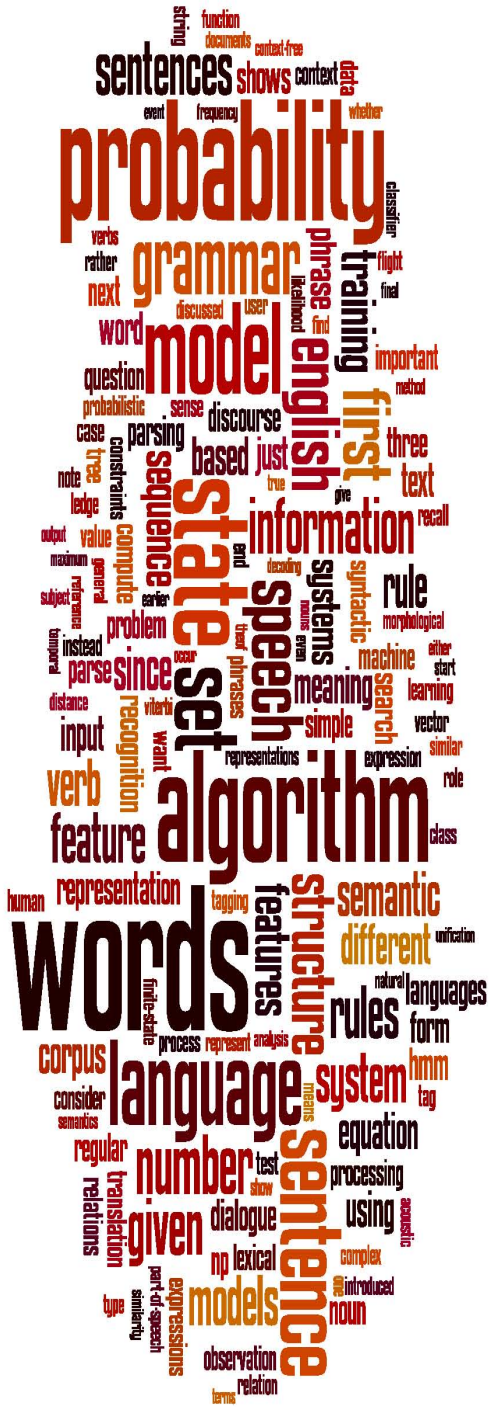


CS 124/LINGUIST 180 From Languages to Information

Dan Jurafsky

Stanford University

Word meaning, Embeddings, and Word2vec

What do words mean?

First thought: look in a dictionary

http://www.oed.com/

Words, Lemmas, Senses, Definitions

lemma **sense** **definition**

pepper, *n.*

Pronunciation: Brit. /ˈpɛpə/ , U.S. /ˈpɛpər/

Forms: OE **peopor** (*rare*), OE **pipcer** (transmission error), OE **pipor**, OE **pipur** (*rare* .

Frequency (in current use):

Etymology: A borrowing from Latin. **Etymon:** Latin *piper*.
< classical Latin *piper*, a loanword < Indo-Aryan (as is ancient Greek *πίπερι*); compare San

I. The spice or the plant.

1.

a. A hot pungent spice derived from the prepared fruits (peppercorns) of the pepper plant, *Piper nigrum* (see sense 2a), used from early times to season food, either whole or ground to powder (often in association with salt). Also (locally, chiefly with distinguishing word): a similar spice derived from the fruits of certain other species of the genus *Piper*; the fruits themselves.

The ground spice from *Piper nigrum* comes in two forms, the more pungent *black pepper*, produced from black peppercorns, and the milder *white pepper*, produced from white peppercorns: see BLACK *adj.* and *n.* Special uses 5a, PEPPERCORN *n.* 1a, and WHITE *adj.* and *n.*[1] Special uses 7b(a).

2.

a. The plant *Piper nigrum* (family Piperaceae), a climbing shrub indigenous to South Asia and also cultivated elsewhere in the tropics, which has alternate stalked entire leaves, with pendulous spikes of small green flowers opposite the leaves, succeeded by small berries turning red when ripe. Also more widely: any plant of the genus *Piper* or the family Piperaceae.

b. Usu. with distinguishing word: any of numerous plants of other families having hot pungent fruits or leaves which resemble pepper (1a) in taste and in some cases are used as a substitute for it.

c. *U.S.* The California pepper tree, *Schinus molle*. Cf. PEPPER TREE *n.*

3. Any of various forms of capsicum, esp. *Capsicum annuum* var. *annuum*. Originally (chiefly with distinguishing word): any variety of the *C. annuum* Longum group, with elongated fruits having a hot, pungent taste, the source of cayenne, chilli powder, paprika, etc., or of the perennial *C. frutescens*, the source of Tabasco sauce. Now frequently (more fully ***sweet pepper***): any variety of the *C. annuum* Grossum group, with large, bell-shaped or apple-shaped, mild-flavoured fruits, usually ripening to red, orange, or yellow and eaten raw in salads or cooked as a vegetable. Also: the fruit of any of these capsicums.

Sweet peppers are often used in their green immature state (more fully ***green pepper***), but some new varieties remain green when ripe.

Lemma pepper

Sense 1: spice from pepper plant

Sense 2: the pepper plant itself

Sense 3: another similar plant (Jamaican pepper)

Sense 4: another plant with peppercorns (California pepper)

Sense 5: *capsicum* (i.e. chili, paprika, bell pepper, etc)

A sense or “concept” is the meaning component of a word

There are relations between senses

Relation: Synonymy

Synonyms have the same meaning in some or all contexts.

- filbert / hazelnut
- couch / sofa
- big / large
- automobile / car
- vomit / throw up
- water / H_20

Relation: Synonymy

Note that there are probably no examples of perfect synonymy.

- Even if many aspects of meaning are identical
- Still may not preserve the acceptability based on notions of politeness, slang, register, genre, etc.

Relation: Synonymy?

water/H_20

big/large

brave/courageous

The Linguistic Principle of Contrast

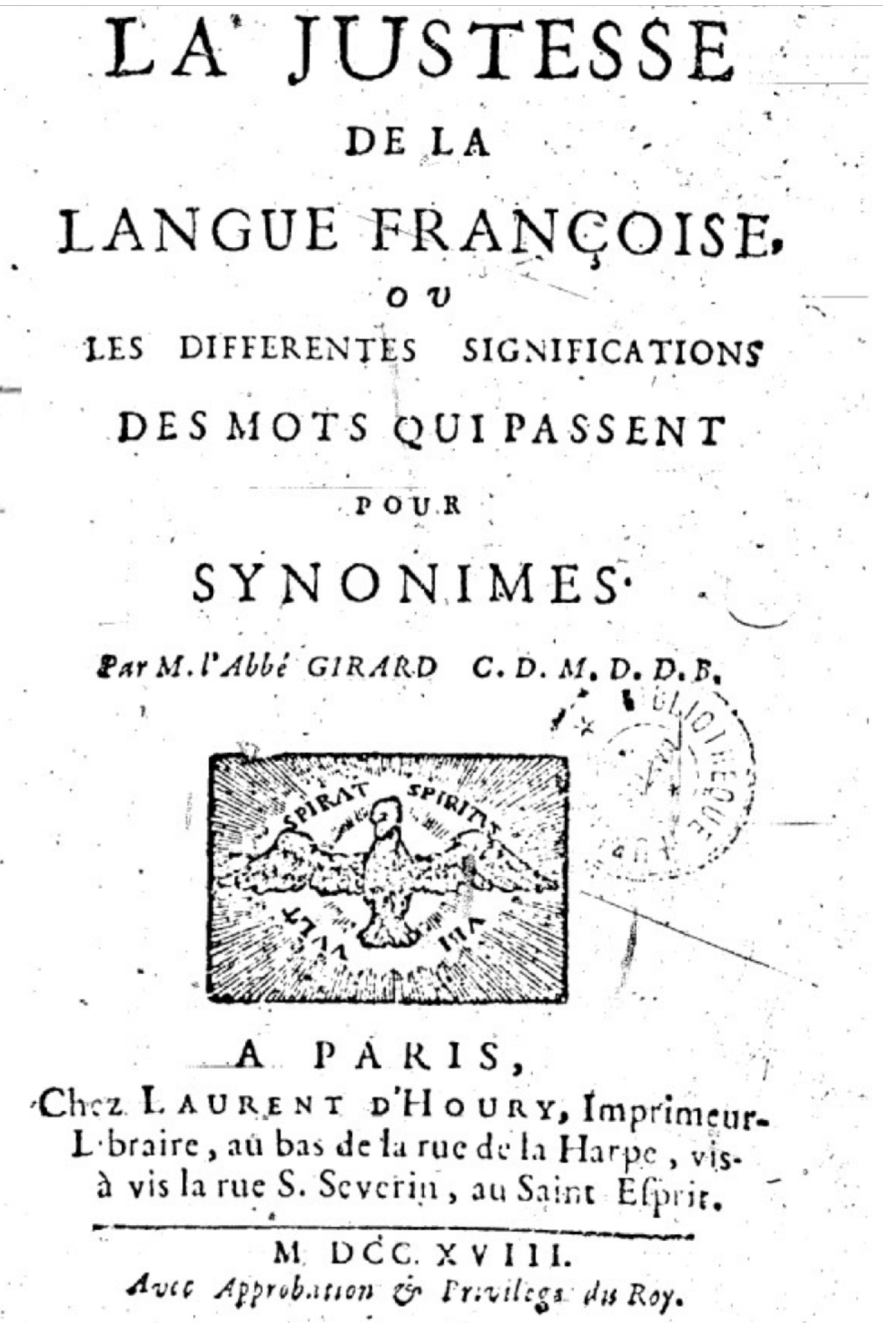
Difference in form -> difference in meaning

Abbé Gabriel Girard 1718

Re: "exact" synonyms

"je ne crois pas qu'il y ait de mot ſynonime dans aucune Langue."

[I do not believe that there is a synonymous word in any language]



LA JUSTESSE
DE LA
LANGUE FRANÇOISE,
OU
LES DIFFERENTES SIGNIFICATIONS
DES MOTS QUI PASSENT
POUR
SYNONIMES.

Par M. l'Abbé GIRARD C. D. M. D. D. B.

A PARIS,
Chez LAURENT D'HOURY, Imprimeur-Libraire, au bas de la rue de la Harpe, vis-à-vis la rue S. Severin, au Saint Eſprit.

M. DCC. XVIII.
Avec Approbation & Privilege du Roy.

Relation: Similarity

Words with similar meanings. Not synonyms, but sharing some element of meaning

`car, bicycle`

`cow, horse`

Ask humans how similar 2 words are

word1	word2	similarity
vanish	disappear	9.8
behave	obey	7.3
belief	impression	5.95
muscle	bone	3.65
modest	flexible	0.98
hole	agreement	0.3

SimLex-999 dataset (Hill et al., 2015)

Relation: Word relatedness

Also called "word association"

Words be related in any way, perhaps via a semantic frame or field

- `car, bicycle`: **similar**
- `car, gasoline`: **related**, not similar

Semantic field

Words that
- cover a particular semantic domain
- bear structured relations with each other.

hospitals

surgeon, scalpel, nurse, anaesthetic, hospital

restaurants

waiter, menu, plate, food, menu, chef

houses

door, roof, kitchen, family, bed

Relation: Antonymy

Senses that are opposites with respect to one feature of meaning

Otherwise, they are very similar!

