Words

9.19: Computational Psycholinguistics
18 September 2023
Roger Levy

How do we learn so many words?

- The average 20-year-old native English speaker knows **42,000 lemmas**
- That is 5.75 lemmas per day, every day!
- The mystery:

The average seventh-grader…must have acquired most of them as a result of reading because (a) the majority of English words are used only in print, (b) she already knew well almost all the words she would have encountered in speech, and (c) she learned less than one word by direct instruction. Studies of children reading grade-school text find that about one word in every 20 paragraphs goes from wrong to right on a vocabulary test. The typical seventh grader would have read less than 50 paragraphs since yesterday, from which she should have learned less than three new words. Apparently, she mastered the meanings of [several] words that she did not encounter.

(Landauer & Dumais, 1997, Psychological Review)

The distributional hypothesis

We saw a cute, hairy wampimuk sleeping behind the tree

- The *Distributional Hypothesis* of Harris (1954): the context in which a word appears carries information about its meaning
- Succinct versions:
  - “You shall know a word by the company it keeps” (Firth, 1957)
  - “…the linguistic meanings which the structure carries can only be due to the relations in which the elements of the structure take part” (Harris, 1968)

More complex examples

The degus was hermetically broamed.

-

(After McDonald & Ramscar, 2001)

Implicit distributional/contextual knowledge

What word can appear in the context of all these words?

Word 1: drown, bathroom, shower, fill, fall, lie, electrocute, toilet, whirlpool, iron, gin

Word 2: eat, fall, pick, slice, peel, tree, throw, fruit, pie, bite, crab, grate

Word 3: advocate, overthrow, establish, citizen, ideal, representative, dictatorship, campaign, bastion, freedom

Word 4: spend, enjoy, remember, last, pass, end, die, happen, brighten, relive

(Slide from Yoav Goldberg)

Implicit distributional/contextual knowledge

What word can appear in the context of all these words?

bathtub

Word 1: drown, bathroom, shower, fill, fall, lie, electrocute, toilet, whirlpool, iron, gin

apple

Word 2: eat, fall, pick, slice, peel, tree, throw, fruit, pie, bite, crab, grate

democracy

Word 3: advocate, overthrow, establish, citizen, ideal, representative, dictatorship, campaign, bastion, freedom

day

Word 4: spend, enjoy, remember, last, pass, end, die, happen, brighten, relive

(*Slide from Yoav Goldberg*)

A more complex case

Word 5: eat, paint, peel, apple, fruit, juice, lemon, blue, grow

(Slide from Yoav Goldberg)

A practical problem for *n*-gram modeling

- Consider the distributions on these contexts:
  - The soup was… 7402
  - The broth was… 1903
  - The chowder was… 231
  - The bisque was… 118
  - The soup will be… 815
  - The broth will be… 122

  Google Web context counts
- *n*-gram models have no built-in ways of leveraging similarity among contexts
- Similar problems exist for conditioning on context for probabilistic grammars

Innovation in multi-word expressions

- What can you *drive someone…*?

mad

crazy

to distraction

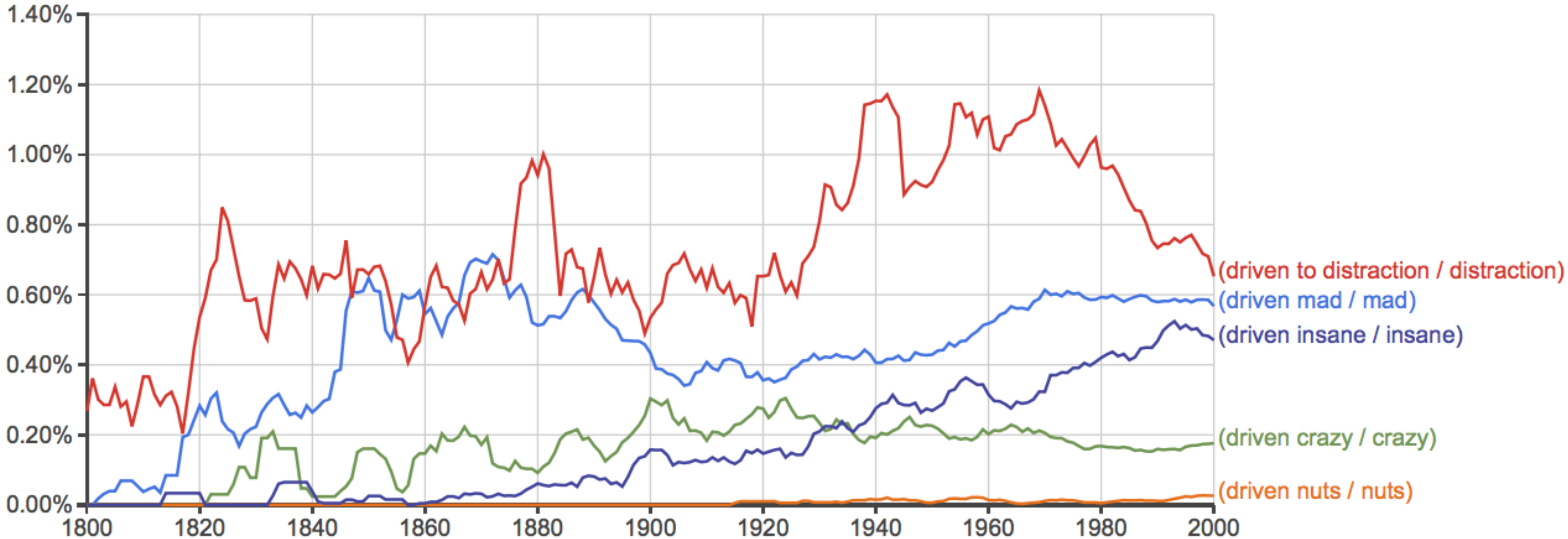
bananas

insane

nuts

(Bybee 2012, Domain-general processes as the basis for grammar)

Innovation in multi-word expressions



- These expressions do not come on the scene independently!
- There is lexical specificity, but innovation also spreads along lines of semantic similarity

Fundamental idea

- We have tens of thousands of words in our lexicon
- But semantic lexical knowledge mostly lives on a lower-dimensional subspace
- By learning that lower-dimensional subspace, we can:
  - Better handle data sparsity in practical NLP applications
  - Resolve the mystery of how we learn so many words so fast
  - Improve our understanding of human conceptual space
  - Better explain the full distribution of linguistic expressions

