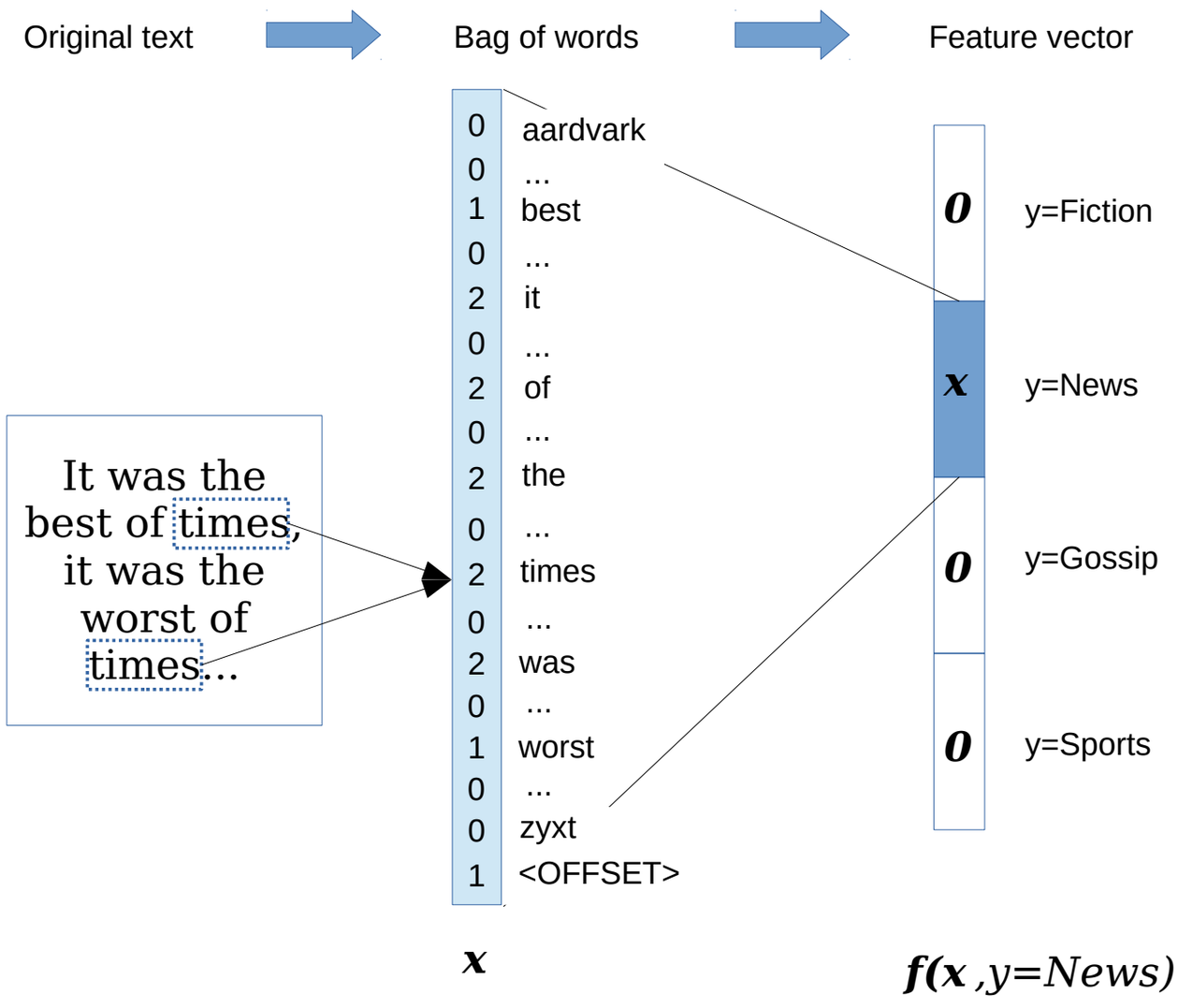
Word Embeddings

CS6120: Natural Language Processing
Northeastern University

David Smith

What Should My Inputs Look Like?



One-Hot Encoding

	time	fruit	flies	like	a	an	arrow	banana
1_{time}	1	0	0	0	0	0	0	0
1_{fruit}	0	1	0	0	0	0	0	0
1_{flies}	0	0	1	0	0	0	0	0
1_{like}	0	0	0	1	0	0	0	0
1_{a}	0	0	0	0	1	0	0	0
1_{an}	0	0	0	0	0	1	0	0
1_{arrow}	0	0	0	0	0	0	1	0
1_{banana}	0	0	0	0	0	0	0	1

Vector Space Model

- Information retrieval model developed by Salton and colleagues in 1968
- Documents and queries are both represented by vectors of term weighs
- Collection thus a matrix of term weights

$$D_i = (d_{i1}, d_{i2}, \ldots, d_{it}) \qquad Q = (q_1, q_2, \ldots, q_t)$$

	$Term_1$	$Term_2$	$\ldots$	$Term_t$
Doc_1	d_{11}	d_{12}	$\ldots$	d_{1t}
Doc_2	d_{21}	d_{22}	$\ldots$	d_{2t}
$\vdots$	$\vdots$			
Doc_n	d_{n1}	d_{n2}	$\ldots$	d_{nt}

Vector Space Model

D_1 Tropical Freshwater Aquarium Fish.
D_2 Tropical Fish, Aquarium Care, Tank Setup.
D_3 Keeping Tropical Fish and Goldfish in Aquariums, and Fish Bowls.
D_4 The Tropical Tank Homepage - Tropical Fish and Aquariums.

Terms	Documents			
	D_1	D_2	D_3	D_4
aquarium	1	1	1	1
bowl	0	0	1	0
care	0	1	0	0
fish	1	1	2	1
freshwater	1	0	0	0
goldfish	0	0	1	0
homepage	0	0	0	1
keep	0	0	1	0
setup	0	1	0	0
tank	0	1	0	1
tropical	1	1	1	2

