Building a Robot Judge: Data Science for the Law

9. Word Embeddings

Elliott Ash

Word Embeddings

- Word embeddings:
  - a hot topic in NLP since arrival of Word2Vec in 2013.
  - refers to a class of statistical models that **represent words or phrases as points in a vector space**.
- The key idea is to represent the meaning of words by the neighboring words – their **contexts**.
- You might hear "word embeddings" and "word2vec" interchangeably, although word2vec technically refers to a particular implementation of a word embedding model.
  - the other well-known implementation is gloVe, which is faster but has similar performance/applications

What is an Embedding?

- An "embedding" is a tool in machine learning that is quite different than anything from applied statistics.
  - It is a vector representation of a categorical variable.
  - where the spatial location of the vector encodes predictive information about the category.
- e.g. U.S. states:
  - instead of including a fifty-dimensional categorical variable, include two-dimensional latitude and longitude
  - or initialize each state to a random two-dimensional vector, and let the model decide where to move the states to improve prediction on your task (e.g. responses to trade shocks).

An embedding layer is just matrix multiplication

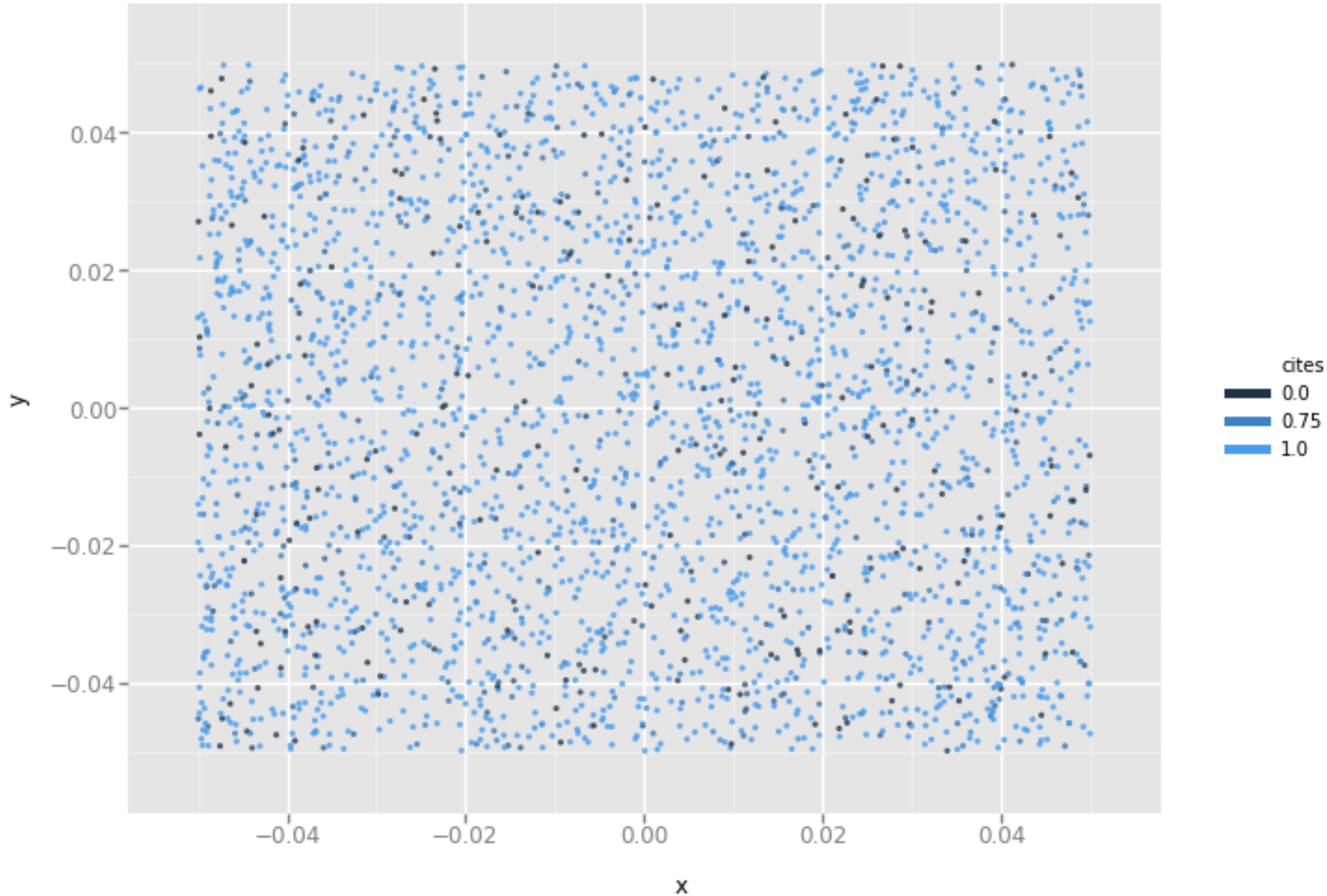
- An embedding layer can be represented as

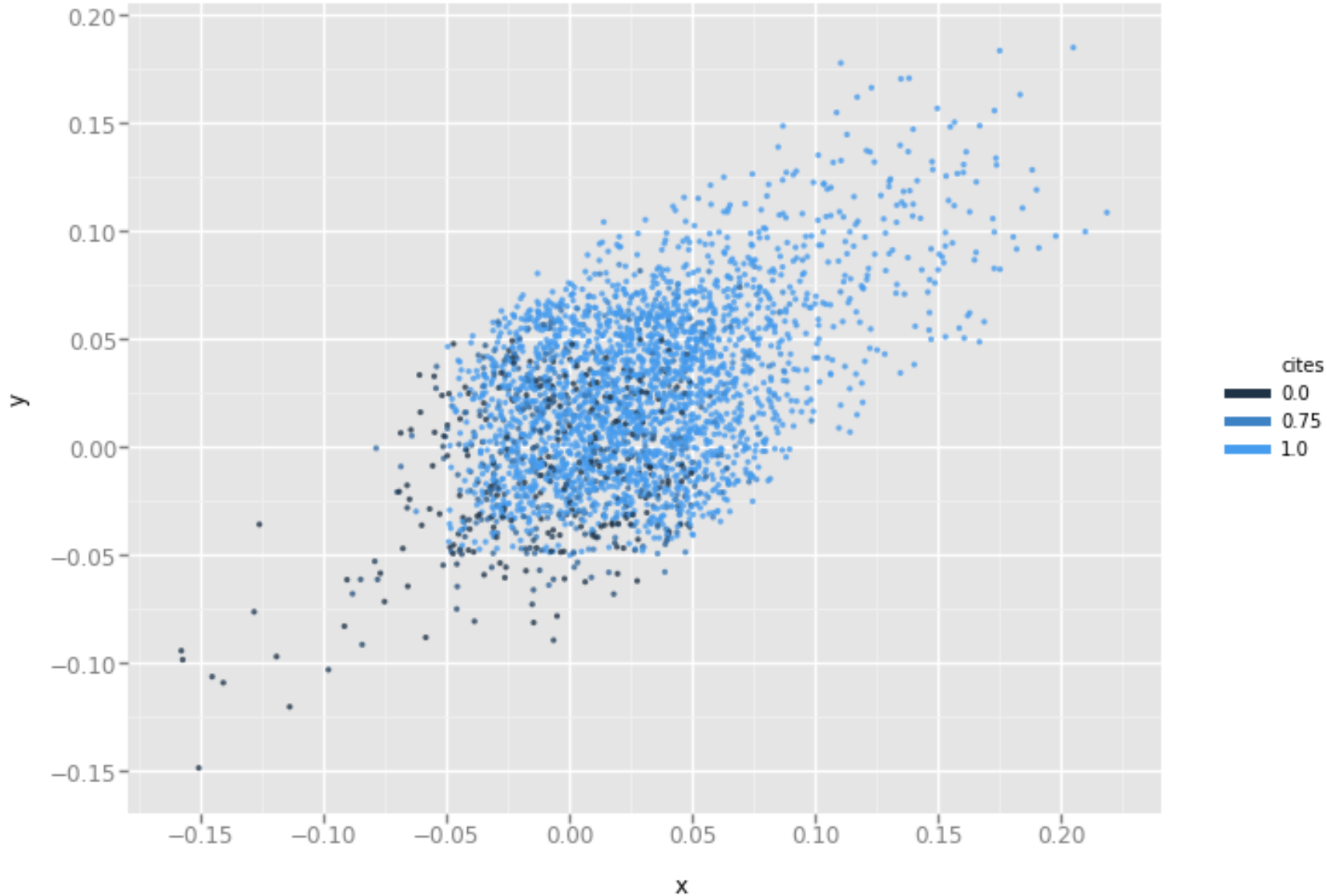
$$\underbrace{x}_{n\times 1} = \underbrace{\Omega'}_{n\times m}\underbrace{w}_{m\times 1}$$