`dark/light` `short/long` `fast/slow` `rise/fal`

`hot/cold` `up/down` `in/out`

More formally: antonyms can

- define a binary opposition
  or be at opposite ends of a scale
  - `long/short, fast/slow`
- Be *reversives*:
  - `rise/fall, up/down`

Relation: Superordinate/ subordinate

One sense is a **subordinate** of another if the first sense is more specific, denoting a subclass of the other

- *car* is a subordinate of *vehicle*
- *mango* is a subordinate of *fruit*

Conversely **superordinate**

- *vehicle* is a superordinate of *car*
- *fruit* is a subodinate of *mango*

Superordinate	vehicle	fruit	furniture
Subordinate	car	mango	chair

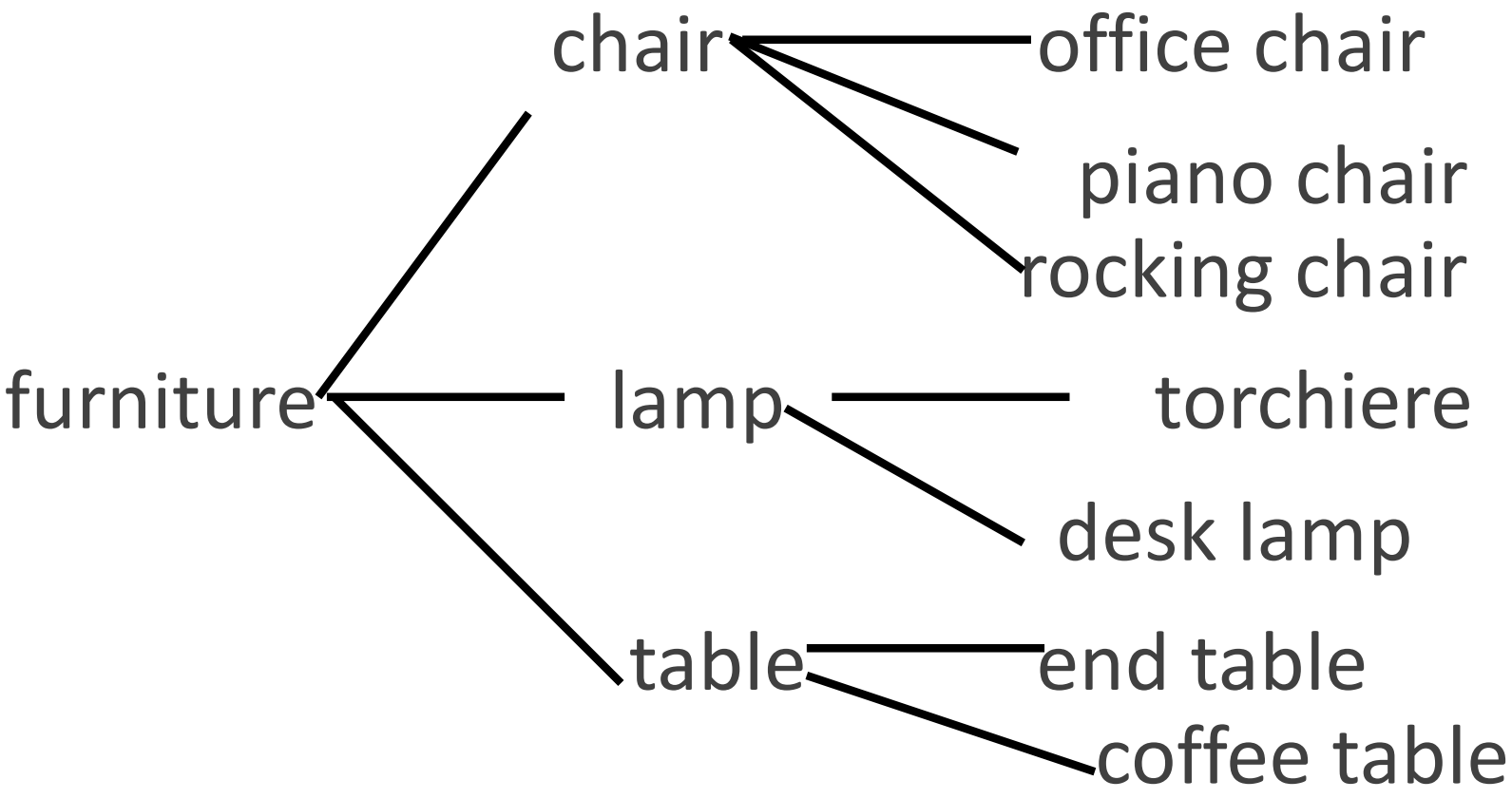
These levels are not symmetric

One level of category is distinguished from the others

The "basic level"

Name these items

Superordinate **Basic** **Subordinate**



Cluster of Interactional Properties

Basic level things are “human-sized”

Consider chairs

- We know how to interact with a chair (sit)
- Not so clear for superordinate categories like furniture
  - “Imagine a furniture without thinking of a bed/table/chair/specific basic-level category”

The basic level

distinctive actions

learned earliest in childhood

names are shortest

names are most frequent

Connotation (sentiment)

Words have **affective** meanings

positive connotations (*happy*)

negative connotations (*sad*)

positive evaluation (*great*, *love*)

negative evaluation (*terrible*, *hate*).

So far

Concepts or word senses

- Have a complex many-to-many association with **words** (homonymy, multiple senses)

Have relations with each other

- Synonymy
- Antonymy
- Similarity
- Relatedness
- Superordinate/subordinate
- Connotation

But how to define a concept?

Classical ("Aristotelian") Theory of Concepts

The meaning of a word:

a concept defined by **necessary** and **sufficient** conditions

A **necessary** condition for being an X is a condition C that X must satisfy in order for it to be an X.

- If not C, then not X
- "Having four sides" is necessary to be a square.

A **sufficient** condition for being an X is condition such that if something satisfies condition C, then it must be an X.

- If and only if C, then X
- The following necessary conditions, jointly, are sufficient to be a square
  - x has (exactly) four sides
  - each of x's sides is straight
  - x is a closed figure
  - x lies in a plane
  - each of x's sides is equal in length to each of the others
  - each of x's interior angles is equal to the others (right angles)
  - the sides of x are joined at their ends

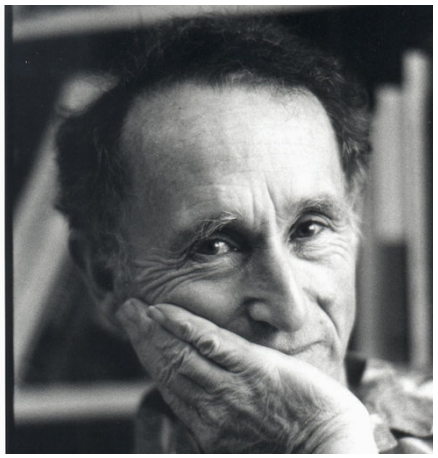
Example from Norman Swartz, SFU

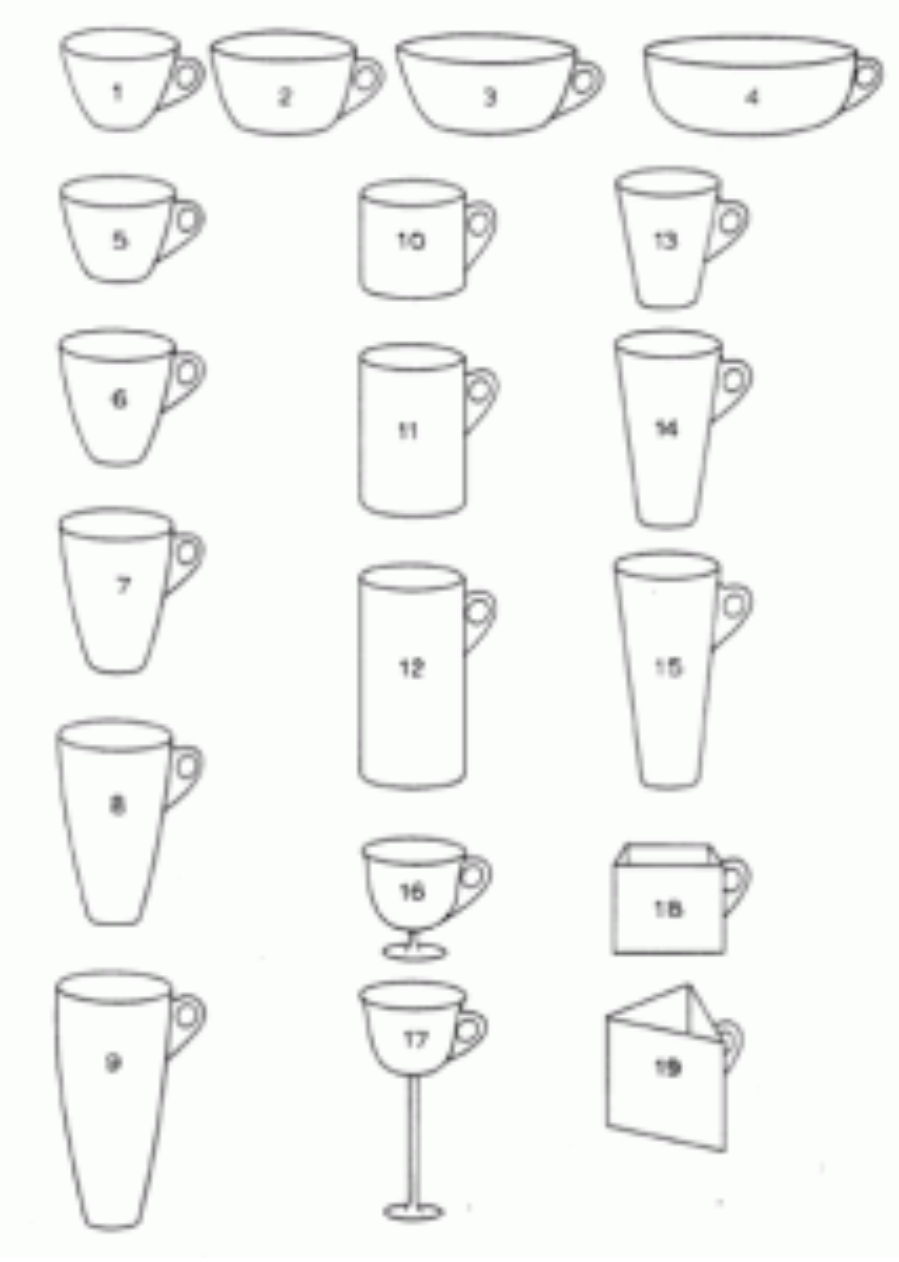
Problem 1: The features are complex and may be context-dependent

William Labov. 1975

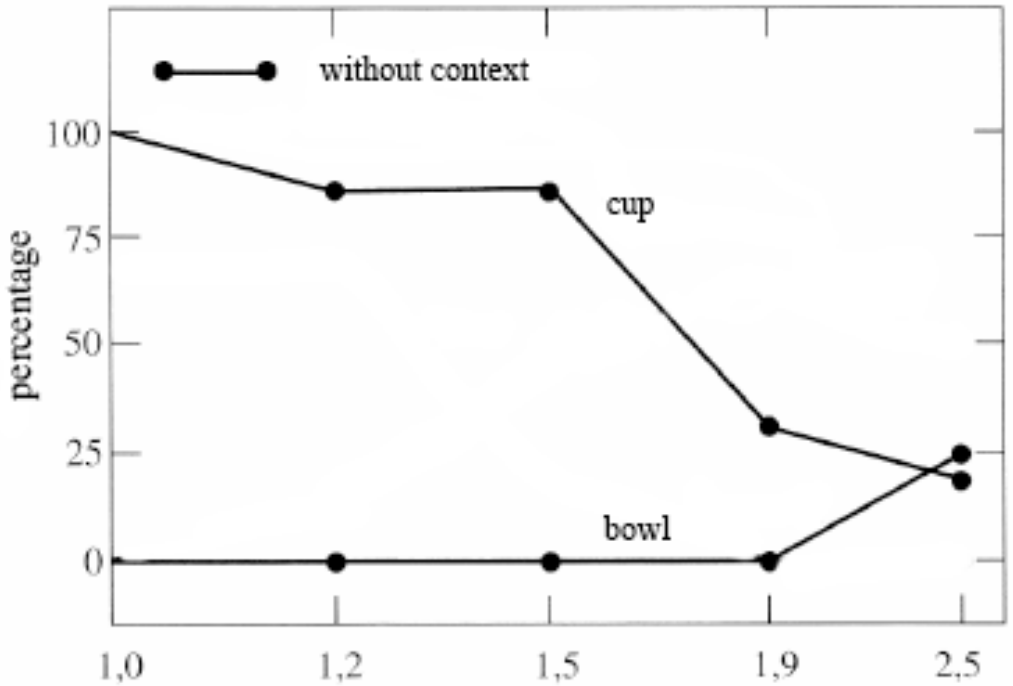
What are these?

Cup or bowl?

