# CSCI 5541: Natural Language Processing

## Lecture 5: Distributional Semantics and Word Embeddings

Dongyeop Kang (DK), University of Minnesota

dongyeop@umn.edu | twitter.com/dongyeopkang | dykang.github.io

computer science
& engineering

MINNESOTA NLP EST. 2021

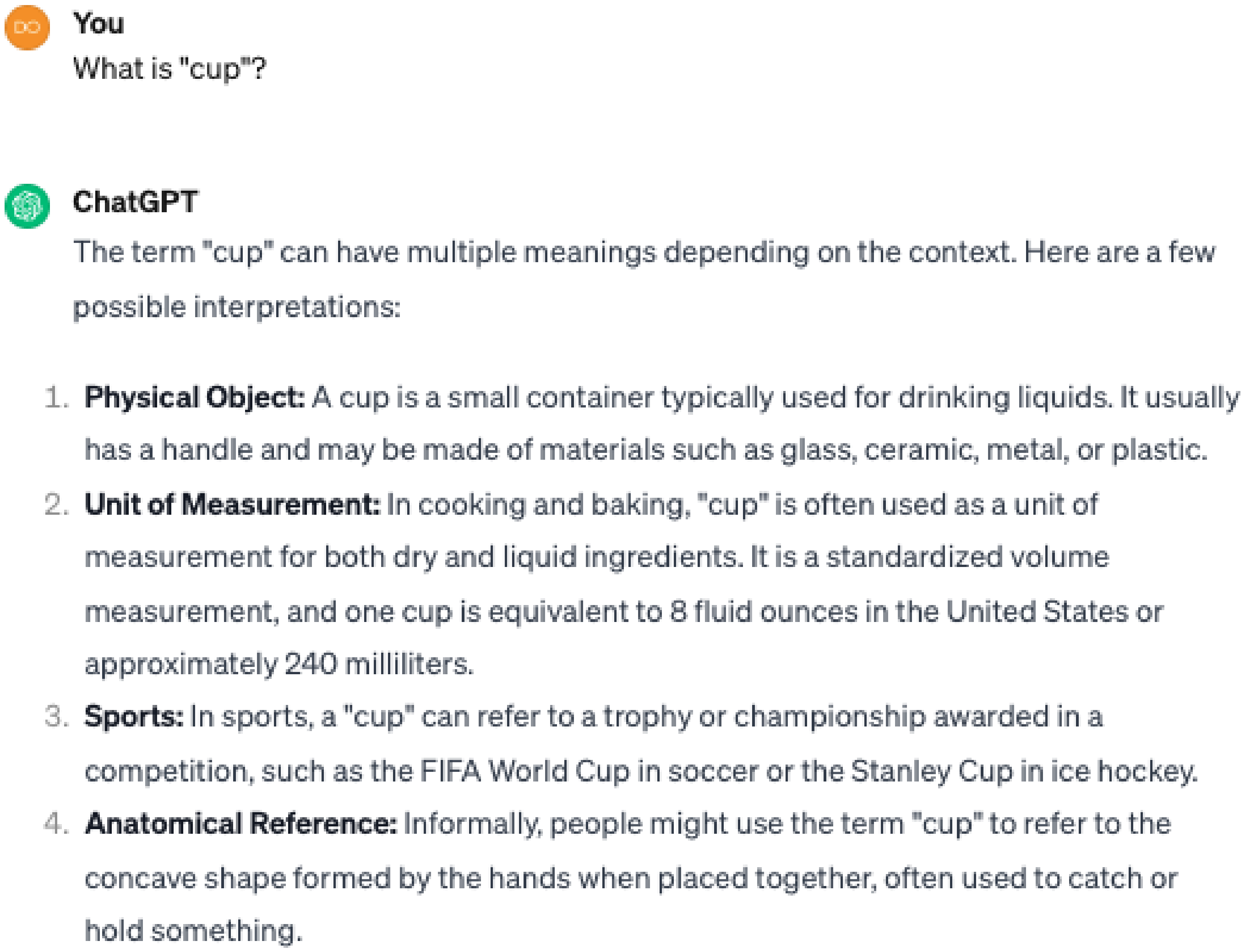
UNIVERSITY OF MINNESOTA
Driven to Discover®