Vector Space Model

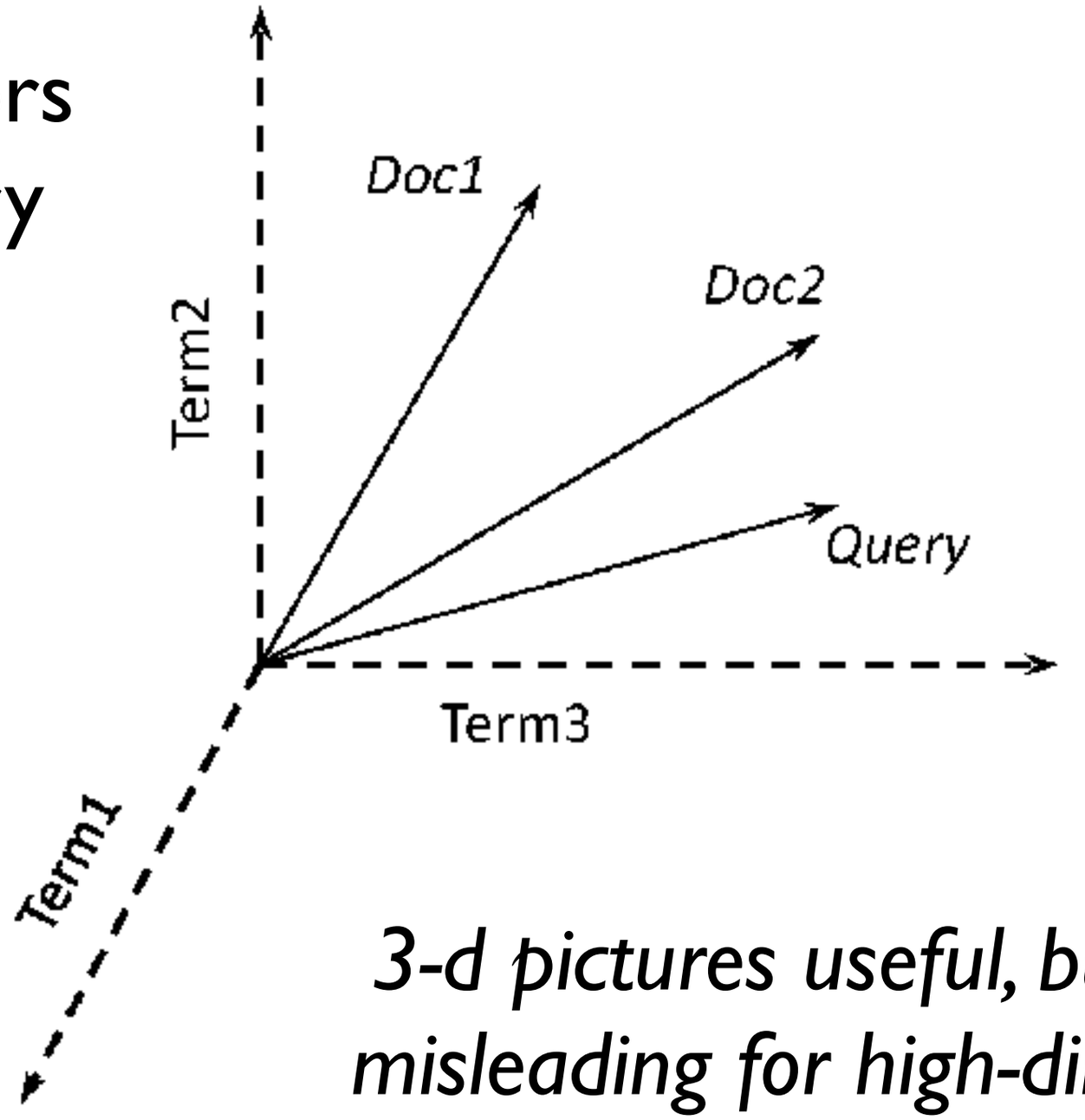
Query: tropical fish

Term	Query
aquarium	0
bowl	0
care	0
fish	**1**
freshwater	0
goldfish	0
homepage	0
keep	0
setup	0
tank	0
tropical	**1**

Usually much sparser than a document!

Vector Space Model

Retrieve vectors near the query



3-d pictures useful, but can be misleading for high-dimensional space

Vector Space Model

- Documents ranked by distance between points representing query and documents
  - *Similarity* measure more common than a distance or *dissimilarity* measure
- e.g. Cosine correlation

$$Cosine(D_i, Q) = \frac{\sum_{j=1}^{t} d_{ij} \cdot q_j}{\sqrt{\sum_{j=1}^{t} {d_{ij}}^2 \cdot \sum_{j=1}^{t} {q_j}^2}}$$

- Note speedup when query is sparse

Term Weights

- Often *tf.idf* weights (Spärck-Jones, 1973)
- Term frequency weight, normalized by all terms j in document i, measures importance in document: $tf_{ik} = \frac{f_{ik}}{\sum_{j=1}^{t} f_{ij}}$
- Inverse document frequency measures importance in collection: $idf_k = \log \frac{N}{n_k}$
- Heuristic combination (note add-1 smoothing to avoid log 0)
- $$d_{ik} = \frac{\log(f_{ik} + 1) \cdot \log(\frac{N}{n_k})}{\sqrt{\left[\sum_{j=1}^{t} \log(f_{ij} + 1) \cdot \log(\frac{N}{n_j})\right]^2}}$$

Standard Word Representation

The vast majority of rule-based **and** statistical NLP work regards words as atomic symbols: hotel, conference, walk

In vector space terms, this is a vector with one 1 and a lot of zeroes

[0 0 0 0 0 0 0 0 0 0 1 0 0 0 0]

Dimensionality: 20K (speech) – 50K (PTB) – 500K (big vocab) – 13M (Google 1T)

We call this a "one-hot" representation. Its problem:

motel [0 0 0 0 0 0 0 0 0 0 1 0 0 0 0] AND
hotel [0 0 0 0 0 0 0 1 0 0 0 0 0 0 0] = 0

Distributional Similarity

You can get a lot of value by representing a word by means of its neighbors

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

(J. R. Firth 1957: 11)

One of the most successful ideas of modern statistical NLP

government debt problems turning into	banking	crises as has happened in
saying that Europe needs unified	banking	regulation to replace the hodgepodge

↖ These words will represent *banking* ↗

You can vary whether you use local or large context to get a more syntactic or semantic clustering

Raw Data: A Concordance

IRISHMAN—an Irishman with.... *Merry Wives*, ii. 2
altogether directed by an Irishman .. *Henry V.* iii. 2
IRISHMEN-against the Irishmen? .2 *Henry VI.* iii. 1
IRK—and yet it irks me..... *As you Like it*, ii. 1
it irks his heart, he cannot .....1 *Henry VI.* i. 4
it irks my very soul .....3 *Henry VI.* ii. 2
IRKSOME—was irksome to me .. *As you Like it*, iii. 5
is an irksome brawling scold .. *Taming of Shrew*, i. 2
irksome is this music to my heart!..2 *Henry VI.* ii. 1
IRON—to wear iron about you.... *Twelfth Night*, iii. 4
my young soldier, put up your iron.. — iv. 1
before barred up with ribs of iron! .. *Much Ado*, iv. 1
runs not this speech like iron through — v. 1
but yet you draw not iron..... *Mid. N.'s Dream*, ii. 2
the iron tongue of midnight hath .... — v. 1
iron may hold with her ..... *Taming of Shrew*, ii. 1
fetch me an iron crow..... *Comedy of Errors*, iii. 1
their iron indignation 'gainst your.. *King John*, ii. 1
with his iron tongue and brazen mouth — iii. 3
heat me these irons hot ..... — iv. 1
must you with hot irons burn (*rep.*).. — iv. 1
none, but in this iron age ..... — iv. 1
stubborn hard than hammered iron? — iv. 1

A Concordance for “party”

from www.webcorp.org.uk

§ thing. She was talking at a party thrown at Daphne's restaurant in
§ have turned it into the hot dinner-party topic. The comedy is the
§ selection for the World Cup party, which will be announced on May 1
§ in the 1983 general election for a party which, when it could not bear to
§ to attack the Scottish National Party, who look set to seize Perth and
§ that had been passed to a second party who made a financial decision
§ the by-pass there will be a street party. "Then," he says, "we are going
§ number-crunchers within the Labour party, there now seems little doubt
§ political tradition and the same party. They are both relatively Anglophilic
§ he told Tony Blair's modernised party they must not retreat into "warm
§ "Oh no, I'm just here for the party," they said. "I think it's terrible
§ A future obliges each party to the contract to fulfil it by
§ be signed by or on behalf of each party to the contract." Mr David N

