

Applying Dynamic Embeddings in Natural Language Processing to track the Evolution of Tech Skills

Maryam Jahanshahi Ph.D.
Research Scientist

tap Recruit.co

The War On/For Talent is Real



Nick Heitzman
 @NickDoesData

Roles and Responsibilities:

- Automate horrible business practices
- Write ad hoc SQL as needed

REQUIRED EXPERIENCE:

- 15 years exp deep learning in Python
- PhD thesis on Bayesian modeling
- NLP experience in 7 languages
- 10 years of creating Hadoop clusters from scratch



I Am Recruitr
 @iamrecruitr

Following

JOB AD:

- * Must have 10+ years Swift experience
- * Must have a PhD in Aqueous Geoochemistry from a top tier University
- * Must be a team player

zero applications later

Founder Blog Post:
The War For Talent is real



Chris @cdubhland · Jul 24

Sees job ad requiring 7+ years of Kubernetes.

Browses to Wikipedia too see it was released 5 years ago.

Begins crafting tweet that ends in

tapRecruit

Job Sync Similar Jobs Draft Large Candidate Pool Search Account

Job will perform poorly

24

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 +

TapRecruit - New York - NY

\$102,100 BETA
\$84,400 \$137,500

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:

Skills and qualifications matter 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**

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:

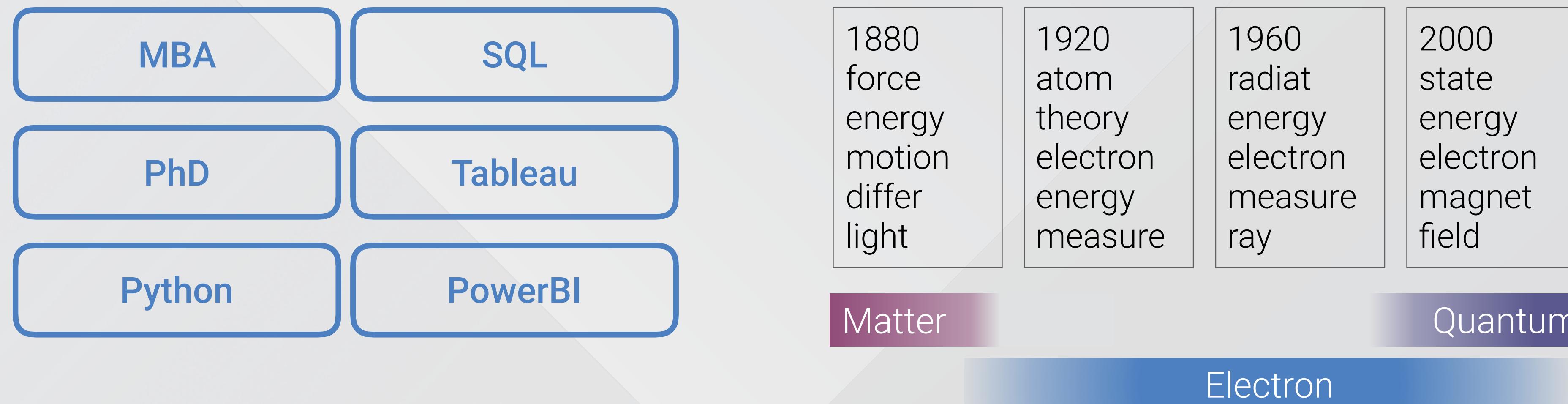
- How do candidates make decisions about which jobs to apply to?
- How do hiring teams make decisions about candidate qualifications?



How have tech skills changed
over the last few years?

Strategies to identify changes among corpora

Traditional approaches do not capture syntactic and semantic shifts



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](#).

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++.

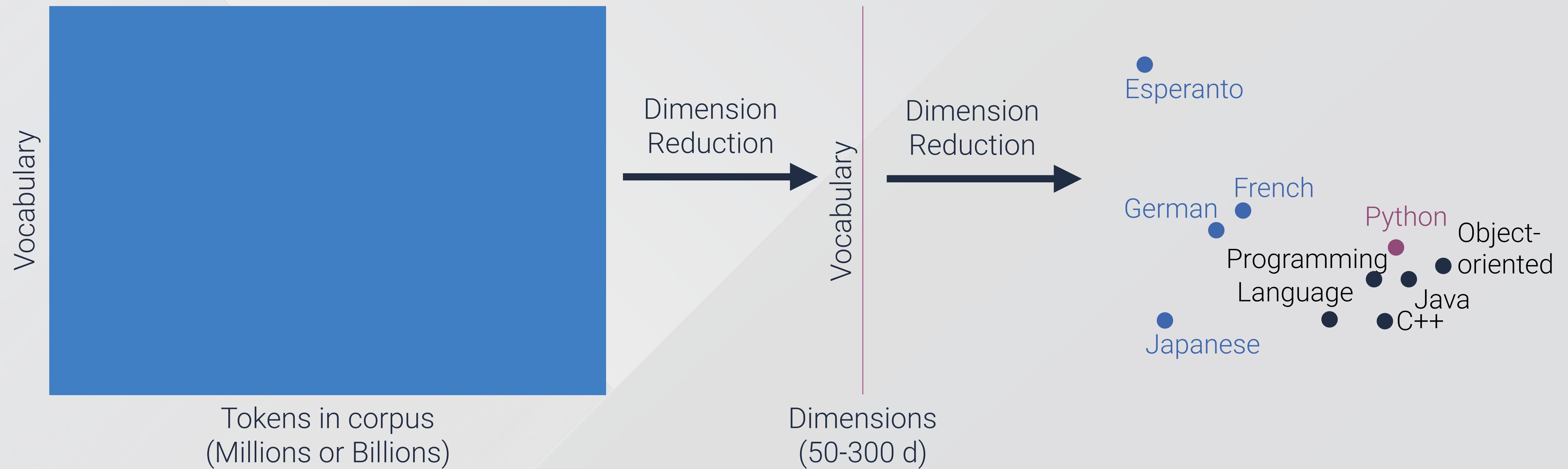
Context

Word

Context

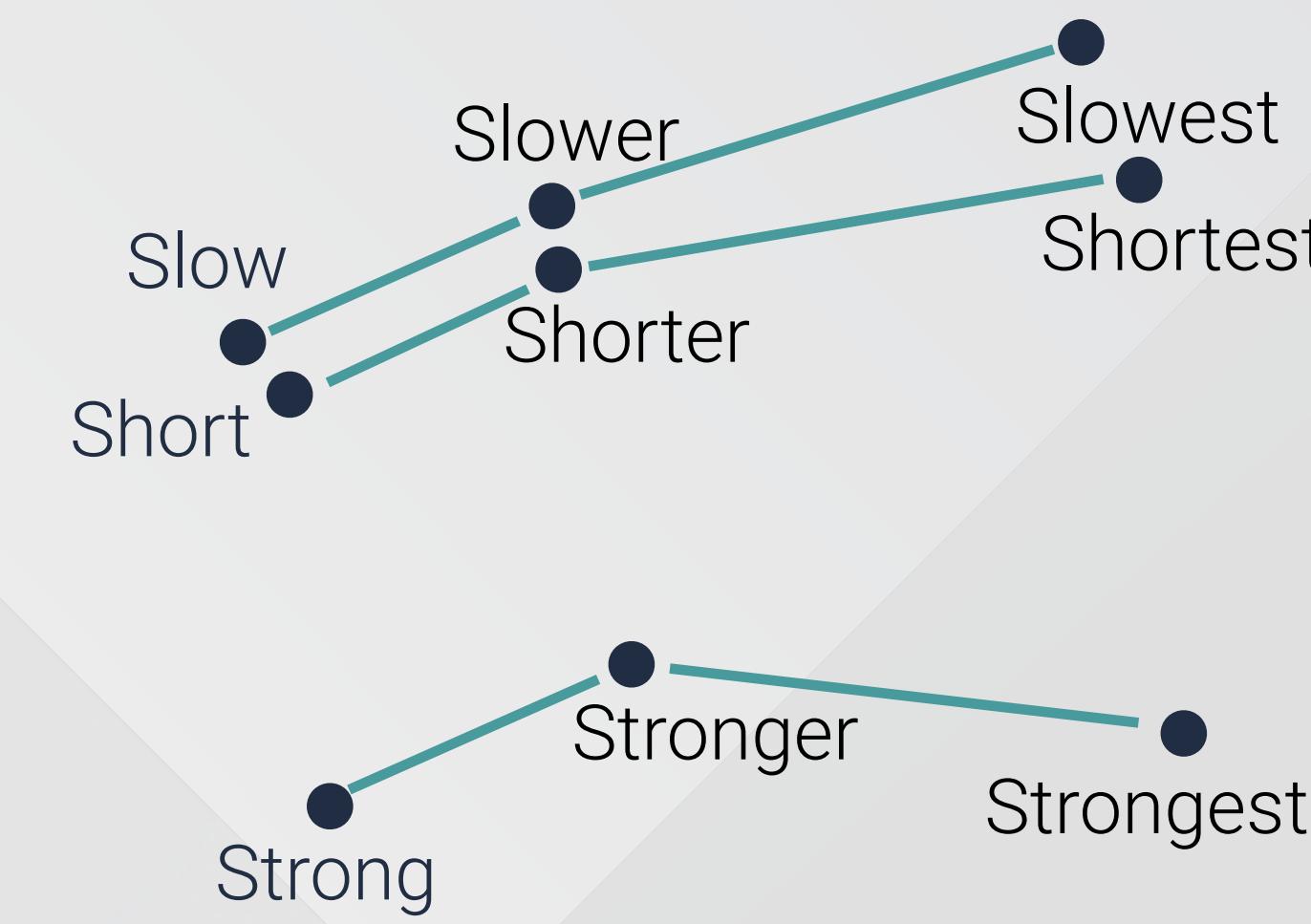
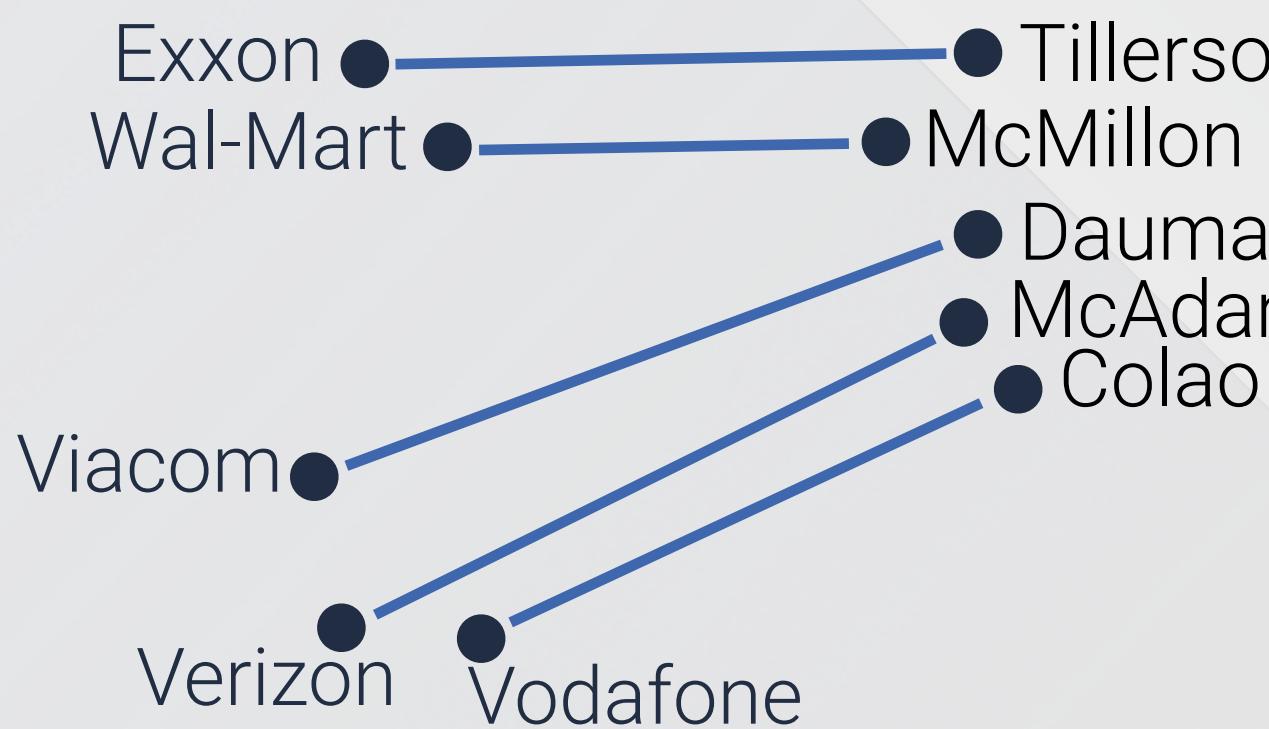
A simplified representation of word embeddings