# Outline

- ❑ What is semantics? How do you define meaning?
- ❑ Three ways of looking at word meaning
  - ○ Decompositional
  - ○ Ontological
  - ○ Distributional
- ❑ Different kinds of encoding “context”
  - ○ Count-based
  - ○ Distributed prediction-based (type) embeddings
  - ○ Distributed contextual (token) embeddings from language models
  - ○ Many more variants

# What is semantics?
# How do you define meaning?

# Define “cup”



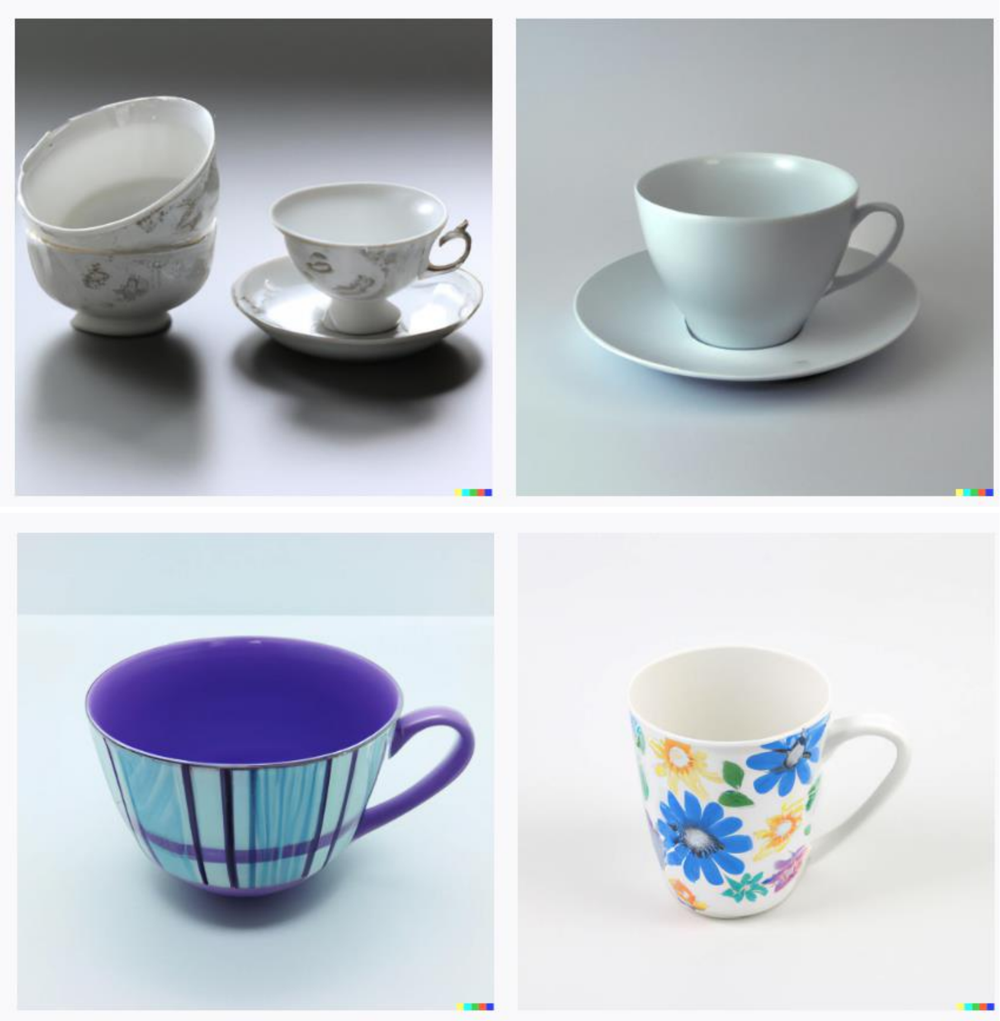
**You**

What is "cup"?