A Concordance for “party”

§ thing. She was talking at a party thrown at Daphne's restaurant in
§ have turned it into the hot dinner-party topic. The comedy is the
§ selection for the World Cup party, which will be announced on May 1
§ in the 1983 general election for a party which, when it could not bear to
§ to attack the Scottish National Party, who look set to seize Perth and
§ that had been passed to a second party who made a financial decision
§ the by-pass there will be a street party. "Then," he says, "we are going
§ number-crunchers within the Labour party, there now seems little doubt
§ political tradition and the same party. They are both relatively Anglophilic
§ he told Tony Blair's modernised party they must not retreat into "warm
§ "Oh no, I'm just here for the party," they said. "I think it's terrible
§ A future obliges each party to the contract to fulfil it by
§ be signed by or on behalf of each party to the contract." Mr David N

A Concordance for “party”

§ thing. She was talking at a party thrown at Daphne's restaurant in
§ have turned it into the hot dinner-party topic. The comedy is the
§ selection for the World Cup party, which will be announced on May 1
§ the by-pass there will be a street party. "Then," he says, "we are going
§ "Oh no, I'm just here for the party," they said. "I think it's terrible

§ in the 1983 general election for a party which, when it could not bear to
§ to attack the Scottish National Party, who look set to seize Perth and
§ number-crunchers within the Labour party, there now seems little doubt
§ political tradition and the same party. They are both relatively Anglophilic
§ he told Tony Blair's modernised party they must not retreat into "warm

§ that had been passed to a second party who made a financial decision
§ A future obliges each party to the contract to fulfil it by
§ be signed by or on behalf of each party to the contract." Mr David N

A Concordance for “party”

§ number-crunchers within the Labour party, there now seems little doubt
§ political tradition and the same party. They are both relatively Anglophilic
§ he told Tony Blair's modernised party they must not retreat into "warm
§ thing. She was talking at a party thrown at Daphne's restaurant in
§ have turned it into the hot dinner-party topic. The comedy is the
§ selection for the World Cup party, which will be announced on May 1
§ the by-pass there will be a street party. "Then," he says, "we are going
§ "Oh no, I'm just here for the party," they said. "I think it's terrible

§ an appearance at the annual awards bash , but feels in no fit state to
§ -known families at a fundraising bash on Thursday night for Learning
§ Who was paying for the bash? The only clue was the name Asprey,
§ Mail, always hosted the annual bash for the Scottish Labour front-
§ popular. Their method is to bash sense into criminals with a short,
§ just cut off people's heads and bash their brains out over the floor,

A Concordance for “party”

§ number-crunchers within the Labour party, there now seems little doubt
§ political tradition and the same party. They are both relatively Anglophilic
§ he told Tony Blair's modernised party they must not retreat into "warm

§ thing. She was talking at a party thrown at Daphne's restaurant in
§ have turned it into the hot dinner-party topic. The comedy is the
§ selection for the World Cup party, which will be announced on May 1
§ the by-pass there will be a street party. "Then," he says, "we are going
§ "Oh no, I'm just here for the party," they said. "I think it's terrible
§ an appearance at the annual awards bash, but feels in no fit state to
§ -known families at a fundraising bash on Thursday night for Learning
§ Who was paying for the bash? The only clue was the name Asprey,
§ Mail, always hosted the annual bash for the Scottish Labour front-

§ popular. Their method is to bash sense into criminals with a short,
§ just cut off people's heads and bash their brains out over the floor,

Words as Vectors

- Represent each word **type** w by a point in V-dimensional space
  - e.g., V is size of vocabulary
  - the 17th coordinate of w represents **strength** of w's association with vocabulary word 17

Words as Vectors

- Represent each word **type** w by a point in V-dimensional space
  - e.g., V is size of vocabulary
  - the 17th coordinate of w represents **strength** of w's association with vocabulary word 17

= party

Words as Vectors

- Represent each word **type** w by a point in V-dimensional space
  - e.g., V is size of vocabulary
  - the 17th coordinate of w represents **strength** of w's association with vocabulary word 17

= party

(0, 0, 3, 1, 0, 7, . . . 1, 0)

Words as Vectors

- Represent each word **type** w by a point in V-dimensional space
  - e.g., V is size of vocabulary
  - the 17th coordinate of w represents **strength** of w's association with vocabulary word 17

= party

aardvark abacus abandoned abbot abduct above zygote zymurgy

(0, 0, 3, 1, 0, 7, . . . 1, 0)

Words as Vectors

- Represent each word **type** w by a point in V-dimensional space
  - e.g., V is size of vocabulary
  - the 17th coordinate of w represents **strength** of w's association with vocabulary word 17

= party

aardvark abacus abandoned abbot abduct above . . . zygote zymurgy

(0, 0, 3, 1, 0, 7, . . . 1, 0)

From corpus:

Arlen Specter **abandoned** the Republican **party**.
There were lots of **abbot**s and nuns dancing at that **party**.
The **party** **above** the art gallery was, **above** all, a laboratory for synthesizing **zygote**s and beer.

Words as Vectors

- Represent each word **type** w by a point in V-dimensional space
  - e.g., V is size of vocabulary
  - the 17th coordinate of w represents **strength** of w's association with vocabulary word 17

= party

aardvark abacus abandoned abbot abduct above count too high (too influential) zygote zymurgy

(0, 0, 3, 1, 0, 7, . . . 1, 0)

From corpus:

Arlen Specter **abandoned** the Republican **party**.
There were lots of **abbot**s and nuns dancing at that **party**.
The **party** **above** the art gallery was, **above** all, a laboratory for synthesizing **zygote**s and beer.

Words as Vectors

- Represent each word **type** w by a point in V-dimensional space
  - e.g., V is size of vocabulary
  - the 17th coordinate of w represents **strength** of w's association with vocabulary word 17

= party

aardvark abacus abandoned abbot abduct above

(0, 0, 3, 1, 0, 7, . . . 1, 0)

count too high (too influential)

zygote zymurgy

count too low

From corpus:

Arlen Specter **abandoned** the Republican **party**.
There were lots of **abbot**s and nuns dancing at that **party**.
The **party** **above** the art gallery was, **above** all, a laboratory for synthesizing **zygote**s and beer.