Dimension reduction is key to all types of embeddings models



Word embeddings capture entity relationships

Dimensionality enables comparison between word pairs along many axes



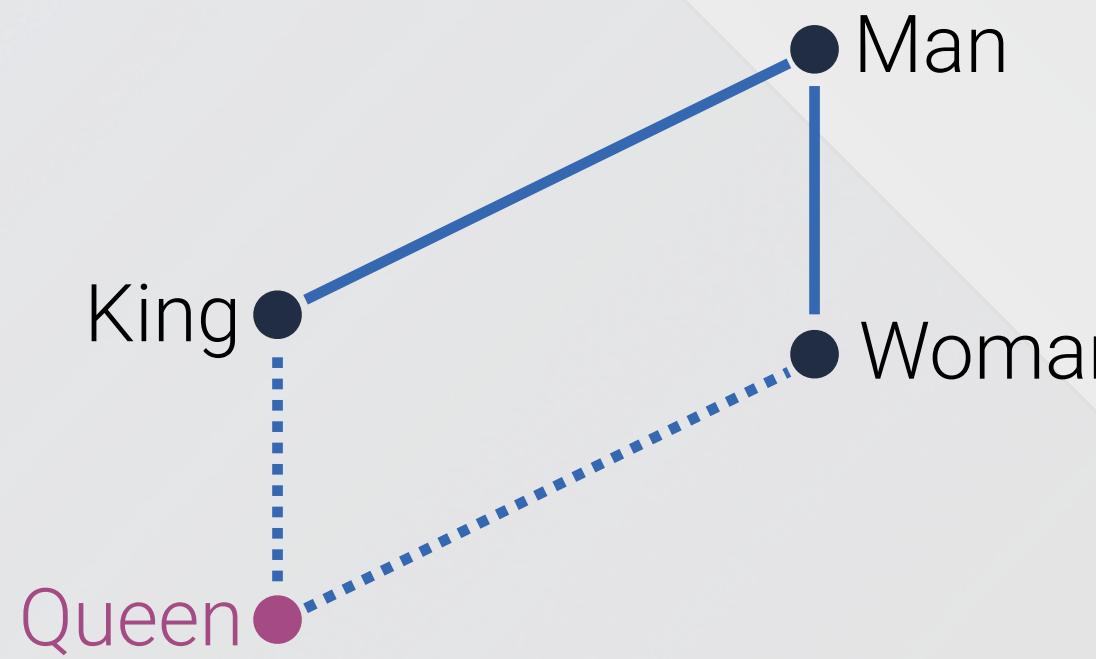
Hierarchies

Comparatives and Superlatives

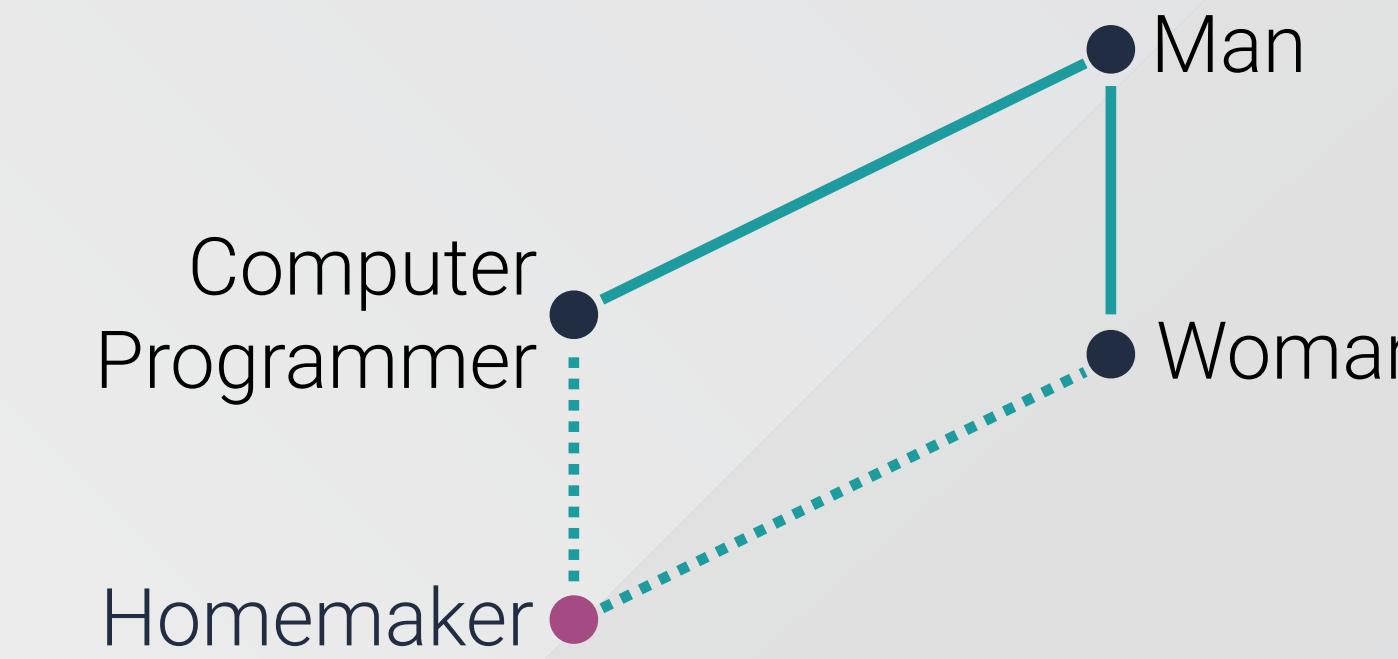
Man :: King as Woman :: ?