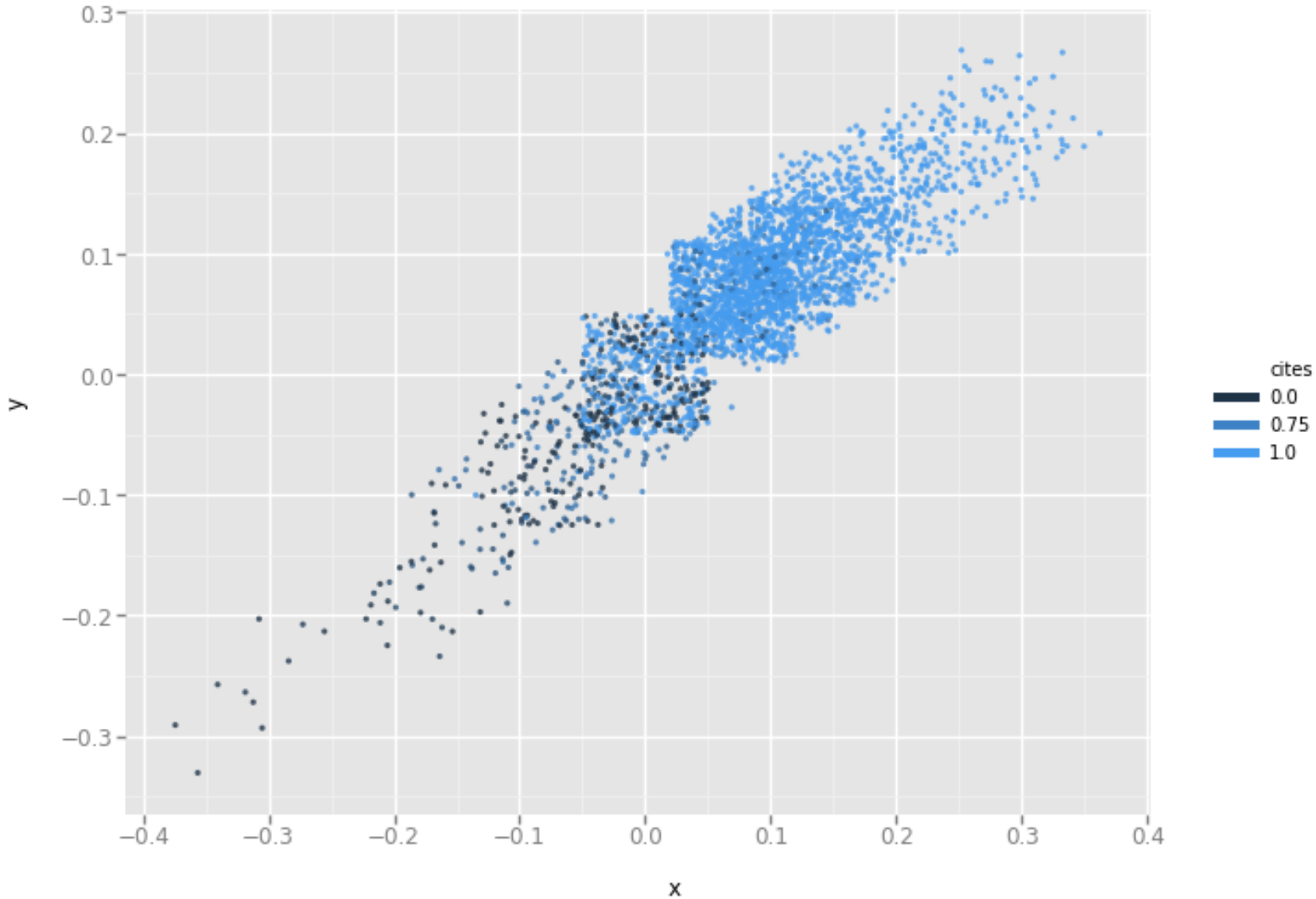
  - w, a categorical variable, a one-hot-encoded vector that is all zeros, with a single item equaling one.
    - The input to the embedding layer.
  - x, a dense representation of the variable.
    - The output of the embedding layer.
  - An embedding matrix Ω.
    - the model learns the weights of this matrix.

The Embedding Matrix Ω

- The model learns the weights of the embedding matrix in the same way that it would learn any model parameters.
- The rows of the matrix correspond to vectors for the n categories.
  - These are the "word vectors" that people talk about when they mention word embeddings or Word2Vec.


0.04
0.02
0.00
−0.02
−0.04
y
−0.04
−0.02
0.00
0.02
0.04
x
cites
0.0
0.75
1.0


0.20
0.15
0.10
0.05
0.00
−0.05
−0.10
−0.15
y
−0.15
−0.10
−0.05
0.00
0.05
0.10
0.15
0.20
x
cites
0.0
0.75
1.0


0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
y
−0.4
−0.3
−0.2
−0.1
0.0
0.1
0.2
0.3
0.4
x
cites
0.0
0.75
1.0

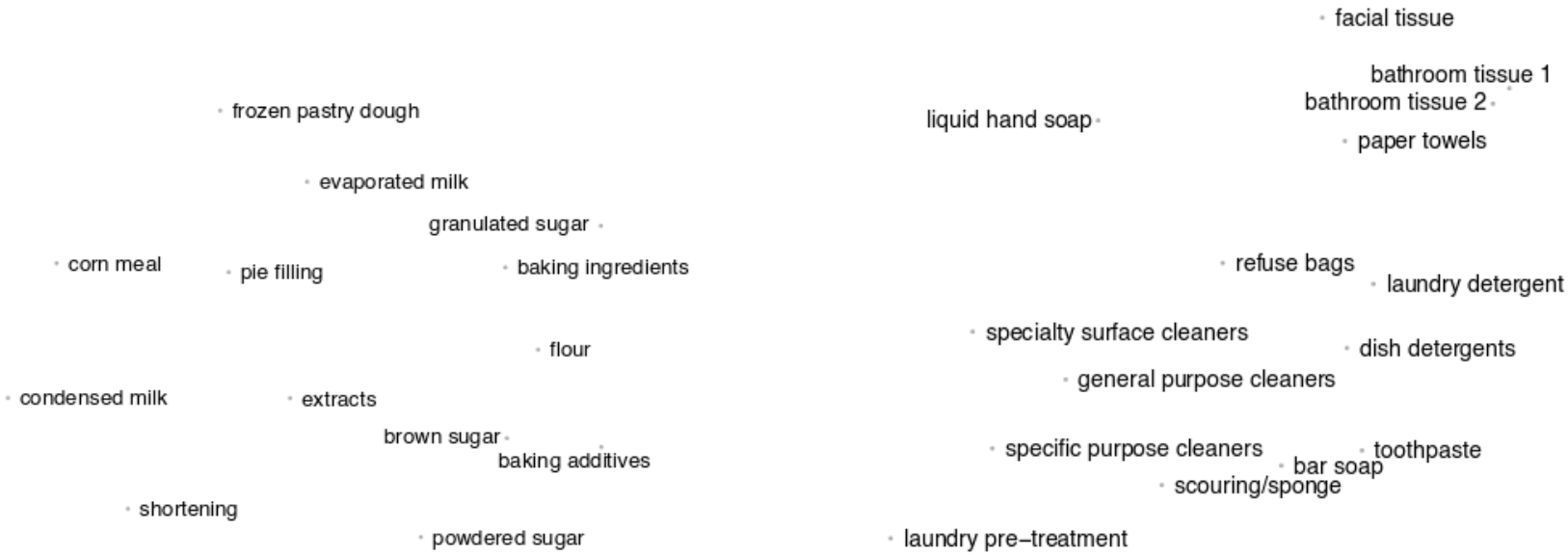
Embedding Layers versus Dense Layers

- A fully-connected dense layer, with sparse data set as input and linear activation, would emulate an embedding layer.
- Why use an embedding layer rather than a dense layer?
  - the main reason is that embedding layers are much faster for this purpose.
  - a related reason is that batch updating with regularization and dropout do not work well on sparse data.

Athey, Blei and Ruiz (2018): Shopping cart embeddings

- Use embeddings for goods in a shopping cart, rather than words in a document.
  - predict co-occuring goods in the basket
  - use embedding layer to represent goods in a geometric space.
- Data set: 2 years of shopping data from a large grocery store
  - 570K baskets, 6M purchases, 5.5K unique items

Embeddings capture features of groceries



Embeddings capture complementarity/substitutability

query items	complementarity score		exchangeability score	
mission tortilla soft taco 1	2.40	taco bell taco seasoning mix	0.05	mission fajita size
	2.26	mcrmck seasoning mix taco	0.07	mission tortilla soft taco 2
	2.24	lawrys taco seasoning mix	0.13	mission tortilla fluffy gordita
private brand hot dog buns	2.99	bp franks meat	0.11	ball park buns hot dog
	2.63	bp franks bun size	0.13	private brand hotdog buns potato 1
	2.37	bp franks beed bun length	0.15	private brand hotdog buns potato 2
private brand mustard squeeze bottle	0.50	private brand hot dog buns	0.15	frenchs mustard classic yellow squeeze
	0.41	private brand cutlery full size forks	0.16	frenchs mustard classic yellow squeezed
	0.24	best foods mayonnaise squeeze	0.21	heinz ketchup squeeze bottle
private brand napkins all occasion	0.78	private brand selection plates 6 7/8 in	0.09	vnty fair napkins all occasion 1
	0.50	private brand selection plates 8 3/4 in	0.11	vnty fair napkins all occasion 2
	0.49	private brand cutlery full size forks	0.12	private brand selection premium napkins