Technical foundations

- We want to go from sparse…

$$[\![dog]\!] = [0,0,\ldots,0,1,0,\ldots,0]$$

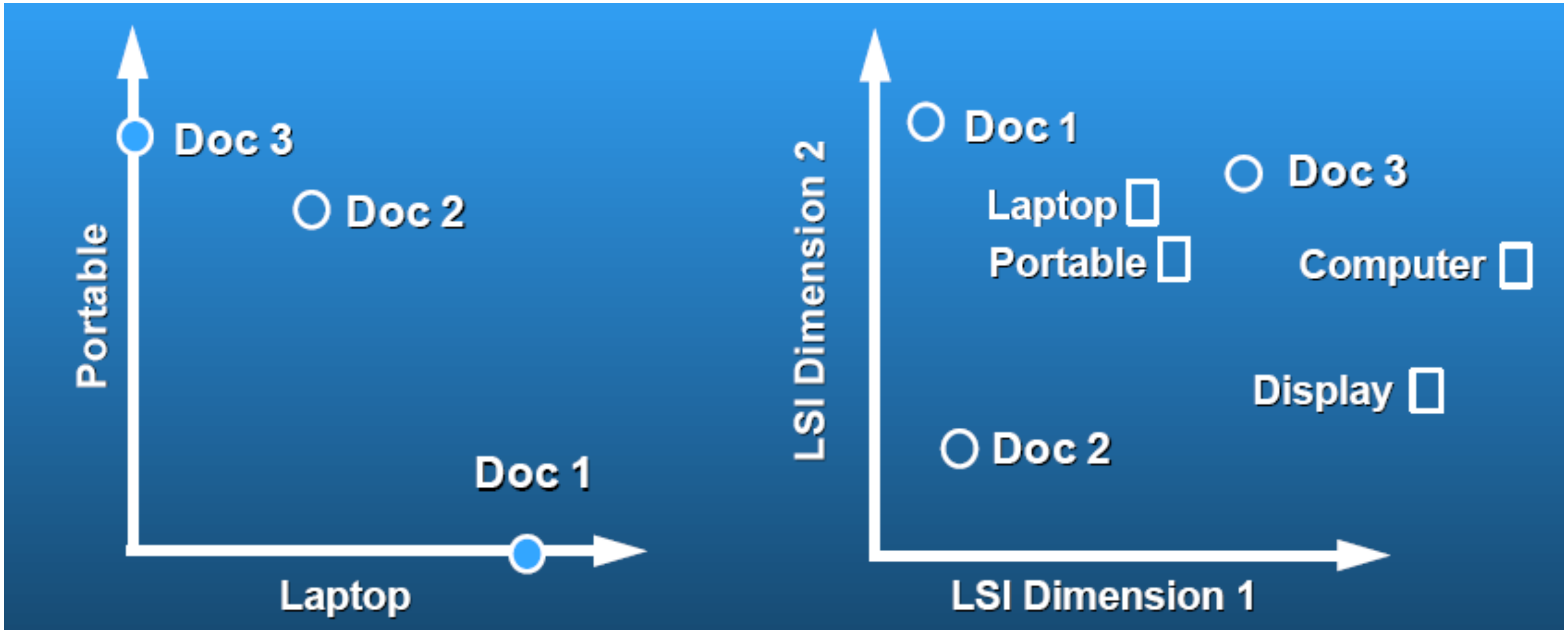
- …to dense:

$$[\![dog]\!] = [-0.11,0.81,\ldots,\ldots,0.58,0.07]$$

- There are many ways proposed to do this!

Low-dimensional word meanings from contexts

- The general goal:



courtesy of Susan Dumais

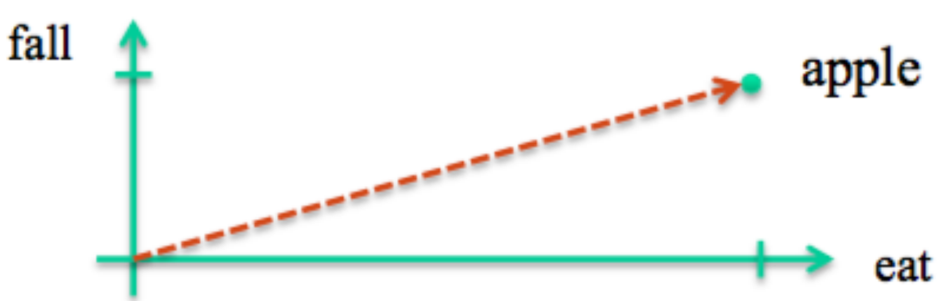
(via Chris Manning & Hinrich Schutze)

How can we compare two context collections in their entirety?

Count how often "apple" occurs close to other words in a large text collection (corpus):

eat	fall	ripe	slice	peel	tree	throw	fruit	pie	bite	crab
794	244	47	221	208	160	145	156	109	104	88

Interpret counts as coordinates:



Every context word becomes a dimension.

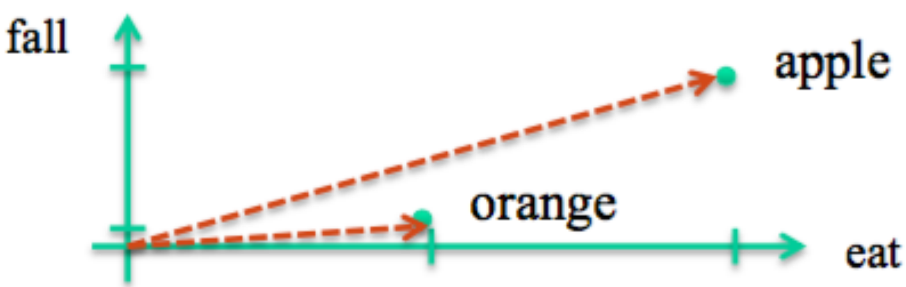
(Slide from Yoav Goldberg)

How can we compare two context collections in their entirety?

Then visualize both count tables as vectors in the same space:

eat	fall	ripe	slice	peel	tree	throw	fruit	pie	bite	crab
794	244	47	221	208	160	145	156	109	104	88

eat	fall	ripe	slice	peel	tree	throw	fruit	pie	bite	crab
265	22	25	62	220	64	74	111	4	4	8

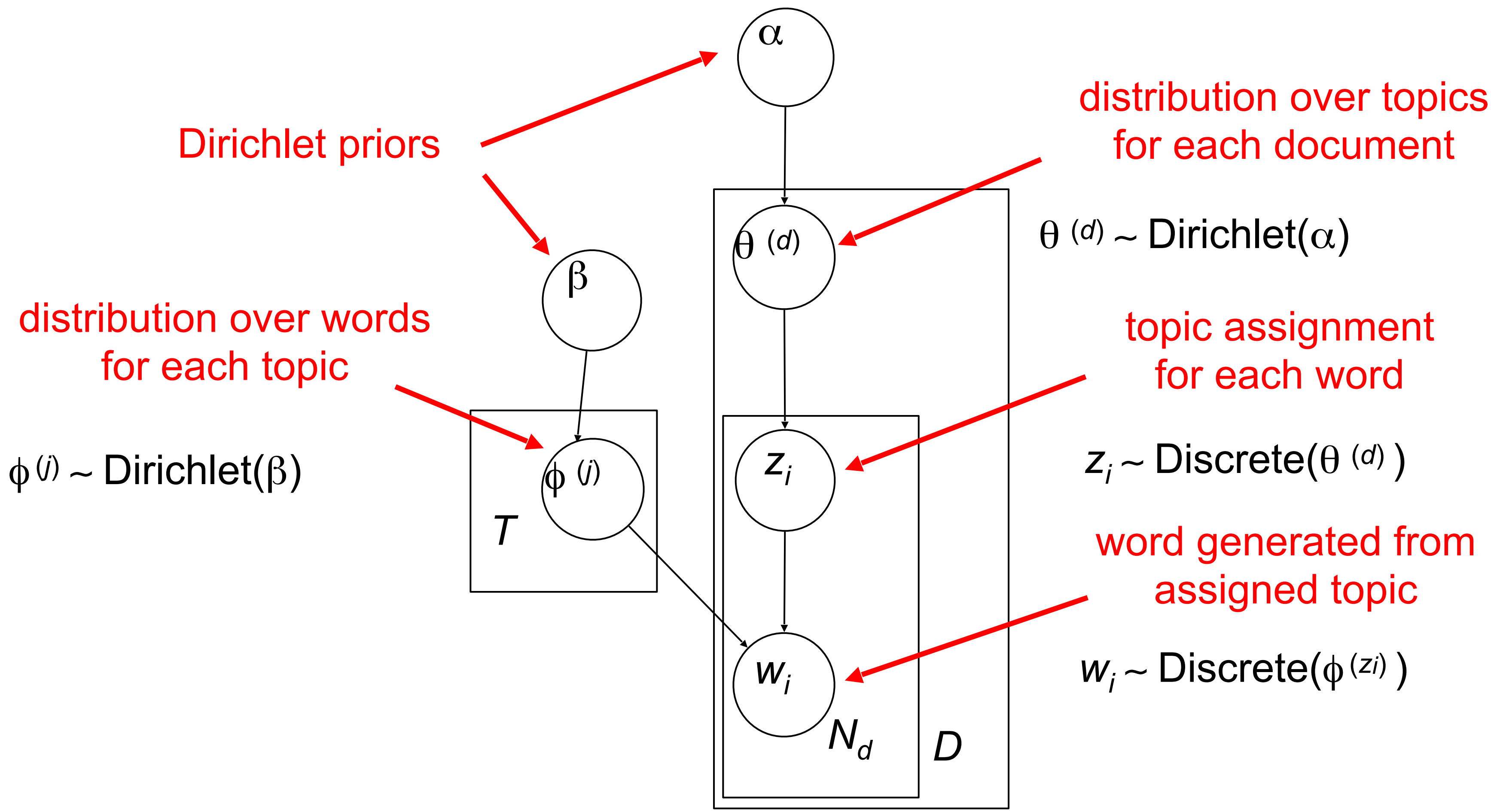


Similarity between two words as proximity in space

(Slide from Yoav Goldberg)