Words as Vectors

- Represent each word **type** w by a point in V-dimensional space
  - e.g., V is size of vocabulary
  - the 17th coordinate of w represents **strength** of w's association with vocabulary word 17

= party

aardvark abacus abandoned abbot abduct above zygote zymurgy

(0, 0, 3, 1, 0, 7, . . . 1, 0)

Words as Vectors

- Represent each word **type** w by a point in V-dimensional space
  - e.g., V is size of vocabulary
  - the 17th coordinate of w represents **strength** of w's association with vocabulary word 17

= party

aardvark abacus abandoned abbot abduct above

(0, 0, 3, 1, 0, 7, . . . 1, 0)

how might you measure this?

zygote zymurgy

Words as Vectors

- Represent each word **type** w by a point in V-dimensional space
  - e.g., V is size of vocabulary
  - the 17th coordinate of w represents **strength** of w's association with vocabulary word 17

= party

aardvark abacus abandoned abbot abduct above

(0, 0, 3, 1, 0, 7, . . . 1, 0)

how might you measure this?

zygote zymurgy

- how often words appear next to each other
- how often words appear near each other

Words as Vectors

- Represent each word **type** w by a point in V-dimensional space
  - e.g., V is size of vocabulary
  - the 17th coordinate of w represents **strength** of w's association with vocabulary word 17

= party

aardvark abacus abandoned abbot abduct above

(0, 0, 3, 1, 0, 7, . . . 1, 0)

how might you measure this?

zygote zymurgy

- how often words appear next to each other
- how often words appear near each other
- how often words are syntactically linked

Words as Vectors

- Represent each word **type** w by a point in V-dimensional space
  - e.g., V is size of vocabulary
  - the 17th coordinate of w represents **strength** of w's association with vocabulary word 17

= party

aardvark abacus abandoned abbot abduct above

(0, 0, 3, 1, 0, 7, . . . 1, 0)

how might you measure this?

zygote zymurgy

- how often words appear next to each other
- how often words appear near each other
- how often words are syntactically linked
- should correct for commonness of word (e.g., "above")

Words as Vectors

- Represent each word **type** w by a point in k-dimensional space
  - e.g., k is size of vocabulary
  - the 17th coordinate of w represents **strength** of w's association with vocabulary word 17

aardvark abacus abandoned abbot abduct above zygote zymurgy

(0, 0, 3, 1, 0, 7, . . . 1, 0)

Words as Vectors

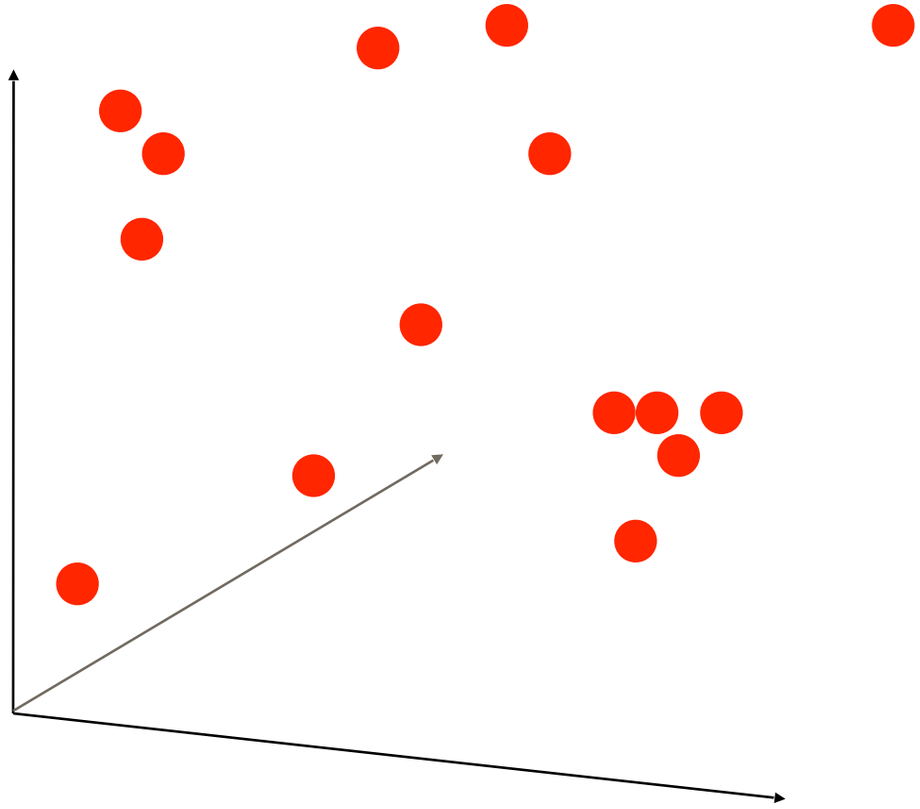
- Represent each word **type** w by a point in k-dimensional space
  - e.g., k is size of vocabulary
  - the 17th coordinate of w represents **strength** of w's association with vocabulary word 17

- Plot all word types in k-dimensional space
- Look for **clusters** of close-together types

Learning Classes by Clustering

- Plot all word types in k-dimensional space
- Look for **clusters** of close-together types

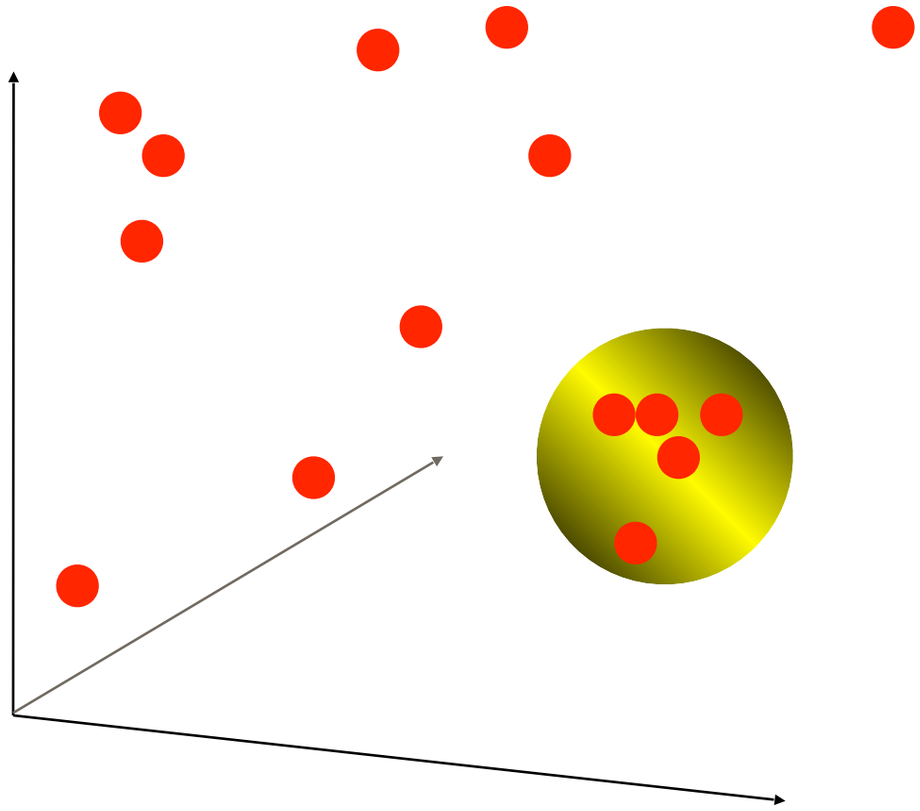
Plot in k dimensions (here k=3)



Learning Classes by Clustering

- Plot all word types in k-dimensional space
- Look for **clusters** of close-together types

