Dan Jurafsky and James Martin

Speech and Language Processing

Chapter 6

Vector Semantics, Part 3

Re-cap: Skip-Gram Training

Training sentence:

... lemon, a tablespoon of **apricot** jam a pinch ...

c1 c2 t c3 c4

positive examples +

t	c
apricot	tablespoon
apricot	of
apricot	preserves
apricot	or

- For each positive example, we'll create *k* negative examples.
- Using *noise* words
- Any random word that isn't *t*

Re-cap: Skip-Gram Training

Training sentence:

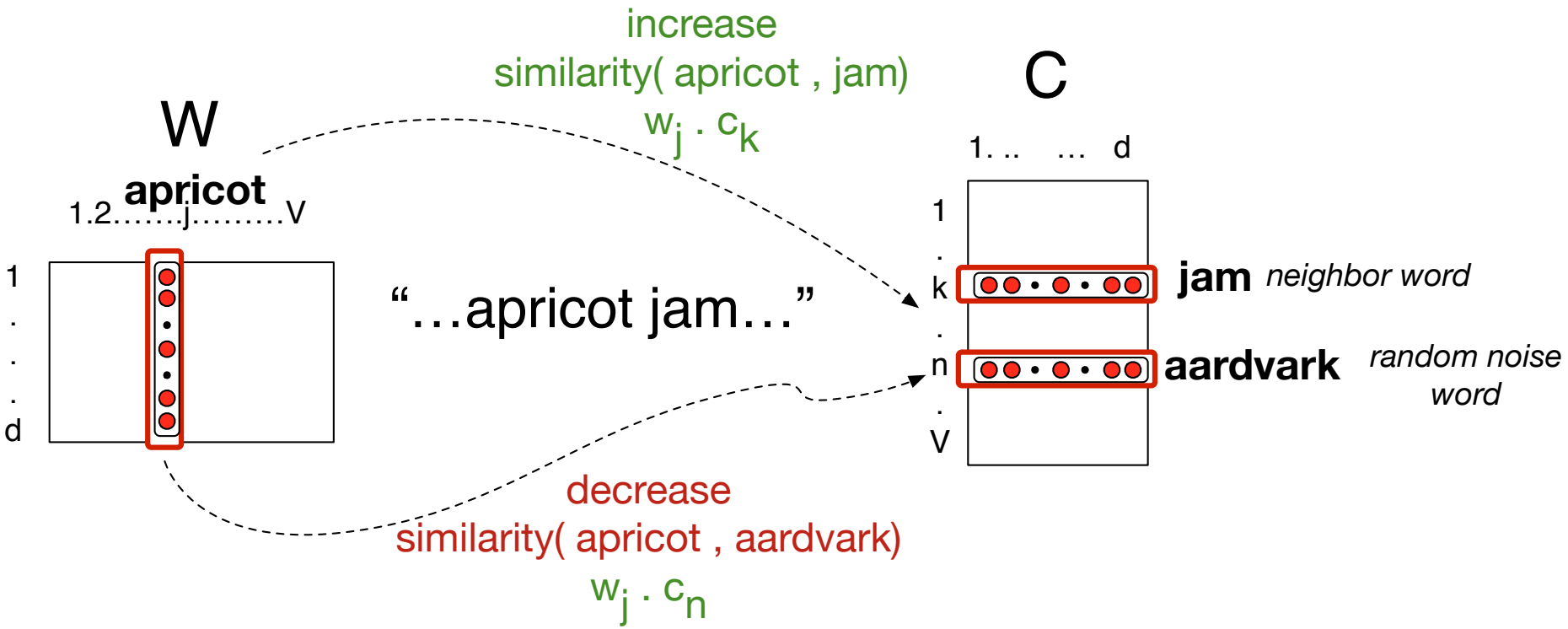
... lemon, a tablespoon of **apricot** jam a pinch ...

c1 c2 t c3 c4

k=2

positive examples +	
t	c
apricot	tablespoon
apricot	of
apricot	preserves
apricot	or

negative examples -			
t	c	t	c
apricot	aardvark	apricot	twelve
apricot	puddle	apricot	hello
apricot	where	apricot	dear
apricot	coaxial	apricot	forever


increase
similarity(apricot , jam)
$w_j \cdot c_k$
W
apricot
1.2.....j.....V
1
.
.
.
d
"…apricot jam…"
C
1. .. ... d
1
.
k
.
n
.
V
jam
neighbor word
aardvark
random noise word
decrease
similarity(apricot , aardvark)
$w_j \cdot c_n$