Hierarchical Bayesian methods

- Latent Dirichlet Allocation (aka Topic Models): Blei, Ng, Jordan (2001,2003)



(Slide from Tom Griffiths)

Interpretable topics

DISEASE	WATER	MIND	STORY	FIELD	SCIENCE	BALL	JOB
BACTERIA	FISH	WORLD	STORIES	MAGNETIC	STUDY	GAME	WORK
DISEASES	SEA	DREAM	TELL	MAGNET	SCIENTISTS	TEAM	JOBS
GERMS	SWIM	DREAMS	CHARACTER	WIRE	SCIENTIFIC	FOOTBALL	CAREER
FEVER	SWIMMING	THOUGHT	CHARACTERS	NEEDLE	KNOWLEDGE	BASEBALL	EXPERIENCE
CAUSE	POOL	IMAGINATION	AUTHOR	CURRENT	WORK	PLAYERS	EMPLOYMENT
CAUSED	LIKE	MOMENT	READ	COIL	RESEARCH	PLAY	OPPORTUNITIES
SPREAD	SHELL	THOUGHTS	TOLD	POLES	CHEMISTRY	FIELD	WORKING
VIRUSES	SHARK	OWN	SETTING	IRON	TECHNOLOGY	PLAYER	TRAINING
INFECTION	TANK	REAL	TALES	COMPASS	MANY	BASKETBALL	SKILLS
VIRUS	SHELLS	LIFE	PLOT	LINES	MATHEMATICS	COACH	CAREERS
MICROORGANISMS	SHARKS	IMAGINE	TELLING	CORE	BIOLOGY	PLAYED	POSITIONS
PERSON	DIVING	SENSE	SHORT	ELECTRIC	FIELD	PLAYING	FIND
INFECTIOUS	DOLPHINS	CONSCIOUSNESS	FICTION	DIRECTION	PHYSICS	HIT	POSITION
COMMON	SWAM	STRANGE	ACTION	FORCE	LABORATORY	TENNIS	FIELD
CAUSING	LONG	FEELING	TRUE	MAGNETS	STUDIES	TEAMS	OCCUPATIONS
SMALLPOX	SEAL	WHOLE	EVENTS	BE	WORLD	GAMES	REQUIRE
BODY	DIVE	BEING	TELLS	MAGNETISM	SCIENTIST	SPORTS	OPPORTUNITY
INFECTIONS	DOLPHIN	MIGHT	TALE	POLE	STUDYING	BAT	EARN
CERTAIN	UNDERWATER	HOPE	NOVEL	INDUCED	SCIENCES	TERRY	ABLE

each column shows words from a single topic, ordered by P(*w*|*z*)

(Slide from Tom Griffiths)

The first neural embedding: `word2vec`

word2vec implements several different algorithms:

Two training methods

- Negative Sampling
- Hierarchical Softmax

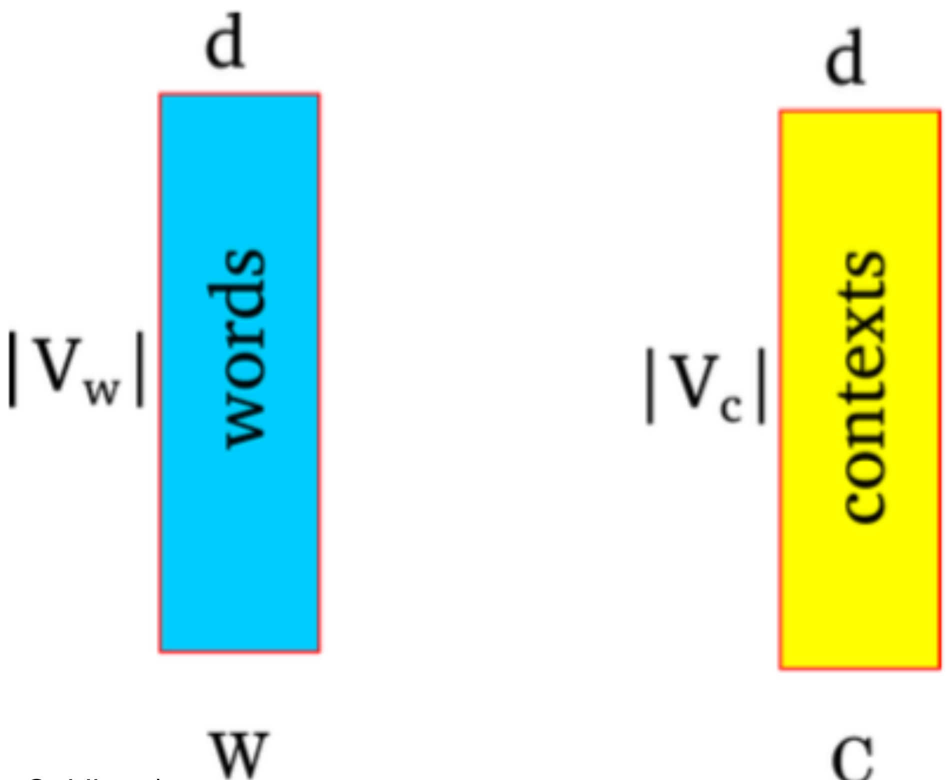
Two context representations

- Continuous Bag of Words (CBOW)
- Skip-grams

(Slide from Yoav Goldberg)

How does word2vec work?

- Represent each word as a *d* dimensional vector.
- Represent each context as a *d* dimensional vector.
- Initalize all vectors to random weights.
- Arrange vectors in two matrices, *W* and *C*.



(Slide from Yoav Goldberg)

How does word2vec work?

While more text:

- Extract a word window:

```
A springer is [ a    cow   or   heifer   close   to   calving  ].
                c1   c2    c3   w        c4      c5   c6
```

- Try setting the vector values such that:

$$\sigma(w \cdot c_1)+\sigma(w \cdot c_2)+\sigma(w \cdot c_3)+\sigma(w \cdot c_4)+\sigma(w \cdot c_5)+\sigma(w \cdot c_6)$$

is **high**

- Create a corrupt example by choosing a random word w'

```
[ a    cow   or   comet   close   to   calving  ]
  c1   c2    c3   w'      c4      c5   c6
```

- Try setting the vector values such that:

$$\sigma(w' \cdot c_1)+\sigma(w' \cdot c_2)+\sigma(w' \cdot c_3)+\sigma(w' \cdot c_4)+\sigma(w' \cdot c_5)+\sigma(w' \cdot c_6)$$

is **low**

(Slide from Yoav Goldberg)

How does word2vec work?

The training procedure results in:

- $w \cdot c$ for **good** word-context pairs is **high**.
- $w \cdot c$ for **bad** word-context pairs is **low**.
- $w \cdot c$ for **ok-ish** word-context pairs is **neither high nor low**.