**ChatGPT**

The term "cup" can have multiple meanings depending on the context. Here are a few possible interpretations:

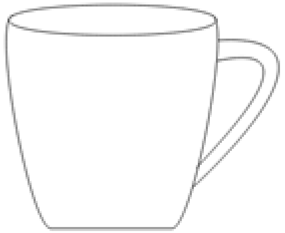
1. **Physical Object:** A cup is a small container typically used for drinking liquids. It usually has a handle and may be made of materials such as glass, ceramic, metal, or plastic.
2. **Unit of Measurement:** In cooking and baking, "cup" is often used as a unit of measurement for both dry and liquid ingredients. It is a standardized volume measurement, and one cup is equivalent to 8 fluid ounces in the United States or approximately 240 milliliters.
3. **Sports:** In sports, a "cup" can refer to a trophy or championship awarded in a competition, such as the FIFA World Cup in soccer or the Stanley Cup in ice hockey.
4. **Anatomical Reference:** Informally, people might use the term "cup" to refer to the concave shape formed by the hands when placed together, often used to catch or hold something.



# Decompositional semantics

**Color**: blue, black, etc

**Shape**:

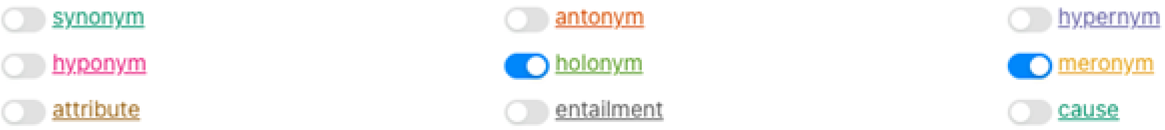


**Texture**: ceramic, wood, glass, clay, etc

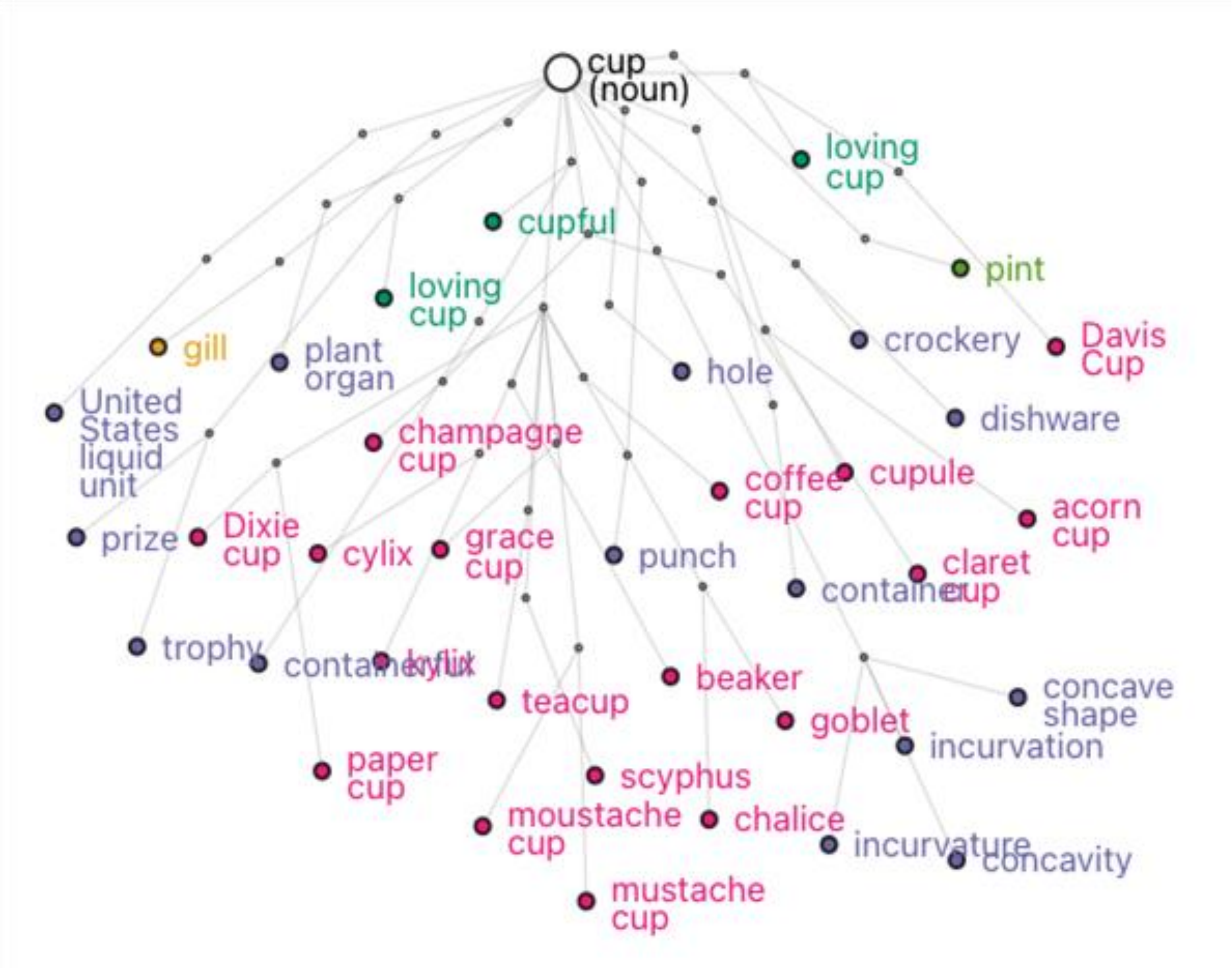
# Three ways of looking at word meaning

- Decompositional
  - What characteristics/components of what the word represents
- Ontological
  - How the meaning of the word relates to the meanings of other words
- Distributional
  - What contexts the word is found in, relative to other words

# Ontological semantics



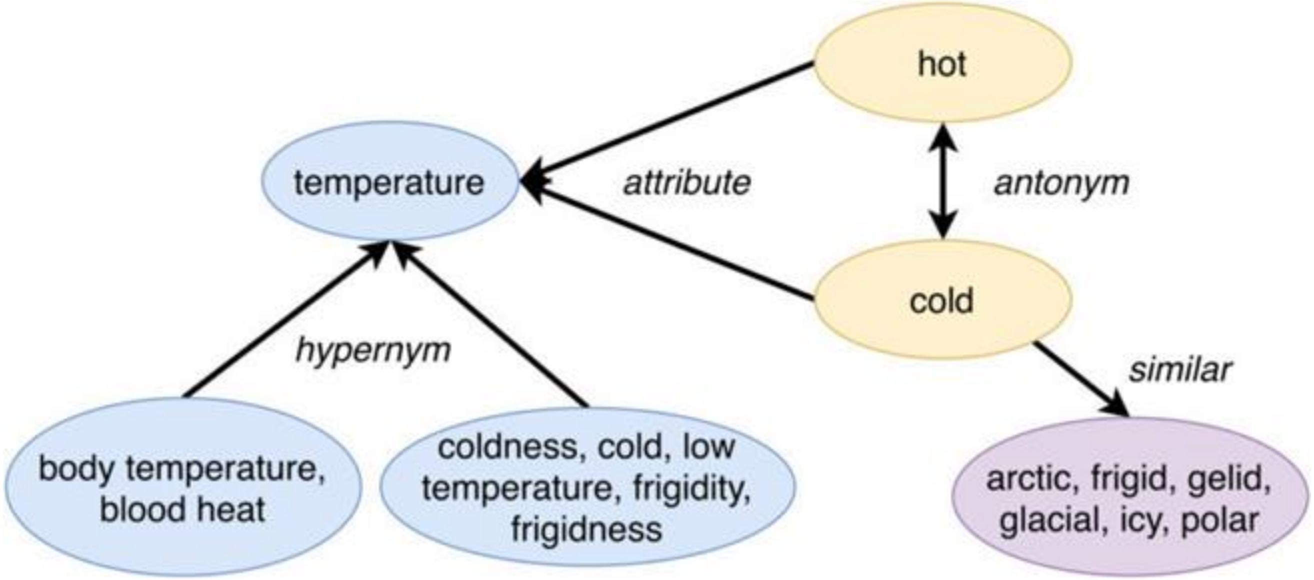
○ synonym
○ antonym
○ hypernym
○ hyponym
● holonym
● meronym
○ attribute
○ entailment
○ cause