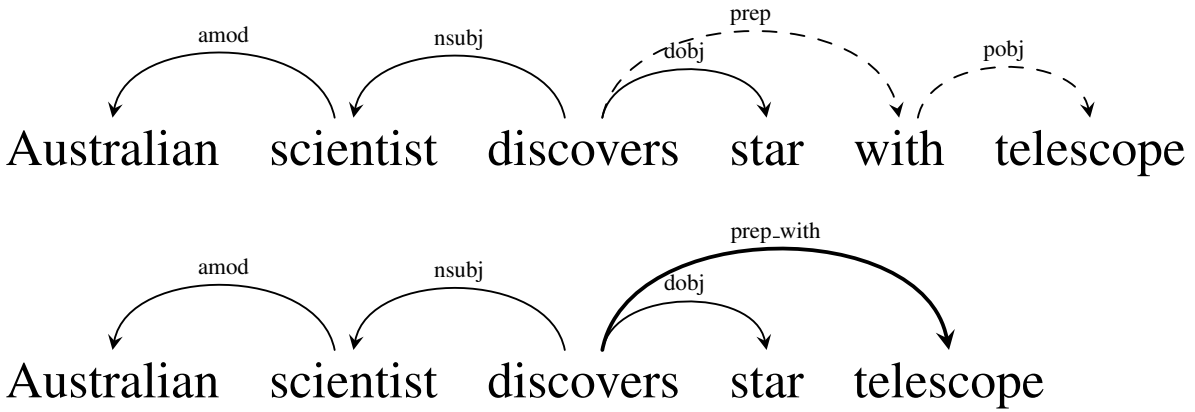
Recap: How to learn word2vec (skip-gram) embeddings

Start with V random 300-dimensional vectors as initial embeddings

Use logistic regression, the second most basic classifier used in machine learning after naïve Bayes

- Take a corpus and take pairs of words that co-occur as positive examples
- Take pairs of words that don't co-occur as negative examples
- Train the classifier to distinguish these by slowly adjusting all the embeddings to improve the classifier performance
- Throw away the classifier code and keep the embeddings.

Dependency-based word embeddings



WORD	CONTEXTS
australian	scientist/amod^{-1}
scientist	australian/amod, discovers/nsubj^{-1}
discovers	scientist/nsubj, star/dobj, telescope/prep_with
star	discovers/dobj^{-1}
telescope	discovers/prep_with^{-1}

Properties of embeddings

Similarity depends on window size C

C = ±2 The nearest words to *Hogwarts:*

- *Sunnydale*
- *Evernight*

C = ±5 The nearest words to *Hogwarts:*

- *Dumbledore*
- *Malfoy*
- *halfblood*

How does context window change word emeddings?

Target Word	BoW5	BoW2	Deps
batman	nightwing aquaman catwoman superman manhunter	superman superboy aquaman catwoman batgirl	superman superboy supergirl catwoman aquaman
hogwarts	dumbledore hallows half-blood malfoy snape	evernight sunnydale garderobe blandings collinwood	sunnydale collinwood calarts greendale millfield
florida	gainesville fla jacksonville tampa lauderdale	fla alabama gainesville tallahassee texas	texas louisiana georgia california carolina
	aspect-oriented	aspect-oriented	event-driven

Solving analogies with embeddings

In a word-analogy task we are given two pairs of words that share a relation (e.g. “man:woman”, “king:queen”).

The identity of the fourth word (“queen”) is hidden, and we need to infer it based on the other three by answering

“*man* is to *woman* as *king* is to — ?”

More generally, we will say **a:a*** as **b:b*.**

Can we solve these with word vectors?

Vector Arithmetic

a:a* as **b:b*.** **b*** is a hidden vector.

b∗ should be similar to the vector b − a + a∗

vector(*'king'*) - vector(*'man'*) + vector(*'woman'*) ≈ vector('queen')

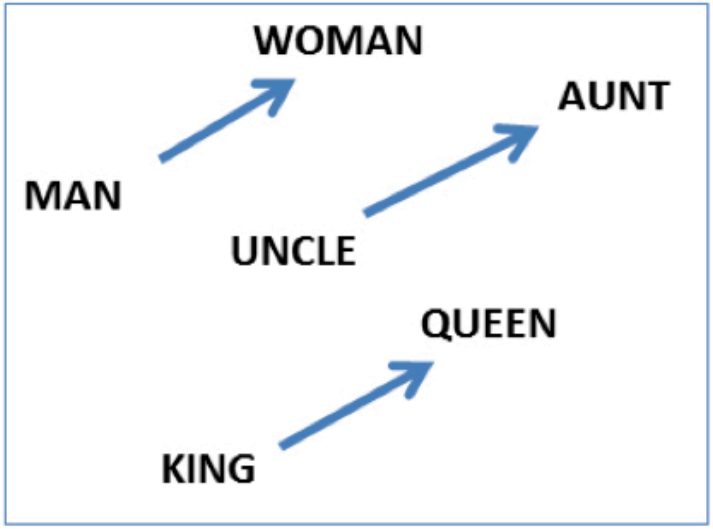
So the analogy question can be solved by optimizing:

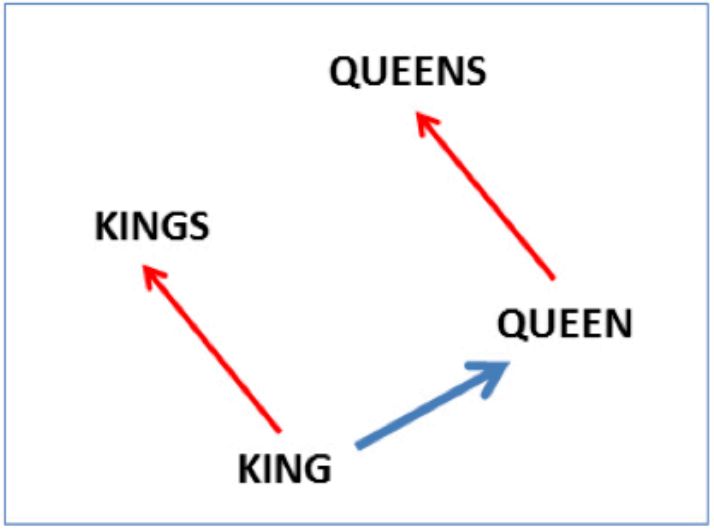
$$\arg\max_{b^* \in V} \left(\cos\left(b^*, b - a + a^*\right)\right)$$

Analogy: Embeddings capture relational meaning!

vector(*'king'*) - vector(*'man'*) + vector(*'woman'*) ≈ vector('queen')

vector(*'Paris'*) - vector(*'France'*) + vector(*'Italy'*) ≈ vector('Rome')





Vector Arithmetic

If all word-vectors are normalized to unit length then

$$\arg\max_{b^* \in V} \left(\cos\left(b^*, b - a + a^*\right)\right)$$

is equivalent to

$$\arg\max_{b^* \in V} \left(\cos\left(b^*, b\right) - \cos\left(b^*, a\right) + \cos\left(b^*, a^*\right)\right)$$