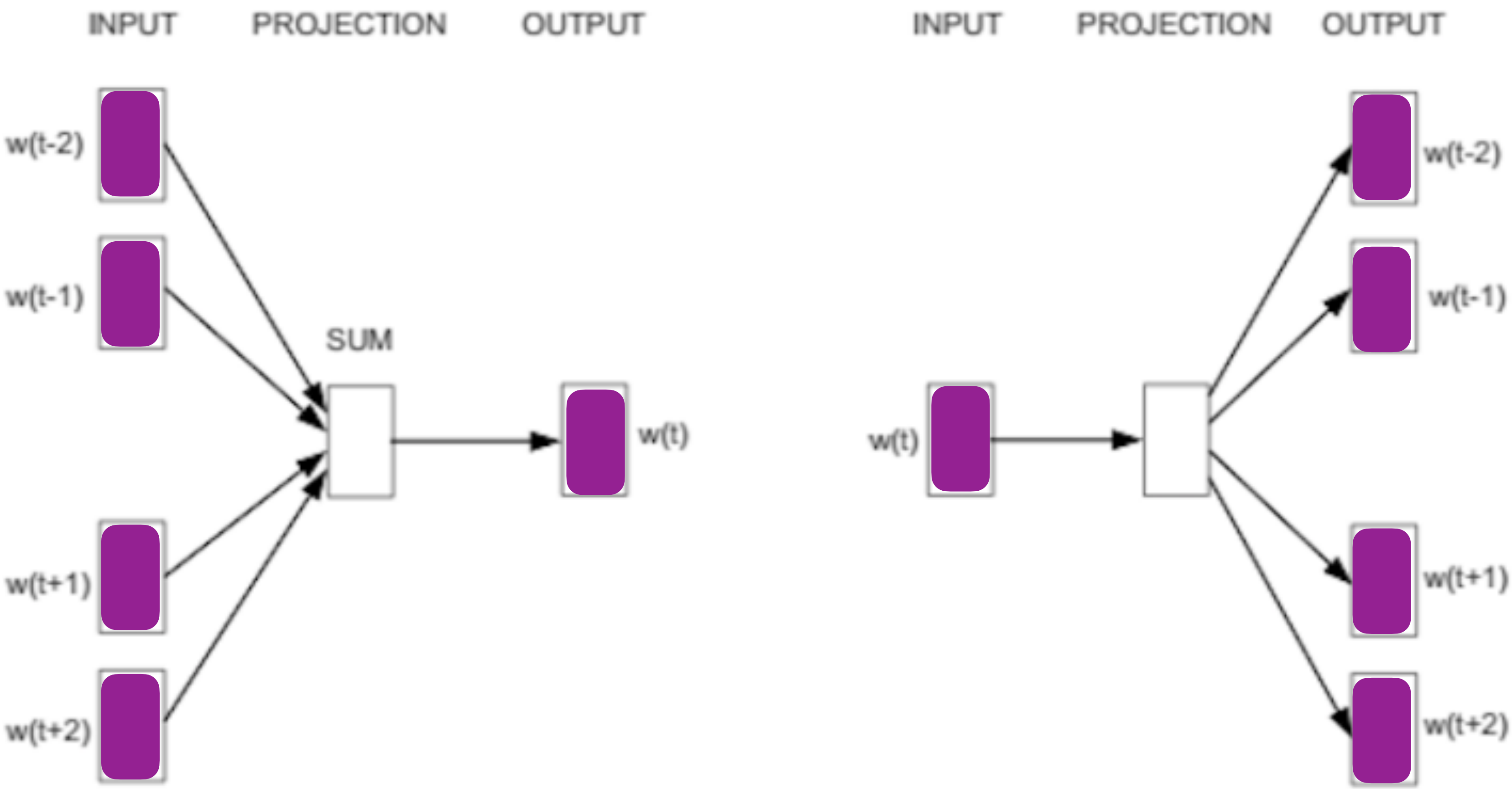
As a result:

- Words that share many contexts get close to each other.
- Contexts that share many words get close to each other.

At the end, word2vec throws away C and returns W.

(Slide from Yoav Goldberg)

word2vec architecture visualized



Continuous bag of words **Skip-gram**

A competitor word embedding: GloVe

- The basic intuition: *ratios of conditional probabilities might give us a handle on meaning components*

Probability and Ratio	$k = solid$	$k = gas$	$k = water$	$k = fashion$
$P(k\|ice)$	1.9×10^{-4}	6.6×10^{-5}	3.0×10^{-3}	1.7×10^{-5}
$P(k\|steam)$	2.2×10^{-5}	7.8×10^{-4}	2.2×10^{-3}	1.8×10^{-5}
$P(k\|ice)/P(k\|steam)$	$\boxed{8.9}$	$\boxed{8.5 \times 10^{-2}}$	$\boxed{1.36}$	$\boxed{0.96}$

*Highly imbalanced; argued to pick out distinctive difference in meaning component of **ice** versus **steam***

*Balanced; argued not to pick out distinctive difference in meaning of **ice** versus **steam***

- Argument: we want our model to optimally approximate

$$\frac{P(w_k|w_i)}{P(w_k|w_j)} = \frac{P(w_k, w_i)}{P(w_j, w_i)} \approx_{RFE} \frac{\boxed{X_{ik}}}{X_{jk}}$$

co-occurrence count of w_i, w_j

(*Pennington et al., 2014*)

Deriving the GloVe word vector model

- We enforce probability ratios to be the ratio of dot-products of the target word to context words

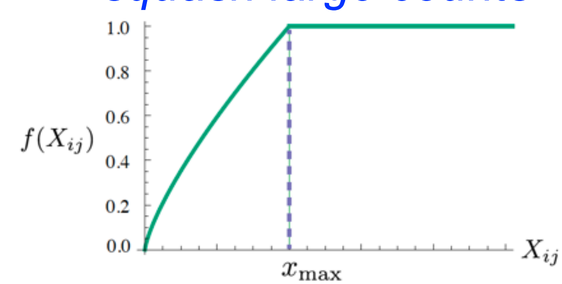
$$F\left((w_i - w_j)^T \tilde{w}_k\right) = \frac{P_{ik}}{P_{jk}}$$

- With a lot more simplification and argumentation we get the following objective function to minimize:

$$J = \sum_{i,j=1}^{V} f\left(X_{ij}\right)\left(\boxed{w_i^T \tilde{w}_j} + \boxed{b_i} + \boxed{\tilde{b}_j} - \log X_{ij}\right)^2$$