Word Embeddings

- Embedding layer mapps word indexes to dense vectors.
- Documents are lists of word indexes $\{w_1, w_2, ..., w_L\}$.
  - equivalently, let w_i be a one-hot vector (dimensionality $m =$ vocab size) where the associate word's index equals one .
  - Normalize all documents to the same length L; shorter documents can be padded with a null token. This requirement can be relaxed with recurrent neural networks.
- The embedding layer replaces the list of sparse one-hot vectors with a list of n-dimensional ($n << m$) dense vectors

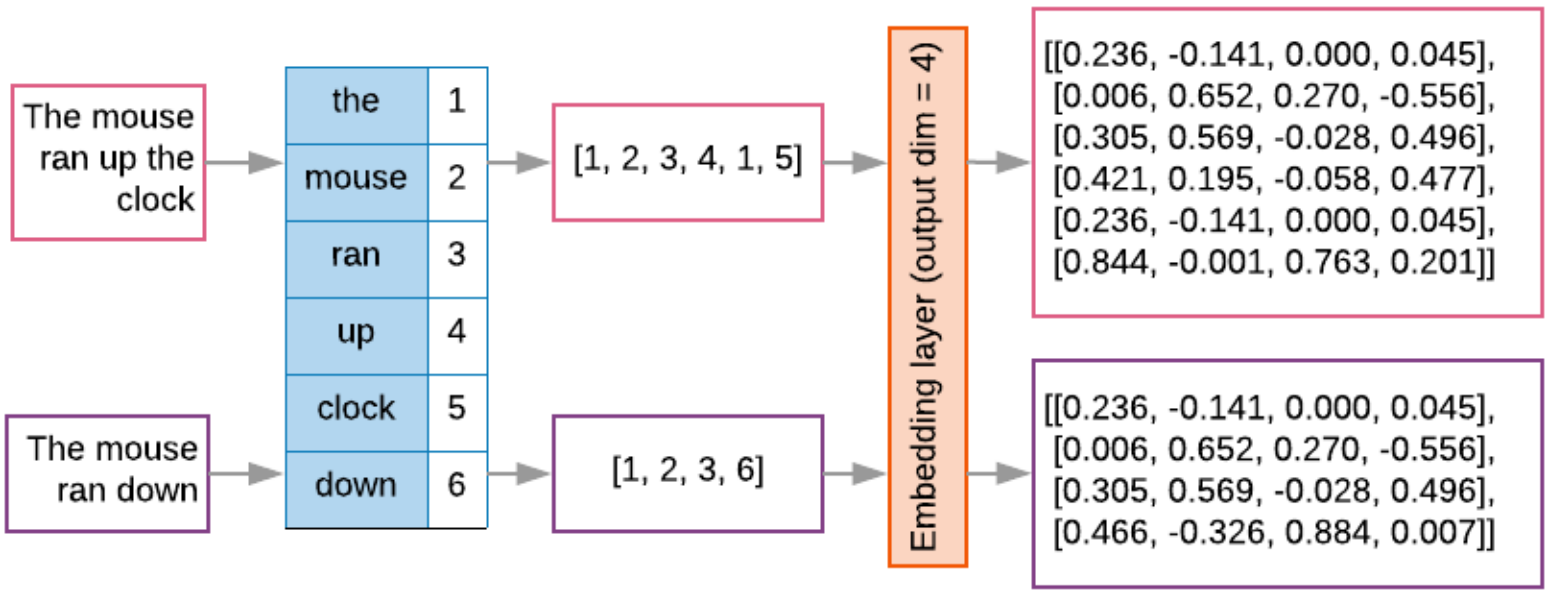
$$\boldsymbol{X} = \begin{bmatrix} x_1 & \dots & x_L \end{bmatrix}$$

where

$$\underbrace{x_i}_{n \times 1} = \underbrace{\Omega'}_{n \times m} \underbrace{w_i}_{m \times 1}$$

- This $\boldsymbol{X}$ matrix is then flattened into a $L * n$ vector for input to the next layer.

Illustration



Examining the Embeddings

- See Jupyter notebook for examples on training and visualizing the embeddings with words as points.
  - Also examples for extracting vectors for words and computing cosine similarity between words.

Word Embeddings and Word2Vec

- Word2Vec , GloVe, and other popular embeddings vectors are trained the same way as the word embeddings we just made for citation counts.
  - rather than predicting some metadata (such as citations) they predict the co-occurence of neighboring words.

Why word vectors?

- Once words are represented as vectors, we can use linear algebra to understand the relationships between words:
  - Words that are geometrically close to each other are similar: e.g. "student" and "pupil."
  - More intriguingly, word2vec algebra can depict conceptual, analogical relationships between words.
    - Consider the analogy: **man is to king as woman is to _____**
    - With word2vec, we have

$$vec(king) - vec(man) + vec(woman) \approx vec(queen)$$

How are word embeddings different from topic models?

- Ben Schmidt:
  - Topic models reduce words to core meanings to understand documents more clearly.
  - Word embedding models ignore information about individual documents to better understand the relationships between words.

Word Function $\longleftrightarrow$ Word Neighbors

- "The meaning of a word is its use in the language"

  - Ludwig Wittgenstein, *Philosophical Investigation*, 1953

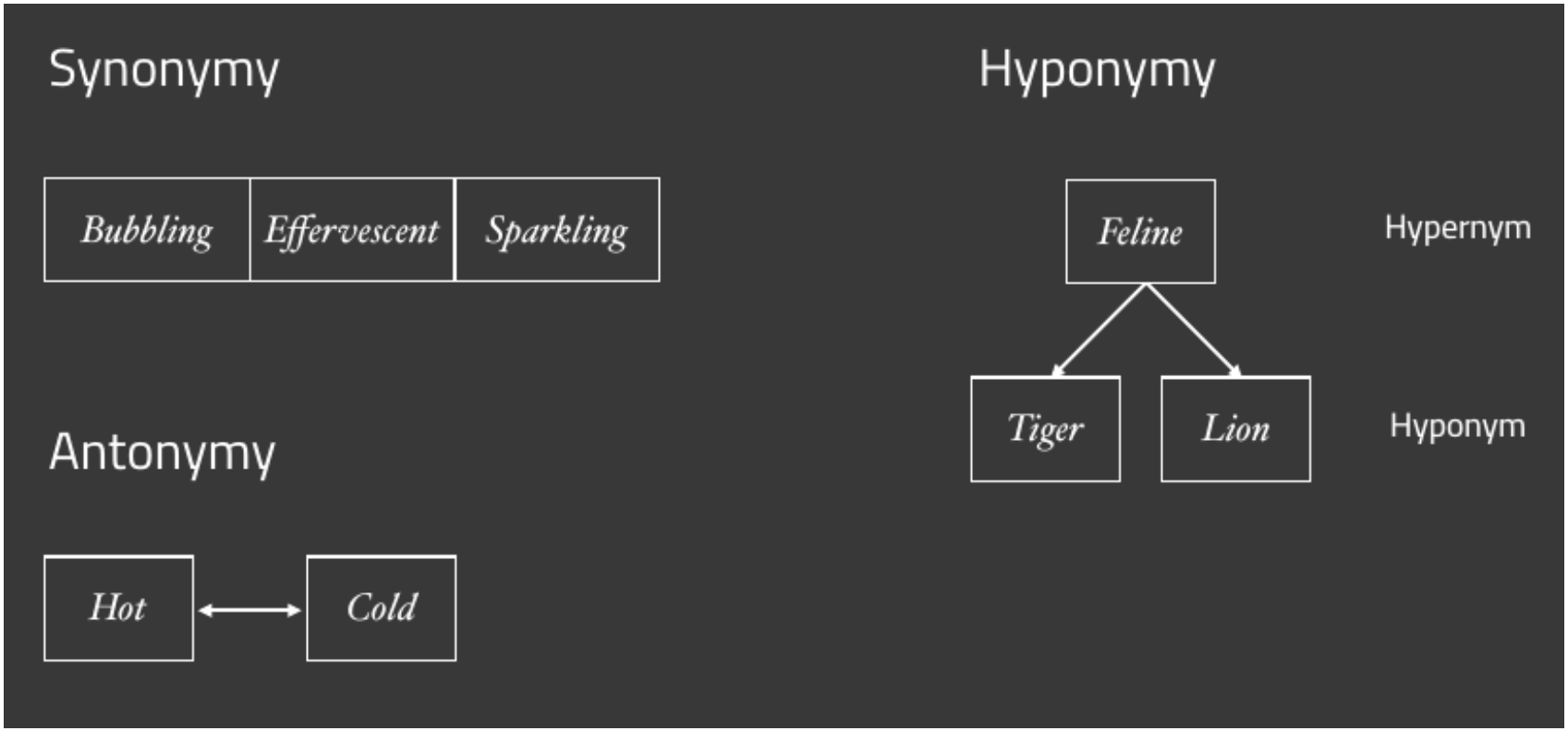
- "You shall know a word by the company it keeps"