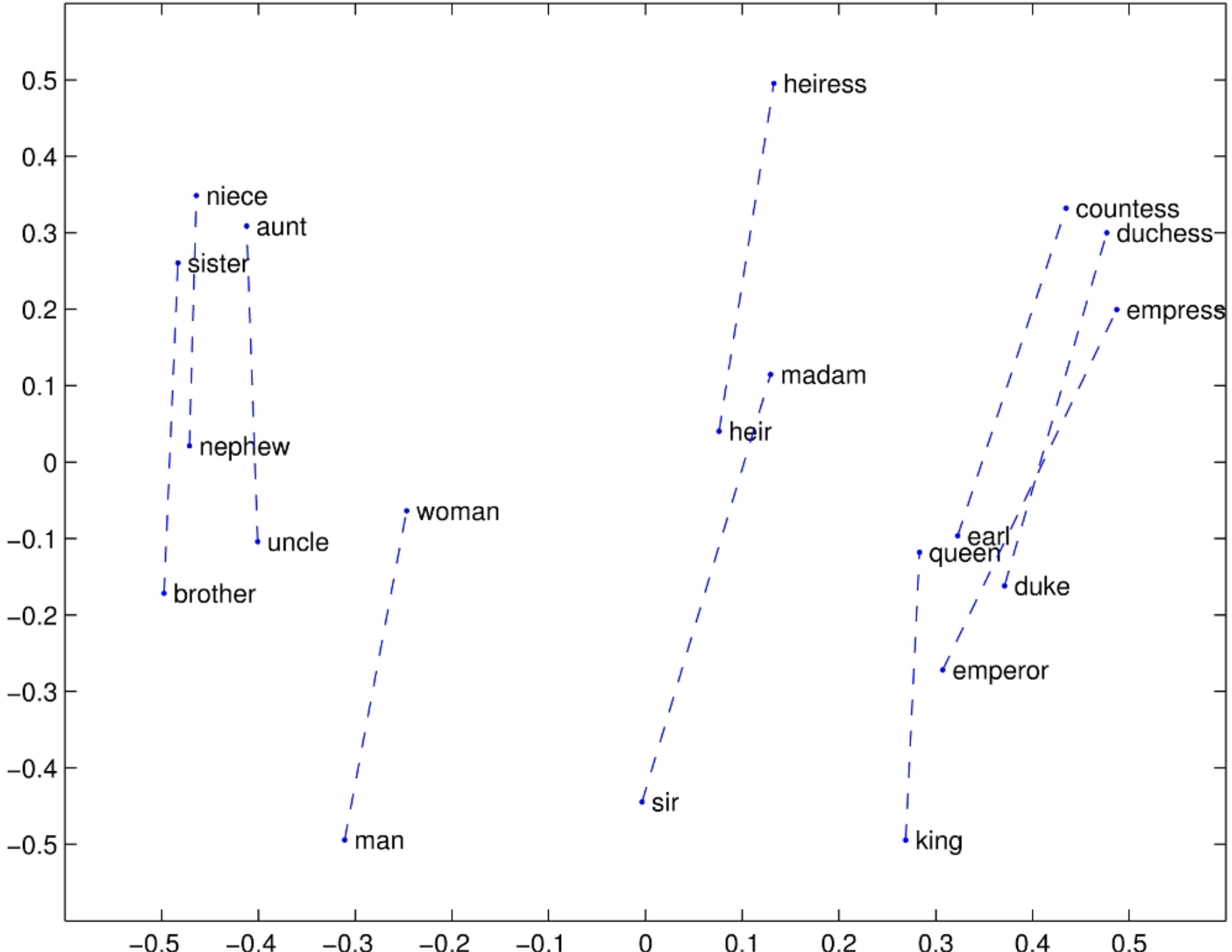
Vector Arithmetic

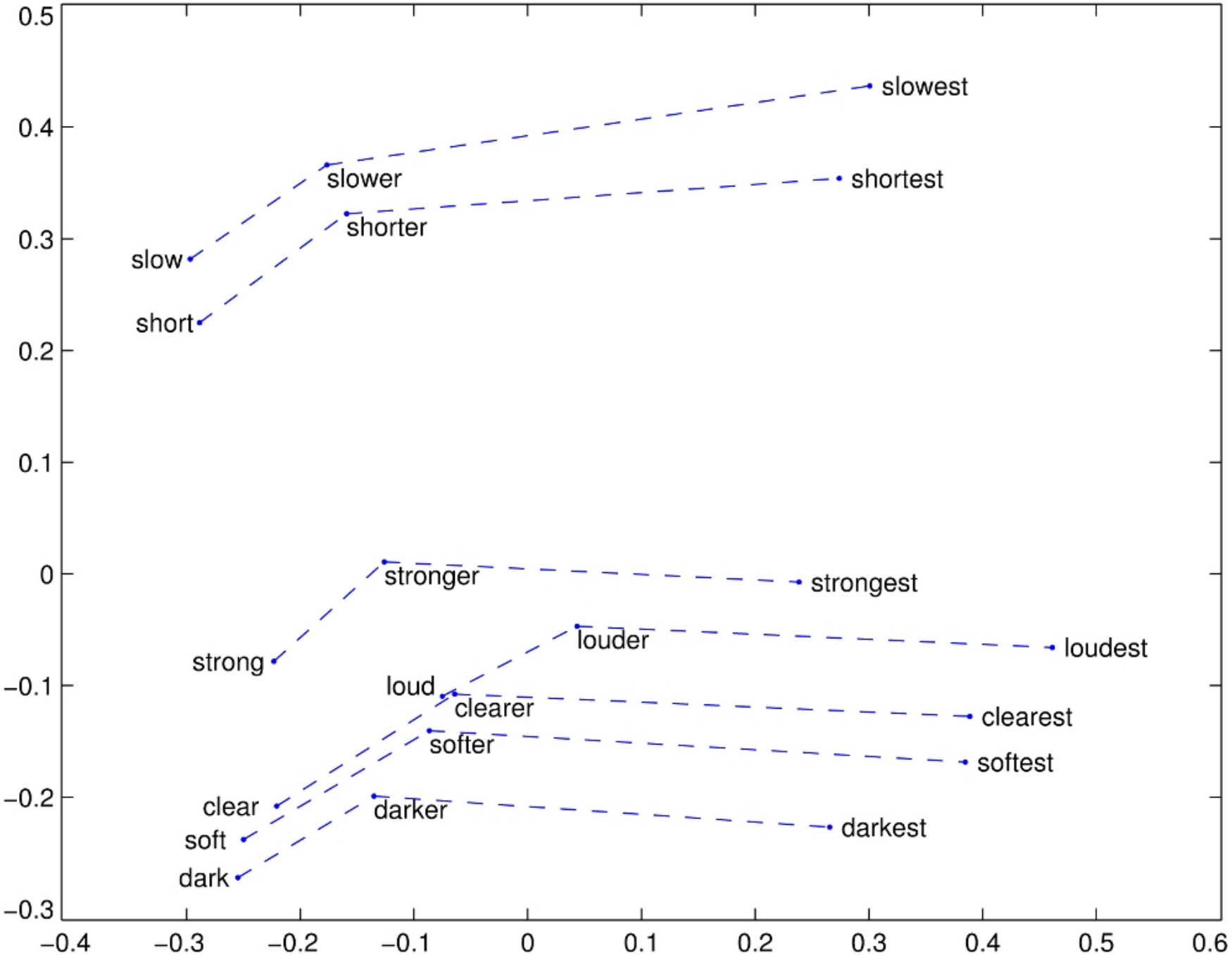
Alternatively, we can require that the direction of the transformation be maintained.

$$\arg\max_{b^* \in V} (\cos(b^*, b - a + a^*))$$

$$\arg\max_{b^* \in V} (\cos(b^* - b, a^* - a))$$

This basically means that b∗ − b **shares the same direction with** a∗ − a, ignoring the distances


0.5
0.4
0.3
0.2
0.1
0
−0.1
−0.2
−0.3
−0.4
−0.5
−0.5
−0.4
−0.3
−0.2
−0.1
0
0.1
0.2
0.3
0.4
0.5
niece
aunt
sister
nephew
uncle
brother
woman
man
heiress
madam
heir
sir
countess
duchess
empress
earl
queen
duke
emperor
king


slowest
slower
shortest
shorter
slow
short
stronger
strongest
louder
loudest
strong
loud
clearer
clearest
softer
softest
clear
darker
darkest
soft
dark
0.5
0.4
0.3
0.2
0.1
0
−0.1
−0.2
−0.3
−0.4
−0.3
−0.2
−0.1
0
0.1
0.2
0.3
0.4
0.5
0.6

