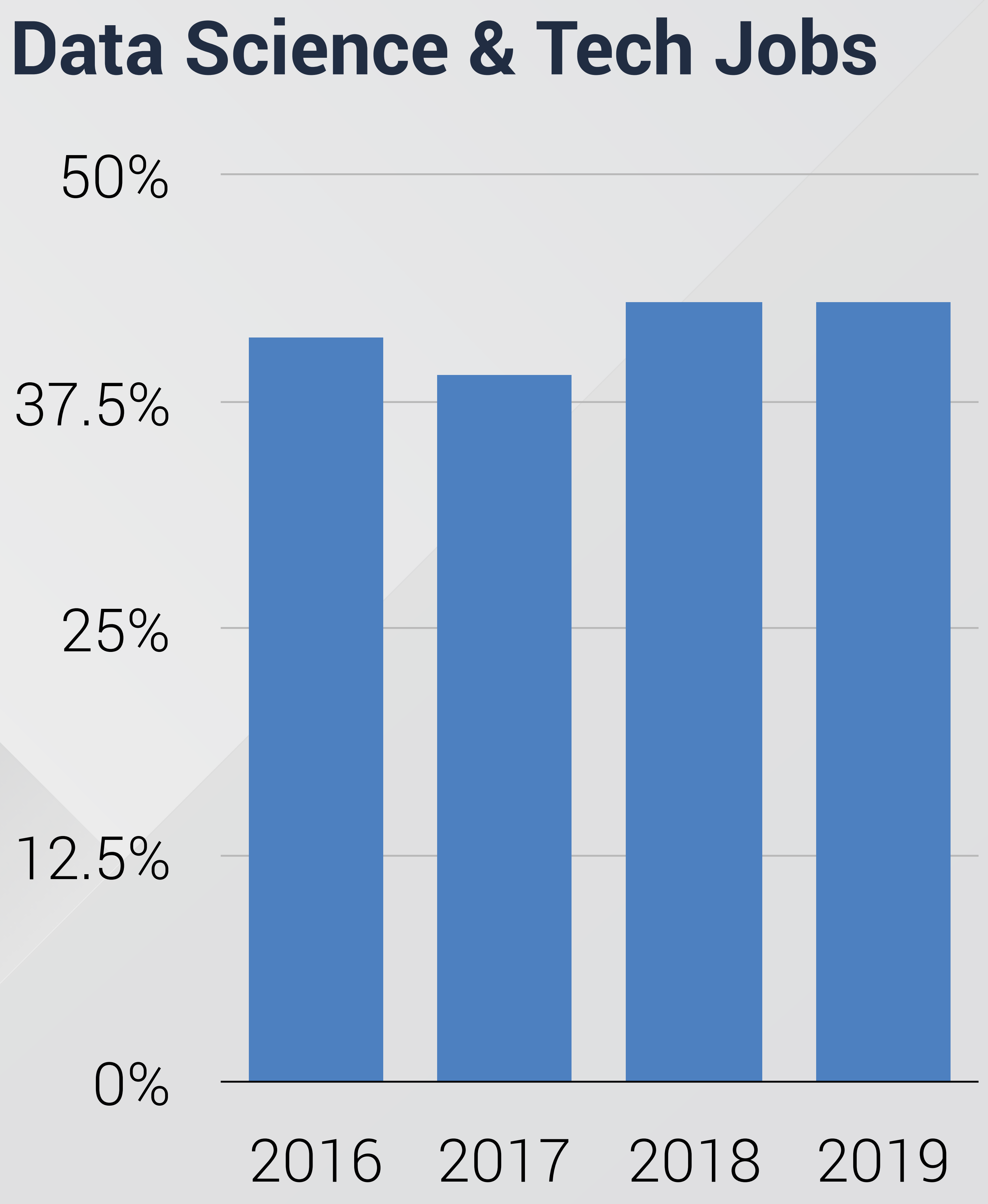
Experiments with dynamic embeddings

	Small Corpus	Large Corpus
Job Types	All US Jobs	All US Jobs
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 Dimensions	100 d	100 d

Large corpus identified SQL as a top drifting word

But no difference in demand for SQL in jobs





tapRecruit.co

Large corpus identified SQL as a top drifting word

Large corpus identified role-type dependent shifts in requirements

SQL requirement increases in specific functions

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

Beyond word2vec

- Flavors of static word embeddings: The Corpus Issue
- Considerations for developing custom embedding models
- Dynamic Embeddings are robust with small datasets

How have tech and data science skills changed?

- Demand for MBAs and PhDs is falling
- Core Skills: DataViz & Scripting Languages
- Commodification of distributed systems impacts demand for Hadoop
- Demand for SQL in a variety of core business functions

Thank you Rev2!

Maryam Jahanshahi Ph.D.
Research Scientist

@mjahanshahi
maryam-j

tapRecruit.co