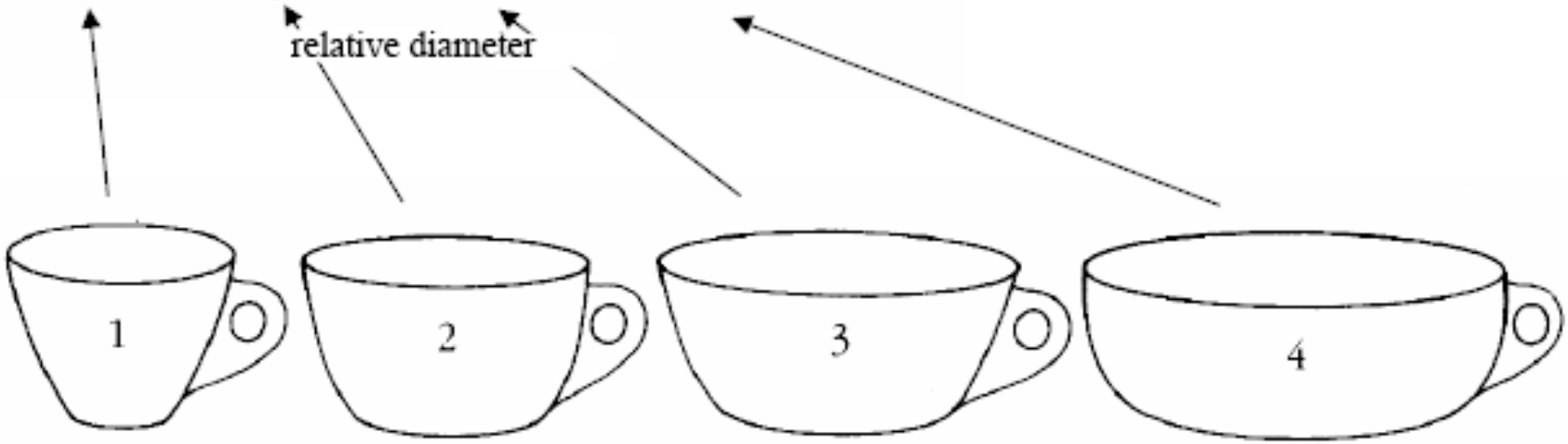




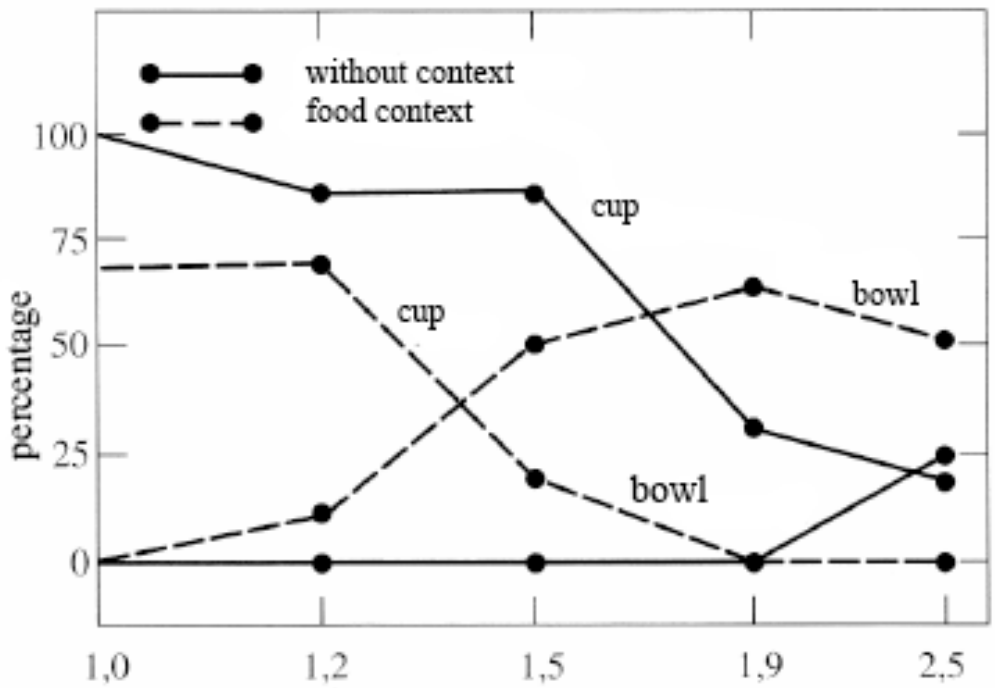
The category depends on complex features of the object (diameter, etc)



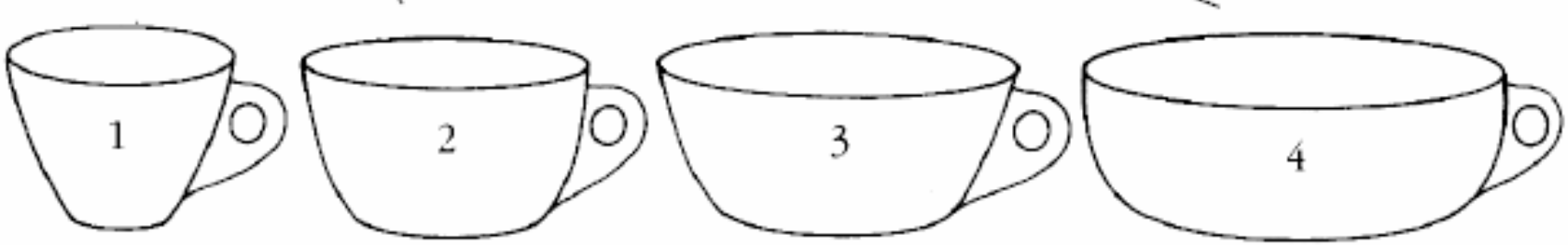
Where does the category „cup“ end?



The category depends on the context! (If there is food in it, it's a bowl)



Boundaries between cups and bowls are context sensitive



Labov's definition of cup

The term *cup* is used to denote round containers with a ratio of depth to width of $1 \pm r$ where $r \leq r_b$, and $r_b = \alpha_1 + \alpha_2 + \ldots \alpha_\upsilon$ and α_1 is a positive quality when the feature i is present and 0 otherwise.

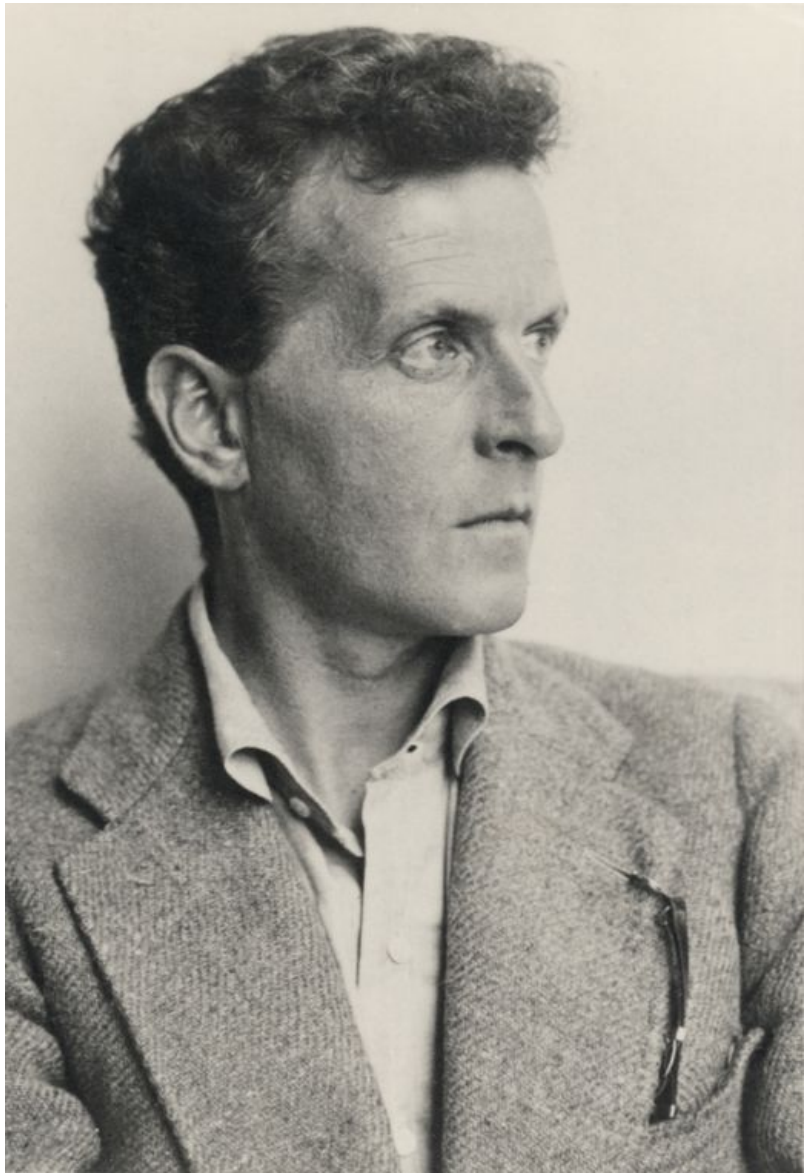
feature 1 = with one handle
2 = made of opaque vitreous material
3 = used for consumption of food
4 = used for the consumption of liquid food
5 = used for consumption of hot liquid food
6 = with a saucer
7 = tapering
8 = circular in cross-section

Cup is used variably to denote such containers with ratios width to depth $1 \pm r$ where $r_b \leq r \leq r_1$ with a probability of $r_1 - r/r_t - r_b$. The quantity $1 \pm r_b$ expresses the distance from the modal value of width to height.

Ludwig Wittgenstein (1889-1951)

Philosopher of language

In his late years, a proponent of studying "ordinary language"



Wittgenstein (1945) *Philosophical Investigations.* Paragraphs 66,67

66. Consider for example the proceedings that we call "games". I mean board-games, card-games, ball-games, Olympic games, and so on. What is common to them all?—Don't say: "There *must* be something common, or they would not be called 'games' "—but *look and see* whether there is anything common to all.—For if you look at them you will not see something that is common to *all*, but similarities, relationships, and a whole series of them at that. To repeat: don't think, but look!—Look for example at board-games, with their multifarious relationships. Now pass to card-games; here you find many correspondences with the first group, but many common features drop out, and others appear. When we pass next to ball-games, much that is common is retained, but much is lost.—Are they all 'amusing'? Compare chess with noughts and crosses. Or is there always winning and losing, or competition between players? Think of patience. In ball games there is winning and losing; but when a child throws his ball at the wall and catches it again, this feature has disappeared. Look at the parts played by skill and luck; and at the difference between skill in chess and skill in tennis. Think now of games like ring-a-ring-a-roses; here is the element of amusement, but how many other characteristic features have disappeared! And we can go through the many, many other groups of games in the same way; can see how similarities crop up and disappear.

And the result of this examination is: we see a complicated network of similarities overlapping and criss-crossing: sometimes overall similarities, sometimes similarities of detail.

67. I can think of no better expression to characterize these similarities than "family resemblances"; for the various resemblances between members of a family: build, features, colour of eyes, gait, temperament, etc. etc. overlap and criss-cross in the same way.—And I shall say: 'games' form a family.

And for instance the kinds of number form a family in the same way. Why do we call something a "number"? Well, perhaps because it has a—direct—relationship with several things that have hitherto been called number; and this can be said to give it an indirect relationship to other things we call the same name. And we extend our concept of number as in spinning a thread we twist fibre on fibre. And the strength of the thread does not reside in the fact that some one fibre runs through its whole length, but in the overlapping of many fibres.

But if someone wished to say: "There is something common to all these constructions—namely the disjunction of all their common properties"—I should reply: Now you are only playing with words. One might as well say: "Something runs through the whole thread—namely the continuous overlapping of those fibres".

What is a game?

Wittgenstein's thought experiment on "What is a game":

PI #66:

"Don't say "there must be something common, or they would not be called `games'"—but *look and see* whether there is anything common to all"

Is it amusing?

Is there competition?

Is there long-term strategy?

Is skill required?

Must luck play a role?

Are there cards?

Is there a ball?

Family Resemblance

Game 1	Game 2	Game 3	Game 4
ABC	BCD	ACD	ABD

"each item has at least one, and probably several, elements in common with one or more items, but no, or few, elements are common to all items" Rosch and Mervis

How about a radically different approach?

Ludwig Wittgenstein

PI #43:

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

Let's define words by their usages

In particular, words are defined by their environments (the words around them)

Zellig Harris (1954): **If A and B have almost identical environments we say that they are synonyms**.

What does ongchoi mean?

Suppose you see these sentences:

- Ong choi is delicious **sautéed with garlic**.
- Ong choi is superb **over rice**
- Ong choi **leaves** with salty sauces

And you've also seen these:

- …spinach **sautéed with garlic over rice**
- Chard stems and **leaves** are **delicious**
- Collard greens and other **salty** leafy greens

Conclusion:

- Ongchoi is a leafy green like spinach, chard, or collard greens

Ong choi: *Ipomoea aquatica "Water Spinach"*

空心菜

kangkong

rau muống

…

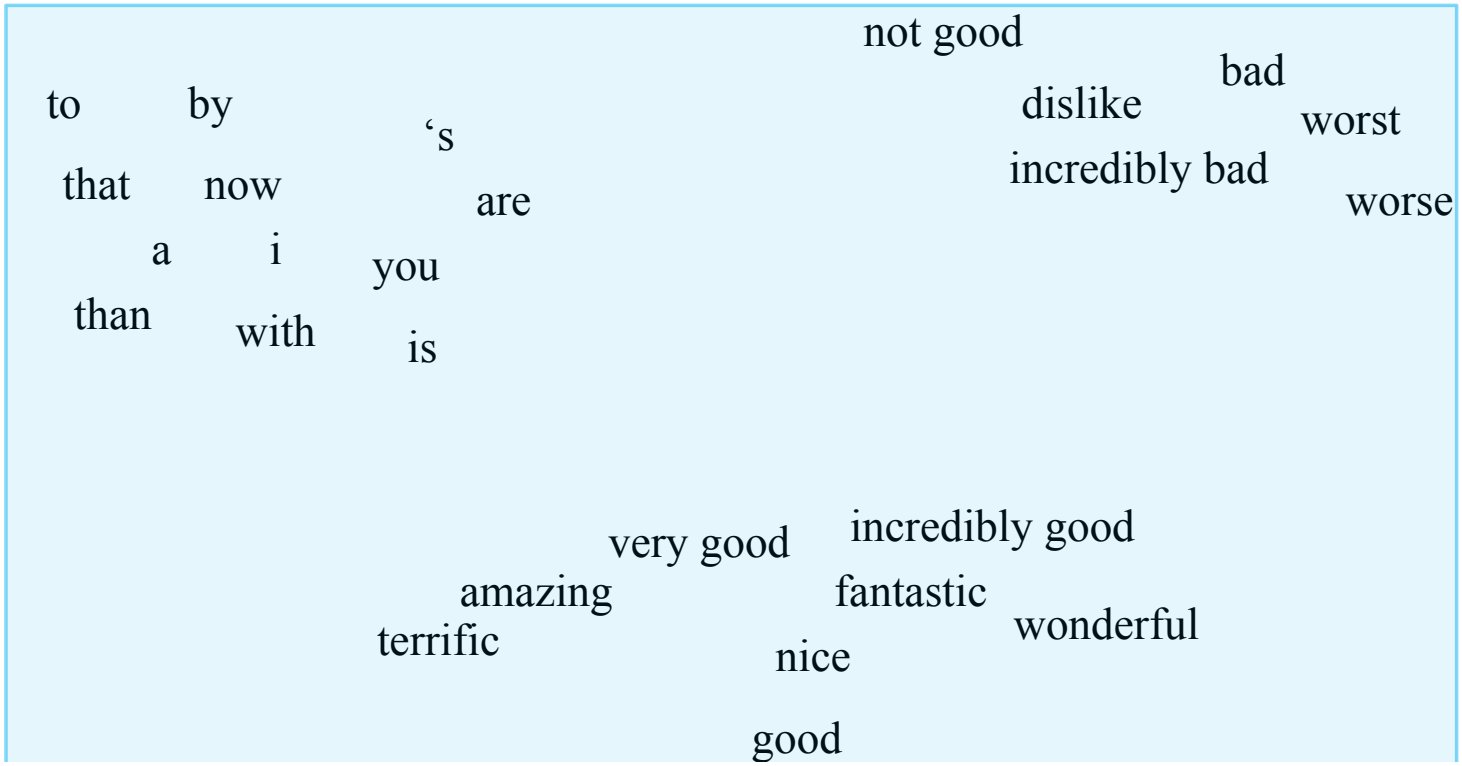
Yamaguchi, Wikimedia Commons, public domain

We'll build a new model of meaning focusing on similarity

Each word = a vector

- Not just "word" or word45.