Vector compositionality

Mikolov et al experiment with using element-wide addition to compose vectors

Czech + currency	Vietnam + capital	German + airlines
koruna	Hanoi	airline Lufthansa
Check crown	Ho Chi Minh City	carrier Lufthansa
Polish zolty	Viet Nam	flag carrier Lufthansa
CTK	Vietnamese	Lufthansa

Russian + river	French + actress
Moscow	Juliette Binoche
Volga River	Vanessa Paradis
upriver	Charlotte Gainsbourg
Russia	Cecile De

Representing Phrases with vectors

Mikolov et al constructed representations for phrases as well as for individual words.

To learn vector representations for phrases, they first find words that appear frequently together but infrequently in other contexts, and represent these n-grams as single tokens.

For example, “New York Times” and “Toronoto Maple Leafs” are replaced by New_York_Times and Toronoto_Maple_Leafs, but a bigram like “this is” remains unchanged.

$$\text{score}(w_i, w_j) = \frac{\text{count}(w_i w_j) - \delta}{\text{count}(w_i) \times \text{count}(w_j)}.$$

Analogical reasoning task for phrases

Newspapers			
New York	New York Times	Baltimore	Baltimore Sun
San Jose	San Jose Mercury News	Cincinnati	Cincinnati Enquirer
NHL Teams			
Boston	Boston Bruins	Montreal	Montreal Canadiens
Phoenix	Phoenix Coyotes	Nashville	Nashville Predators
NBA Teams			
Detroit	Detroit Pistons	Toronto	Toronto Raptors
Oakland	Golden State Warriors	Memphis	Memphis Grizzlies
Airlines			
Austria	Austrian Airlines	Spain	Spainair
Belgium	Brussels Airlines	Greece	Aegean Airlines
Company executives			
Steve Ballmer	Microsoft	Larry Page	Google
Samuel J. Palmisano	IBM	Werner Vogels	Amazon

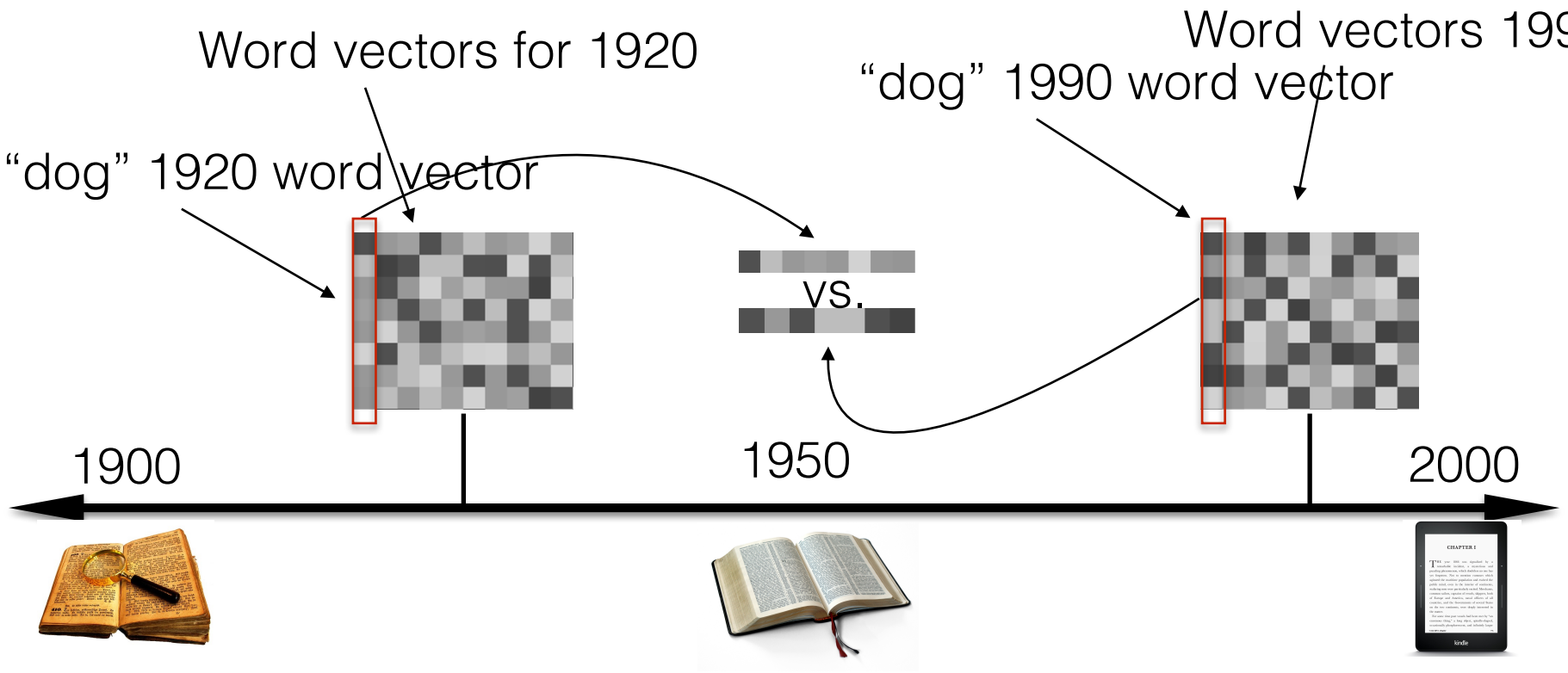
Embeddings can help study word history!