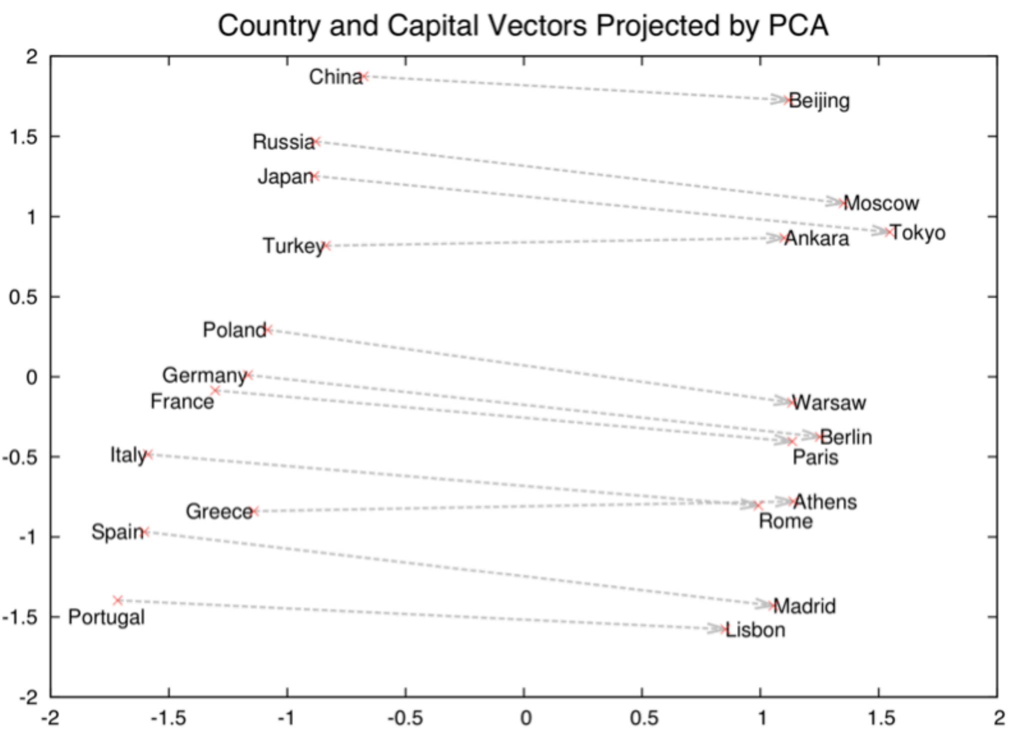
word vector   *bias vectors (ignore)*

f(X_ij) chosen to squash large counts

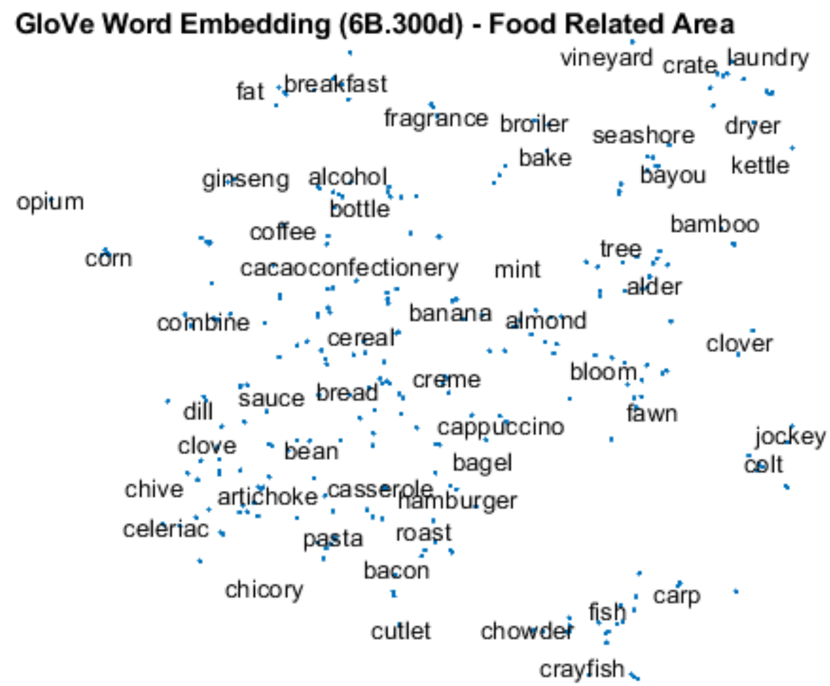


Word meanings reflected in embeddings

word2vec



(Mikolov et al., 2013)



(Pennington et al., 2014)

Exploring GLoVE meaning spaces & analogies

Try playing with this fun visualization tool!

https://lamyiowce.github.io/word2viz/

word2vec embeddings over time

Figure 1: Two-dimensional visualization of semantic change in English using SGNS vectors.[2] **a,** The word *gay* shifted from meaning "cheerful" or "frolicsome" to referring to homosexuality. **b,** In the early 20th century *broadcast* referred to "casting out seeds"; with the rise of television and radio its meaning shifted to "transmitting signals". **c,** *Awful* underwent a process of pejoration, as it shifted from meaning "full of awe" to meaning "terrible or appalling" (Simpson et al., 1989).

(Hamilton, Leskovic, & Jurafsky, 2016 ACL)

Application: is bias embedded in our language?

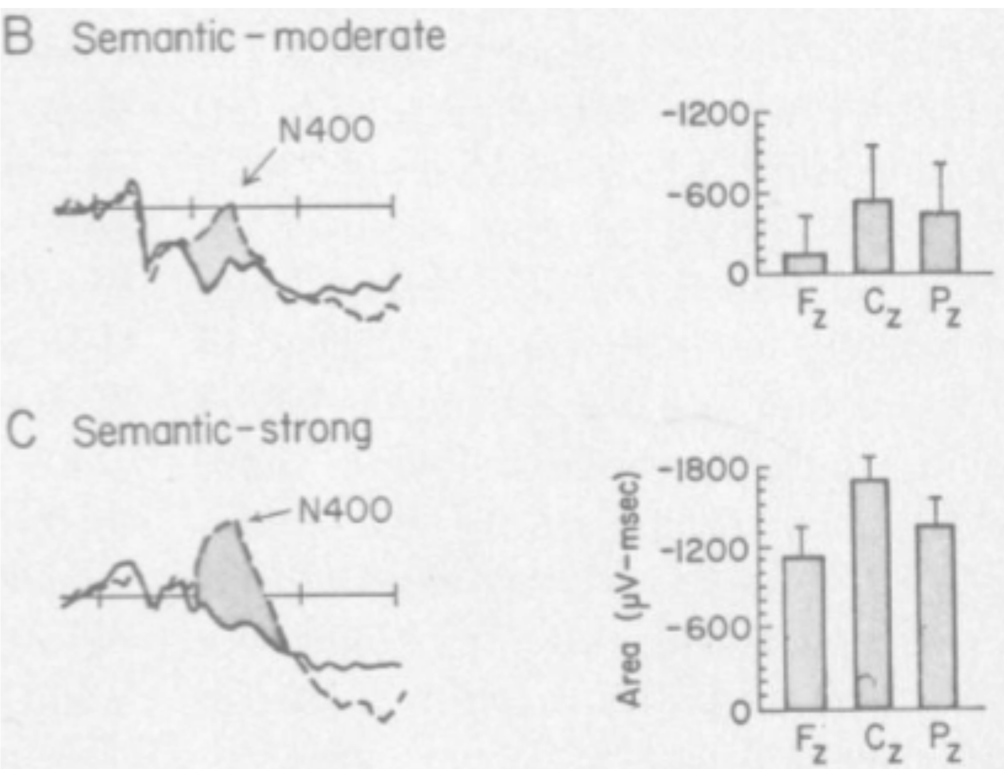
How do we bring this question into our scientific reach?

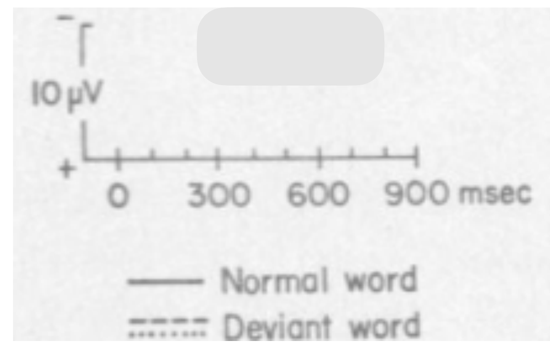
Electrophysiological responses

Rapid Serial Visual Presentation

The N400 in language comprehension

- Differing degrees of semantic congruity:
    - He took a sip from the *drink*. (normal)
    - He took a sip from the *waterfall*. (moderate incongruity)
    - He took a sip from the *transmitter*. (strong incongruity)