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

Drawbacks of pretrained word embeddings

Casing

Abbreviations vs Words
e.g. IT vs it

Out of Vocabulary Words

Domain Specific Words & Acronyms

Polysemy

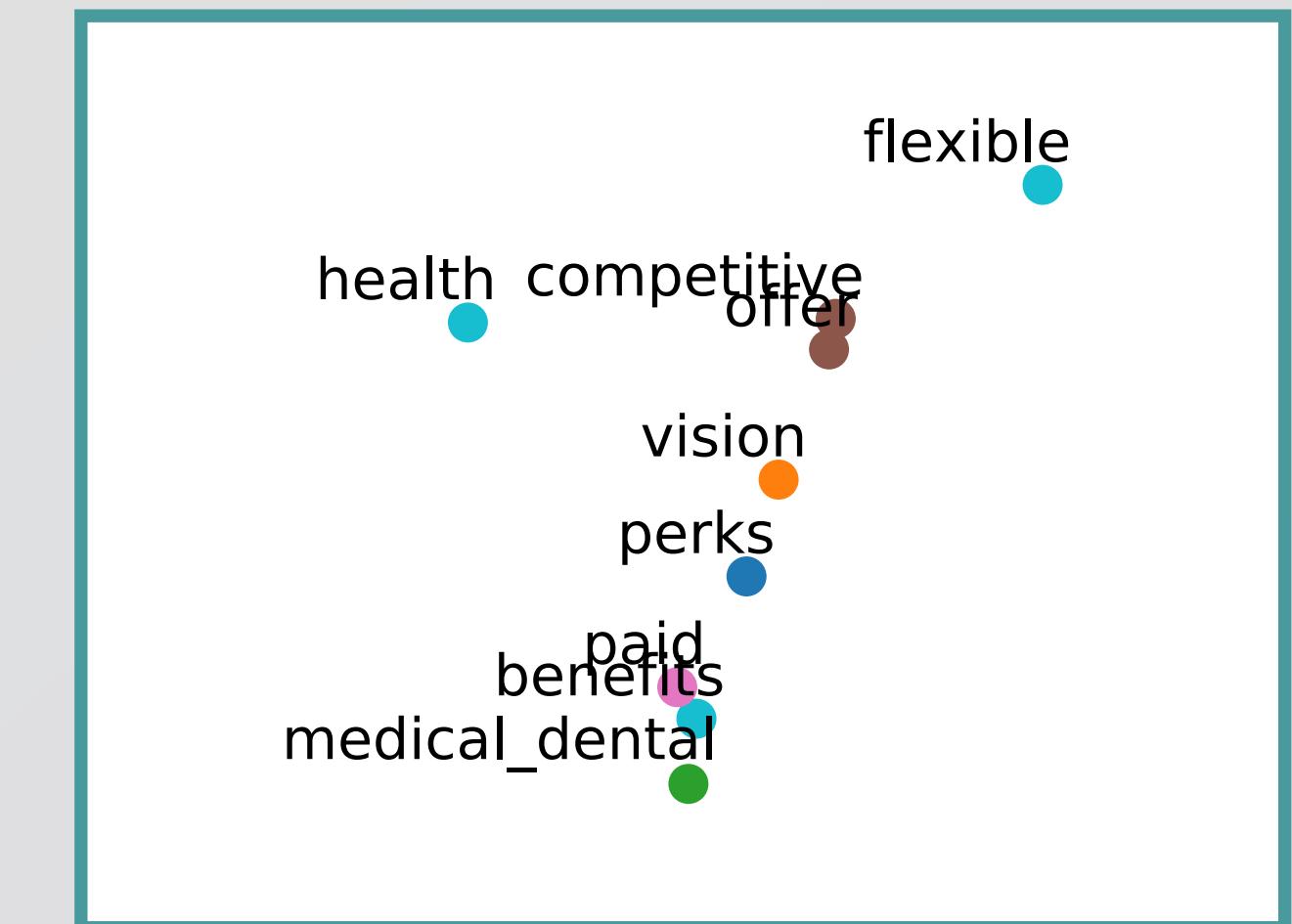
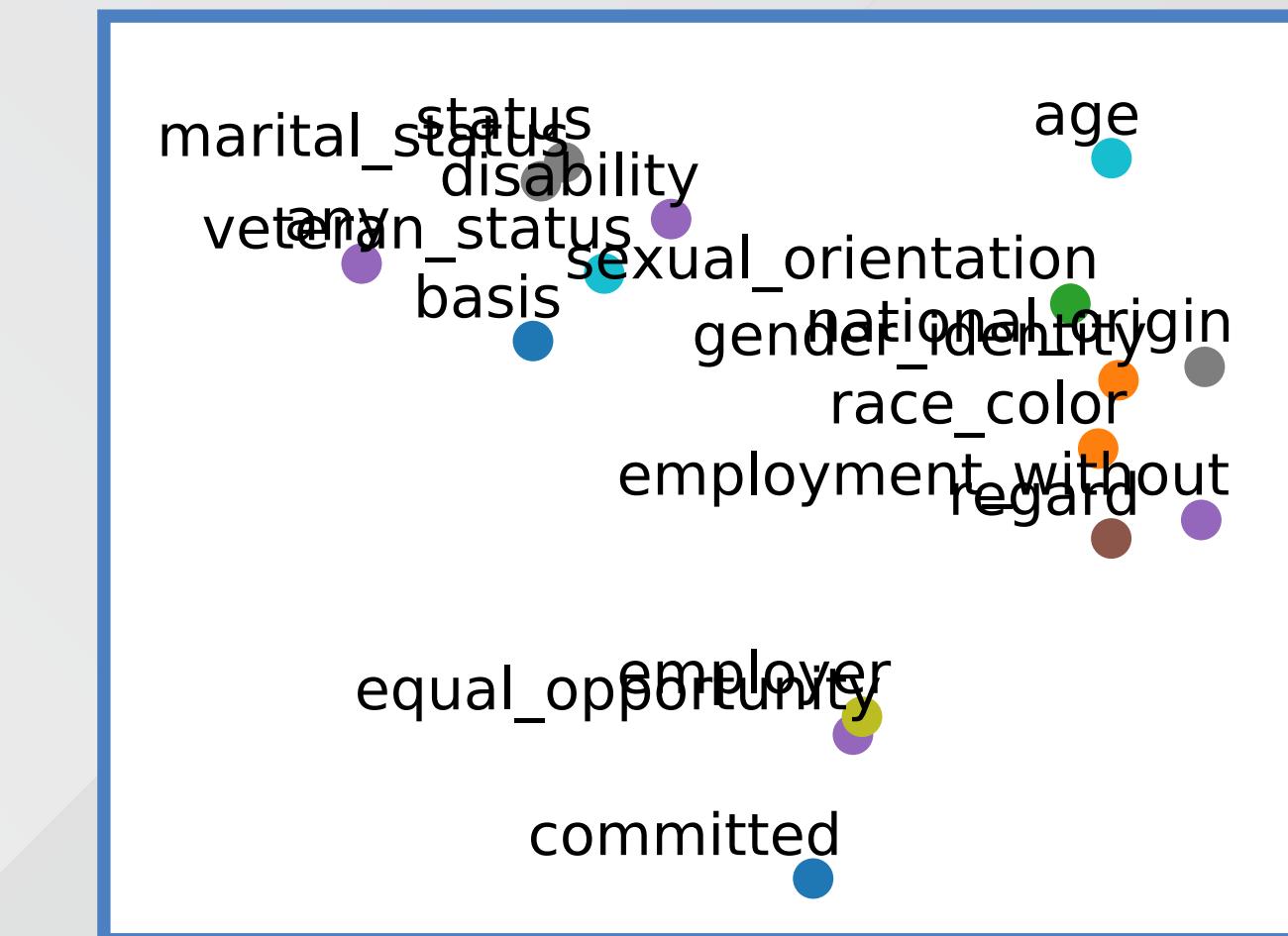
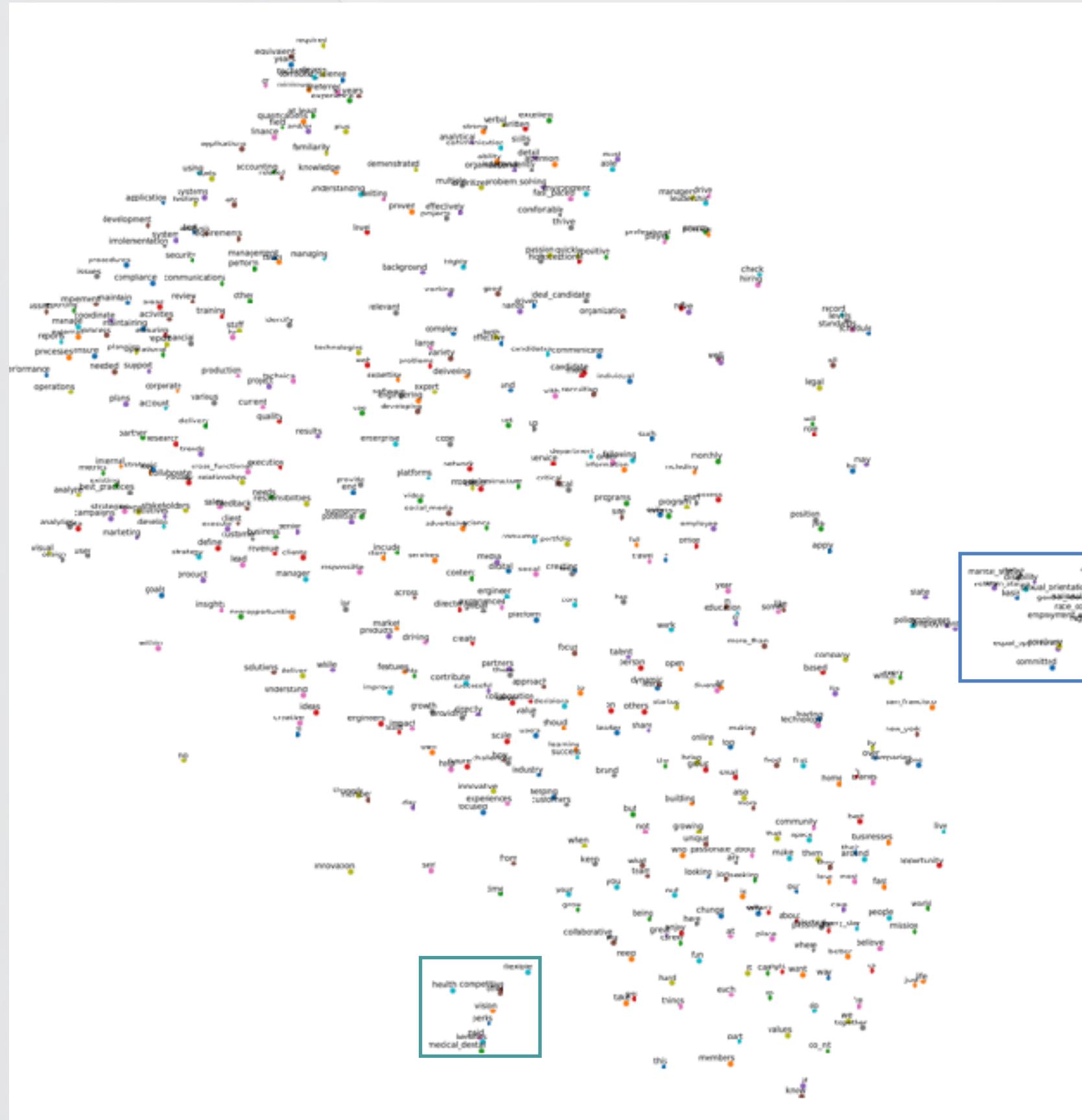
Words with multiple meanings
e.g. drive (a car) vs drive (results)
e.g. Chef (the job) vs Chef (the language)