Train embeddings on old books to study changes in word meaning!!

Dan Jurafsky

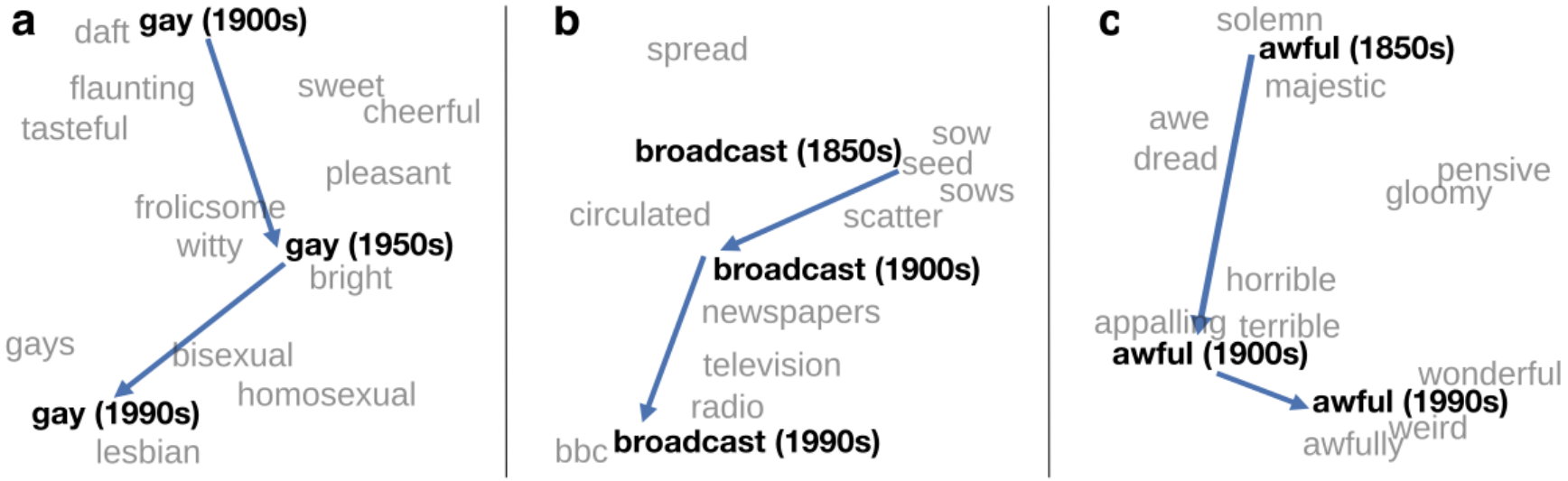
Will Hamilton

Diachronic word embeddings for studying language change!



Visualizing changes

Project 300 dimensions down into 2



~30 million books, 1850-1990, Google Books data

gay | gā |

adjective (**gayer**, **gayest**)

1 (of a person) homosexual (used especially of a man): *that friend of yours, is he gay?*
- relating to or used by homosexuals: *a gay bar* | *the gay vote can decide an election*.