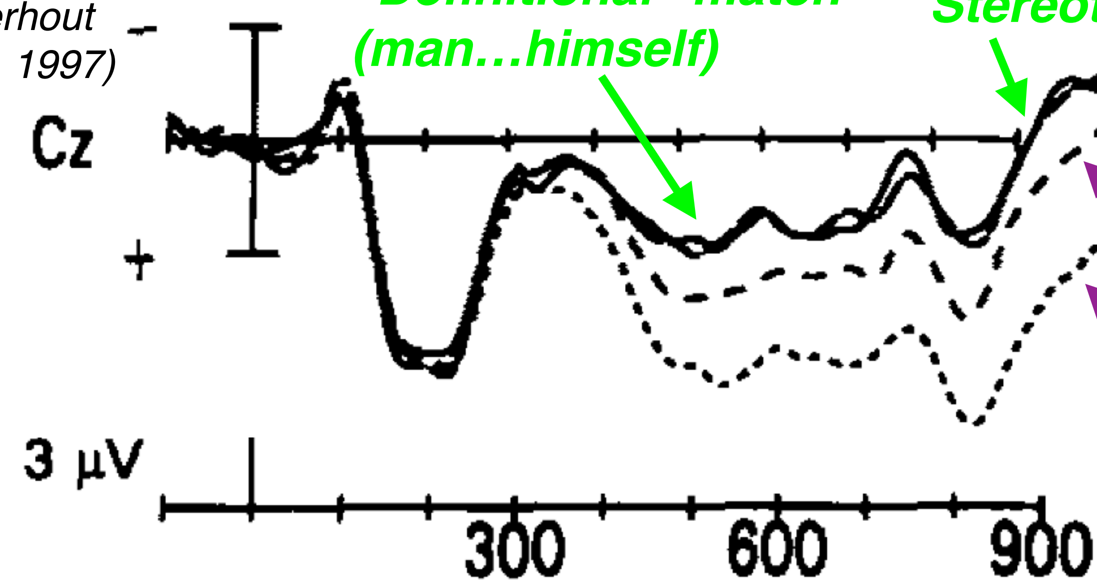
(Kutas & Hillyard, 1980, 1984)

Categorical & stereotypical semantic knowledge

- Mismatches to lexically specified (*definitional**) semantic properties induce measurable expectation violations

  The man prepared herself for the interview.

(Osterhout et al., 1997)

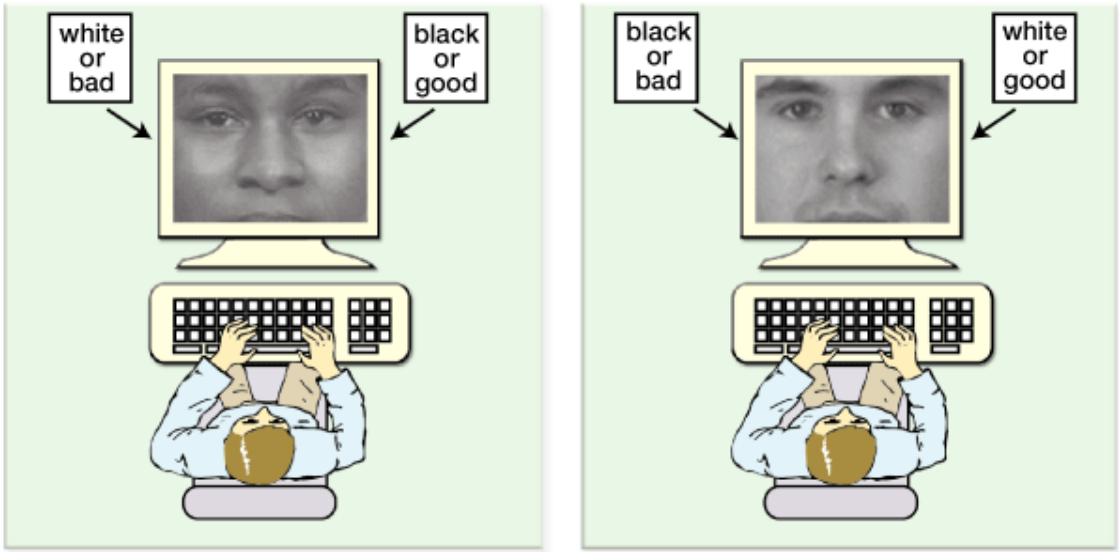


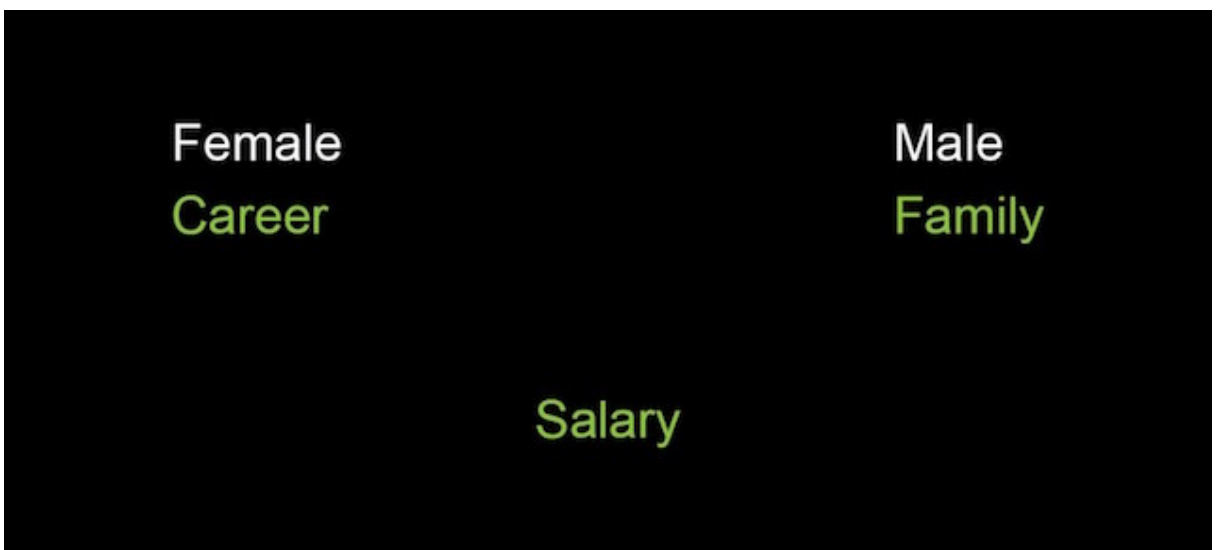
- Mismatches to *stereotypical* semantic properties induce similar violations

  The nurse prepared himself for the operation.

(Osterhout et al., 1997; see also reading time studies by Sturt, 2003; Duffy & Keir, 2004, inter alia)

Stereotypes as implicit associations among concepts





Might stereotypes manifest in distributed linguistic representations too, *biasing* them?

How could we tell?

Quantifying embedding bias

Group 1 words
Mean vector: v_1

he, son, his, him, father, man, boy, himself, male, brother, sons, fathers, men, boys, males, brothers, uncle, uncles, nephew, nephews

Group 2 words
Mean vector: v_2

she, daughter, hers, her, mother, woman, girl, herself, female, sister, daughters, mothers, women, girls, females, sisters, aunt, aunts, niece, nieces