Multi-word Expressions

Phrases that have new meanings
e.g. Front-end vs front + end

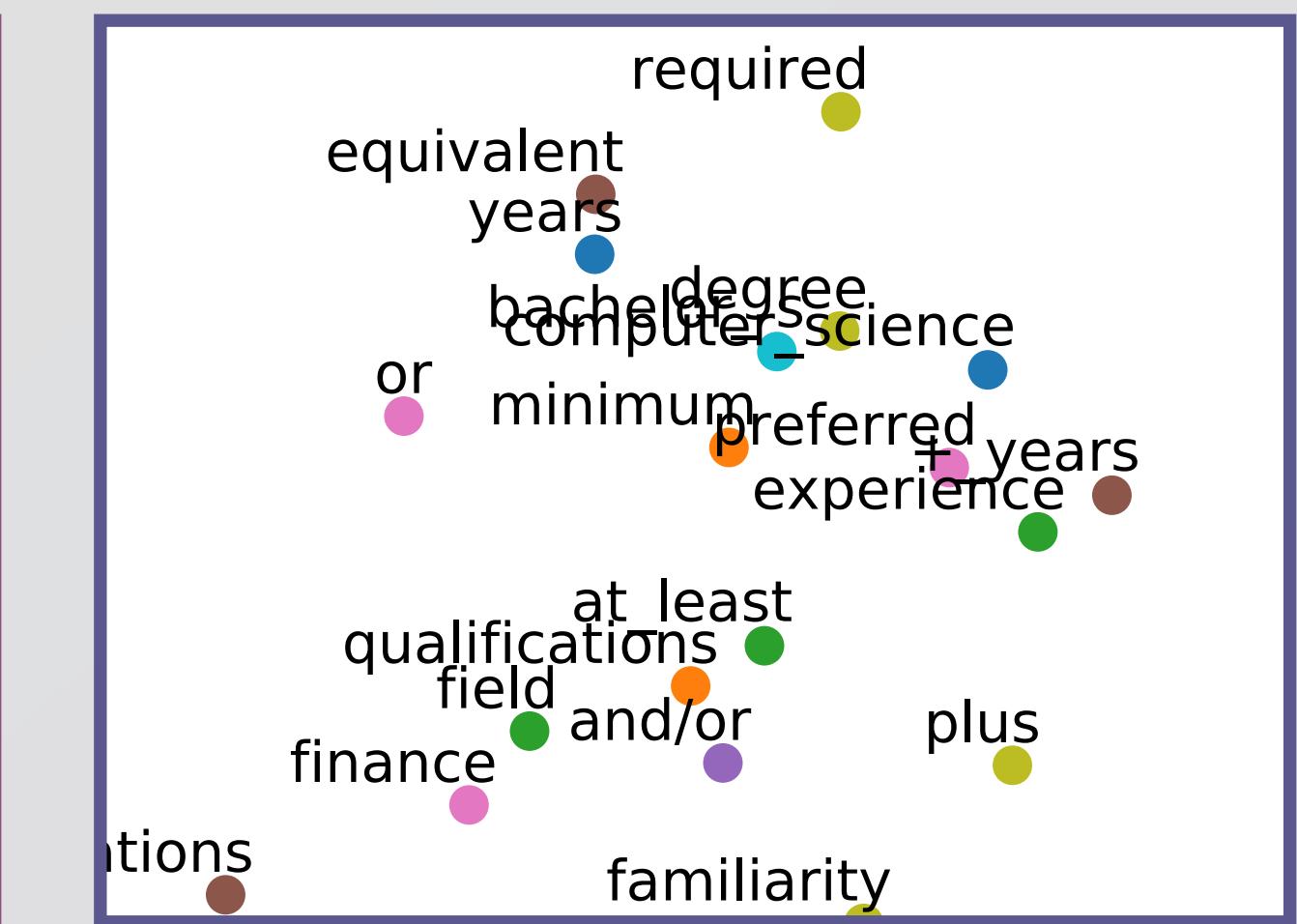
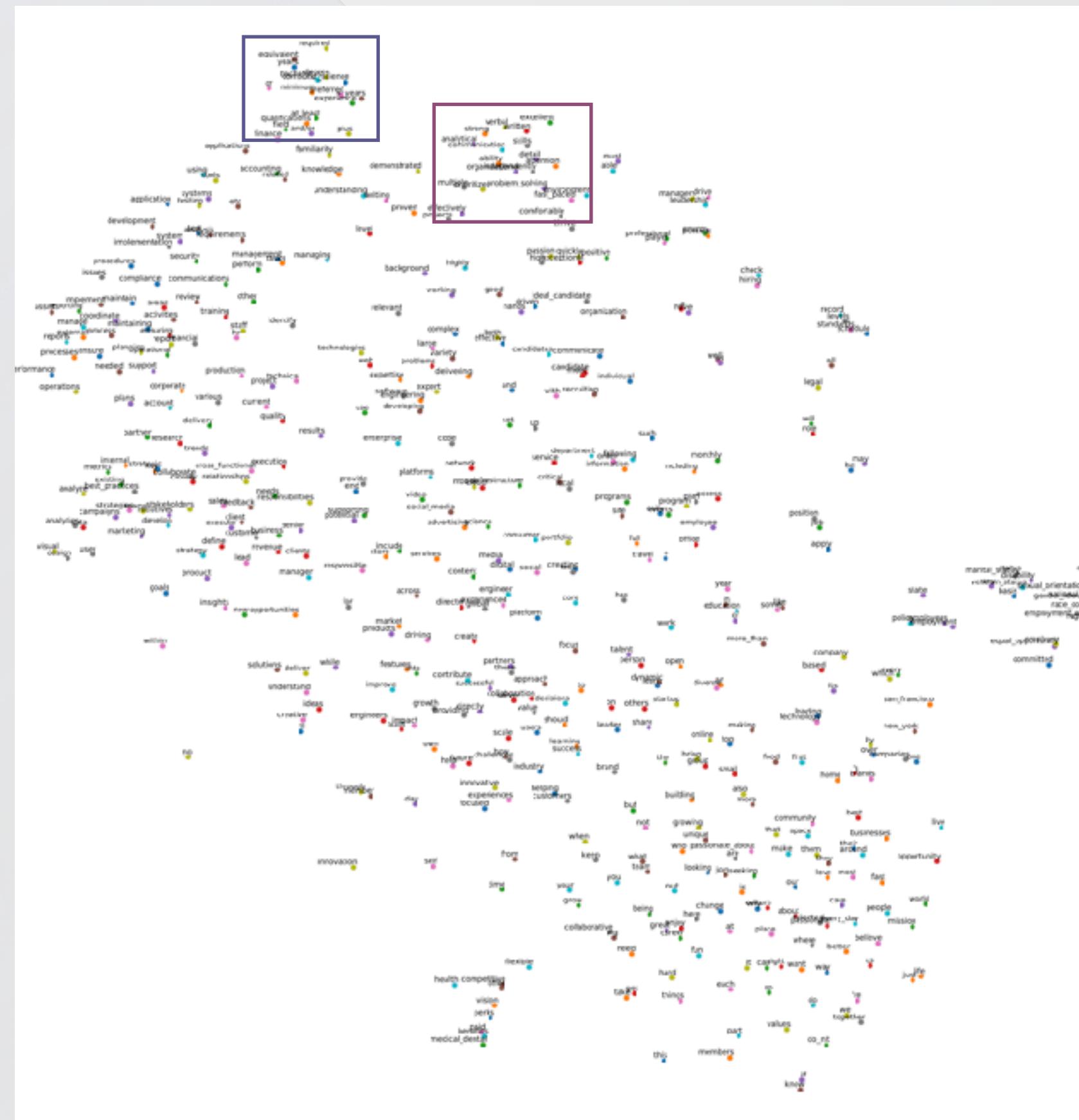
Career language embedding model

Identified equal opportunity and perks language



Career language embedding model

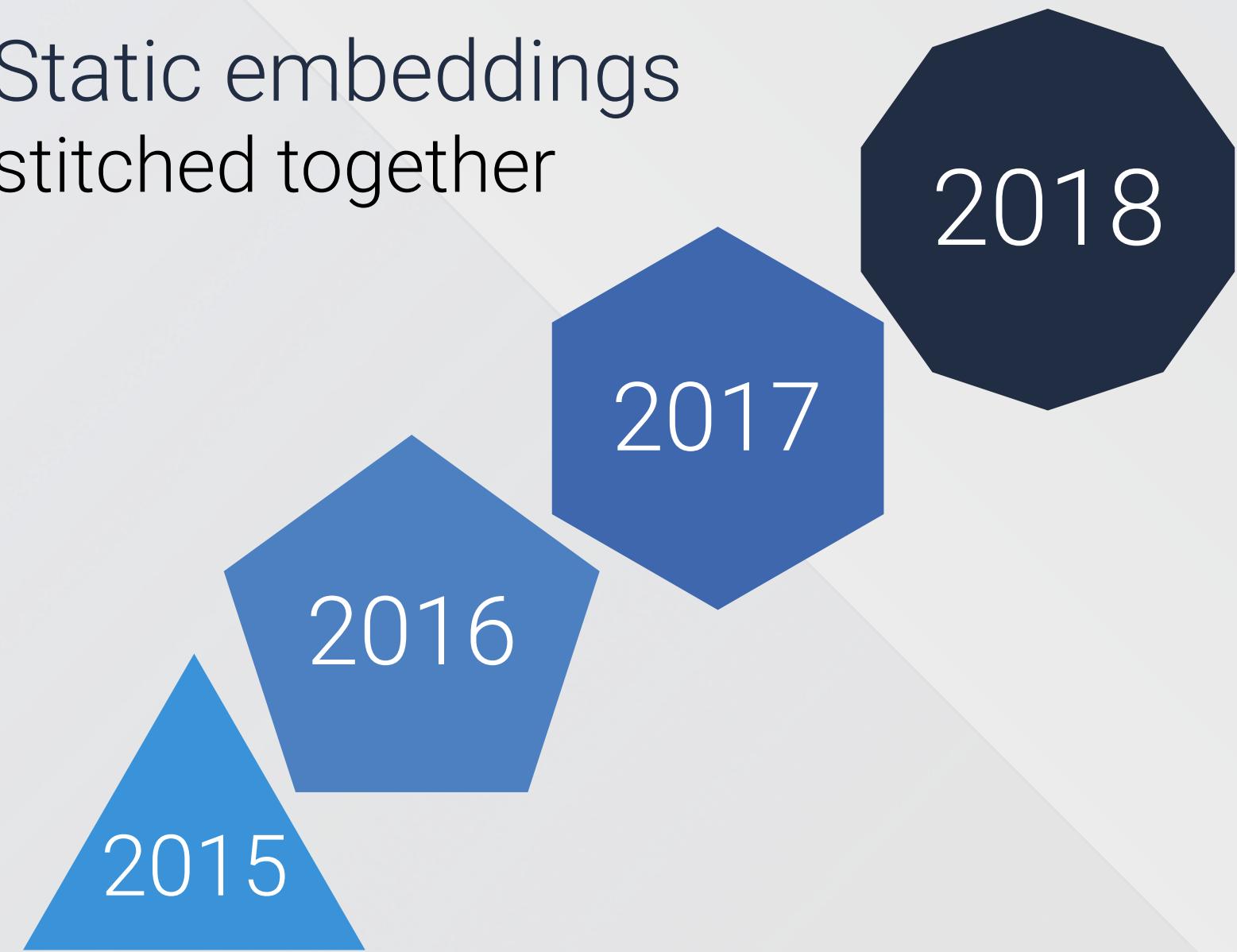
Identified 'soft' skills and language around experience