Similar words are "nearby in space"



We define a word as a vector

Called an "embedding" because it's embedded into a space

The standard way to represent meaning in NLP

Fine-grained model of meaning for similarity

Intuition: why vectors?

Consider sentiment analysis:

- Using **words**, a feature is a word identity
  - 'The previous word was "terrible"'
  - requires **same** word to be in training and test
- With **embeddings**:
  - Feature is a word vector
  - 'The previous word was vector [35,22,17…]
  - OK if **similar** words occurred!!!

2 kinds of embeddings

tf-idf

- Information Retrieval workhorse you already know!
- A common baseline model
- **Sparse** vectors
- Words are represented by (a simple function of) the **counts** of nearby words

Word2vec

- **Dense** vectors
- Representation is created by training a classifier to **predict** whether a word is likely to appear nearby

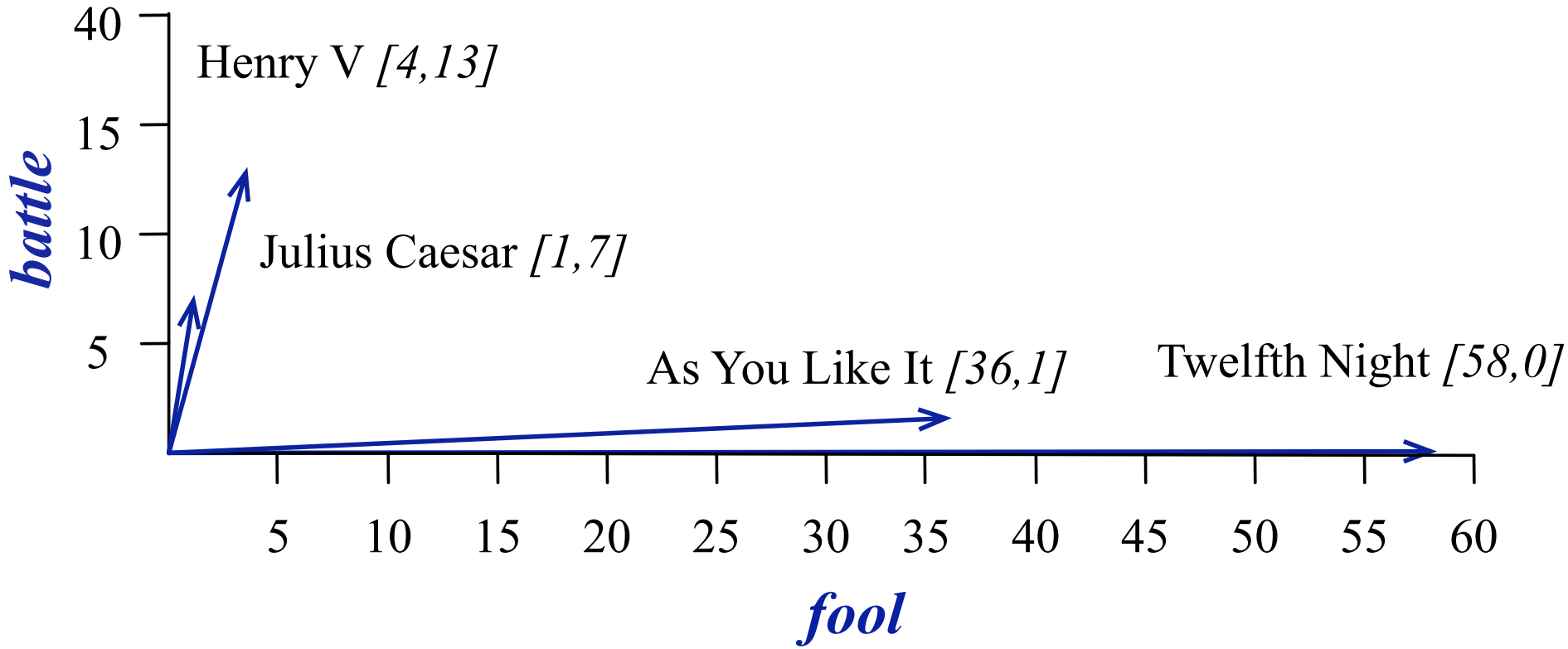
Review: words, vectors, and co-occurrence matrices

Term-document matrix

Each document is represented by a vector of words

	As You Like It	Twelfth Night	Julius Caesar	Henry V
battle	1	0	7	13
good	114	80	62	89
fool	36	58	1	4
wit	20	15	2	3

Visualizing document vectors



Vectors are the basis of information retrieval

	As You Like It	Twelfth Night	Julius Caesar	Henry V
battle	1	0	7	13
good	114	80	62	89
fool	36	58	1	4
wit	20	15	2	3

Vectors are similar for the two comedies
Different than the history

Comedies have more *fools* and *wit* and fewer *battles*.

New idea for word meaning: Words can be vectors too!!!

	As You Like It	Twelfth Night	Julius Caesar	Henry V
battle	1	0	7	13
good	114	80	62	89
fool	36	58	1	4
wit	20	15	2	3

battle is "the kind of word that occurs in Julius Caesar and Henry V"

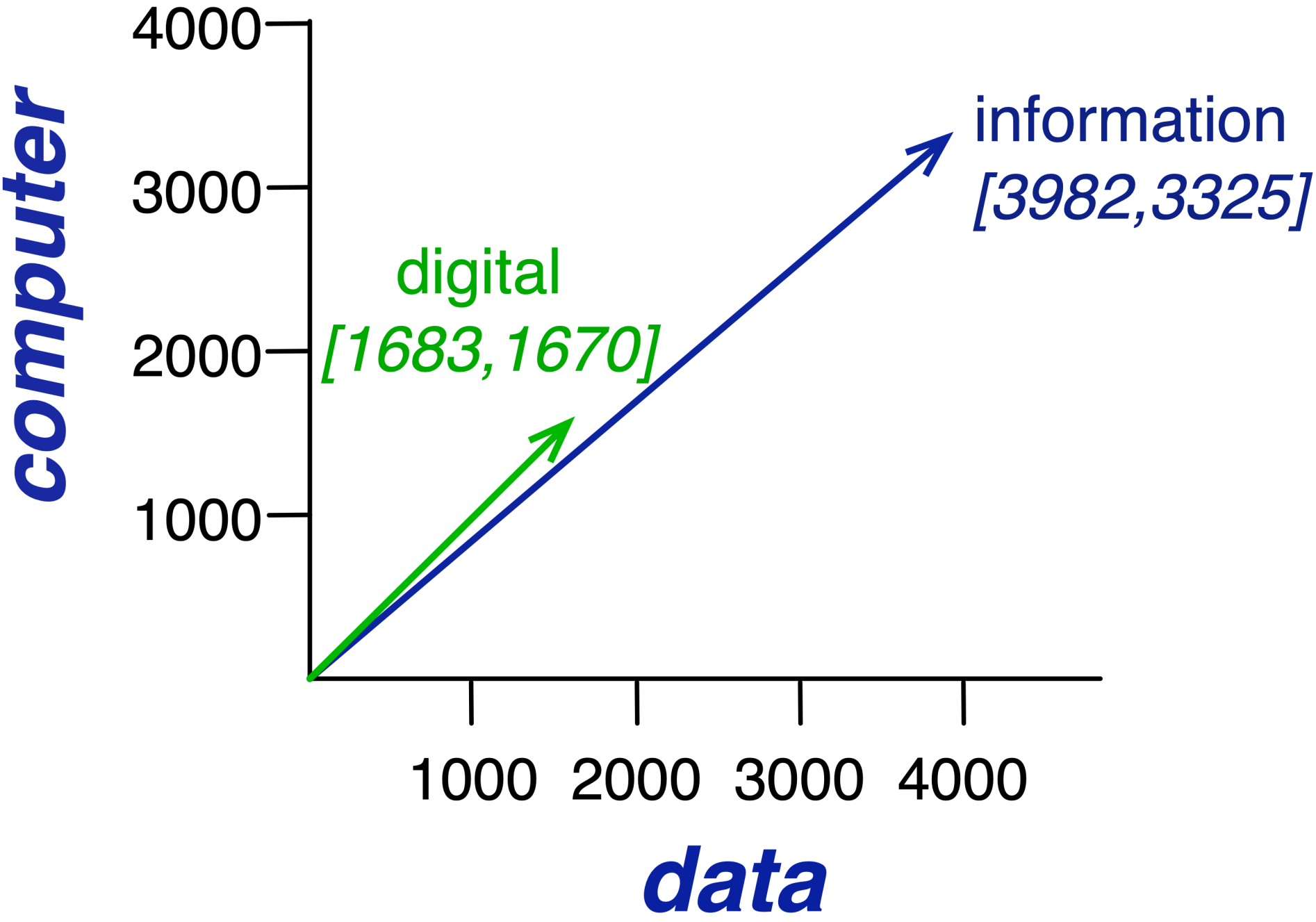
fool is "the kind of word that occurs in comedies, especially Twelfth Night"

More common: word-word matrix (or "term-context matrix")

Two **words** are similar in meaning if their context vectors are similar

is traditionally followed by	**cherry**	pie, a traditional dessert
often mixed, such as	**strawberry**	rhubarb pie. Apple pie
computer peripherals and personal	**digital**	assistants. These devices
a computer. This includes	**information**	available on the internet

	aardvark	...	computer	data	result	pie	sugar	...
cherry	0	...	2	8	9	442	25	
strawberry	0	...	0	0	1	60	19	
digital	0	...	1670	1683	85	5	4	
information	0	...	3325	3982	378	5	13	


4000
3000
2000
1000
computer
information
[3982,3325]
digital
[1683,1670]
1000
2000
3000
4000
data

Cosine for computing word similarity

$$\text{cosine}(\vec{v},\vec{w}) = \frac{\vec{v}\cdot\vec{w}}{|\vec{v}||\vec{w}|} = \frac{\sum_{i=1}^{N} v_i w_i}{\sqrt{\sum_{i=1}^{N} v_i^2}\sqrt{\sum_{i=1}^{N} w_i^2}}$$

v_i is the count for word v in context i

w_i is the count for word w in context i.