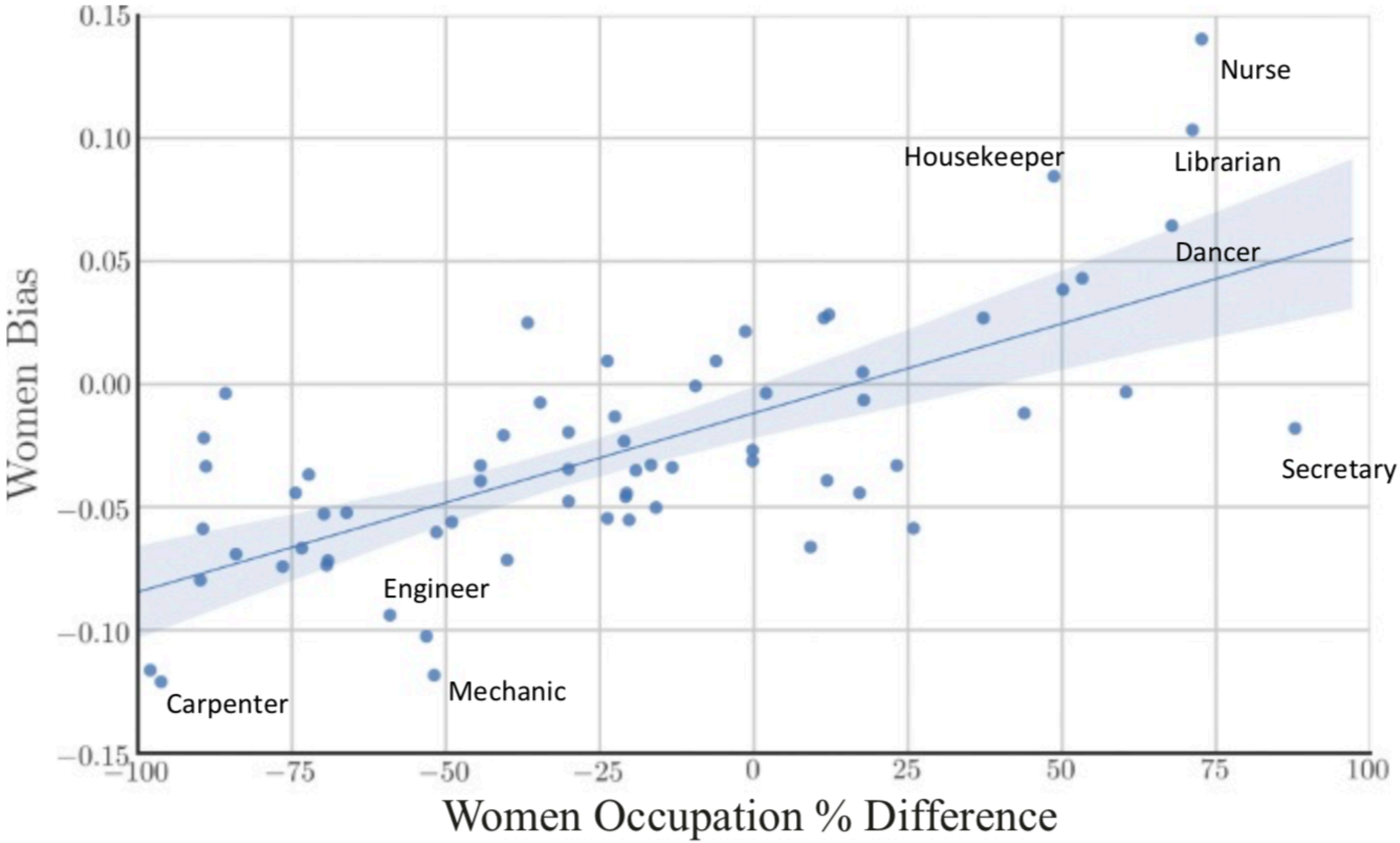
Group "M" words
$\{w_m\}$

janitor, statistician, midwife, bailiff, auctioneer, photographer, geologist, shoemaker, athlete, cashier, dancer, housekeeper, accountant, physicist, gardener, dentist, weaver, blacksmith, psychologist, supervisor, mathematician, surveyor, tailor, designer, economist, mechanic, laborer, postmaster, broker, chemist, librarian, attendant, clerical, musician, porter, scientist, carpenter, sailor, instructor, sheriff, pilot, inspector, mason, baker, administrator, architect, collector, operator, surgeon, driver, painter, conductor, nurse, cook, engineer, retired, sales, lawyer, clergy, physician, farmer, clerk, manager, guard, artist, smith, official, police, doctor, professor, student, judge, teacher, author, secretary, soldier

$$\text{Gender Bias}(w_m) = \text{Dist}(v_1, w_m) - \text{Dist}(v_2, w_m)$$

(Garg et al., 2018)

Word embeddings vs. ground truth



(Garg et al., 2018)

Tracking bias over time

Group 1 words
Mean vector: v_1

baptism, messiah, catholicism, resurrection, christianity, salvation, protestant, gospel, trinity, jesus, christ, christian, cross, catholic, church

Group 2 words
Mean vector: v_2

allah, ramadan, turban, emir, salaam, sunni, koran, imam, sultan, prophet, veil, ayatollah, shiite, mosque, islam, sheik, muslim, muhammad

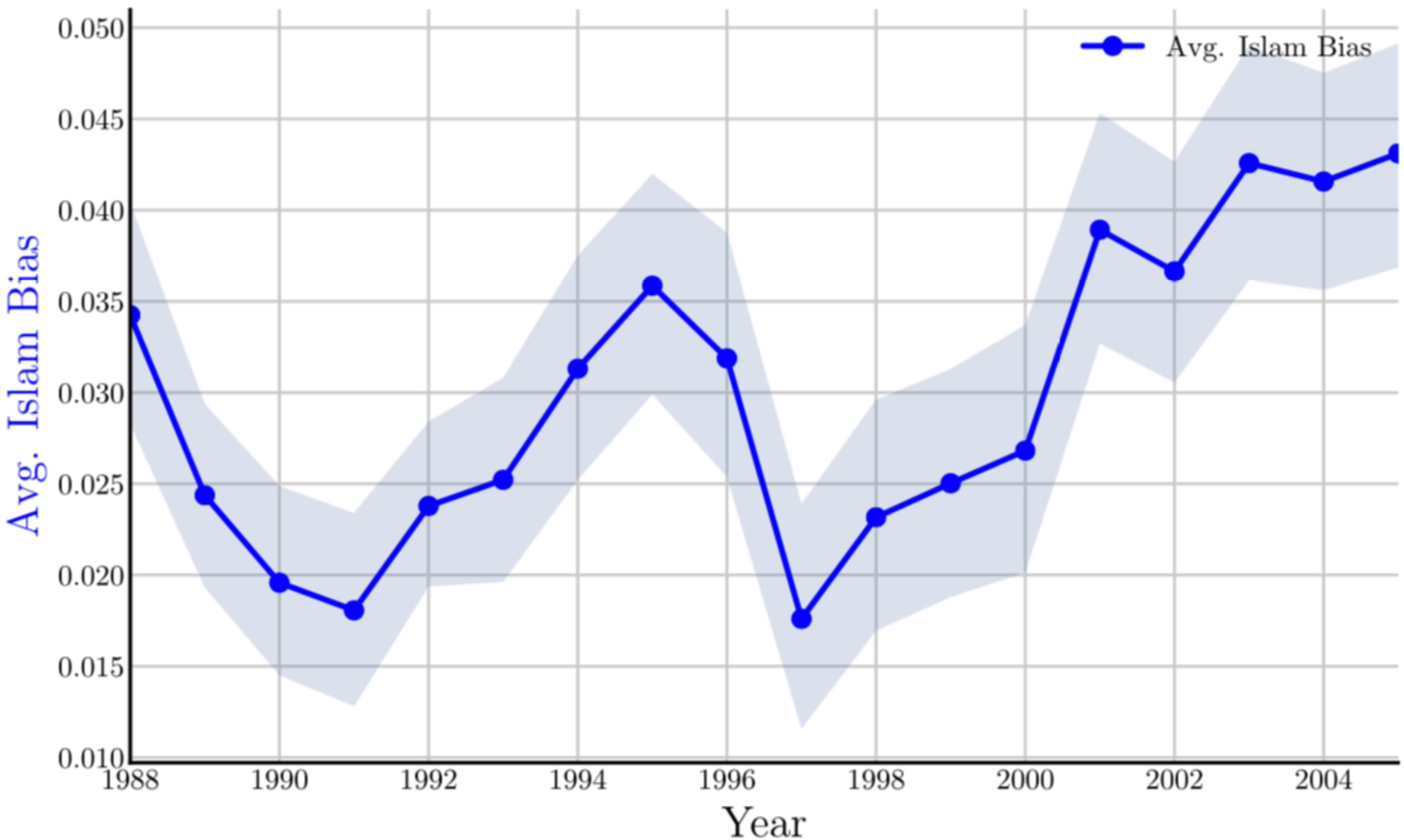
Group "M" words
$\{w_m\}$

terror, terrorism, violence, attack, death, military, war, radical, injuries, bomb, target, conflict, dangerous, kill, murder, strike, dead, violence, fight, death, force, stronghold, wreckage, aggression, slaughter, execute, overthrow, casualties, massacre, retaliation, proliferation, militia, hostility, debris, acid, execution, militant, rocket, guerrilla, sacrifice, enemy, soldier, terrorist, missile, hostile, revolution, resistance, shoot

$$\text{Overall Bias} = \sum_{w_m} \text{Dist}(v_1, w_m) - \text{Dist}(v_2, w_m)$$