I've got 300 dimensions...
but time ain't one

Two approaches to connect embeddings

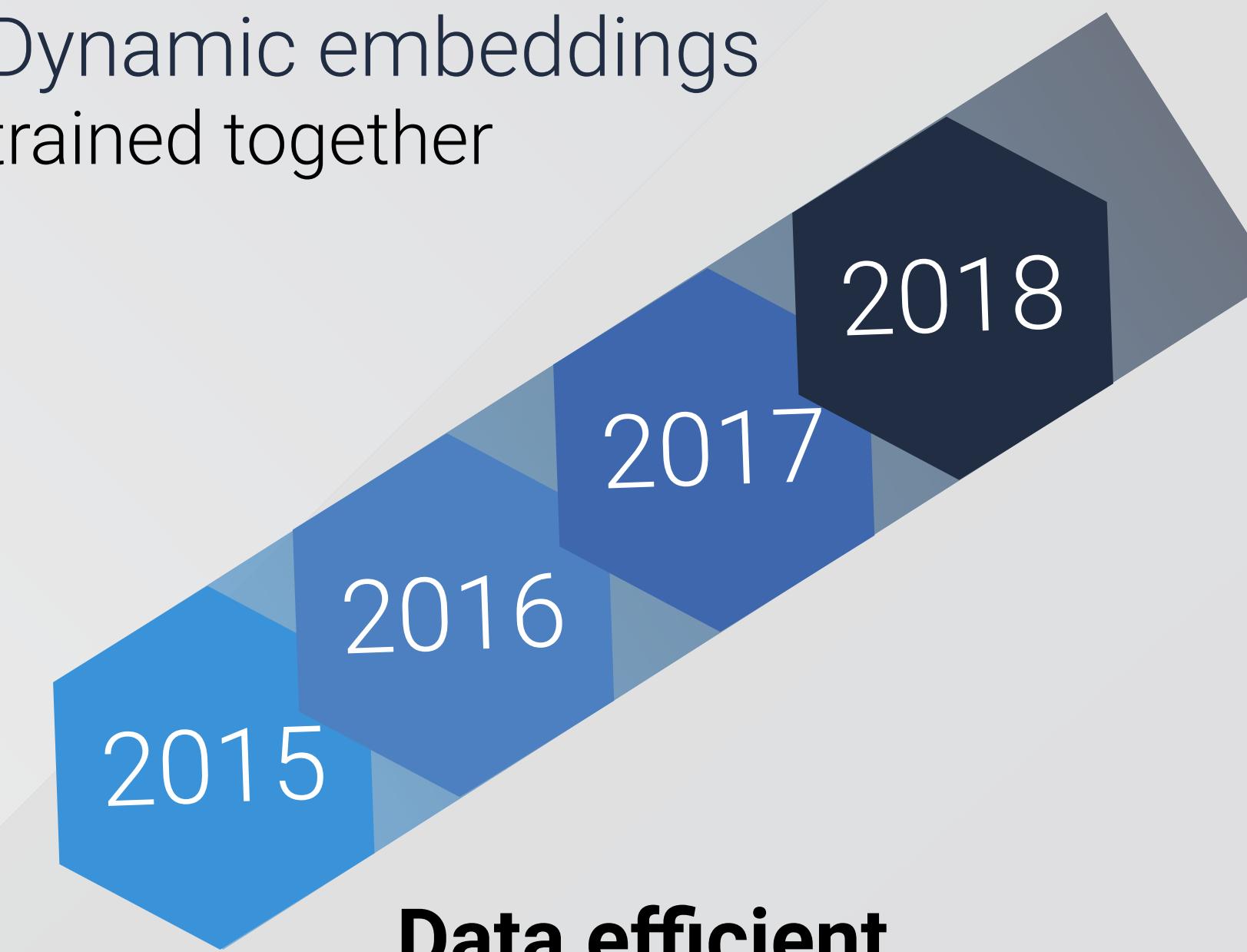
Static embeddings
stitched together



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](#)

Dynamic Bernoulli embeddings

Outputs facilitate quick analysis of trends

Absolute drift

Identifies top words whose usage changes over time course

Embedding neighborhoods

Extract semantic changes by nearest neighbors of drifting words

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

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

tap Recruit.co

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

Small corpus identified MBAs and PhDs

Reduced requirement for advanced degrees in many jobs

Demand for MBAs is Falling in US Roles

MBAs in All Jobs
-35%

MBAs in DS Jobs
-15%

MBAs in Tech Jobs
+30%

MBAs in All Jobs
-10%

Demand for PhDs is Falling in US Roles

PhDs in All Jobs
-35%

PhDs in DS Jobs
-20%

PhDs in ML Jobs
-30%

PhDs in DS Jobs
+200%

PhDs in ML Jobs
+100%

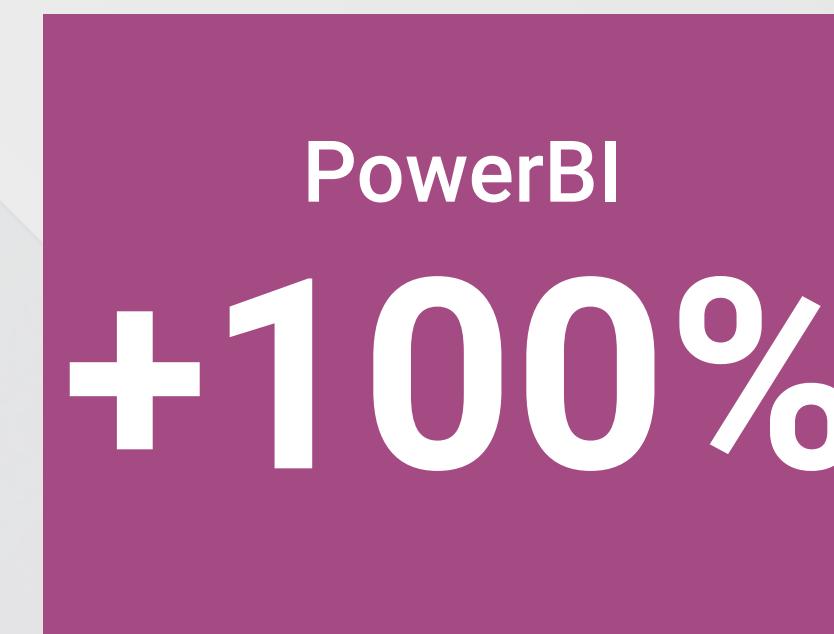
and in Australian Roles

but not in Australian Roles

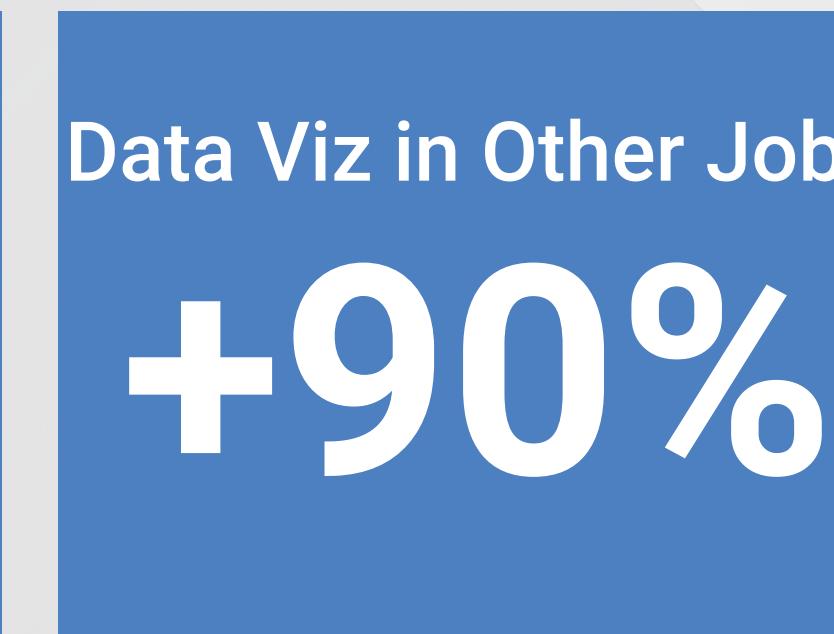
Small corpus identified skill demands

Data Viz is up in lots of different roles

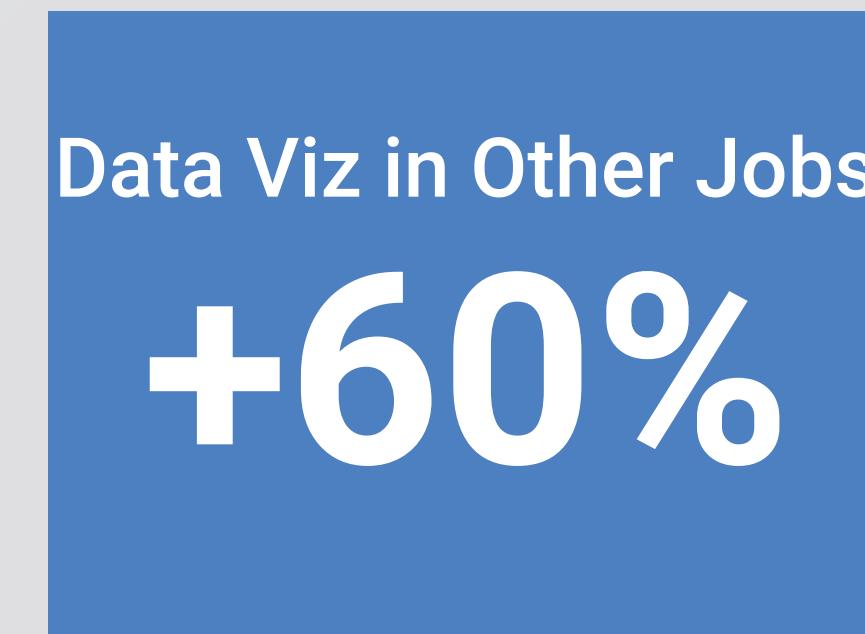
Demand for Data Visualization tools is up