  - J.R. Firth, Papers in Linguistics,1957

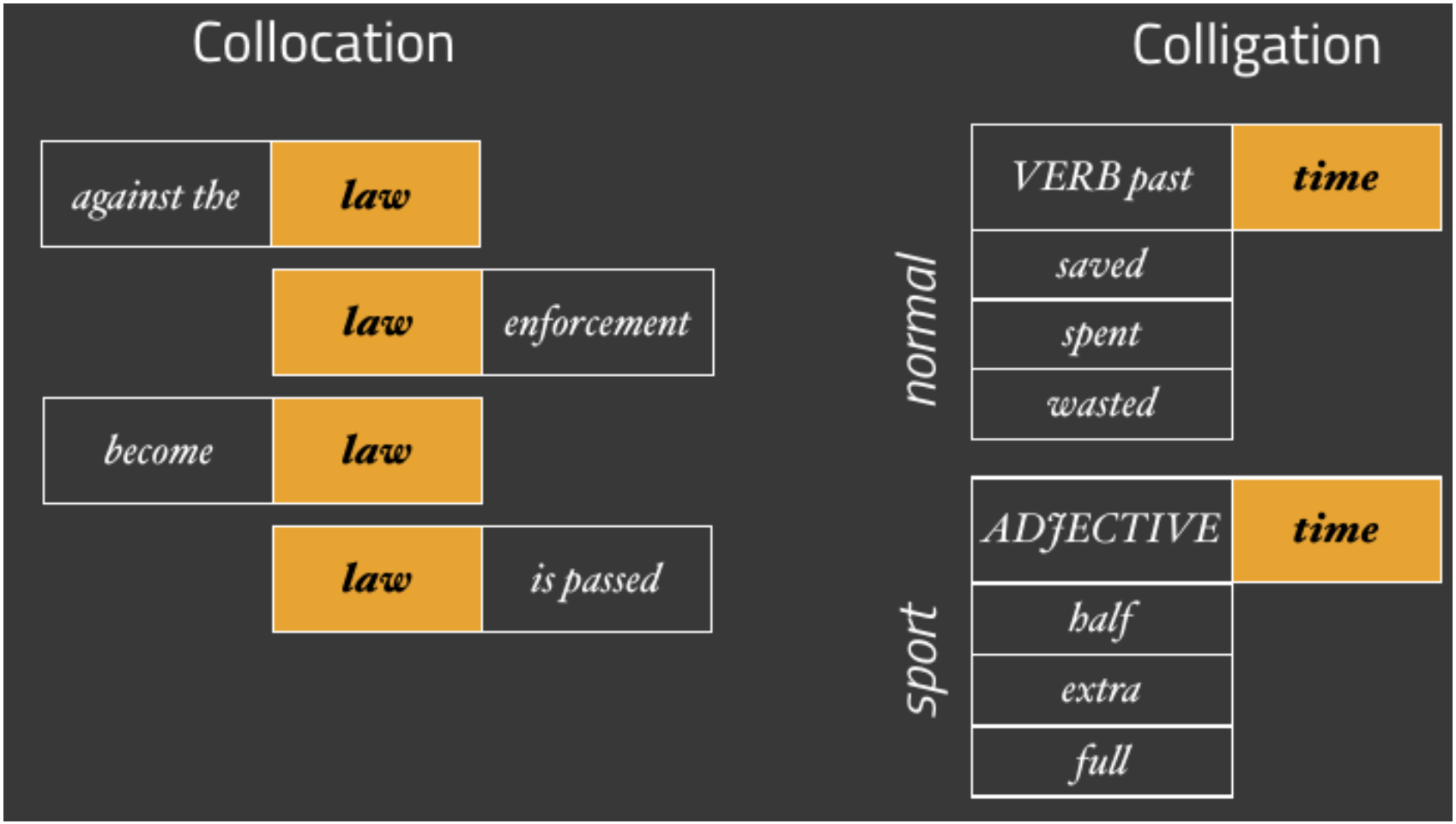
I've never seen this word before, but...

- He filled the **wampimuk**, passed it around and we all drunk some
- We found a little, hairy **wampimuk** sleeping behind the tree

Linguistic Relations



Collocational Relations

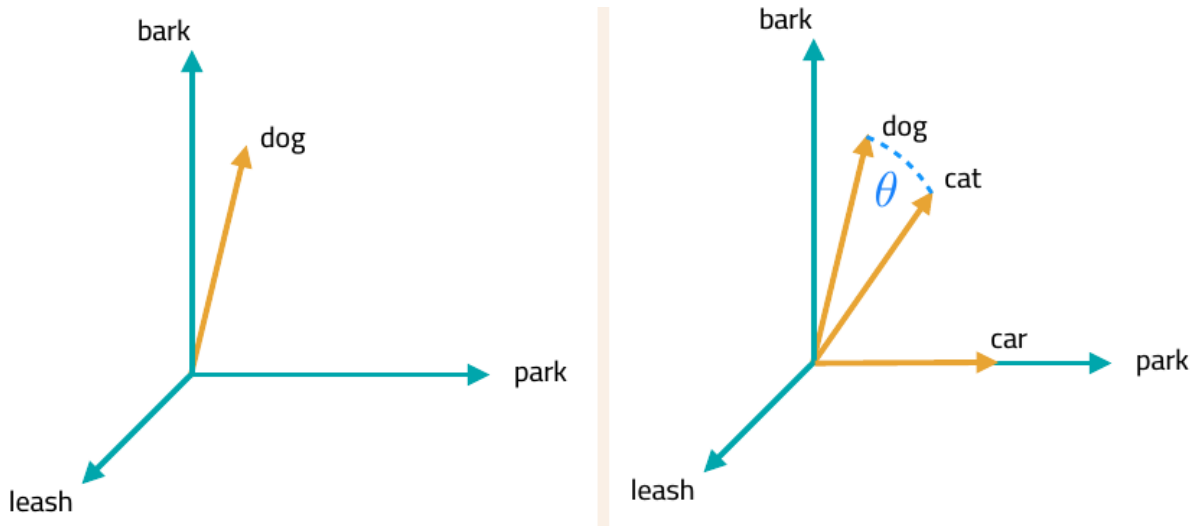


Similarity vs. Relatedness

- Semantic **similarity**: words sharing salient attributes / features
  - synonymy (car / automobile)
  - hypernymy (car / vehicle)
  - co-hyponymy (car / van / truck)
- Semantic **relatedness**: words semantically associated without necessarily being similar
  - function (car / drive)
  - meronymy (car / tire)
  - location (car / road)
  - attribute (car / fast)

(Budansky and Hirst, 2006)

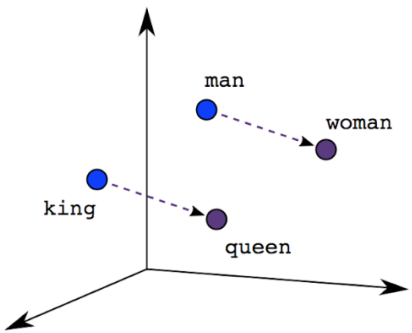
Words as Vectors



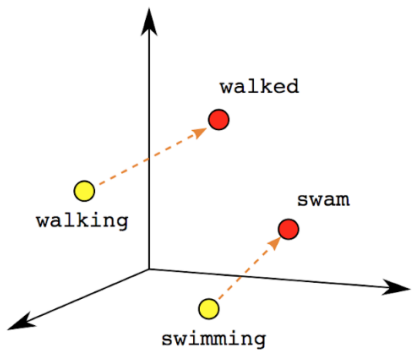
- Use cosine similarity as a measure of relatedness:

$$\cos\theta = \frac{v_1 \cdot v_2}{||v_1||||v_2||}$$

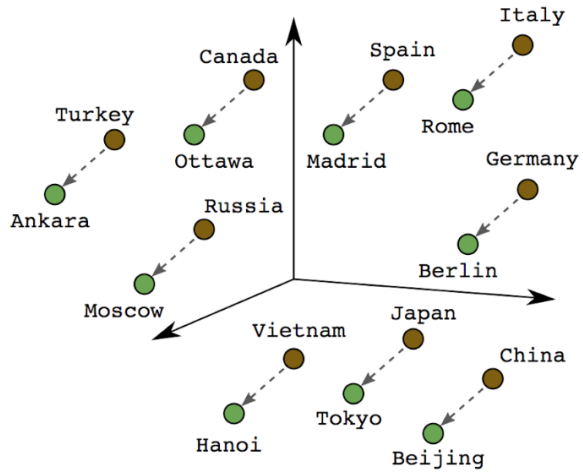
Vector Directions $\leftrightarrow$ Meaning