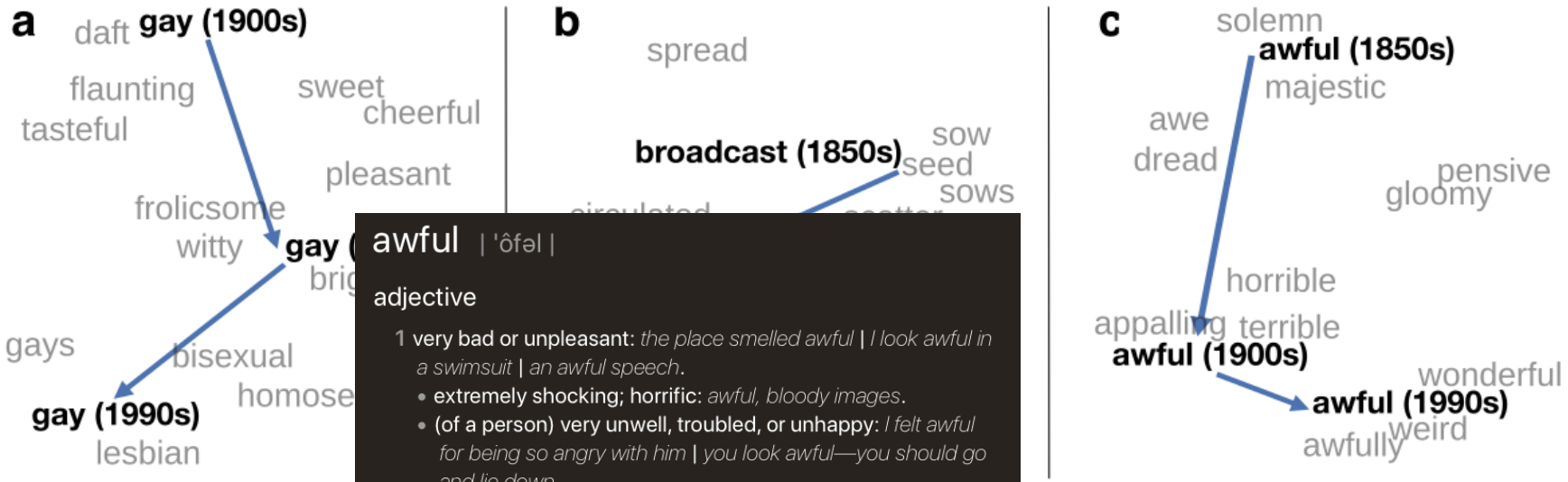
2 *dated* lighthearted and carefree: *Nan had a gay disposition and a very pretty face*.
- brightly colored; showy; brilliant: *a gay profusion of purple and pink sweet peas*.

broadcast | 'brôd,kast |

verb (past and past participle **broadcast**) *[with object]*

1 transmit (a program or some information) by radio or television: *the announcement was broadcast live* | (as noun **broadcasting**) *: the 1920s saw the dawn of broadcasting*.
- *[no object]* take part in a radio or television transmission: *the station broadcasts 24 hours a day*.
- tell (something) to many people; make widely known: *we don't want to broadcast our unhappiness to the world*.

2 scatter (seeds) by hand or machine rather than placing in drills or rows.



awful | 'ôfəl |

adjective

1 very bad or unpleasant: *the place smelled awful* | *I look awful in a swimsuit* | *an awful speech*.
- extremely shocking; horrific: *awful, bloody images*.
- (of a person) very unwell, troubled, or unhappy: *I felt awful for being so angry with him* | *you look awful—you should go and lie down*.

2 *[attributive]* used to emphasize the extent of something, especially something unpleasant or negative: *I've made an awful fool of myself*.

3 *archaic* inspiring reverential wonder or fear.

~30 million boo

Embeddings and bias

Embeddings reflect cultural bias

Bolukbasi, Tolga, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam T. Kalai. "Man is to computer programmer as woman is to homemaker? debiasing word embeddings." In *Advances in Neural Information Processing Systems*, pp. 4349-4357. 2016.

Ask “Paris : France :: Tokyo : x”

- x = Japan

Ask “father : doctor :: mother : x”

- x = nurse

Ask “man : computer programmer :: woman : x”

- x = homemaker

Measuring cultural bias

Implicit Association test (Greenwald et al 1998): How associated are

- concepts (*flowers*, *insects*) & attributes (*pleasantness*, *unpleasantness*)?
- Studied by measuring timing latencies for categorization.

Psychological findings on US participants:

- African-American names are associated with unpleasant words (more than European-American names)
- Male names associated more with math, female names with arts
- Old people's names with unpleasant words, young people with pleasant words.

Embeddings reflect cultural bias

Aylin Caliskan, Joanna J. Bruson and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. Science 356:6334, 183-186.

Caliskan et al. replication with embeddings:

- African-American names (*Leroy, Shaniqua*) had a higher GloVe cosine with unpleasant words (*abuse, stink, ugly*)
- European American names (*Brad, Greg, Courtney*) had a higher cosine with pleasant words (*love, peace, miracle*)

Embeddings reflect and replicate all sorts of pernicious biases.

Directions

Debiasing algorithms for embeddings

- Bolukbasi, Tolga, Chang, Kai-Wei, Zou, James Y., Saligrama, Venkatesh, and Kalai, Adam T. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems*, pp. 4349–4357.

Use embeddings as a historical tool to study bias

Embeddings as a window onto history

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou, (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, *115*(16), E3635–E3644

Use the Hamilton historical embeddings

The cosine similarity of embeddings for decade X for occupations (like teacher) to male vs female names

- Is correlated with the actual percentage of women teachers in decade X

History of biased framings of women

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou, (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, *115*(16), E3635–E3644

Embeddings for competence adjectives are biased toward men

- *Smart, wise, brilliant, intelligent, resourceful, thoughtful, logical, etc.*

This bias is slowly decreasing

Princeton Trilogy experiments

Study 1: Katz and Braley (1933)

Investigated whether traditional social stereotypes had a cultural basis

Ask 100 male students from Princeton University to choose five traits that characterized different ethnic groups (for example Americans, Jews, Japanese, Negroes) from a list of 84 word

84% of the students said that Negroes were superstitious and 79% said that Jews were shrewd. They were positive towards their own group.