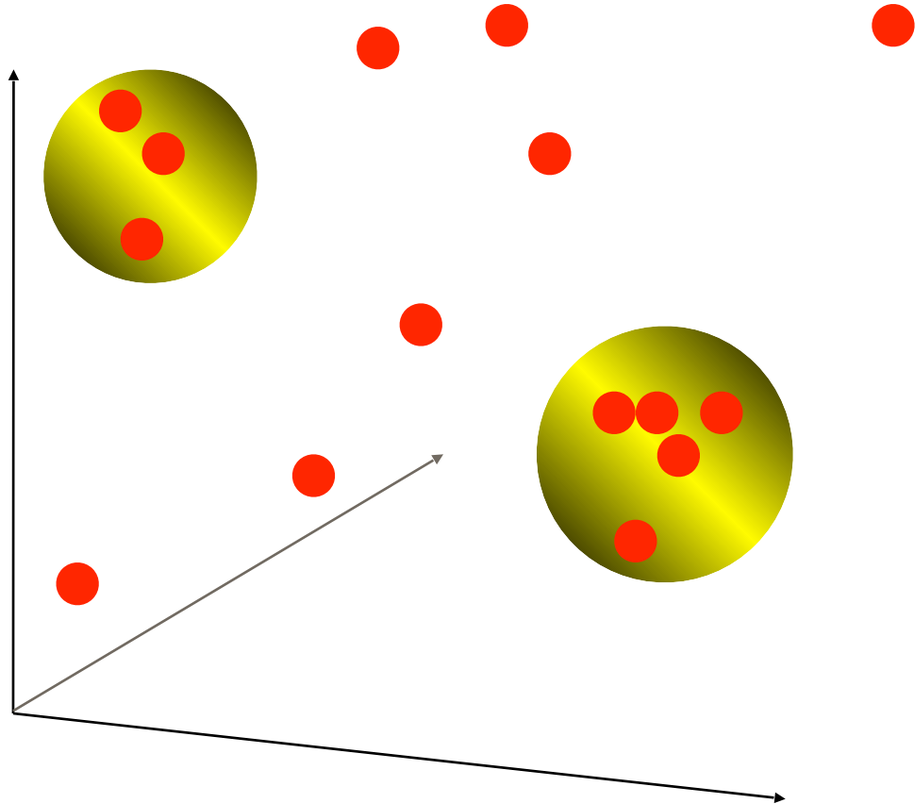
Plot in k dimensions (here k=3)



Learning Classes by Clustering

- Plot all word types in k-dimensional space
- Look for **clusters** of close-together types

Plot in k dimensions (here k=3)



Hard/Soft Clustering

Class based models learn word classes of similar words based on distributional information (~ class HMM)

- Brown clustering (Brown et al. 1992)
- Exchange clustering (Martin et al. 1998, Clark 2003)
- Desparsification and great example of unsupervised pre-training

Soft clustering models learn for each cluster/topic a distribution over words of how likely that word is in each cluster

- Latent Semantic Analysis (LSA/LSI), Random projections
- Latent Dirichlet Analysis (LDA), HMM clustering

Distributed Representation

Similar idea

Combine vector space semantics with the prediction of probabilistic models (Bengio et al. 2003, Collobert & Weston 2008, Turian et al. 2010)

In all of these approaches, including deep learning models, a word is represented as a dense vector

$$\textit{linguistics} = \begin{pmatrix} 0.286 \\ 0.792 \\ -0.177 \\ -0.107 \\ 0.109 \\ -0.542 \\ 0.349 \\ 0.271 \end{pmatrix}$$

Word2vec

- Instead of **counting** how often each word w occurs near 'party'
- Train a classifier on a binary **prediction** task:
  - Is w likely to show up near 'party'?
- We don't do prediction for its own sake
  - Treat learned classifier weights as **word embeddings**

Key Insight: *Auxiliary Tasks*

- A word w near 'party' acts as the label for the binary classification question:
  - "Is word w likely to show up near *party*?"
- Automatically generate lots of training data!
- Similar ideas in fixed-context feedforward language modeling and domain adaptation
  - Bengio et al., 2003; Collobert et al., 2011

Word2vec: *Skip-gram task*

- Common option for auxiliary task:
- Skip-gram w/negative sampling (SGNS)
  - Predict each of *n* words near target word
  - True context=positive training sample
  - Approximate distribution over alternate contexts by small (negative) sample of other other words

Word2vec: *Skip-gram task*

In LMs, we normally predict each word in sequence from its context u.

$$\log p(w_1^M) \approx \sum_{m=1}^{M} \log p(w_m \mid u) = \sum_{m=1}^{M} \log \frac{\exp u \cdot v_{w_m}}{\sum_{w' \in \mathcal{V}} \exp u \cdot v_{w'}}$$

In the skip-gram model, we predict each context word from a *target* word (which is weird because each word is in the context of *multiple* words).

$$\log p(w_1^M) \approx \sum_{m=1}^{M} \sum_{n \in [-h_m, h_m] \setminus 0} \log p(w_{m+n} \mid w_m)$$

Word2vec: *Skip-gram task*

In the skip-gram model, we predict each context word from a *target* word (which is weird because each word is in the context of *multiple* words).

$$\log p(w_1^M) \approx \sum_{m=1}^{M} \sum_{n \in [-h_m, h_m] \setminus 0} \log p(w_{m+n} \mid w_m)$$

Represent each context word by a K-dimensional embedding vector *u*.
Represent each target word by a K-dimensional embedding vector *v*.

$$\log p(w_1^M) \approx \sum_{m=1}^{M} \sum_{n \in [-h_m, h_m] \setminus 0} \log \frac{\exp u_{w_{m+n}} \cdot v_{w_m}}{\sum_{w' \in \mathcal{V}} \exp u_{w'} \cdot v_{w_m}}$$

$$= \sum_{m=1}^{M} \sum_{n \in [-h_m, h_m] \setminus 0} u_{w_{m+n}} \cdot v_{w_m} - \log \sum_{w' \in \mathcal{V}} u_{w'} \cdot v_{w_m}$$

Word2vec: *SGNS*

Represent each context word by a K-dimensional embedding vector *u*.
Represent each target word by a K-dimensional embedding vector *v*.

$$\log p(w_1^M) \approx \sum_{m=1}^{M} \sum_{n \in [-h_m, h_m] \setminus 0} u_{w_{m+n}} \cdot v_{w_m} - \log \sum_{w' \in \mathcal{V}} \exp u_{w'} \cdot v_{w_m}$$

Instead of one multi-class problem over the whole vocabulary,
train one positive and several negative binary classification problems
for each context word.

$$\log p(w_1^M) \approx \sum_{m=1}^{M} \sum_{n \in [-h_m, h_m] \setminus 0} \log \sigma(u_{w_{m+n}} \cdot v_{w_m}) + \sum_{w' \in \mathcal{W}_{\text{neg}}} \log(1 - \sigma(u_{w'} \cdot v_{w_m}))$$

Skip-Gram Training

Training sentence:

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

c1 c2 t c3 c4

Training data: input/output pairs centering on *apricot*

Assume a +/- 2 word window

Skip-Gram Training

Training sentence:

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

Skip-Gram Training

Training sentence:

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

c1 c2 t c3 c4

k=2

positive examples +		**negative examples -**			
t	c	t	c	t	c
apricot	tablespoon	apricot	aardvark	apricot	twelve
apricot	of	apricot	puddle	apricot	hello
apricot	preserves	apricot	where	apricot	dear
apricot	or	apricot	coaxial	apricot	forever

Negative Sampling

- Could choose w by unigram frequency
- Effective to do *multiplicative* smoothing
  - $p_\alpha(w) = \frac{count(w)^\alpha}{\sum_{w'} count(w')^\alpha}$
- Mikolov et al. (2013) suggest $\alpha = \frac{3}{4}$
  - 5–20 samples for small data
  - 2–5 samples for large data

SGNS Summary

- Start with 2*V* random *K*-dimensional vectors as initial word and context embeddings
- Train binary logistic regression models to distinguish true context words from negative samples

Evaluating Embeddings

- Compare to human scores on word similarity-type tasks:
  - WordSim-353 (Finkelstein et al., 2002)
  - SimLex-999 (Hill et al., 2015)
  - Stanford Contextual Word Similarity (SCWS) dataset (Huang et al., 2012)
  - TOEFL dataset: *Levied* is closest in meaning to: *imposed, believed, requested, correlated*

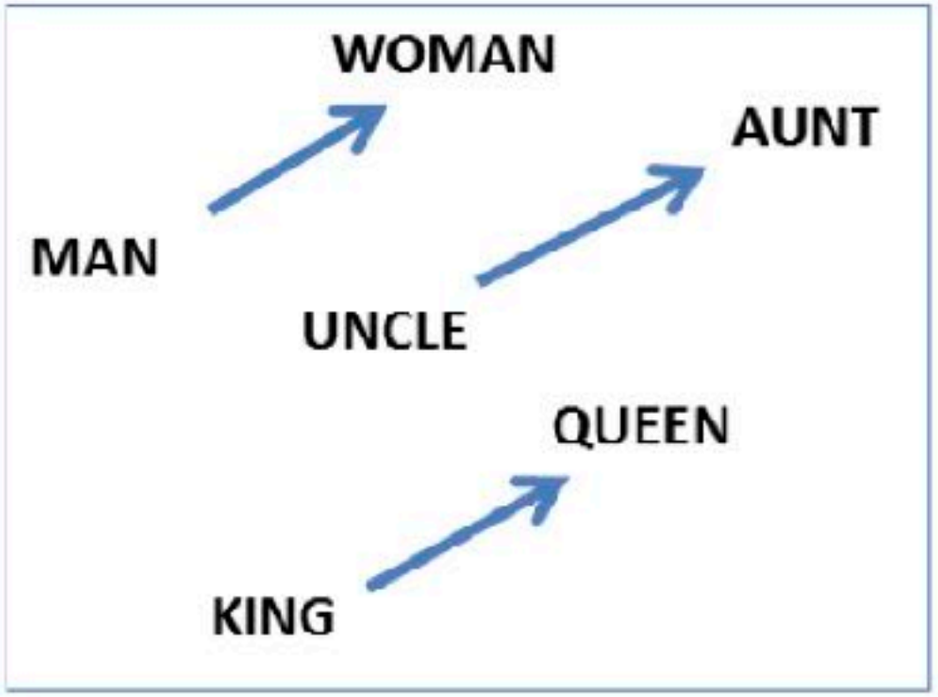
Properties of Embeddings

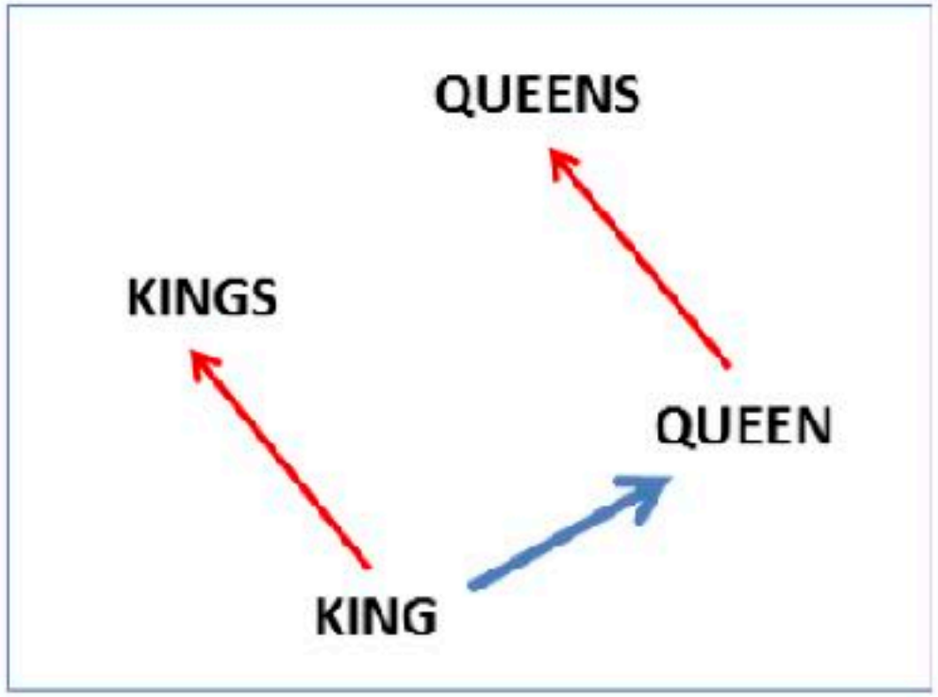
- Similarity depends on window size *h*
- *h* = ±2 The nearest words to *Hogwarts*:
  - Sunnydale
  - Evernight
- *h* = ±5 The nearest words to *Hogwarts*:
  - Dumbledore
  - Malfoy
  - halfblood