Male-Female



Verb Tense



Country-Capital

Evaluation of Word Embeddings

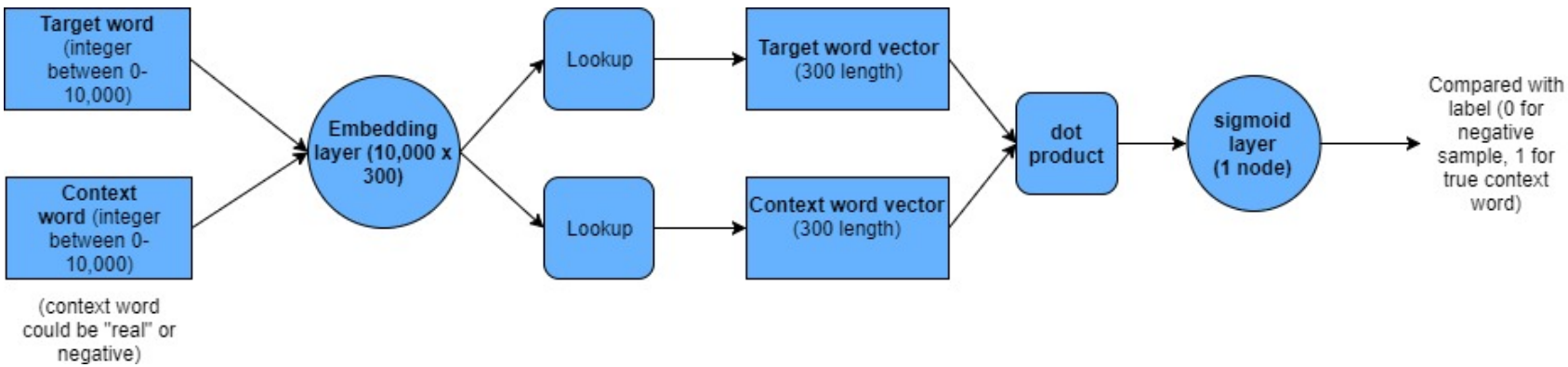
- Intrinsic:
  - evaluate word-pairs similarities $\rightarrow$ compare with similarity judgments given by humans
  - evaluate on analogy tasks ("Paris is to France as Tokyo is to ____")
- Extrinsic:
  - use the vectors in a downstream task (classification, translation, ...) and evaluate the final performance on the task

SGNS: Skip-gram with negative sampling

- When people mention "word2vec", they are usually talking about SGNS: "skip gram with negative sampling."
  - This is a particular word-embedding model with good performance on a range of analogy and prediction tasks.
- How does it learn the meaning of the word "fox":

  The quick brown **fox** jumps over the lazy dog
- Word2Vec reads in every example of the word "fox" and tries to predict what other words will be in the context window.
  - the prediction weights on these other words (after dimension reduction) are the word vectors

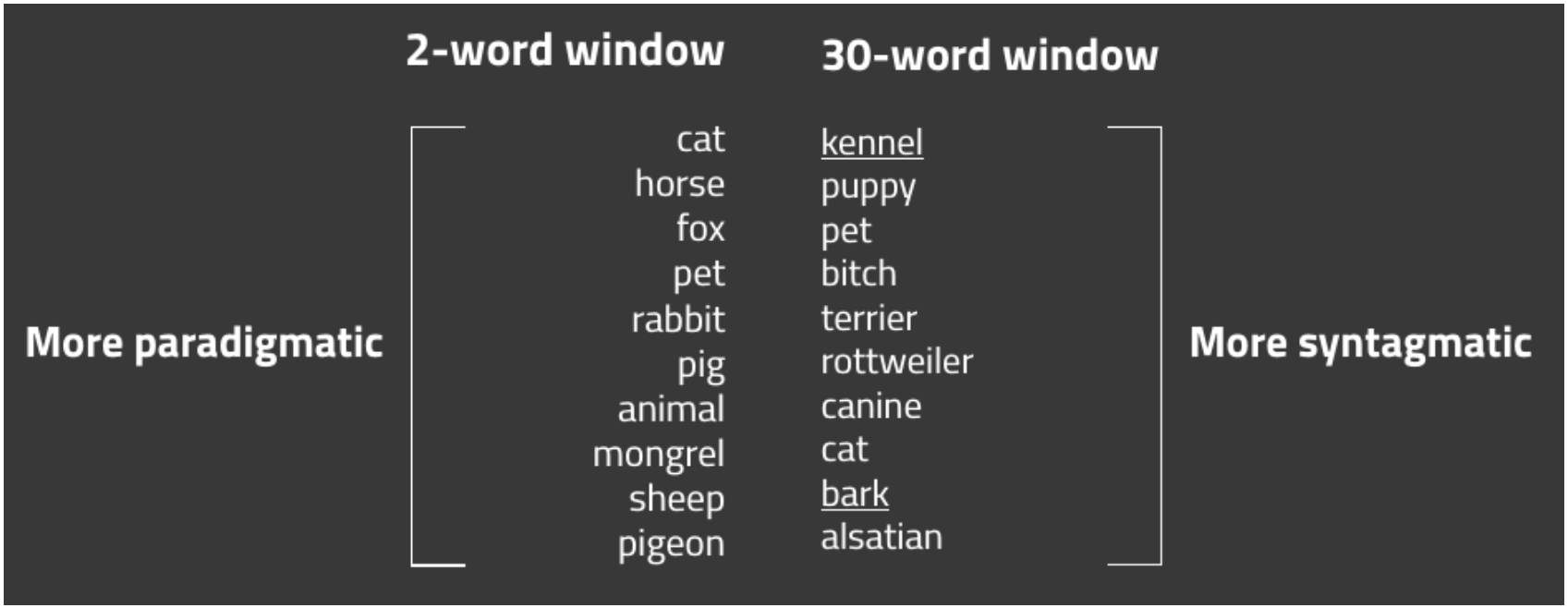
Word2Vec Schema



gensim implemention

- gensim's implementation of word2vec has the same benefits as the LDA implementation:
  - intuitive, streaming, and fast/parallellized
  - main parameter is window size.
- Then can easily get similarities between words or groups of words.

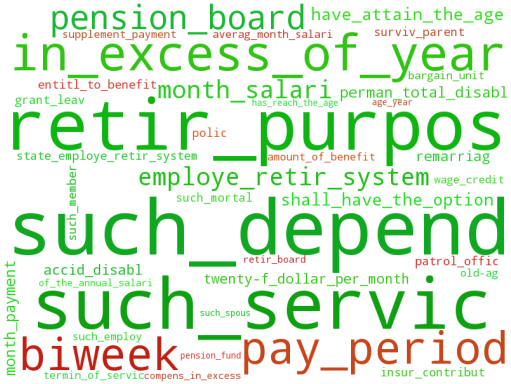
Most similar words to dog, depending on window size

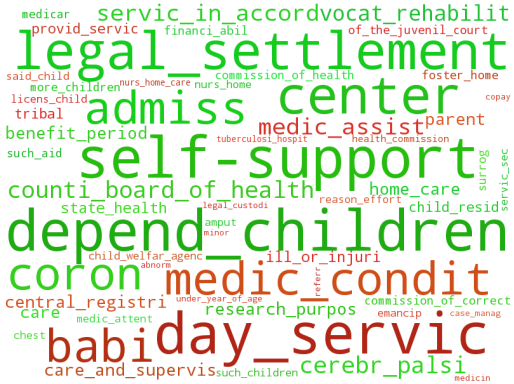


- Small windows pick up substitutable words; large windows pick up topics.

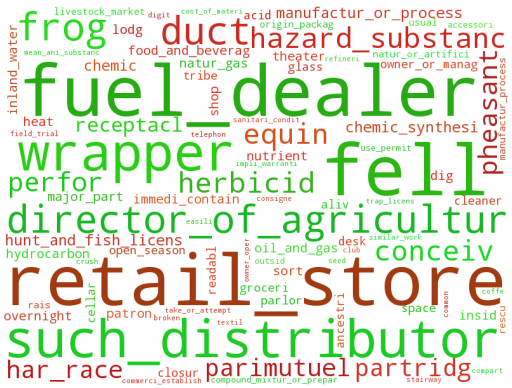
K-means clustering with Word Embeddings

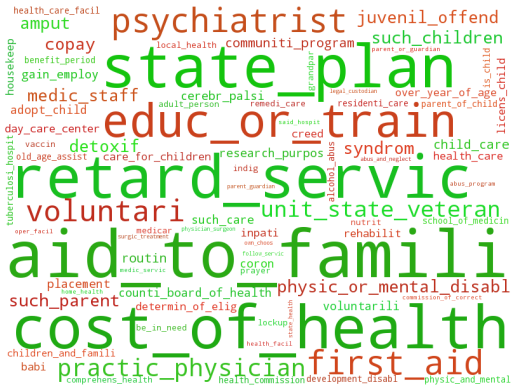
Income Tax (Pensions Topic and Health Care Topic)





Sales Tax (Retail Topic and Health Care Topic)





- Clustered phrases affecting tax revenues (Ash 2018); Green words tend to increase revenues; red words tend to decrease revenues.

Can robots rewrite the tax code?

- Example statutory substitution recommended by estimates (Virginia Code § 54-377): "person, firm, or corporation" should be replaced with "such person, firm, or corporation"; "failure" should be replaced with "such failure"
  - Adding "such" improves clarity, is considered good legal writing (e.g. Osbeck 2012).
  - Predicted revenue impact: +$16.2 million.

Version 1:

(a) If any person, firm or corporation shall continue after the expiration of a license previously issued without obtaining a new license, person, firm or corporation shall, if failure to obtain a new license continues for one month, be subject to a penalty of ten percentum of the amount of the license tax ... If failure to obtain a new license be continued for a longer period than one month, person, firm or corporation shall be guilty or a misdemeanor. Such conviction shall not relieve any person, firm or corporation from the payment of any license tax and penalty imposed by this article.
(b) If any person, firm or corporation shall commence any business, employment or profession in the city for which a license tax is required under this article, without first obtaining such license, person, firm or corporation shall be guilty of a misdemeanor ... If such violation of law be continued for one month, person, firm or corporation shall be subject to a penalty of ten percentum of the license tax which was due...