Data Viz growth in US Non-DS Roles



and Australian Non-DS Roles

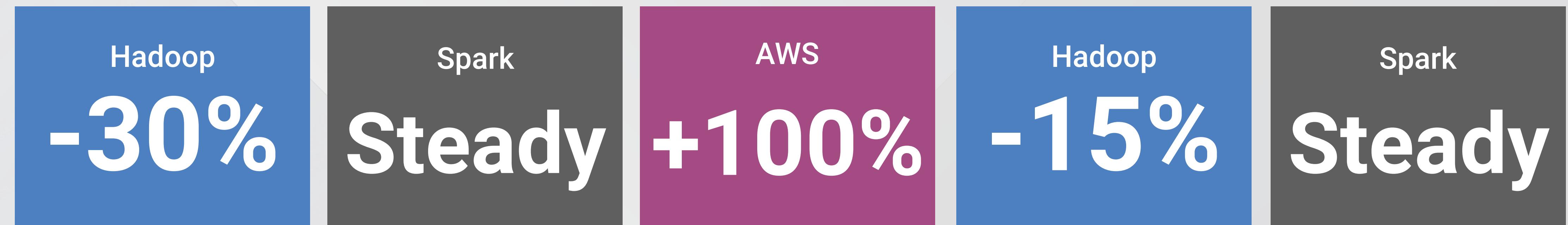


Blue boxes indicate phrases identified from top drifting words analysis.
Grey and pink boxes indicate 'control' skills.

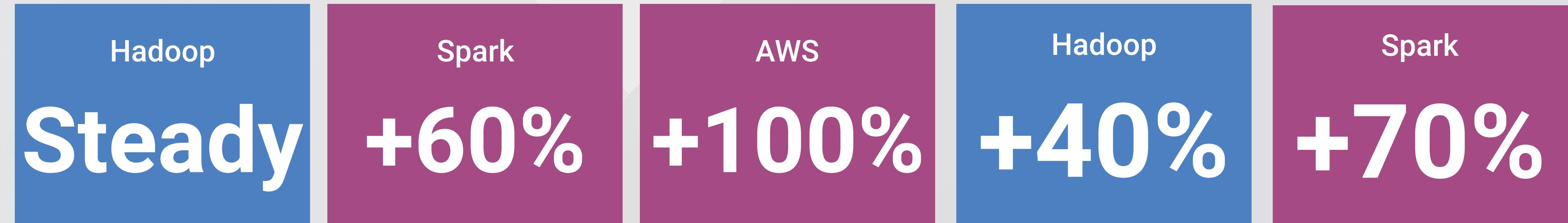
Small corpus identified skill demands

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

US-based Data Science Jobs



US-based (non-DS) Tech Jobs

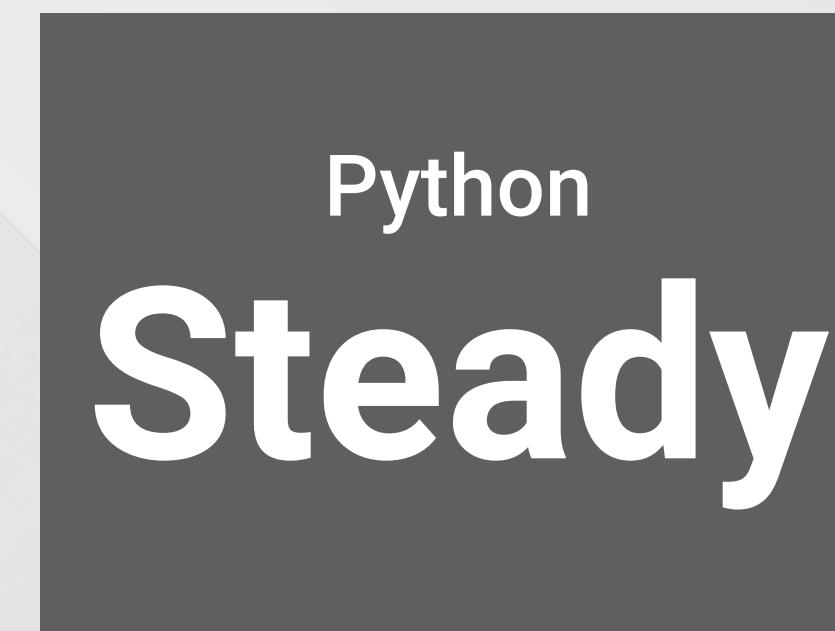


Blue boxes indicate phrases identified from top drifting words analysis.
Grey and pink boxes indicate 'control' skills.

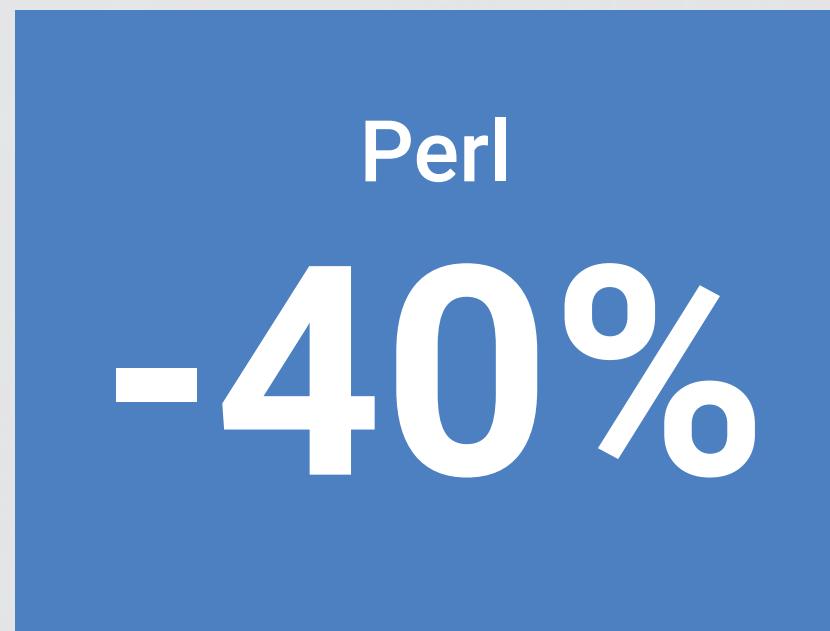
Battle of the languages: Supply vs Demand

Demand for Perl is down in Data Science Roles

US-based Data Science Jobs



Aussie-based Data Science Jobs



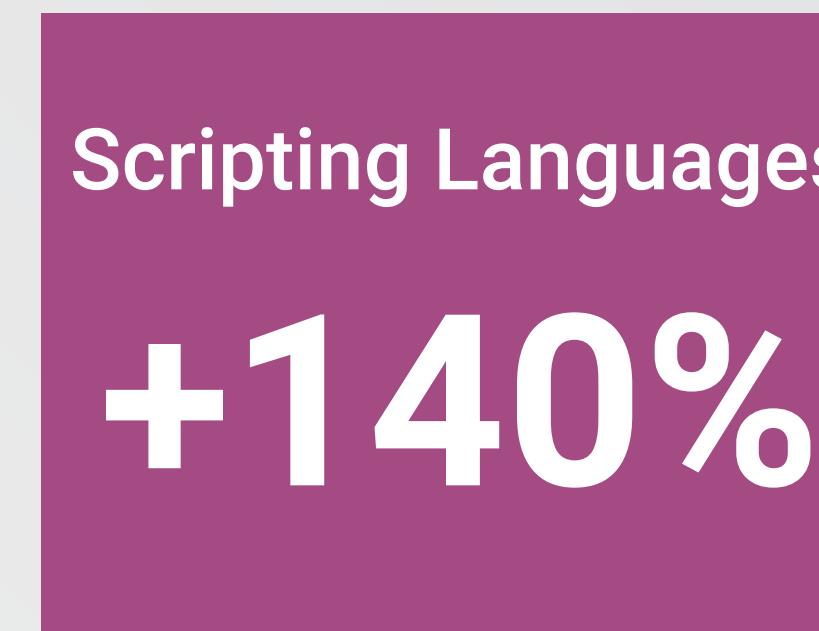
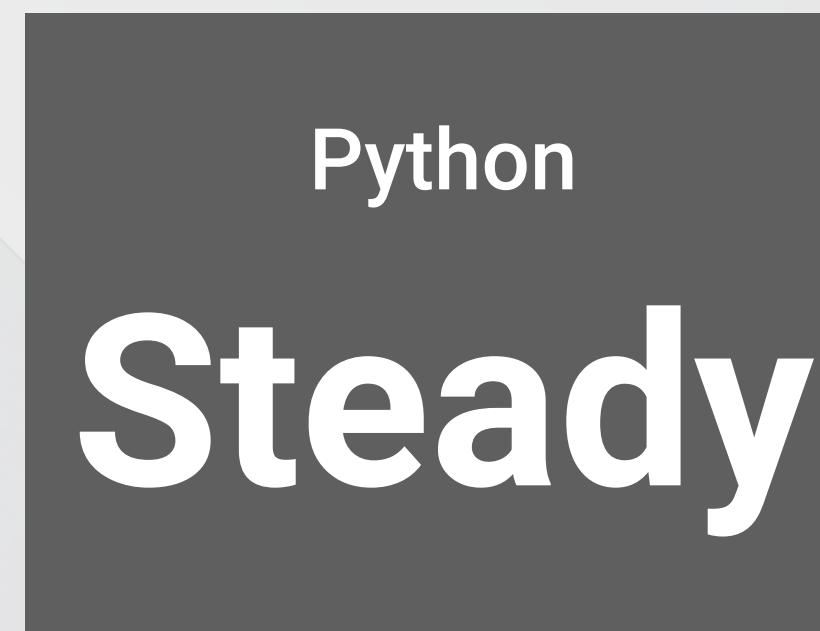
- Python, the fastest-growing major programming language, has risen in the ranks of programming languages in our survey yet again, edging out Java this year and standing as the second most loved language (behind Rust).