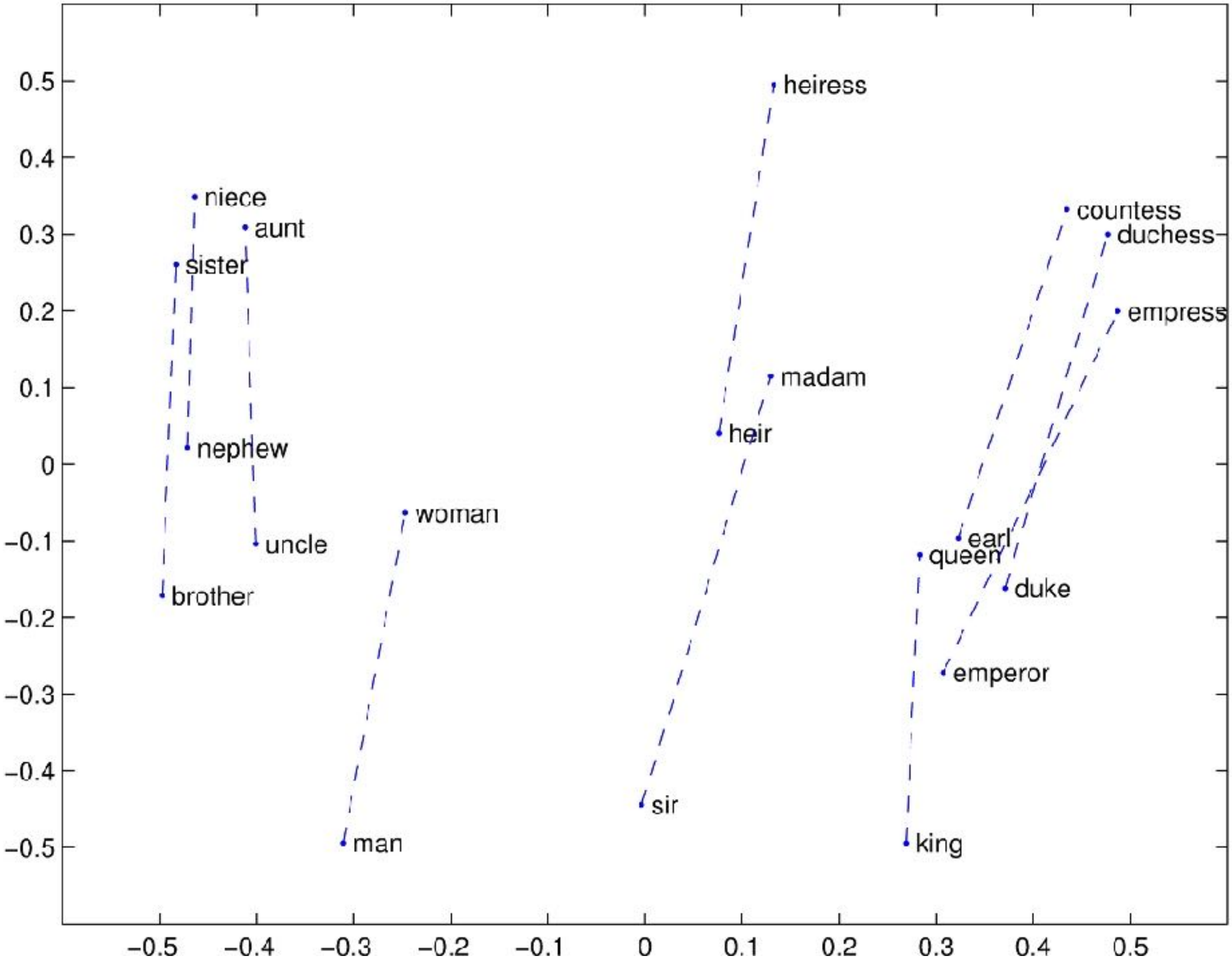
Vector Analogies!

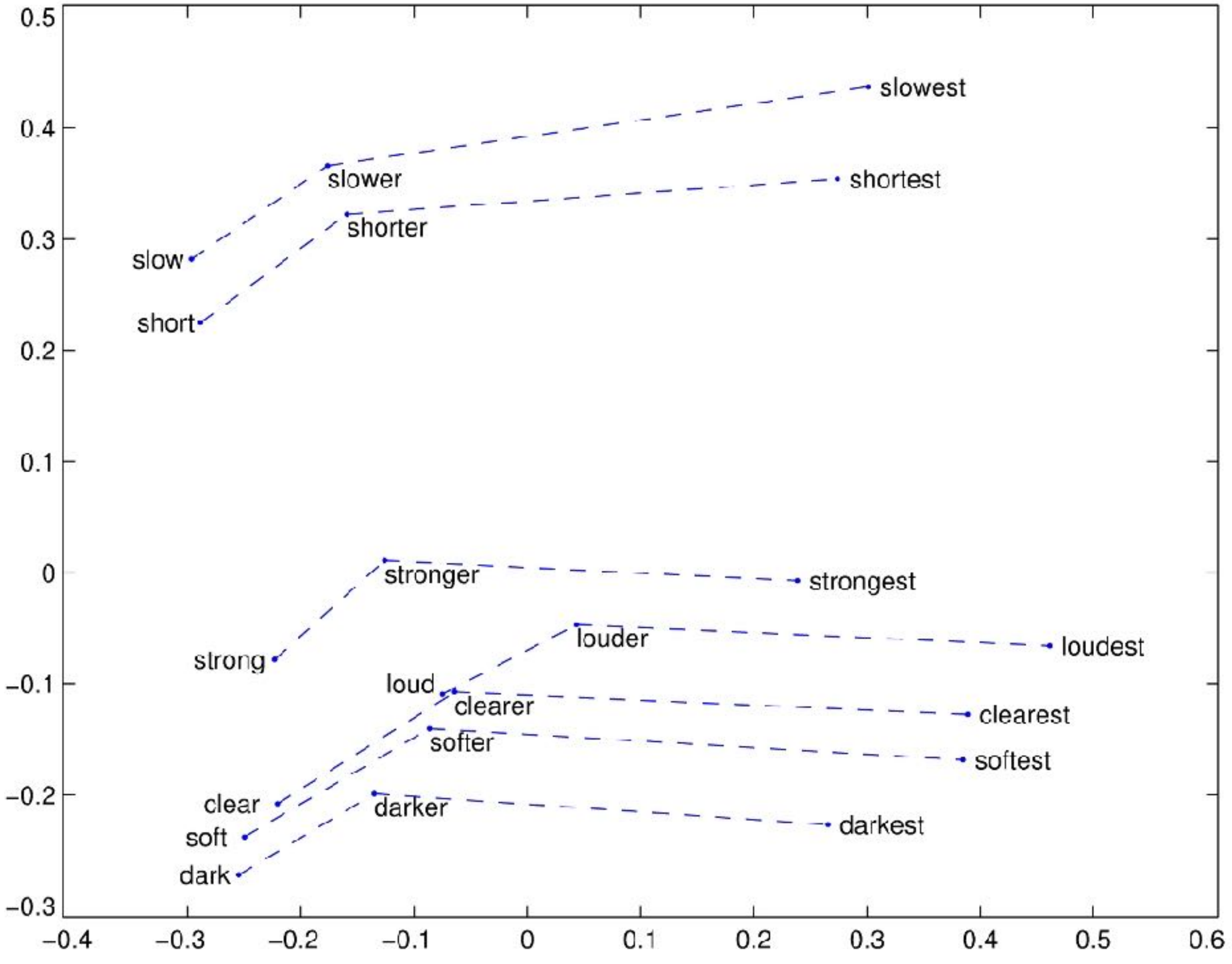
vector(*'king'*) - vector(*'man'*) + vector(*'woman'*) ≈ vector('queen')