Cos(v,w) is the cosine similarity of v and w

$$\vec{a}\cdot\vec{b} = |\vec{a}||\vec{b}|\cos\theta$$

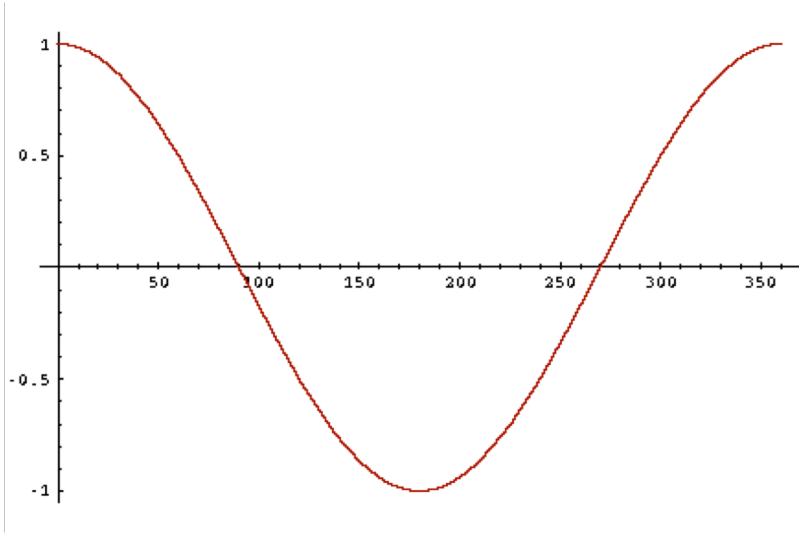
$$\frac{\vec{a}\cdot\vec{b}}{|\vec{a}||\vec{b}|} = \cos\theta$$

Cosine as a similarity metric

-1: vectors point in opposite directions

+1: vectors point in same directions

0: vectors are orthogonal



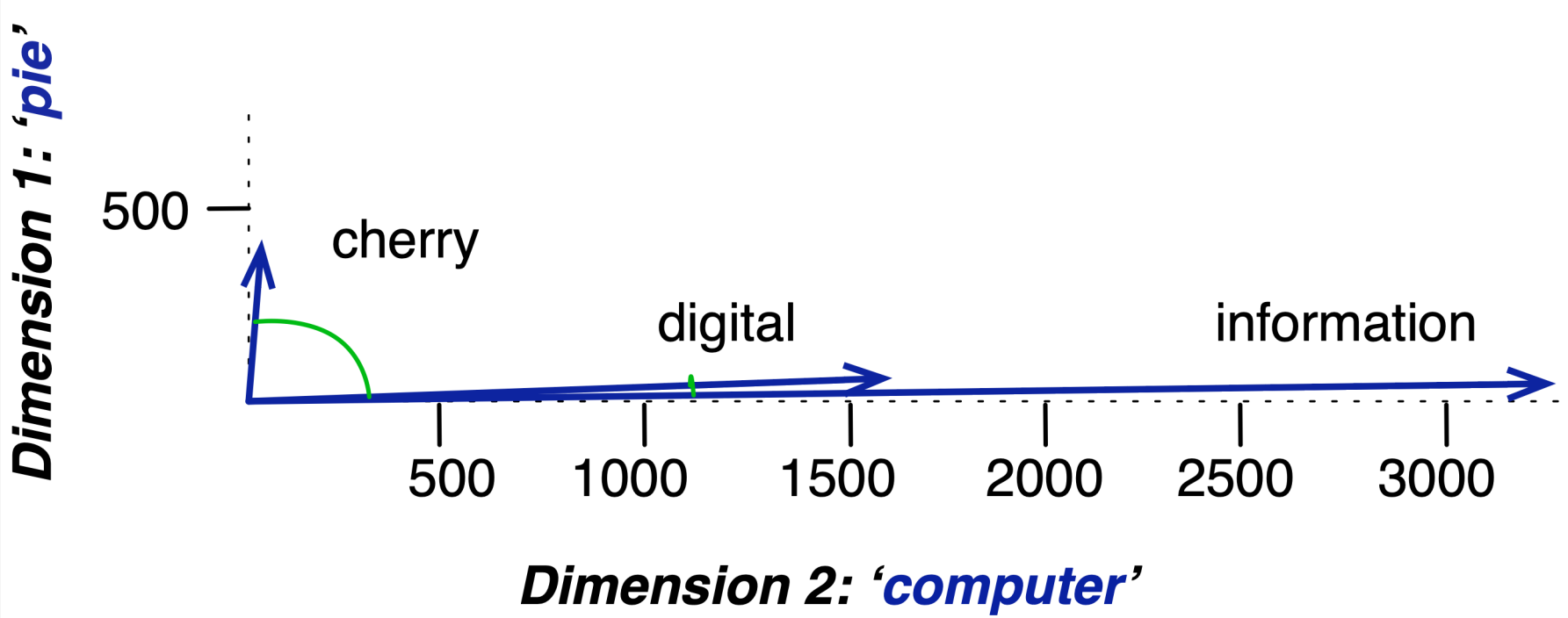
	pie	data	computer
cherry	442	8	2
digital	5	1683	1670
information	5	3982	3325

$$\cos(\vec{v},\vec{w}) = \frac{\vec{v}\bullet\vec{w}}{|\vec{v}||\vec{w}|} = \frac{\vec{v}}{|\vec{v}|}\bullet\frac{\vec{w}}{|\vec{w}|} = \frac{\sum_{i=1}^{N} v_i w_i}{\sqrt{\sum_{i=1}^{N} v_i^2}\sqrt{\sum_{i=1}^{N} w_i^2}}$$

$$\cos(\text{cherry},\text{information}) = \frac{442*5+8*3982+2*3325}{\sqrt{442^2+8^2+2^2}\sqrt{5^2+3982^2+3325^2}} = .017$$

$$\cos(\text{digital},\text{information}) = \frac{5*5+1683*3982+1670*3325}{\sqrt{5^2+1683^2+1670^2}\sqrt{5^2+3982^2+3325^2}} = .996$$

Visualizing cosines (well, angles)



But raw frequency is a bad representation

- Frequency is clearly useful; if *sugar* appears a lot near *apricot*, that's useful information.
- But overly frequent words like *the*, *it,* or *they* are not very informative about the context
- Need a function that resolves this frequency paradox!

Many functions can help in reweighting the counts

tf-idf:

tf-idf value for word t in document d:

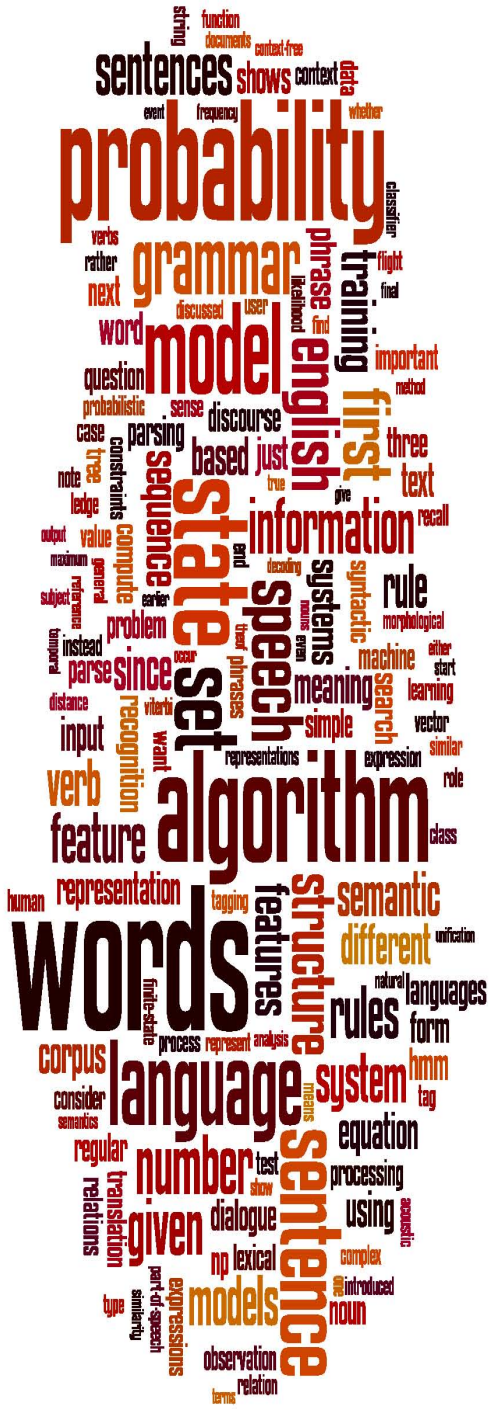
$$w_{t,d} = \mathrm{tf}_{t,d} \times \mathrm{idf}_t$$

Words like "the" or "good" have very low idf

Pointwise mutual information

- $\mathrm{PMI}(w_1, w_2) = log \frac{p(w_1,w_2)}{p(w_1)p(w_2)}$

See if words like "good" appear more often with "great" than we would expect by chance



Vector Semantics

Dense Vectors

Sparse versus dense vectors

- tf-idf vectors are
  - **long** (length $|V| = 20{,}000$ to $50{,}000$)
  - **sparse** (most elements are zero)
- Alternative: learn vectors which are
  - **short** (length 50-1000)
  - **dense** (most elements are non-zero)

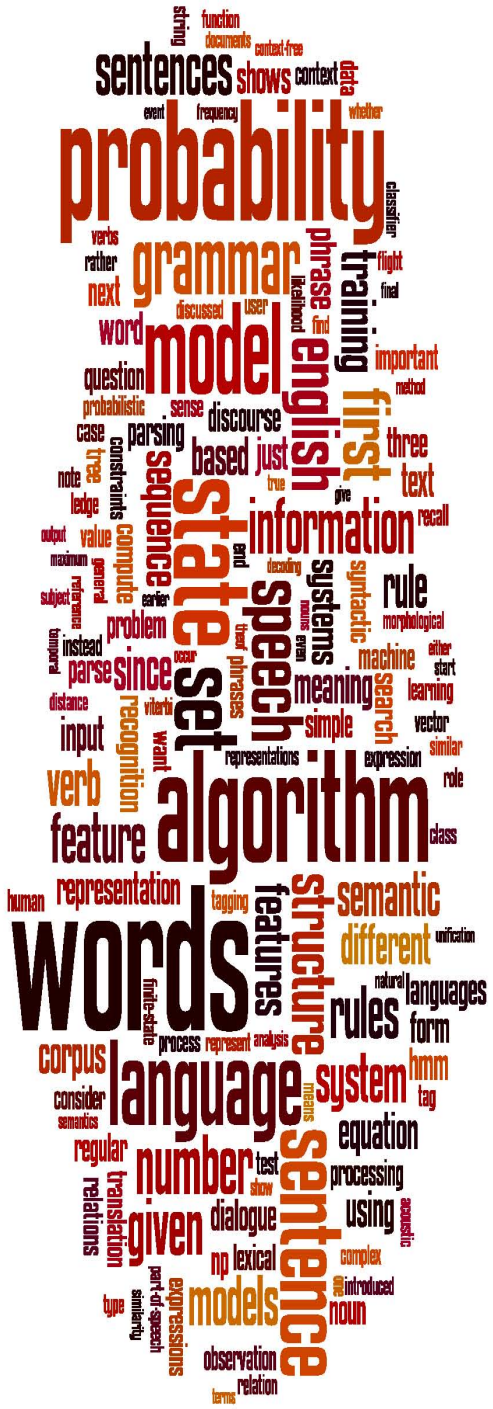
Sparse versus dense vectors

Why dense vectors?

- Short vectors may be easier to use as **features** in machine learning (less weights to tune)
- Dense vectors may **generalize** better than storing explicit counts
- They may do better at capturing synonymy:
  - *car* and *automobile* are synonyms; but are distinct dimensions
    - a word with *car* as a neighbor and a word with *automobile* as a neighbor should be similar, but aren't
- **In practice, they work better**

Two common methods for getting short dense vectors

- "Neural Language Model"-inspired models
  - Word2vec (skip-grams, CBOW), Glove
- Singular Value Decomposition (SVD)
  - A special case of this is called LSA – Latent Semantic Analysis



Vector Semantics

Embeddings inspired by neural language models: word2vec

Embeddings you can download!

- Word2vec (Mikolov et al)
- https://code.google.com/archive/p/word2vec/

- Glove (Pennington, Socher, Manning)
- http://nlp.stanford.edu/projects/glove/

Word2vec

- Popular embedding method
- Very fast to train
- Code available on the web
- Idea: **predict** rather than **count**

Word2vec

- Instead of **counting** how often each word *w* occurs near "*apricot*"
- Train a classifier on a binary **prediction** task:
  - Is *w* likely to show up near "*apricot*"?

- We don’t actually care about this task
  - But we'll take the learned classifier weights as the word embeddings

Word2Vec: Skip-Gram Task

Word2vec provides a variety of options. Let's do

"skip-gram with negative sampling" (SGNS)

Approach: predict if candidate word *c* is a "neighbor"

- Treat the target word *t* and a neighboring context word *c* as **positive examples**.
- Randomly sample other words in the lexicon to get negative examples
- Use logistic regression to train a classifier to distinguish those two cases
- Use the regression weights as the embeddings

Skip-Gram Training Data

Assume a +/- 2 word window, given training sentence:

…lemon, a [tablespoon of apricot jam, a] pinch…

c1 c2 [target] c3 c4

Skip-Gram Training data

…lemon, a [tablespoon of apricot jam, a] pinch…

c1 c2 [target] c3 c4

positive examples +

t	c
apricot	tablespoon
apricot	of
apricot	jam
apricot	a

Skip-Gram Training data

…lemon, a [tablespoon of apricot jam, a] pinch…

c1 c2 [target] c3 c4

positive examples +

t	c
apricot	tablespoon
apricot	of
apricot	jam
apricot	a

For each positive example we'll grab k negative examples, sampling by frequency

Skip-Gram Training data

…lemon, a [tablespoon of apricot jam, a] pinch…

c1 c2 [target] c3 c4

positive examples +

t	c
apricot	tablespoon
apricot	of
apricot	jam
apricot	a

negative examples -

t	c	t	c
apricot	aardvark	apricot	seven
apricot	my	apricot	forever
apricot	where	apricot	dear
apricot	coaxial	apricot	if

Word2vec: how to learn vectors

Let's represent words as vectors of some length (say 300), randomly initialized.