https://lexical-graph.herokuapp.com/

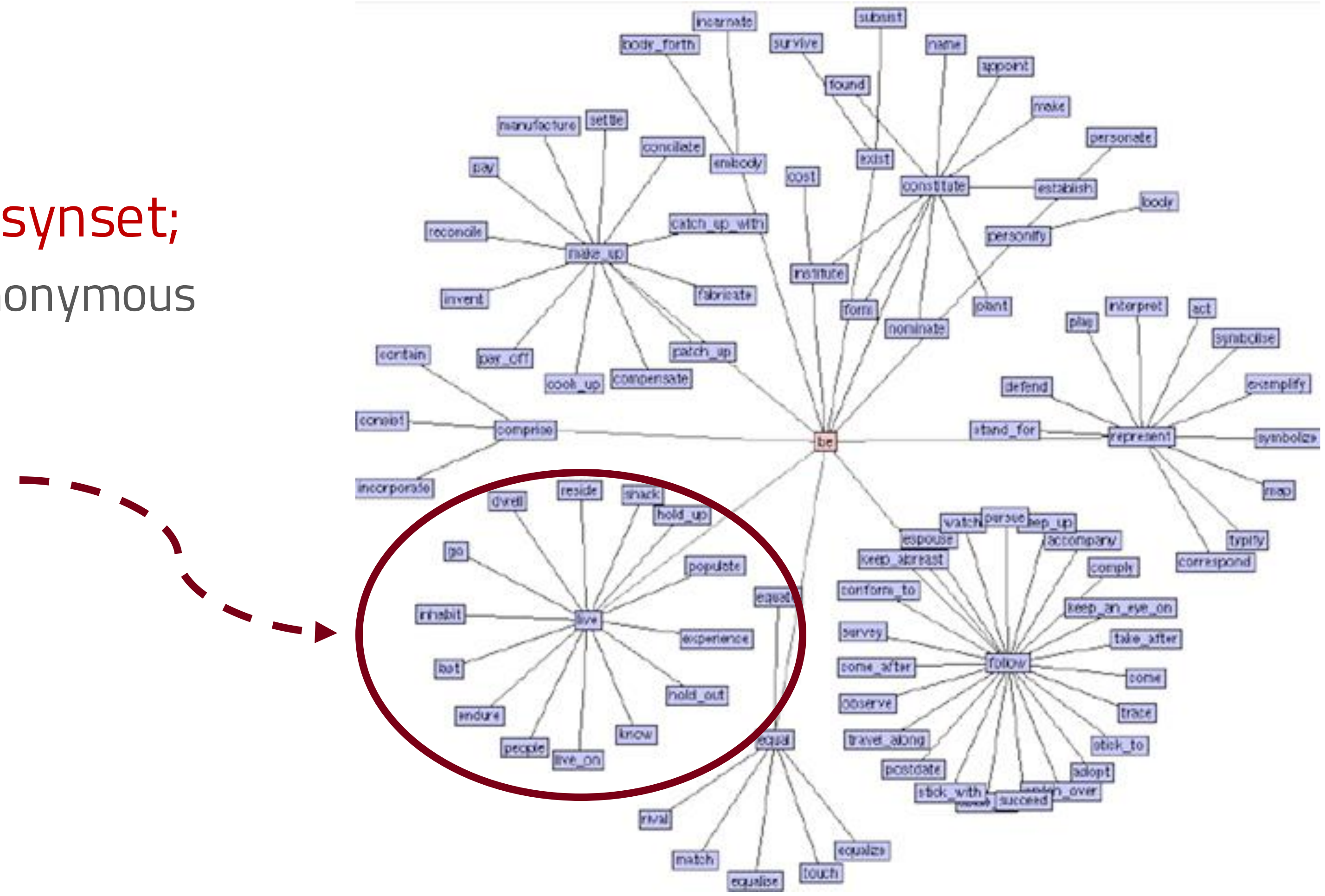
# Semantic relations

- Synonymy — equivalence
  - <small, little>
- Antonymy — opposition
  - <small, large>
- Meronymy — part-of relation
  - <liver, body>
- Holonymy — has-a relation
  - <body, liver>
- **Hyponymy** — subset; is-a relation
  - <dog, mammal>
- **Hypernymy** — superset
  - <mammal, dog>


hot
temperature
*attribute*
*antonym*
cold
*hypernym*
*similar*
body temperature, blood heat
coldness, cold, low temperature, frigidity, frigidness
arctic, frigid, gelid, glacial, icy, polar