Version 2:

(a) If any person, firm or corporation shall continue after the expiration of a license previously issued without obtaining a new license, such person, firm or corporation shall, if such failure to obtain a new license continues for one month, be subject to a penalty of ten percentum of the amount of the license tax ... If such failure to obtain a new license be continued for a longer period than one month, such person, firm or corporation shall be guilty or a misdemeanor. Such conviction shall not relieve any such person, firm or corporation from the payment of any license tax and penalty imposed by this article.
(b) If any person, firm or corporation shall commence any business, employment or profession in the city for which a license tax is required under this article, without first obtaining such license, such person, firm or corporation shall be guilty of a misdemeanor ... If such violation of law be continued for one month, such person, firm or corporation shall be subject to a penalty of ten percentum of the license tax which was due...

Pre-trained word embeddings

- For some NLP tasks, you might not need to train your own vectors.
  - or your corpus might be too small to do so.
- In these cases, you can use a pre-trained model, like spaCy:
  - one million vocabulary entries
  - 300-dimensional vectors
  - trained on the Common Crawl corpus
- Can initialize prediction model using pre-trained embeddings.

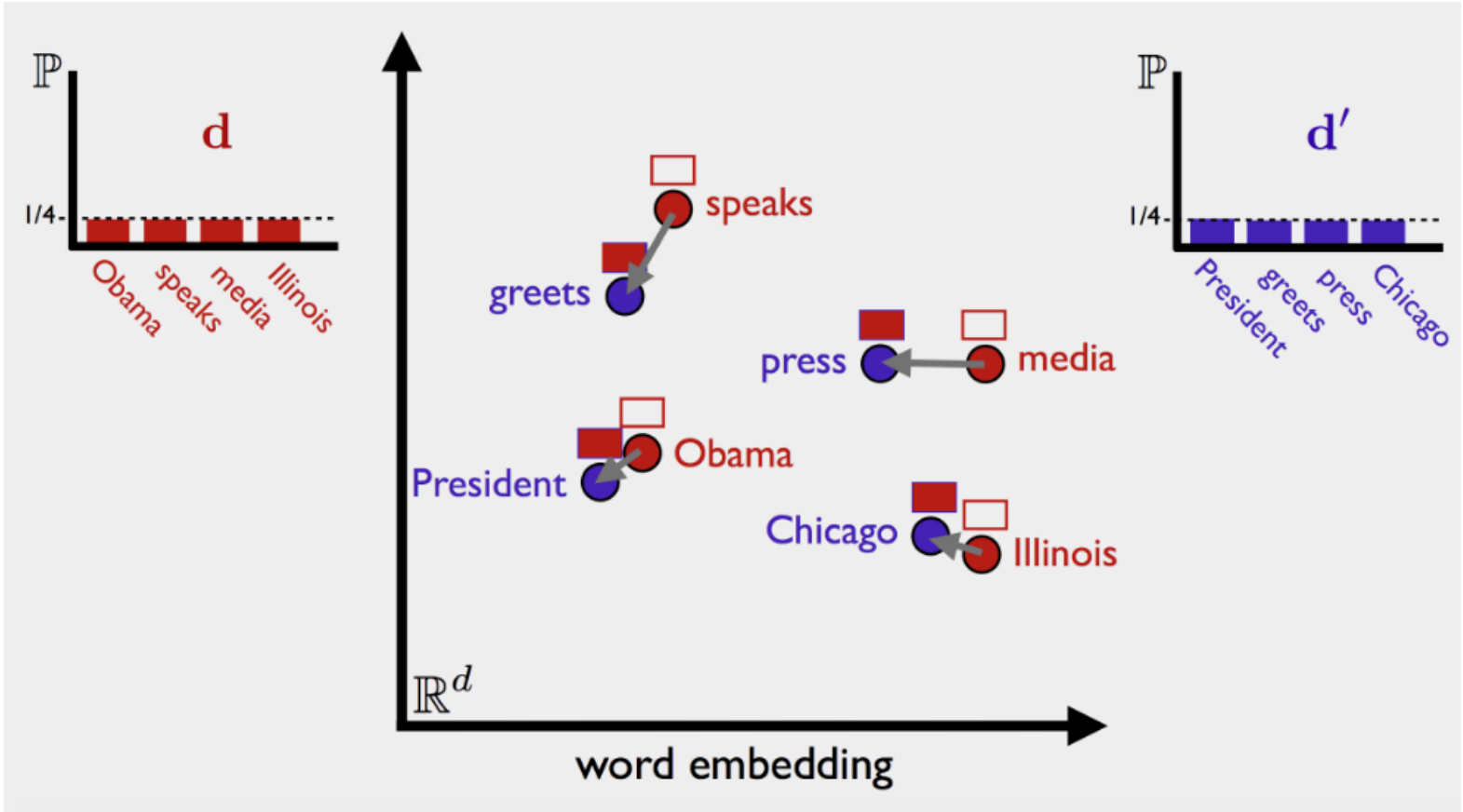
Tips for using pre-trained embeddings

- Split training in two steps:
  - in first run, train the model with the first layer (the pre-trained embeddings) frozen.
  - in second run, un-freeze the embedding layer for fine tuning.

Word Mover Distance

- Cosine distance treats synonyms as just as close as totally unrelated words.
- Word mover distance between two texts is given by:
  - total amount of "mass" needed to move words from one side into the other
  - multiplied by the distance the words need to move
  - Kusner, Sun, Kolkin, and Weinberger (ICML 2015)
- Requires measure of distance between words (word embeddings).
  - see wmd package in Python.

Illustration



- d (obama speaks media illinois) is orthogonal to d' (president greets press chicago):
  - cosine similarity is zero
  - Word mover distance sums the shortest distances between the words in the documents.

Rudolph and Blei (2017)

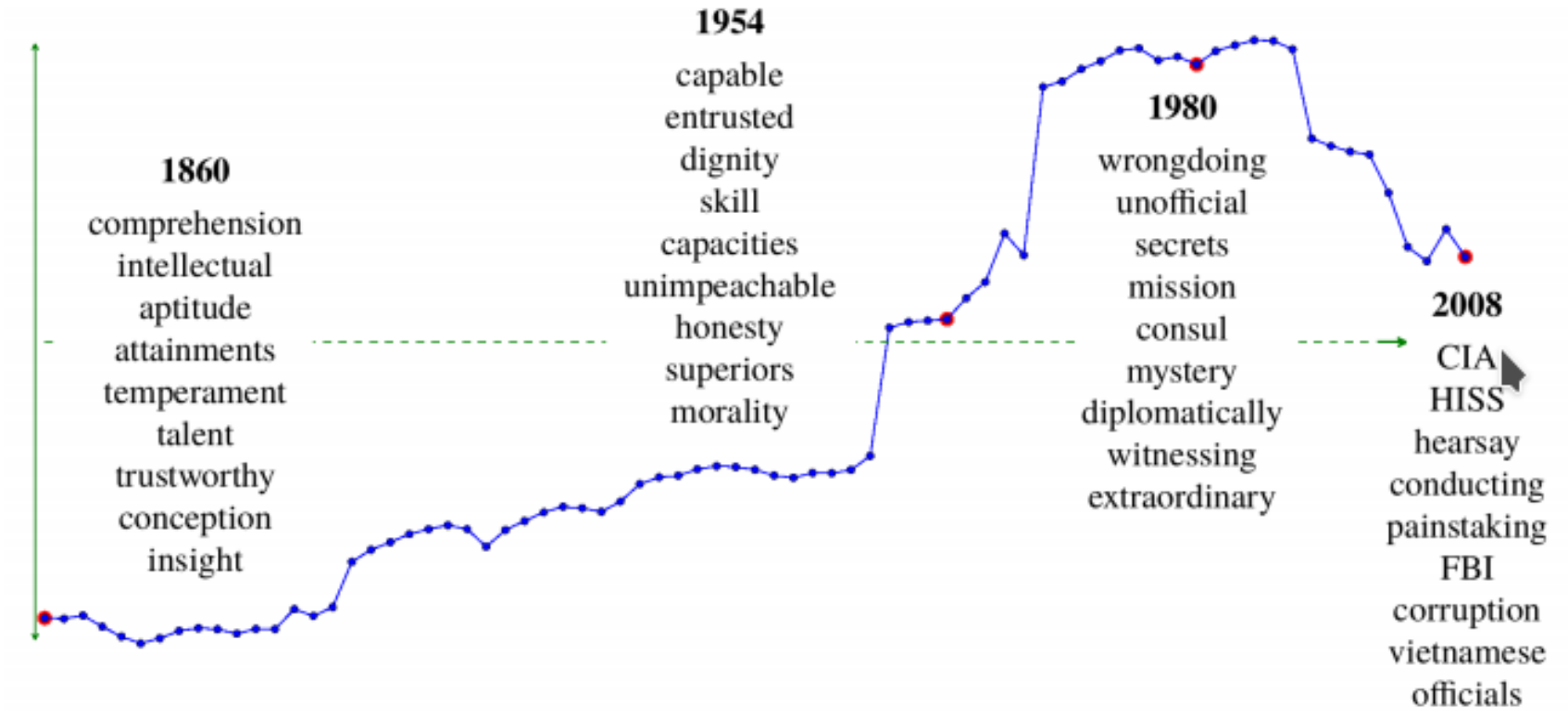
- Train word embeddings on the U.S. Congressional Record, 1858-2009.
- Dynamic word embeddings model:
  - Captures how the meaning of words evolves over time.
  - The innovation is to include "year" in the embedding model, and allow word vectors to drift over time (following a random walk).
    - Doing this with standard word2vec would require you to train a different model in each year, so no information could be shared across years.