So we start with 300 * V random parameters

Over the entire training set, we'd like to adjust those word vectors such that we

- **Maximize** the similarity of the target word, context word pairs (t,c) drawn from the positive data
- **Minimize** the similarity of the (t,c) pairs drawn from the negative data.

The classifier's goal (learning objective)

- **Maximize** the similarity of (t,c) pairs in +
- **Minimize** the similarity of (t,c) pairs in -

$$L(\theta) = \sum_{(t,c)\in+} \log P(+|t,c) + \sum_{(t,c)\in-} \log P(-|t,c)$$

SGNS classifier's goal (learning objective)

- **Maximize** the similarity of (t,c) pairs in +
  - maximizing P(+)
- **Minimize** the similarity of (t,c) pairs in –
  - Maximizing P(-)

$$L(\theta) = \sum_{(t,c)\in+} \log P(+|t,c) + \sum_{(t,c)\in-} \log P(-|t,c)$$

Similarity is computed from dot product

- Remember: two vectors are similar if they have a high dot product
  - Cosine is just a normalized dot product
- So:
  - Similarity(t,c) $\propto t \cdot c$
- We'll need to normalize to get a probability
  - (cosine isn't a probability either)

Turning dot products into probabilities

- $\text{Sim(t,c)} = t \cdot c$
- To turn this into a probability.
- We'll use the sigmoid from logistic regression:

$$P(+|t,c) = \frac{1}{1+e^{-sim(t,c)}}$$

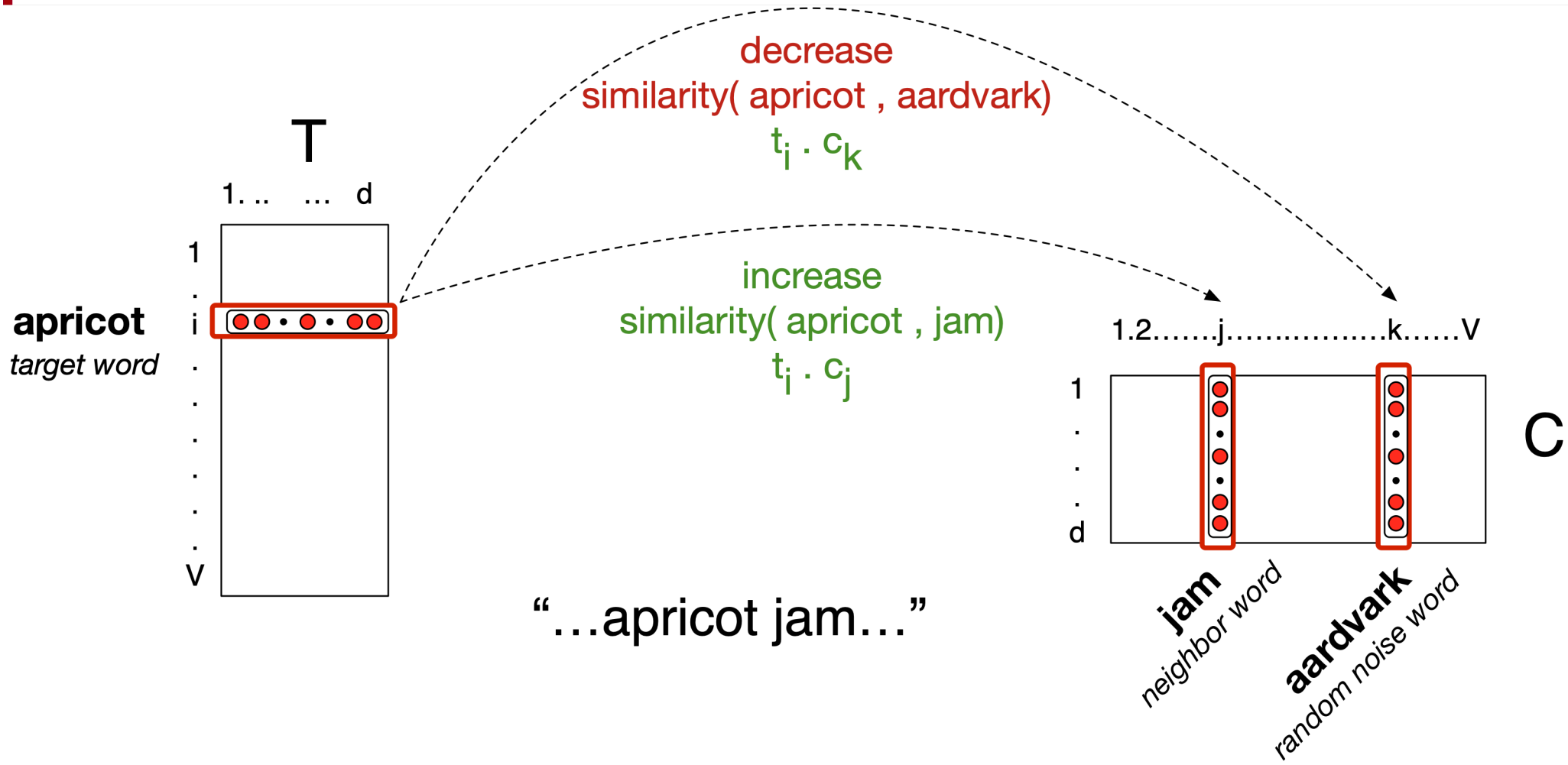
SGNS Classifier's goal

$$
\begin{aligned}
L(\theta) &= \log P(+|t,c) + \sum_{i=1}^{\kappa} \log P(-|t,n_i) \\
&= \log \sigma(c \cdot t) + \sum_{i=1}^{k} \log \sigma(-n_i \cdot t) \\
&= \log \frac{1}{1+e^{-c \cdot t}} + \sum_{i=1}^{k} \log \frac{1}{1+e^{n_i \cdot t}}
\end{aligned}
$$

Learning the classifier

- How to learn?
  - Stochastic gradient descent!

- We'll adjust the word weights to
  - make the positive pairs more likely
  - and the negative pairs less likely,
  - over the entire training set.


decrease
similarity(apricot , aardvark)
$t_i \cdot c_k$
T
1. ..
... d
1
i
V
apricot
target word
increase
similarity(apricot , jam)
$t_i \cdot c_j$
1.2.....j.....k.....V
1
d
C
jam
neighbor word
aardvark
random noise word
"…apricot jam…"

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

- Start with V random 300-dimensional vectors as initial embeddings
- Use logistic regression:
  - 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.

Or in other words

- Start with some initial embeddings (e.g., random)
- iteratively make the embeddings for a word
  - more like the embeddings of its neighbors
  - less like the embeddings of other words.

An embedding's neighbors are similar/related words!

Hogwarts nearest neighbors(C= +/- 5)

Dumbledore

Half-blood

Malfoy

Related words
In same semantic field
(Harry Potter)

Hogwarts nearest neighbors (C = +/- 2)