(Garg et al., 2018)

Tracking bias over time



(Garg et al., 2018)

An alternative bias-quantifying method

Group 1 words $\{w_1\}$ — *he, son, his, him, father, man, boy, himself, male, brother, sons, fathers, men, boys, males, brothers, uncle, uncles, nephew, nephews*

Group 2 words $\{w_2\}$ — *she, daughter, hers, her, mother, woman, girl, herself, female, sister, daughters, mothers, women, girls, females, sisters, aunt, aunts, niece, nieces*

Group A words $\{w_A\}$ — *career, office, salary, …*

Group B words $\{w_B\}$ — *family, home, children, …*

Overall Bias:

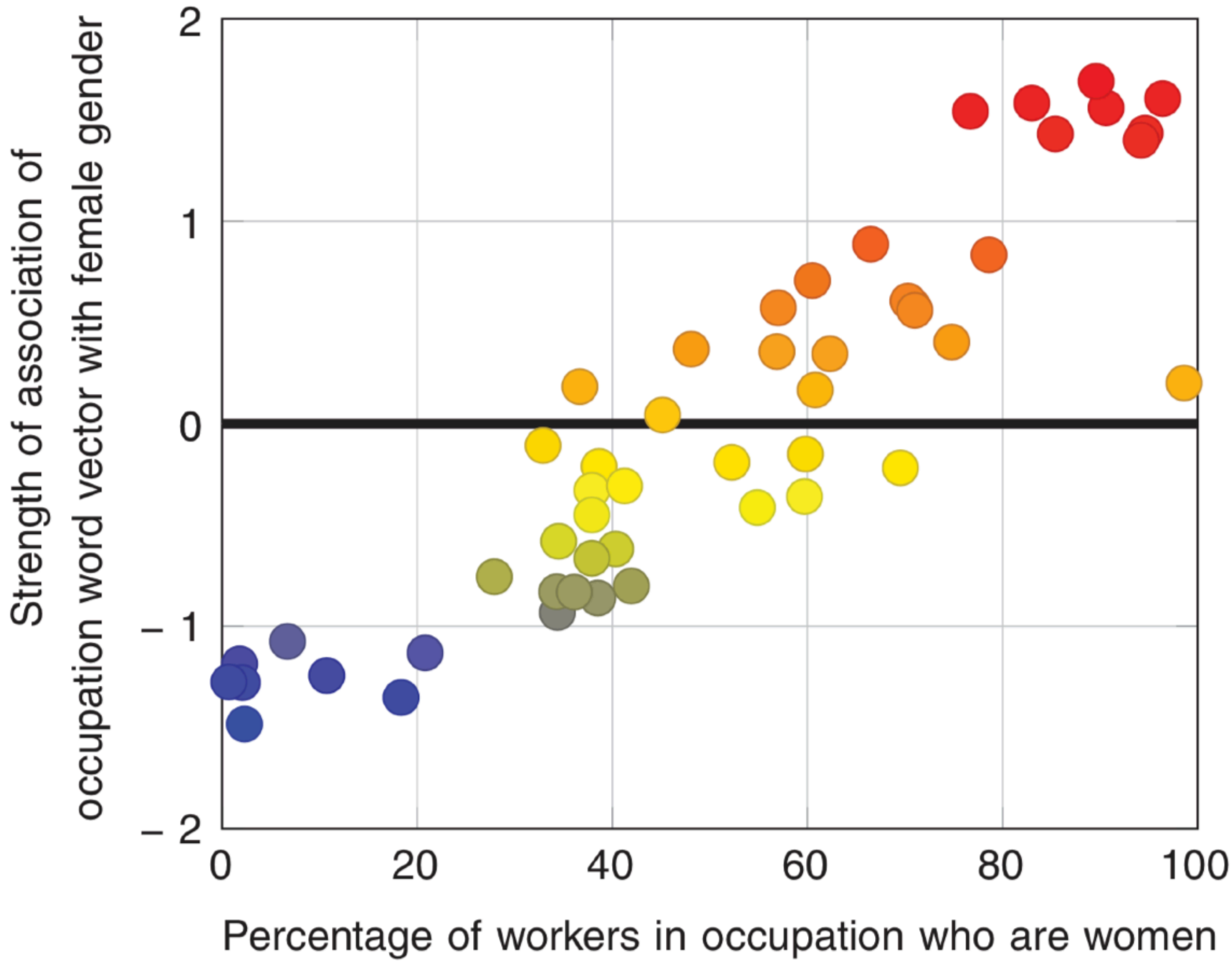
$$\text{Average}_{w_1, w_2, w_A, w_B}\left[\text{Dist}(v_1, w_A) - \text{Dist}(v_2, w_A) - \text{Dist}(v_1, w_B) + \text{Dist}(v_2, w_B)\right]$$

GloVe cosine similarities:

	career	***children***
woman	0.29	0.42
man	0.32	0.27

(Caliskan et al., 2017; "WEFAT")

WEFAT Results



(Caliskan et al., 2017)

WEFAT results: many stereotypes

Target words	Attribute words	Original finding				Our finding			
		Ref.	*N*	*d*	*P*	N_T	N_A	*d*	*P*
Flowers vs. insects	Pleasant vs. unpleasant	(*5*)	32	1.35	10^{-8}	25 × 2	25 × 2	1.50	10^{-7}
Instruments vs. weapons	Pleasant vs. unpleasant	(*5*)	32	1.66	10^{-10}	25 × 2	25 × 2	1.53	10^{-7}
European-American vs. African-American names	Pleasant vs. unpleasant	(*5*)	26	1.17	10^{-5}	32 × 2	25 × 2	1.41	10^{-8}
European-American vs. African-American names	Pleasant vs. unpleasant from (*5*)	(*7*)	Not applicable			16 × 2	25 × 2	1.50	10^{-4}
European-American vs. African-American names	Pleasant vs. unpleasant from (*9*)	(*7*)	Not applicable			16 × 2	8 × 2	1.28	10^{-3}
Male vs. female names	Career vs. family	(*9*)	39k	0.72	$<10^{-2}$	8 × 2	8 × 2	1.81	10^{-3}
Math vs. arts	Male vs. female terms	(*9*)	28k	0.82	$<10^{-2}$	8 × 2	8 × 2	1.06	.018
Science vs. arts	Male vs. female terms	(*10*)	91	1.47	10^{-24}	8 × 2	8 × 2	1.24	10^{-2}
Mental vs. physical disease	Temporary vs. permanent	(*23*)	135	1.01	10^{-3}	6 × 2	7 × 2	1.38	10^{-2}
Young vs. old people's names	Pleasant vs. unpleasant	(*9*)	43k	1.42	$<10^{-2}$	8 × 2	8 × 2	1.21	10^{-2}

Summary

- **Embed** size-V vocabulary in a D-dimensional space; $D<<V$
- Word embedding representations are **dense** numeric vectors
- Embeddings are learned to predict word--word co-occurrence statistics in large corpora
- Because the **distributional hypothesis** holds, a word's embedding representation reflects features of its meaning
- Words with **similar** meanings are **closer** in embedding space
- Perhaps remarkably, many **features** of word meaning turn out to be **linearly separable** in the embedding space
- This enables embedding-based **analogical reasoning**
- Since corpus statistics reflect the world, word embeddings implicitly encode **biases**
- **Open question:** do these biases simply reflect information about the world, or does language present distorted representations of that information?