Meaning Changes

computer	
1858	**1986**
computer	computer
draftsman	software
draftsmen	computers
copyist	copyright
photographer	technological
computers	innovation
copyists	mechanical
janitor	hardware
accountant	technologies
bookkeeper	vehicles

bush	
1858	**1990**
bush	bush
barberry	cheney
rust	nonsense
bushes	nixon
borer	reagan
eradication	george
grasshoppers	headed
cancer	criticized
tick	clinton
eradicate	blindness

Drift in word “intelligence”



Drift in word “prostitution”

prostitution				
1930	**1945**	**1962**	**1988**	**1990**
prostitution	prostitution	prostitution	harassment	prostitution
punishing	indecent	indecent	intimidation	servitude
immoral	vile	harassment	prostitution	harassment
bootlegging	immoral	intimidation	counterfeit	intimidation
riotous	induces	sexual	illegal	trafficking
forbidden	incite	vile	trafficking	harassing
anarchists	abortion	counterfeit	indecent	apprehended
assemblage	forbid	anarchists	disregard	killings
forbid	harboring	mobs	anarchists	labeled
abet	assemblage	lawbreakers	punishing	naked

Caliskan, Bryson, and Narayanan (*Science* 2017)

- "We replicated a spectrum of known biases, as measured by the Implicit Association Test, using a widely used, purely statistical machine-learning model trained on a standard corpus of text from the World Wide Web. . . "

Word Embedding Association Test

- Target words:
  - programmer, engineer, scientist, ...
  - nurse, teacher, librarian, ...
- Attribute words:
  - man, male, ...
  - woman, female, ...
- WEAT Test:
  - Compute similarities between all target words and all attribute words
  - Compute mean target-attribute clustering

Example Stimuli

- Targets:
  - **Flowers**: aster, clover, hyacinth, marigold, poppy, azalea, crocus, iris, orchid, rose, bluebell, daffodil, lilac, pansy, tulip, buttercup, daisy, lily, peony, violet, carnation, gladiola, magnolia, petunia, zinnia.
  - **Insects**: ant, caterpillar, flea, locust, spider, bedbug, centipede, fly, maggot, tarantula, bee, cockroach, gnat, mosquito, termite, beetle, cricket, hornet, moth, wasp, blackfly, dragonfly, horsefly, roach, weevil.
- Attributes:
  - **Pleasant**: caress, freedom, health, love, peace, cheer, friend, heaven, loyal, pleasure, diamond, gentle, honest, lucky, rainbow, diploma, gift, honor, miracle, sunrise, family, happy, laughter, paradise, vacation.
  - **Unpleasant**: abuse, crash, filth, murder, sickness, accident, death, grief, poison, stink, assault, disaster, hatred, pollute, tragedy, divorce, jail, poverty, ugly, cancer, kill, rotten, vomit, agony, prison.

Results

- Pleasant vs. Unpleasant?
  - Flowers vs. Insects
  - Musical instruments vs. weapons.
  - European-American names vs. African-American names
- Male names vs. Female names?
  - Career words (e.g. professional, corporation, ...) vs. family words (e.g. home, children, ...)
  - Math/science words vs arts words

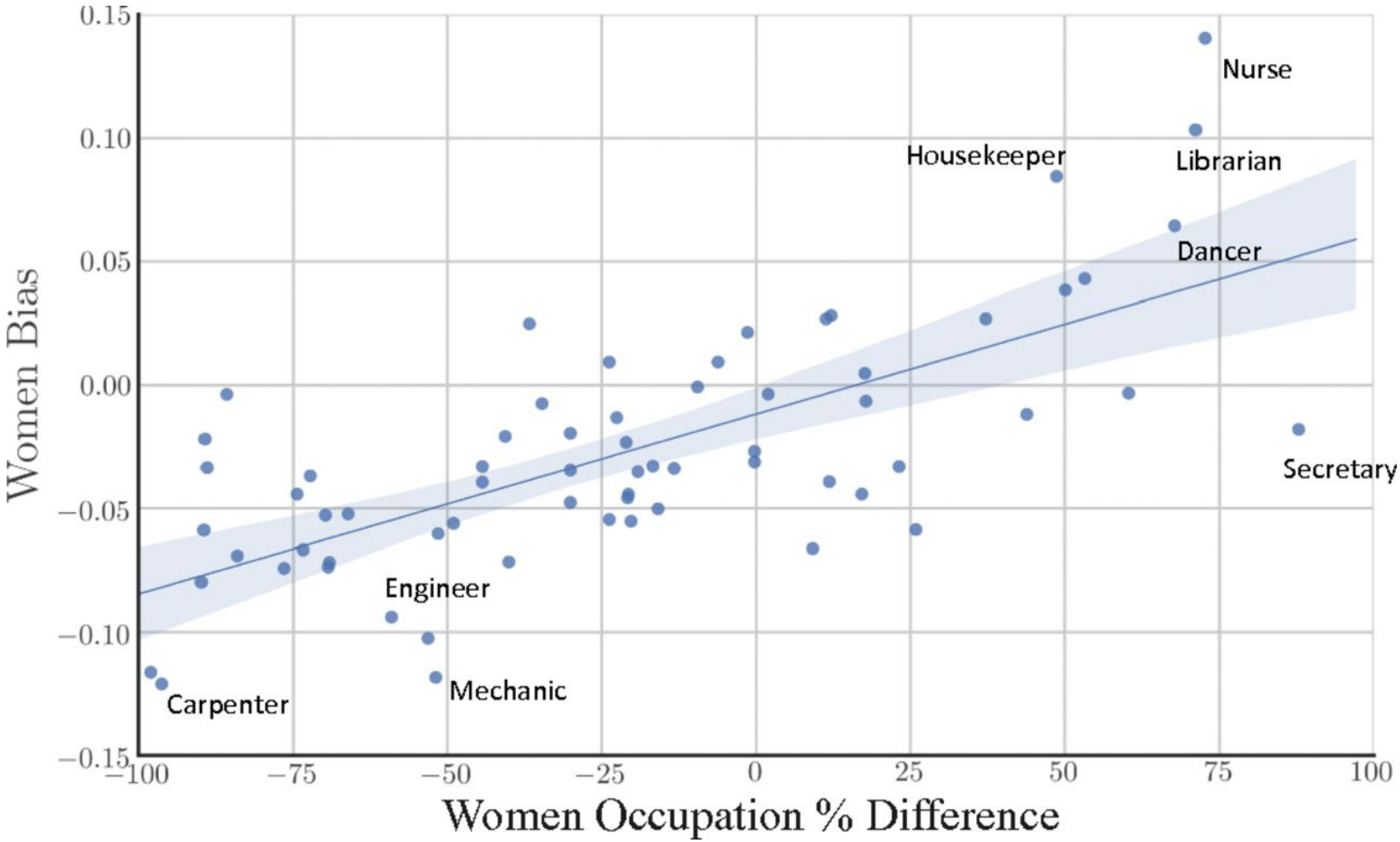
Bolukbasi et al (NIPS 2016)

- "Geometrically, **gender bias is first shown to be captured by a direction in the word embedding**."
- "Second, gender neutral words are shown to be linearly separable from gender definition words in the word embedding."
- "Using these properties, we provide a methodology for modifying an embedding to remove gender stereotypes, such as the association between the words receptionist and female, while maintaining desired associations such as between the words queen and female."

Gonen and Goldberg (2019)