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





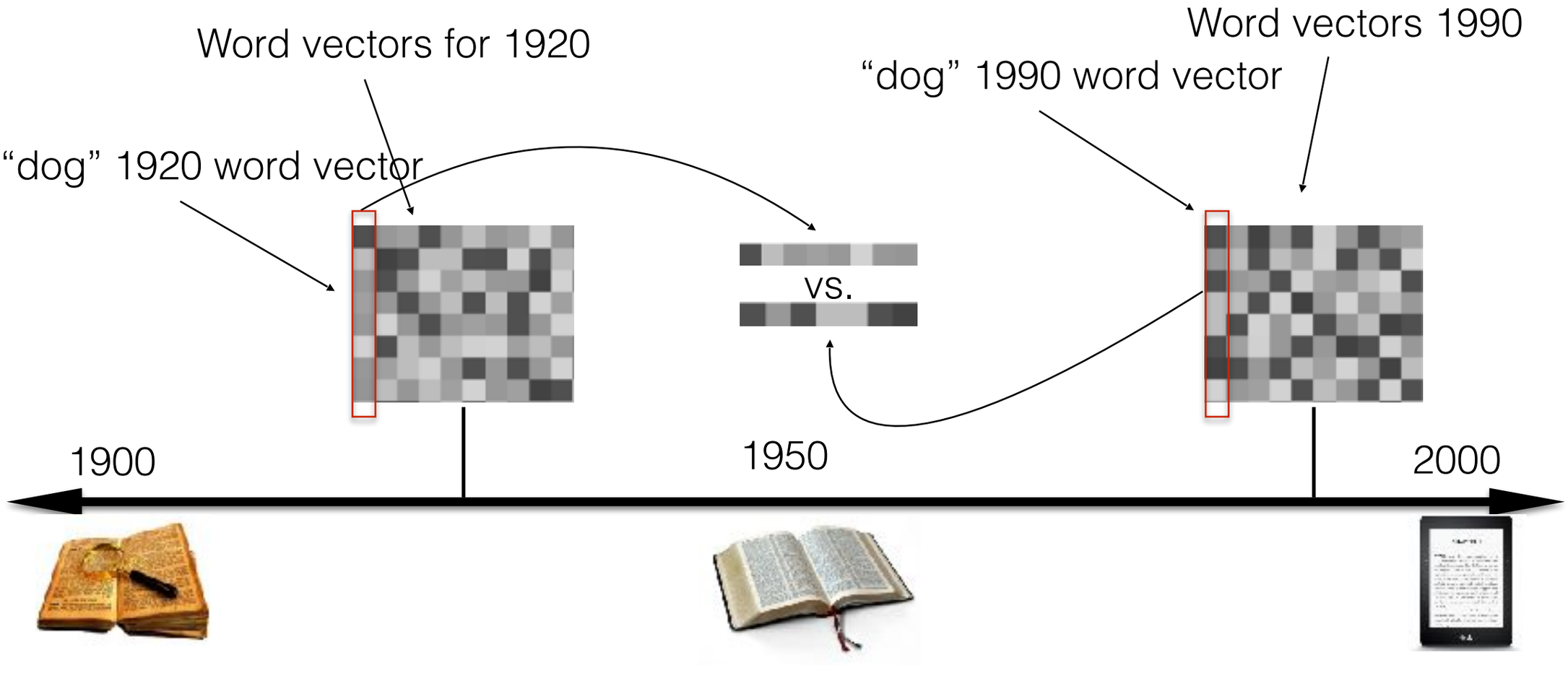

niece
aunt
sister
nephew
uncle
brother
woman
man
heiress
madam
heir
sir
countess
duchess
empress
earl
queen
duke
emperor
king
0.5
0.4
0.3
0.2
0.1
0
−0.1
−0.2
−0.3
−0.4
−0.5
−0.5
−0.4
−0.3
−0.2
−0.1
0
0.1
0.2
0.3
0.4
0.5


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
slow
slower
slowest
short
shorter
shortest
strong
stronger
strongest
loud
louder
loudest
clear
clearer
clearest
soft
softer
softest
dark
darker
darkest

Caveats

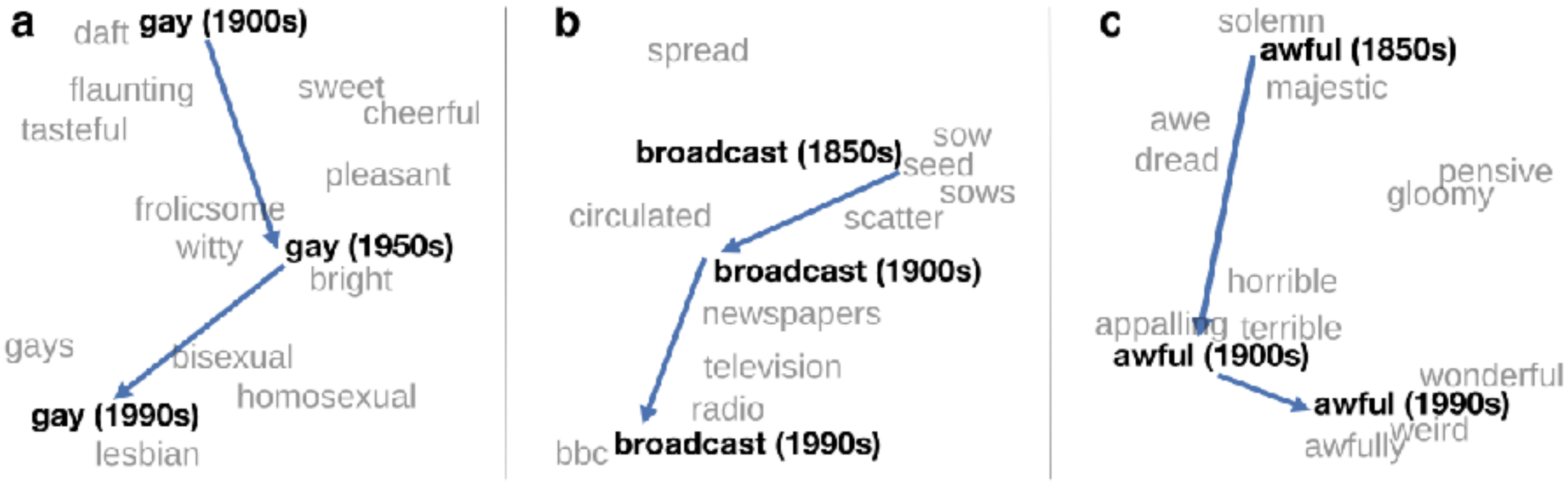
- "Parallelogram analogies" only seem to work for frequent words, small distances and certain relations (relating countries to capitals, or parts of speech), but not others. (Linzen 2016, Gladkova et al. 2016, Ethayarajh et al. 2019a)
- Understanding analogy is an open area of research (Peterson et al. 2020)

Diachronic Embeddings



Diachronic Embeddings

Project 300 dimensions down into 2



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

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 *NeurIPS*, 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
- Inferences might thus lead to bias in hiring…

Historical Embeddings to Study Bias

Garg, N., Schiebinger, L., Jurafsky, D., and Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. Proceedings of the National Academy of Sciences 115(16), E3635–E3644.

- Compute a **gender** or **ethnic bias** for each adjective: e.g., how much closer the adjective is to "woman" synonyms than "man" synonyms, or names of particular ethnicities
  - Embeddings for **competence** adjective (*smart, wise, brilliant, resourceful, thoughtful, logical*) are biased toward men, a bias slowly decreasing 1960-1990
  - Embeddings for **dehumanizing** adjectives (*barbaric, monstrous, bizarre*) were biased toward Asians in the 1930s, bias decreasing over the 20th century.
- These match the results of old surveys done in the 1930s

Embeddings as Features

Compared to a method like LSA, neural word embeddings can become more meaningful through adding supervision from one or multiple tasks

"Discriminative fine-tuning"

For instance, sentiment is usually not captured in unsupervised word embeddings but can be in neural word vectors

We can build representations for large linguistic units