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

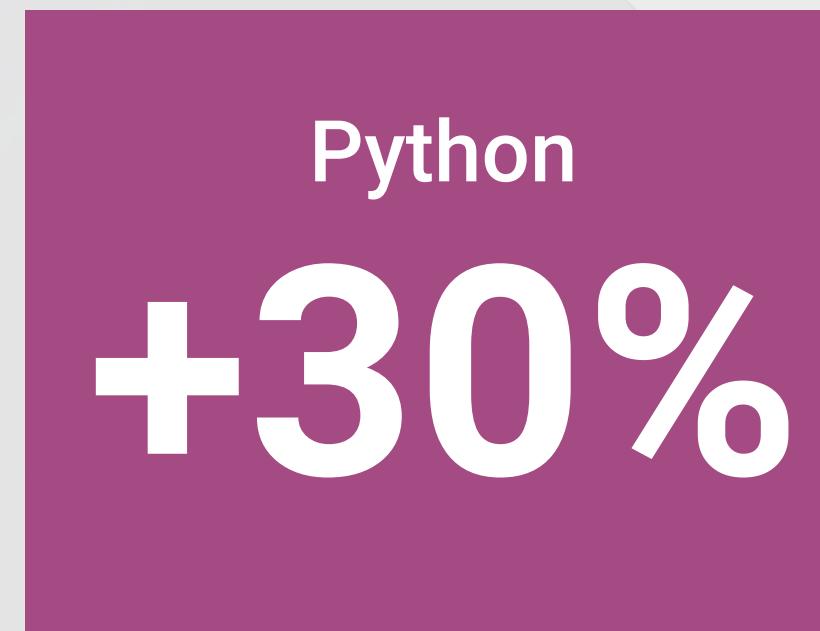
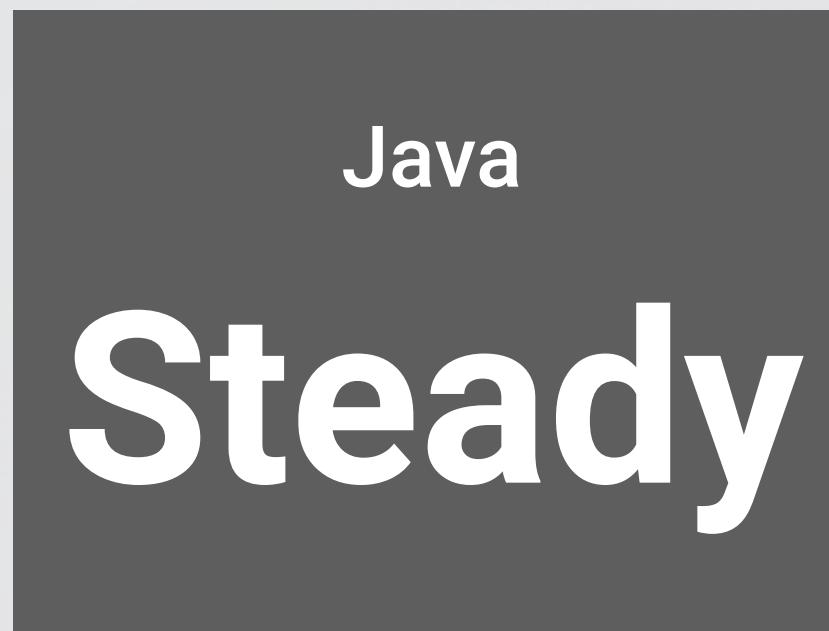
Battle of the languages: Supply vs Demand

Scripting Languages absorbing changes

Data Science Jobs



Tech Jobs (non-DS)



- 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.

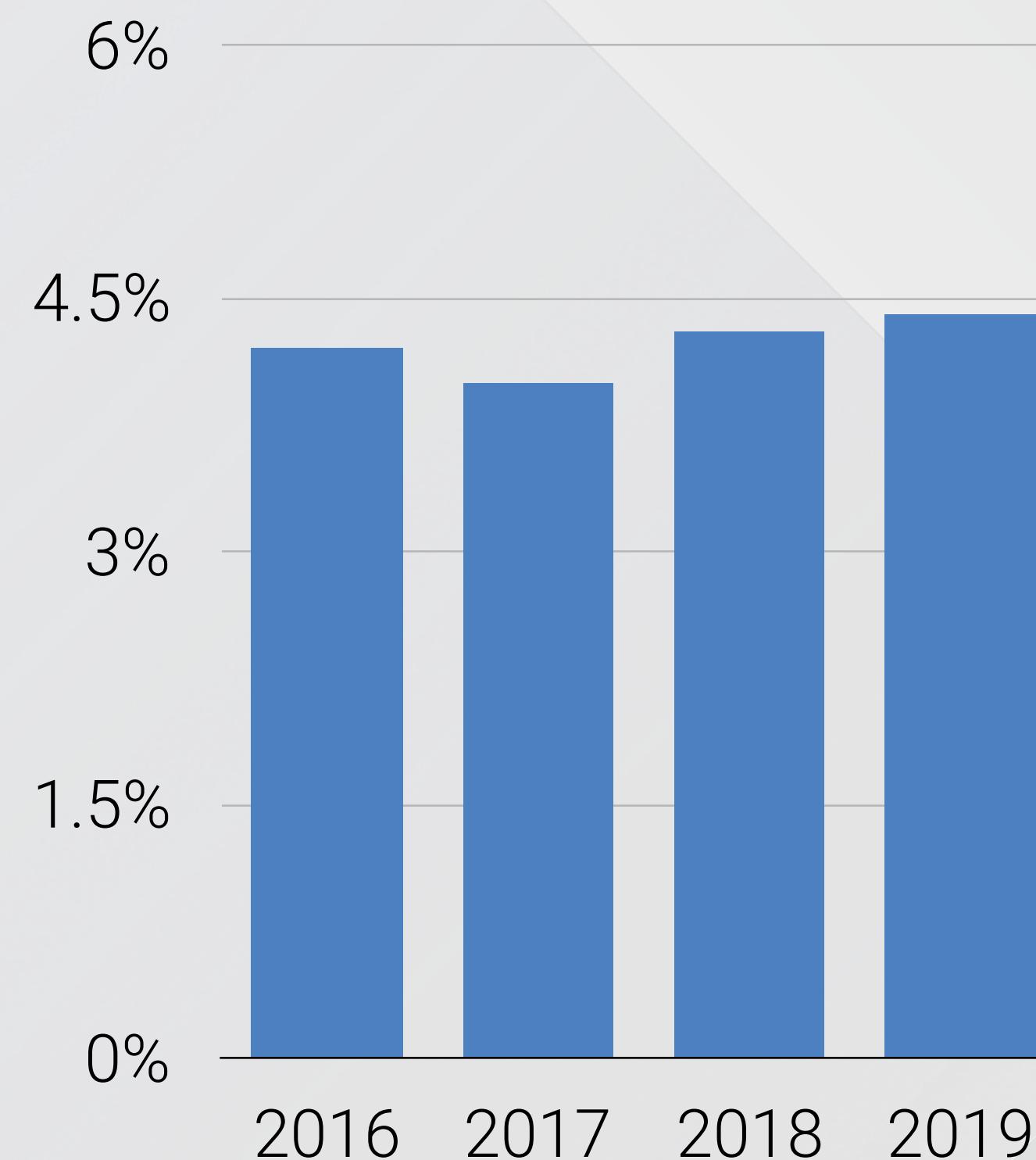
Experiments with dynamic embeddings

	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

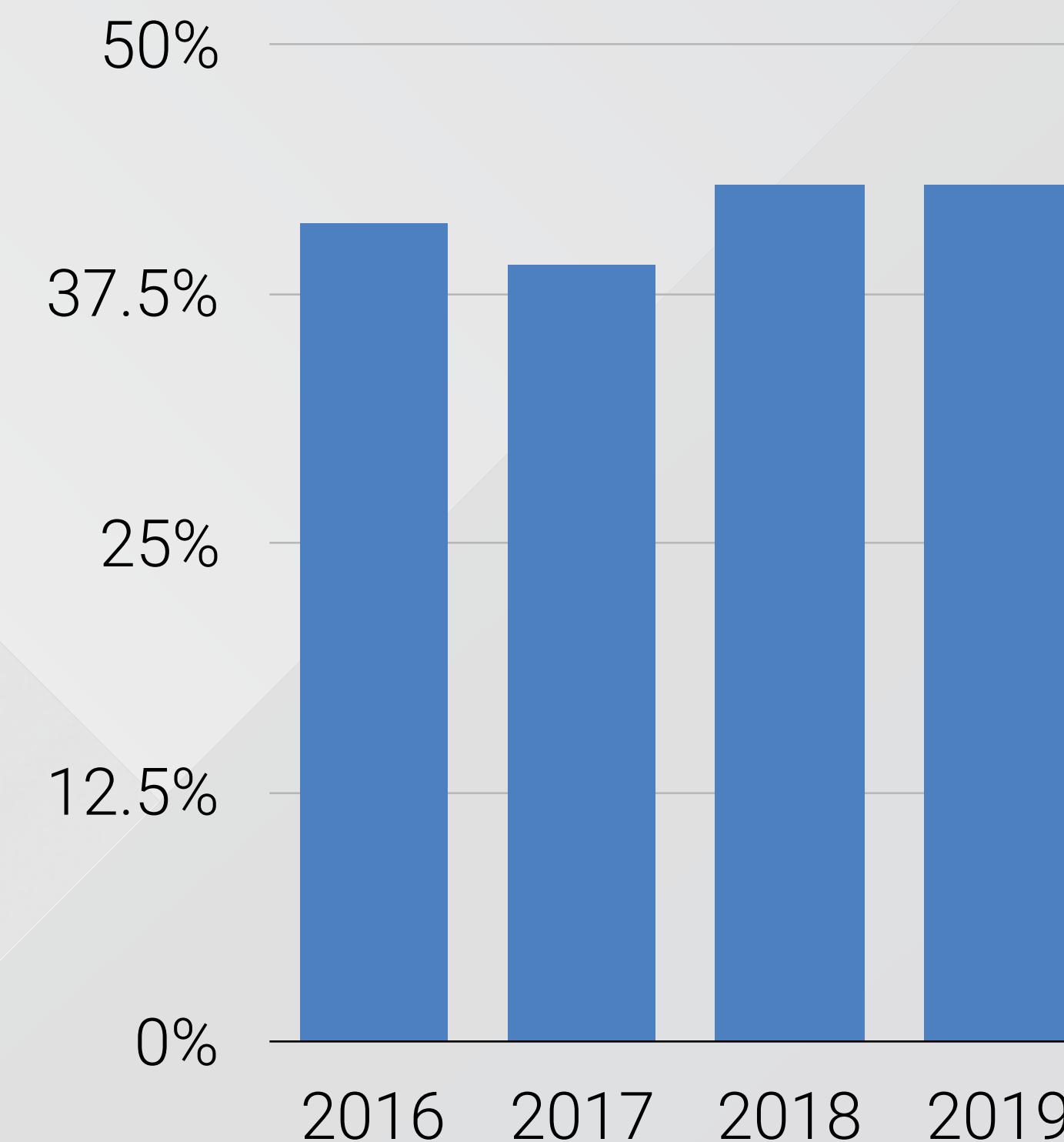
Large corpus identified SQL as a top drifting word