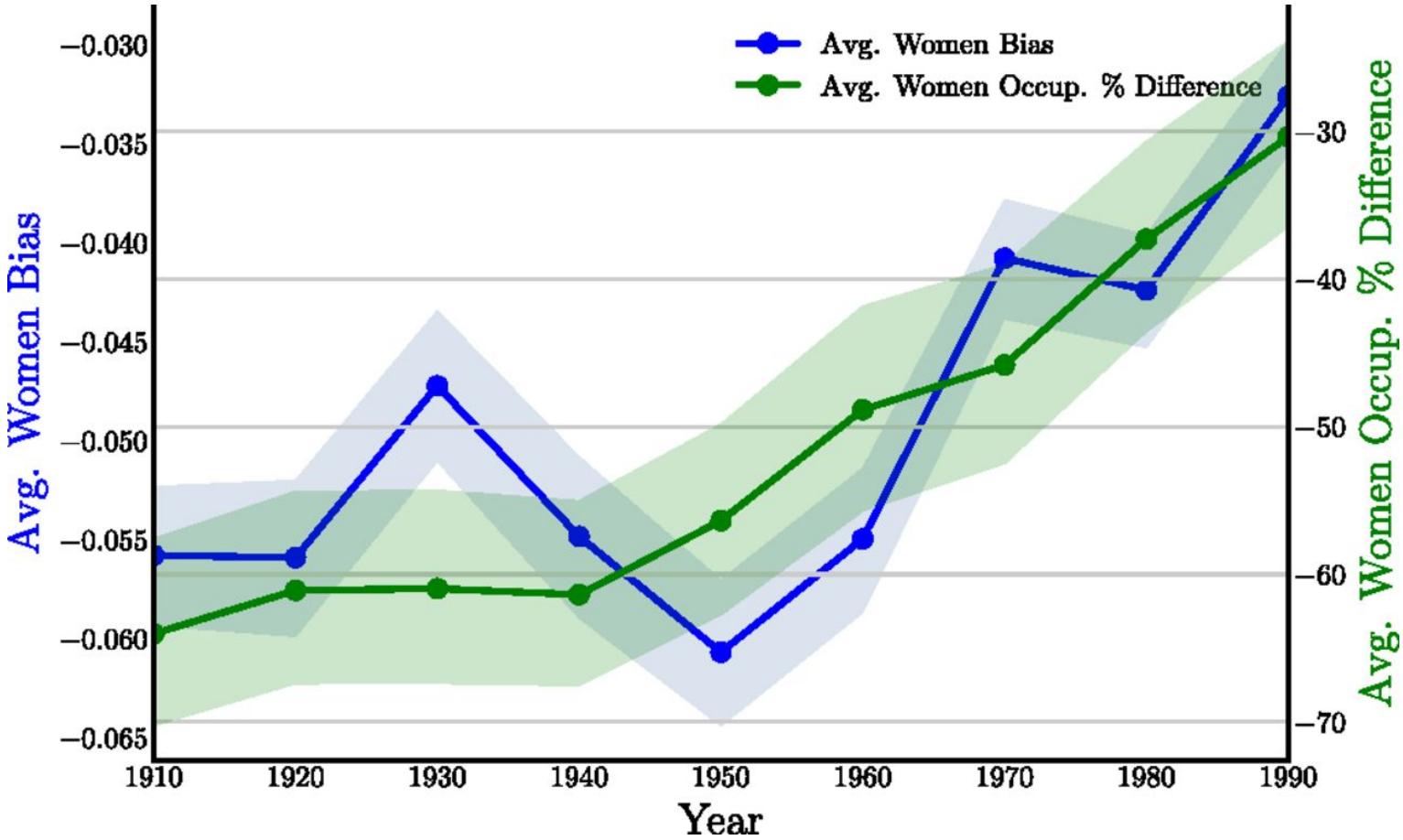
"... we argue that this removal is superficial. While the bias is indeed substantially reduced according to the provided bias definition, the actual effect is mostly hiding the bias, not removing it. The gender bias information is still reflected in the distances between 'gender-neutralized' words in the debiased embeddings, and can be recovered from them..."

Garg, Schiebinger, Jurafsky, and Zou (2018)



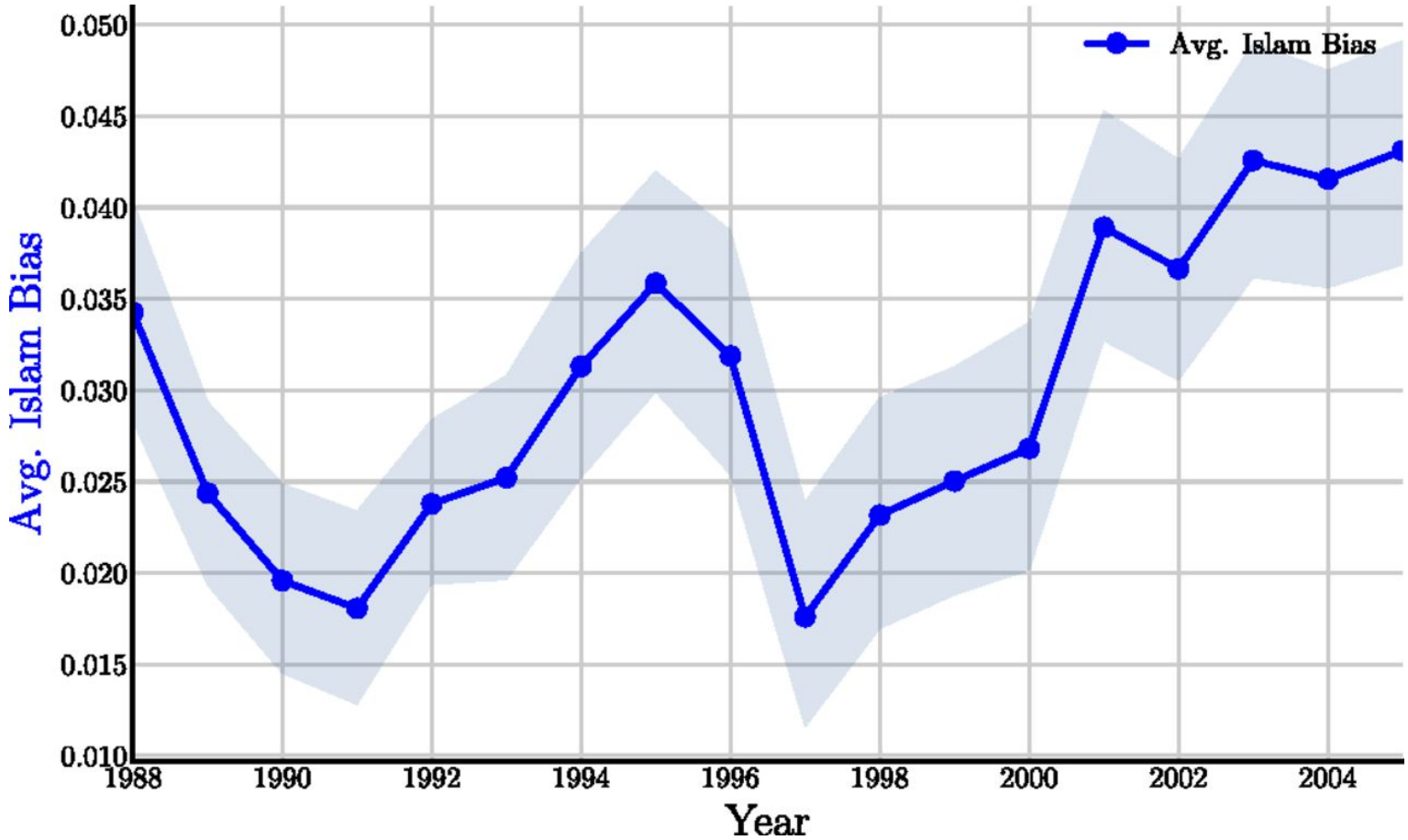
Women's occupation relative percentage vs. embedding bias in Google News vectors.

Garg, Schiebinger, Jurafsky, and Zou (2018)



Average gender bias score over time in COHA embeddings in occupations vs. the average percentage of difference. More positive means a stronger association with women. In blue is relative bias toward women in the embeddings, and in green is the average percentage of difference of women in the same occupations.

Garg et al: Islam↔Terrorism



Religious (Islam vs. Christianity) bias score over time for words related to terrorism in New York Times data.

Garg et al: Ethnic groups $\leftrightarrow$ Occupations

Hispanic	Asian	White
Housekeeper	Professor	Smith
Mason	Official	Blacksmith
Artist	Secretary	Surveyor
Janitor	Conductor	Sheriff
Dancer	Physicist	Weaver
Mechanic	Scientist	Administrator
Photographer	Chemist	Mason
Baker	Tailor	Statistician
Cashier	Accountant	Clergy
Driver	Engineer	Photographer

The top 10 occupations most closely associated with each ethnic group in the Google News embedding.

Garg et al: Female-Associated Words Over Time

1910	1950	1990
Charming	Delicate	Maternal
Placid	Sweet	Morbid
Delicate	Charming	Artificial
Passionate	Transparent	Physical
Sweet	Placid	Caring
Dreamy	Childish	Emotional
Indulgent	Soft	Protective
Playful	Colorless	Attractive
Mellow	Tasteless	Soft
Sentimental	Agreeable	Tidy

Kozlowski, Evans, and Taddy (2018)

- Data-set: Google 5-grams.
  - n-grams of length five for U.S. and U.K. publications.
  - provides counts for each year
- Extract language dimensions:
  - get the complete list of WordNet antonym pairs (e.g, "weak/strong", "tall/short")
    - filter on document frequency to 428 pairs.
  - map the dimensional shifts between the antonyms.
  - compare this vector shift to the one between men and women.

Mapping gender, class, and race

Gender	Class	Race†
man – woman	rich – poor	black – white
men – women	richer – poorer	blacks – whites
he – she	richest – poorest	Blacks – Whites
him – her	affluence – poverty	Black – White
his – her	affluent – impoverished	African – European
his – hers	expensive – inexpensive	African – Caucasian
boy – girl	luxury – cheap	
boys – girls	opulent – needy	
male – female		
masculine – feminine		

Matching antonyms to gender/class

Gender dimension nearest neighbors	
1. rugged–delicate	.219 (.213, .224)
2. soft–loud	-.209 (-.216, -.201)
3. tender–tough	-.202 (-.210, -.197)
4. timid–bold	-.181 (-.186, -.174)
5. soft–hard	-.161 (-.168, -.158)

Class dimension nearest neighbors	
1. weak-strong	-.292 (-.301, -.287)
2. fortunate-unfortunate	.291 (.286, .297)
3. unhappy-happy	-.259 (-.266, -.254)
4. beautiful-ugly	.242 (.238, .245)
5. potent_impotent	.234 (.227, .244)

Mapping musical genres to race/class