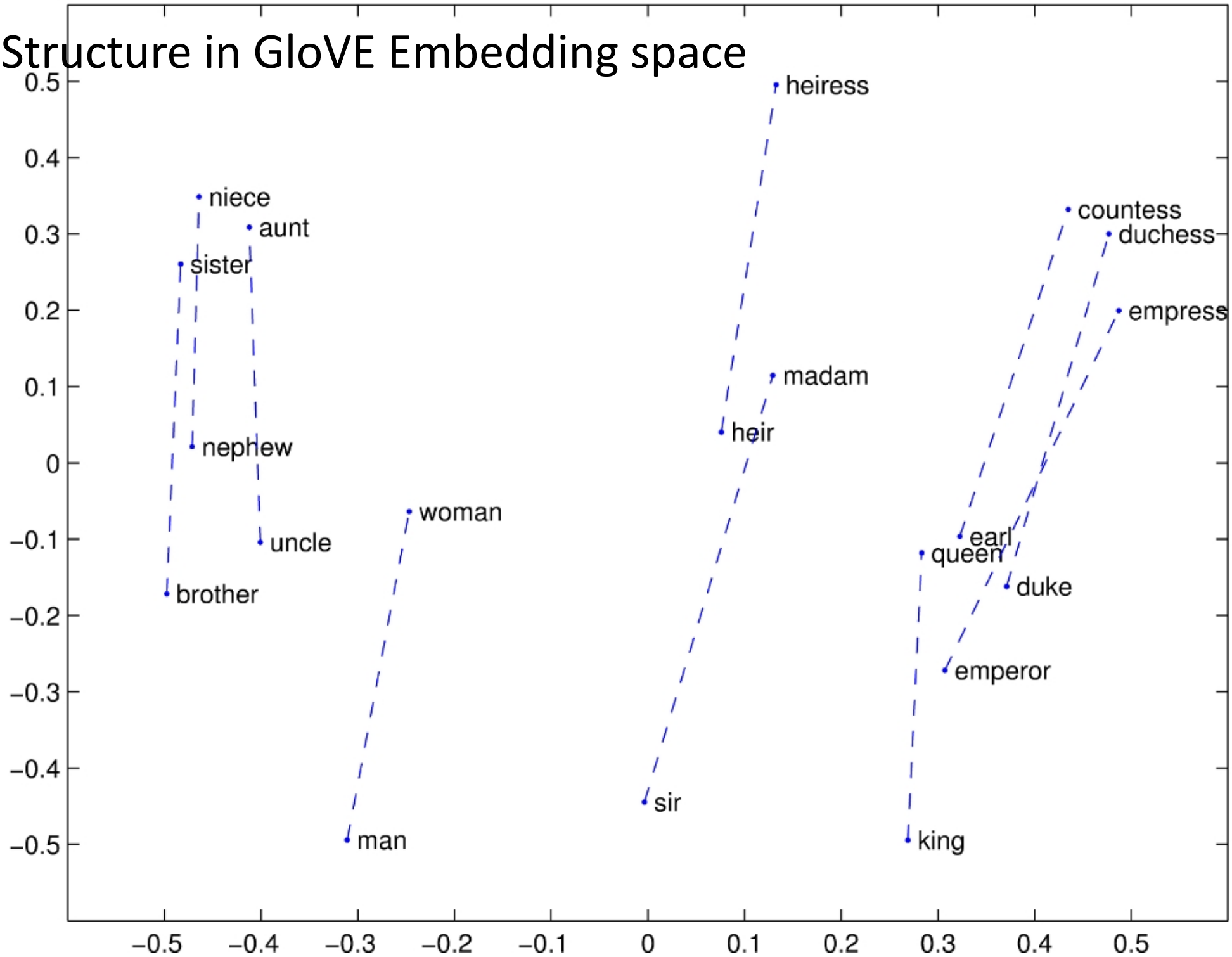
Sunnydale

Evernight

Blandings

Similar Nouns
Buildings/schools
Words in same Taxonomy

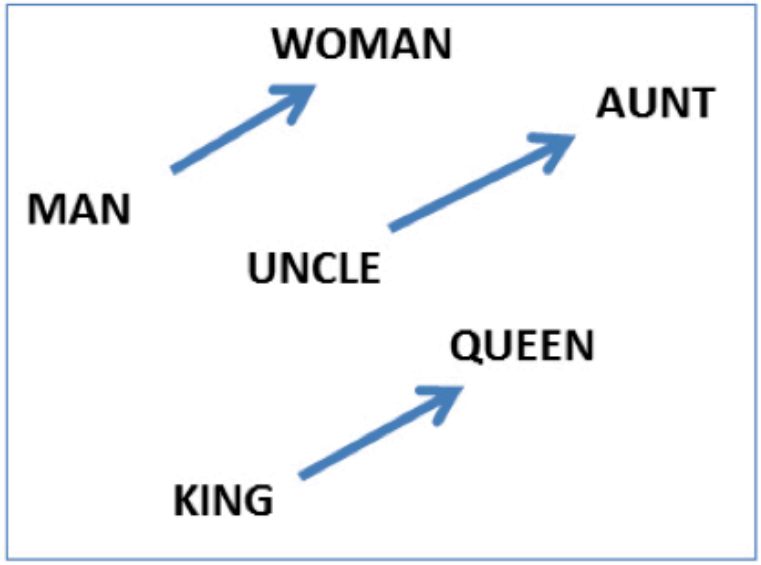
Structure in GloVE Embedding space

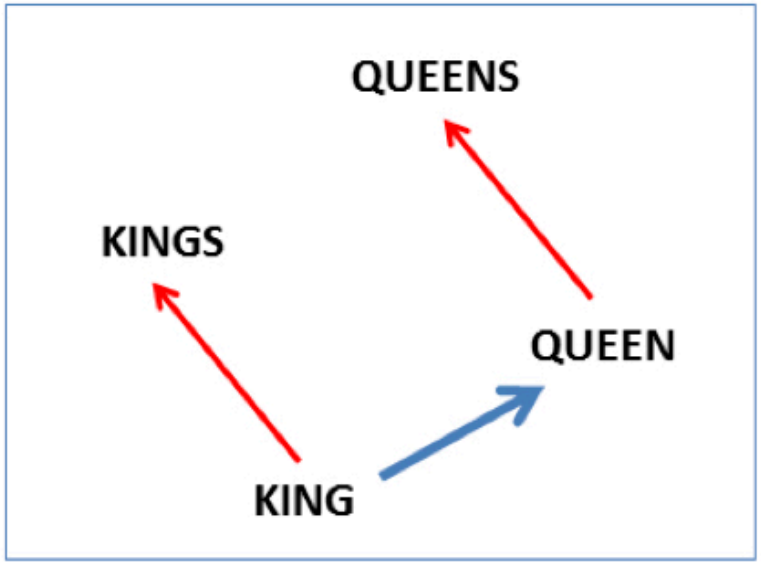


Analogies! Embeddings may capture relational meaning

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

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





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

Two things we can do about this cultural bias problem

1. Find ways to debias word embeddings
2. Use the embeddings to **study** cultural bias!

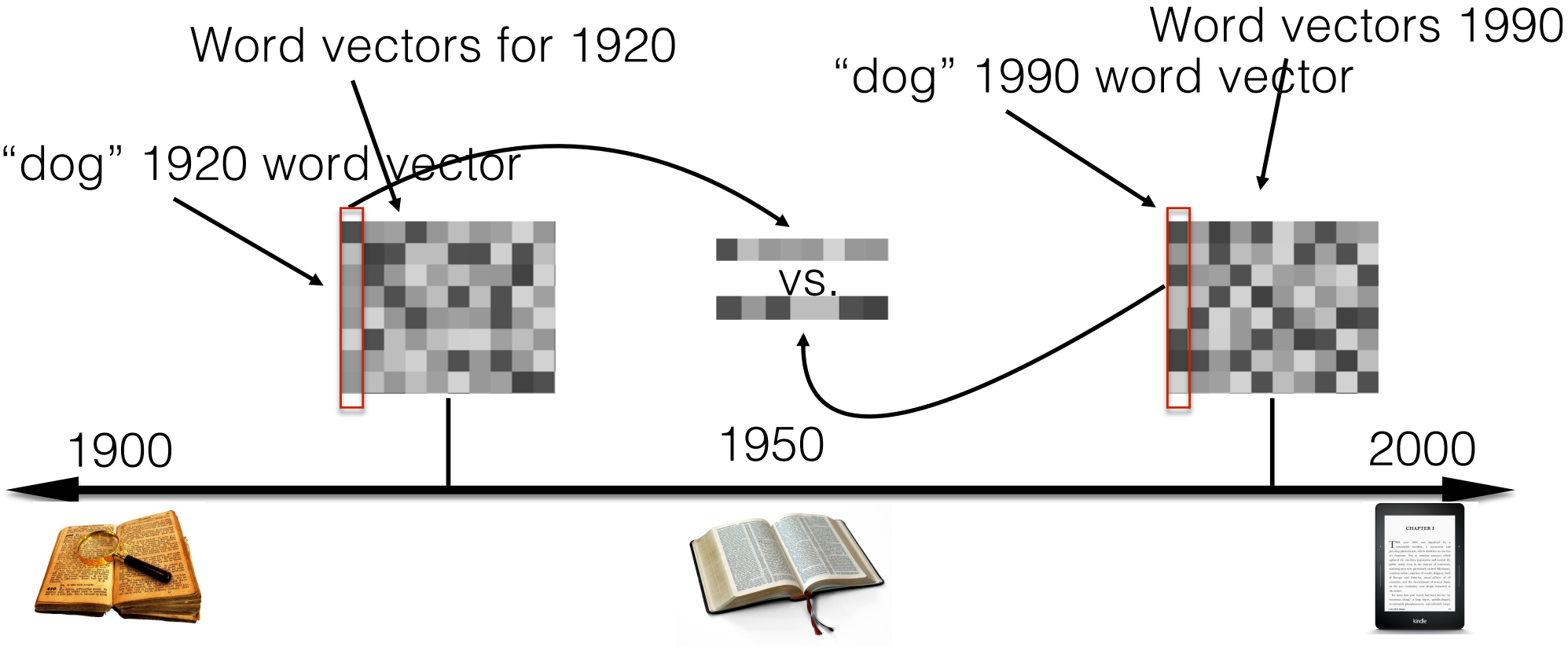
First: embeddings as a window on history

- William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change. Proceedings of ACL 2016.
- William L. Hamilton, Jure Leskovec, Dan Jurafsky. 2016. Cultural Shift or Linguistic Drift? Comparing Two Computational Models of Semantic Change. Proceedings of EMNLP 2016.
- William L. Hamilton, Kevin Clark, Jure Leskovec, Dan Jurafsky. 2016. Inducing Domain-Specific Sentiment Lexicons from Unlabeled Corpora. Proceedings of EMNLP 2016.

Will Hamilton

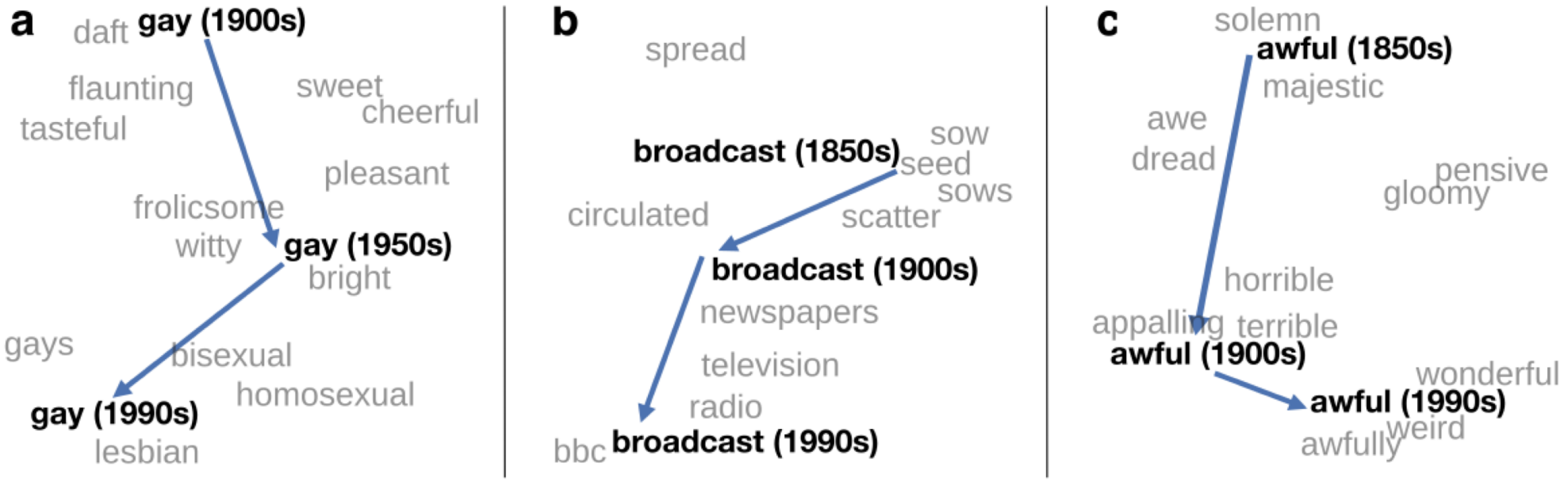
Jure Leskovec

Train embeddings on old books to study changes in word meaning



Embeddings over time: meaning shift

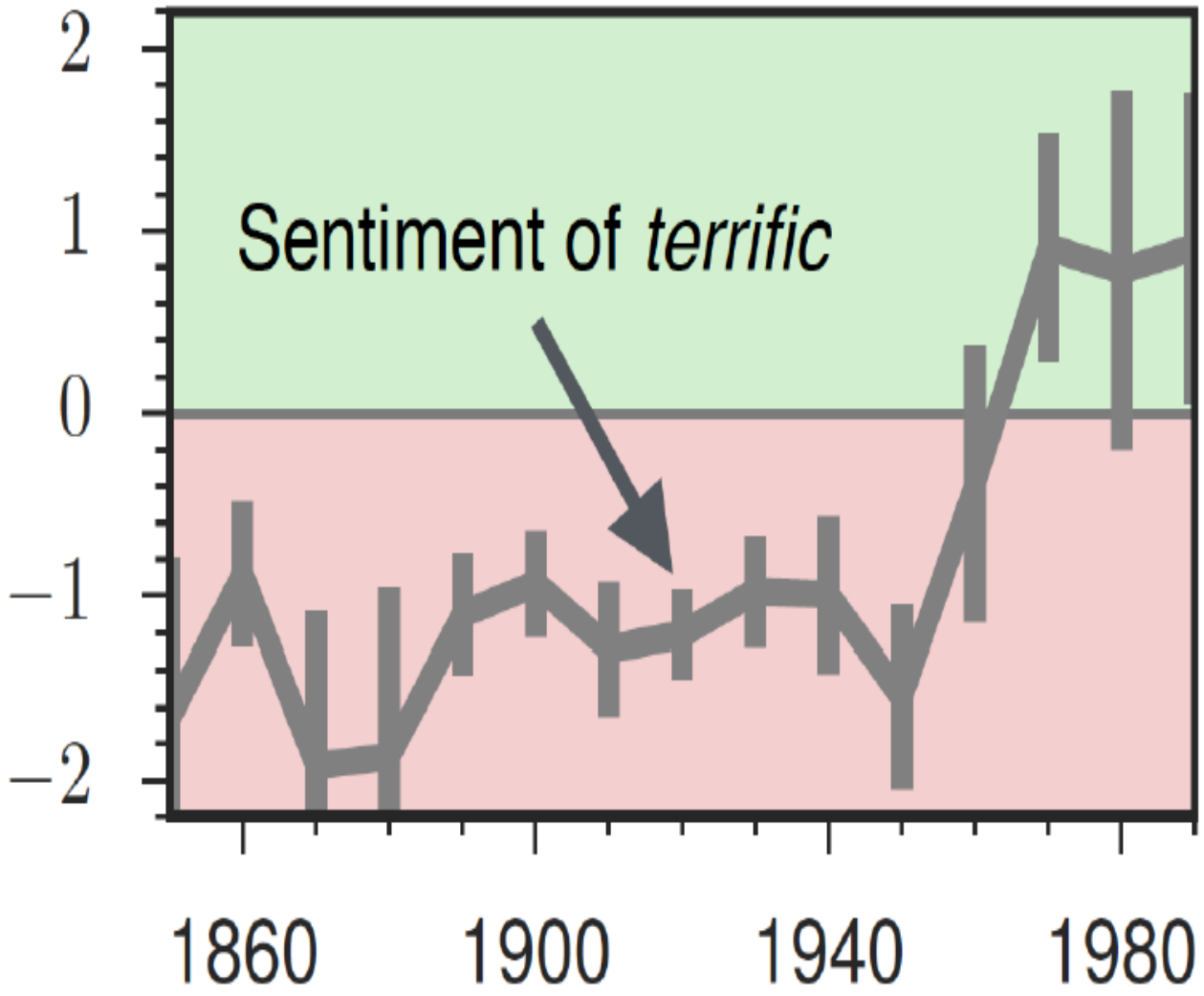
Project 300 dimensions down into 2
("I see well in many dimensions as long as the dimensions are around two." Martin Shubek)



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

The evolution of sentiment words

Negative words change faster than positive words



Use historical embedding as a tool to investigate cultural biases

PNAS

Word embeddings quantify 100 years of gender and ethnic stereotypes

Nikhil Garg[a,1], Londa Schiebinger[b], Dan Jurafsky[c,d], and James Zou[e,f,1]

[a]Department of Electrical Engineering, Stanford University, Stanford, CA 94305; [b]Department of History, Stanford University, Stanford, CA 94305; [c]Department of Linguistics, Stanford University, Stanford, CA 94305; [d]Department of Computer Science, Stanford University, Stanford, CA 94305; [e]Department of Biomedical Data Science, Stanford University, Stanford, CA 94305; and [f]Chan Zuckerberg Biohub, San Francisco, CA 94158

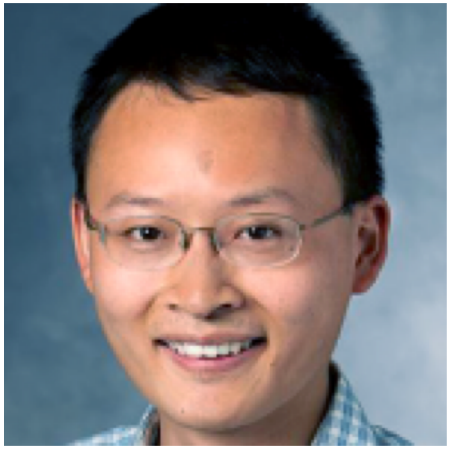
Edited by Susan T. Fiske, Princeton University, Princeton, NJ, and approved March 12, 2018 (received for review November 22, 2017)

Word embeddings are a powerful machine-learning framework that represents each English word by a vector. The geometric relationship between these vectors captures meaningful semantic relationships between the corresponding words. In this paper, we develop a framework to demonstrate how the temporal dynamics

in the large corpora of training texts (20–23). For example, the vector for the adjective honorable would be close to the vector for man, whereas the vector for submissive would be closer to woman. These stereotypes are automatically learned by the embedding algorithm and could be problematic if the embedding is then used