Study 2: Gilbert (1951)
Less uniformity of agreement about unfavorable traits than in 1933.

Study 3: Karlins et al. (1969)

Many students objected to the task but this time there was greater agreement on the stereotypes assigned to the different groups compared with the 1951 study. Interpreted as a re-emergence of social stereotyping but in the direction more favorable stereotypical images.

Embeddings reflect ethnic stereotypes over time

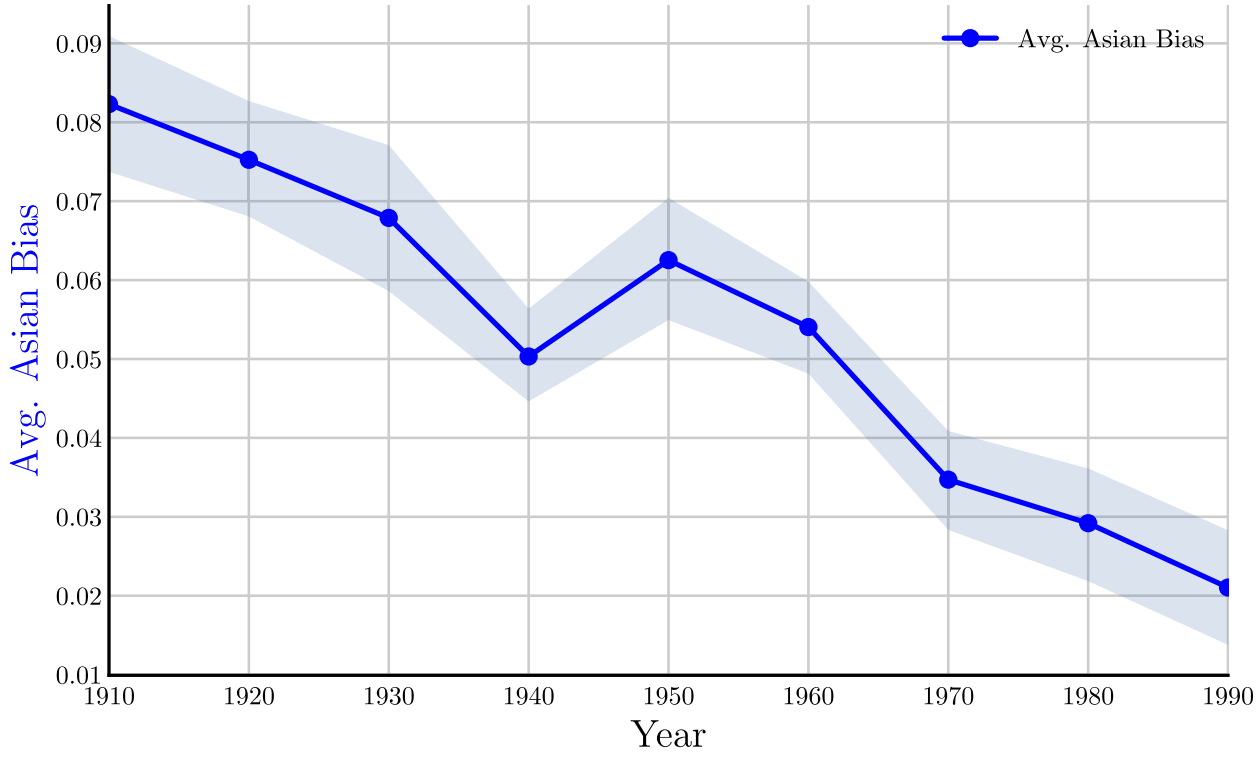
Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou, (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, *115*(16), E3635–E3644

- Princeton trilogy experiments
- Attitudes toward ethnic groups (1933, 1951, 1969) scores for adjectives
  - *industrious, superstitious, nationalistic*, etc
- Cosine of Chinese name embeddings with those adjective embeddings correlates with human ratings.

Change in linguistic framing 1910-1990

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou, (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, *115*(16), E3635–E3644

Change in association of Chinese names with adjectives framed as "othering" (*barbaric*, *monstrous*, *bizarre*)



Changes in framing: adjectives associated with Chinese

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou, (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, *115*(16), E3635–E3644

1910	1950	1990
Irresponsible	Disorganized	Inhibited
Envious	Outrageous	Passive
Barbaric	Pompous	Dissolute
Aggressive	Unstable	Haughty
Transparent	Effeminate	Complacent
Monstrous	Unprincipled	Forceful
Hateful	Venomous	Fixed
Cruel	Disobedient	Active
Greedy	Predatory	Sensitive
Bizarre	Boisterous	Hearty

Conclusion

Embeddings = vector models of meaning

- More fine-grained than just a string or index
- Especially good at modeling similarity/analogy
  - Just download them and use cosines!!
- Can use sparse models (tf-idf) or dense models (word2vec, GLoVE)

- **Useful in practice but know they encode cultural stereotypes**