But no difference in demand for SQL in jobs

All Jobs



Data Science & Tech Jobs



Large corpus identified SQL as a top drifting word

Large corpus identified role-type dependent shifts in requirements

SQL requirement increases in specific functions

FinTech Roles
Steady

Sales Roles
Steady

Marketing Roles
Steady

FP&A Roles
+70%

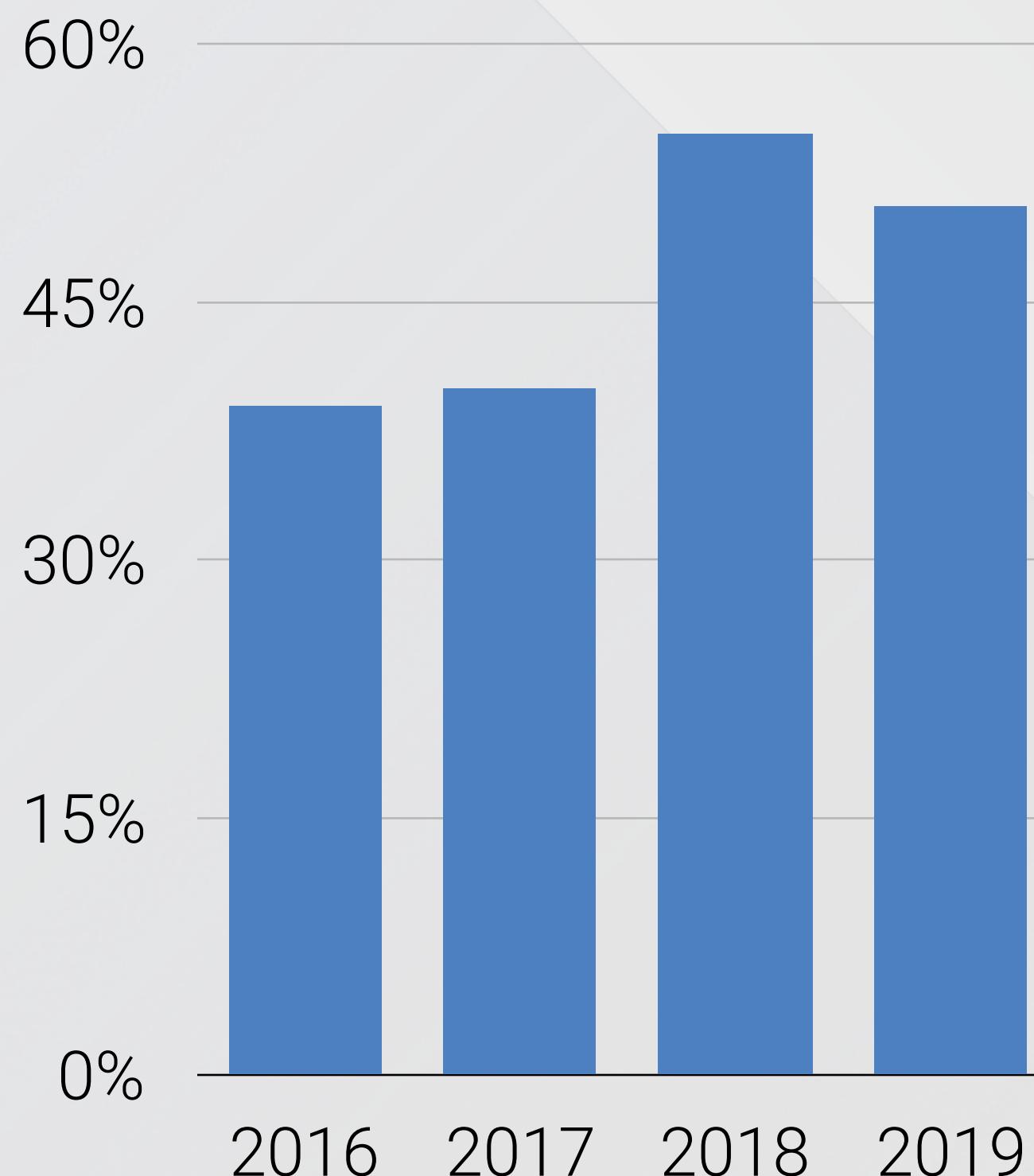
BizDev Roles
+50%

HR Roles
+25%

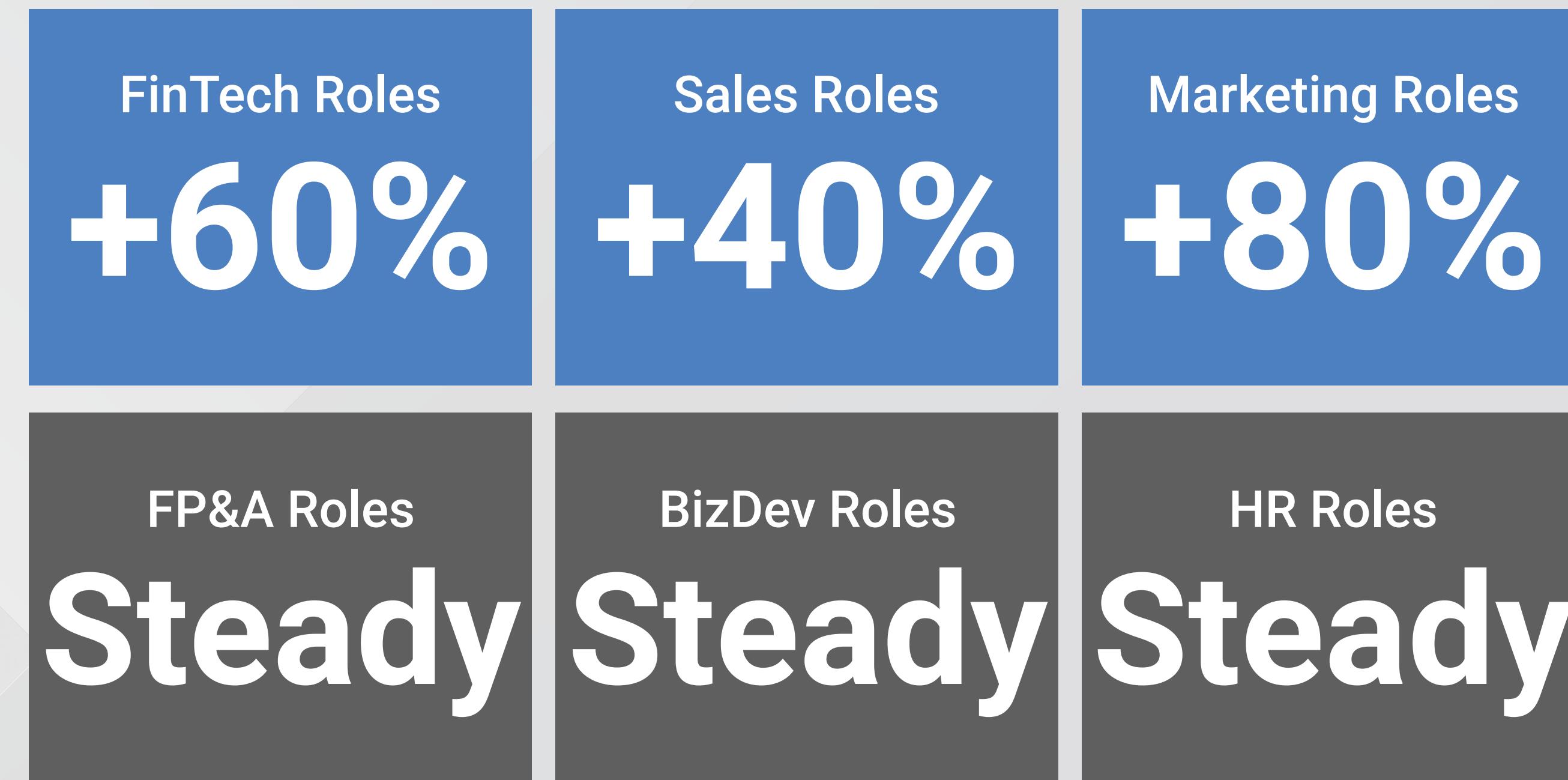
Large corpus identified SQL as a top drifting word

Requirement for SQL hit a high in 2018 in Australian jobs

Aussie Data Science & Tech Jobs



SQL requirement is inverted compared to US Roles



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 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

Using Dynamic Embeddings without Time



- Dynamic Embeddings are not intrinsically based on time units
- They can be used for any form of ordinal data
- Our experiments show you need a minimum of 3 groups to get robust and reproducible insights

Thank you YOW! Data

Maryam Jahanshahi Ph.D.
Research Scientist

 @mjahanshahi
 maryam-j

tap Recruit.co

The Battle for the Aussie DevOps Market

GCP is a racing ahead in the Australian DevOps, Azure is not.

Tech requirements in Aussie DevOps Jobs



Before 2017, GCP and AWS were mentioned in a similar number of DevOps jobs. After 2018, GCP has seen significantly more growth.

Fewer DevOps roles in Australia required familiarity with Azure. This has been steadily declining over the last few years.

Caveat emptor: This analysis reflects requirements from a sampling of job descriptions. Please do not use this to make important life decisions!