Nikhil Garg

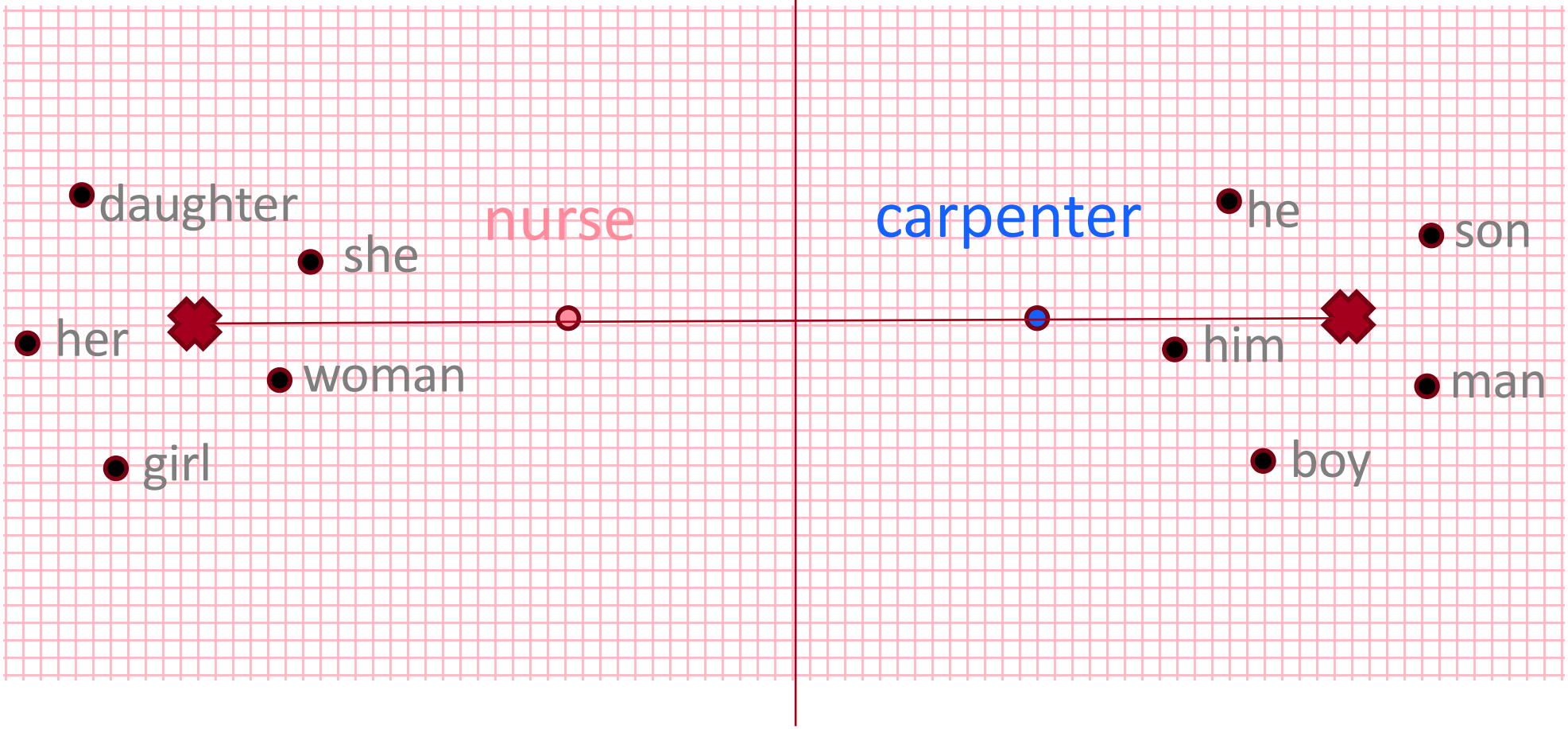
Londa Schiebinger



James Zou

Computing the gender bias of a word

How much closer a word is to "woman" synonyms than "man" synonyms?



Now use historical embedding as a tool to investigate cultural biases

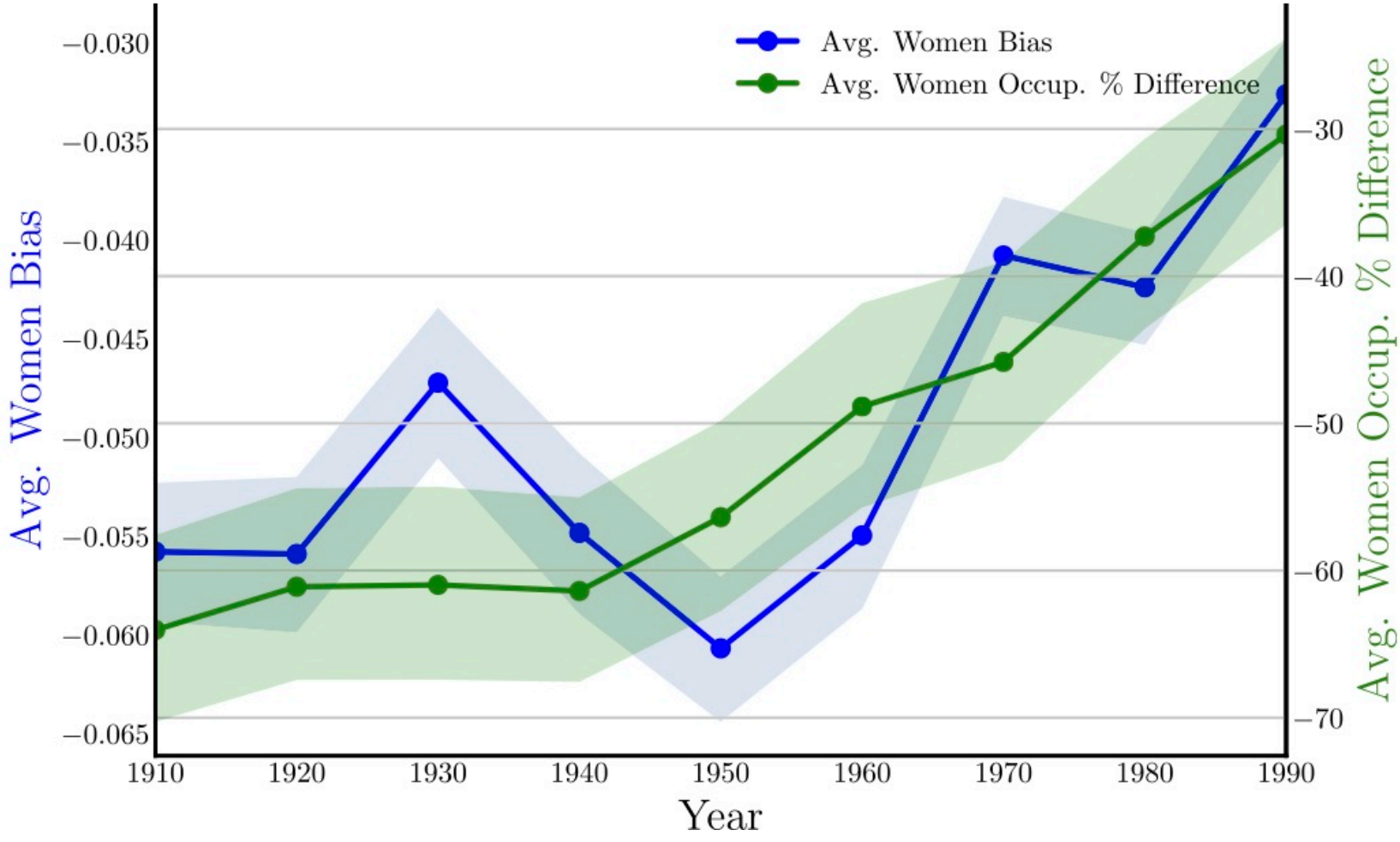
- From the historical embeddings for each decade

1. Compute historical biases of each word:
   - **Gender bias**: how much closer a word is to "woman" synonyms than "man" synonyms.
   - **Chinese bias**: how much closer a word is to Chinese names (Wang, Yang, Chen) than Anglo names
2. Correlate with occupational data from historical census
3. Look at how all these change over time

Embedding bias correlates with actual occupation data

Is "nurse" closer to "man" than "woman"?

Embeddings reflects gender bias in occupations across time (1910-1990)



Embeddings also reflect **framings** of women over time

Embeddings for **competence** adjectives are biased toward men

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

This bias is slowly decreasing 1960-1990

Embeddings reflect ethnic stereotypes over time

- Famous "Princeton Trilogy" experiments 1933/1951/1969
- Tested attitudes toward ethnic groups
  - Chinese, Japanese, Italians, Germans
  - *How "industrious, superstitious, nationalistic"*, etc.

Result:

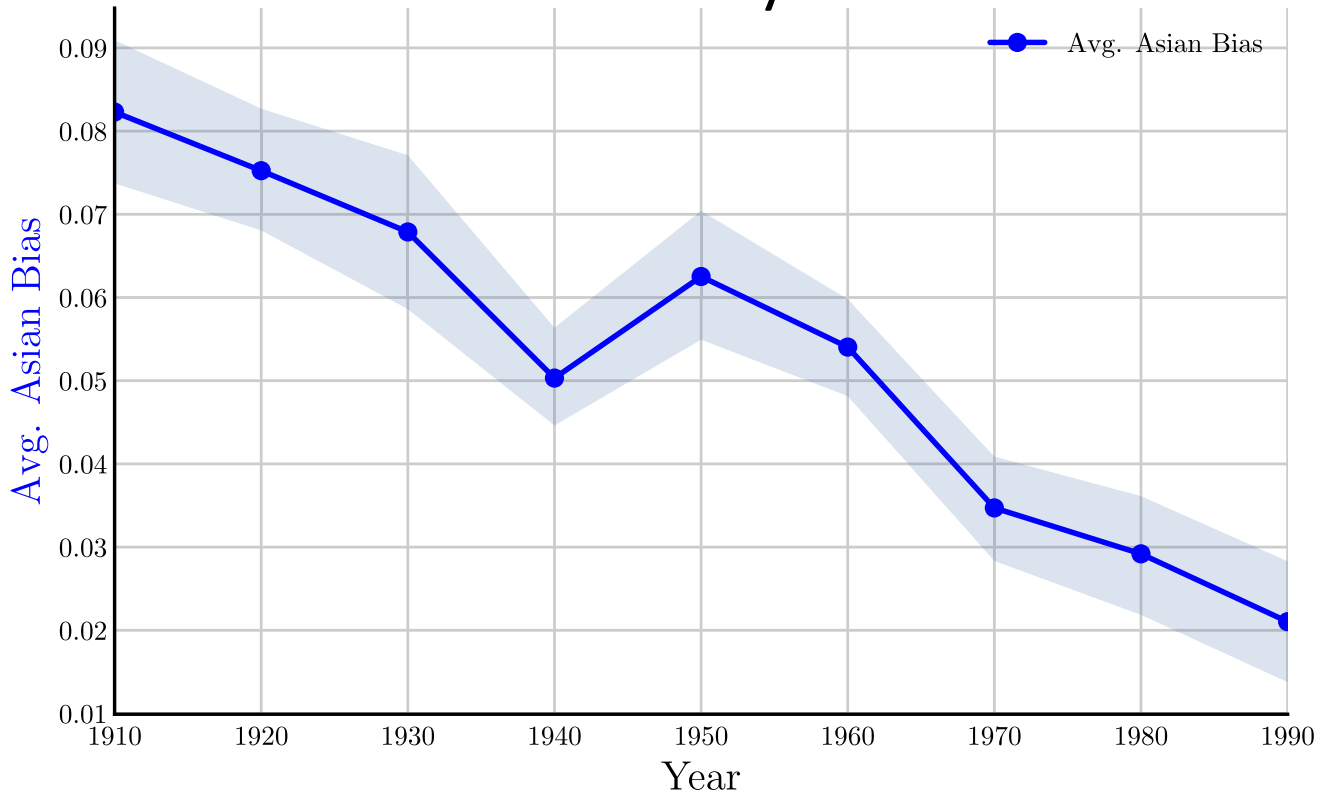
We can match these attitudes just by looking at text!

Adjectives used for Asians over time

1910

Irresponsible
Envious
Barbaric
Aggressive
Transparent
Monstrous
Hateful
Cruel
Greedy
Bizarre

Decreased "Dehumanizing" of Asians over the 20th century



We can time travel by using NLP on old texts!

We can **run social science experiments in the past** to find people's implicit attitudes!

Summary

- **Concepts** or word senses
  - Have a complex many-to-many association with **words** (homonymy, multiple senses)
  - Have relations with each other
    - Synonymy, Antonymy, Superordinate
  - But are hard to define formally (necessary & sufficient conditions)

Summary

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!!
- Useful in practice but also encode cultural stereotypes
  - Can **debias** embeddings, and use them to **study bias**

Embeddings in classes

- CS224U: Next Quarter
- CS224N: Winter 2019