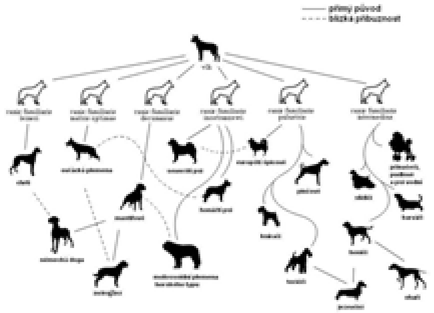
# WordNet

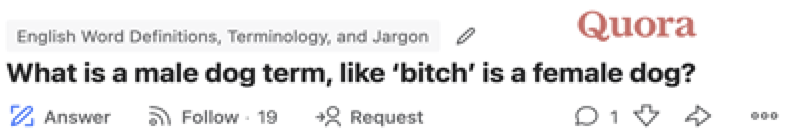
- ❑ Each sense is associated with a synset;
  - ○ a set of words that are roughly synonymous for a particular sense



Synset

# Synsets for *dog* (n)





English Word Definitions, Terminology, and Jargon

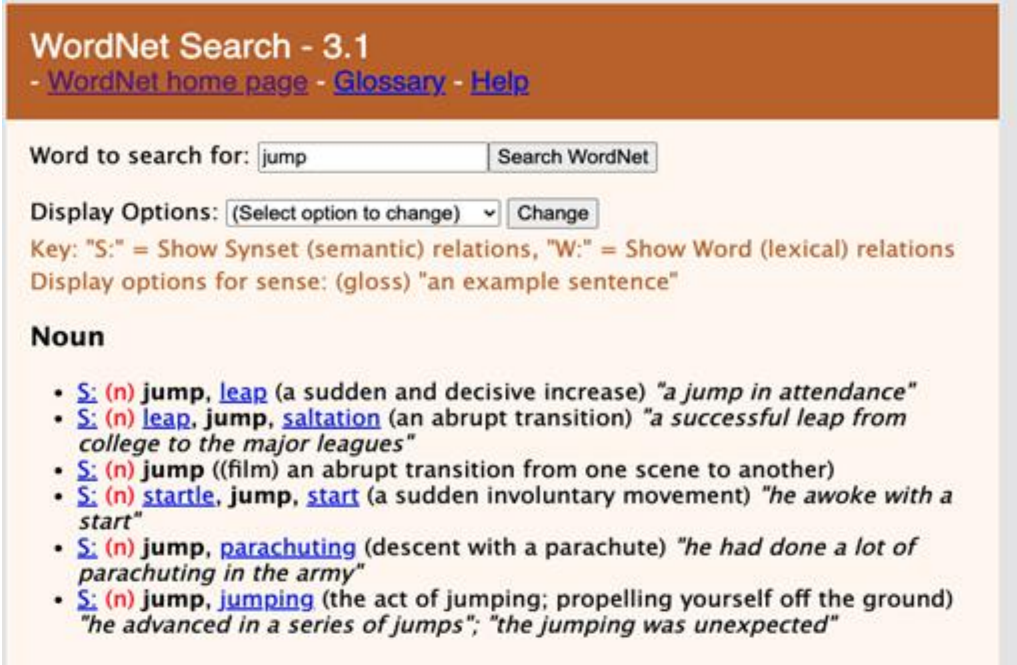
Quora

What is a male dog term, like 'bitch' is a female dog?

- S: (n) dog, domestic dog, Canis familiaris (a member of the genus Canis (probably descended from the common wolf) that has been domesticated by man since prehistoric times; occurs in many breeds) "the dog barked all night"
- S: (n) dog (informal term for a man) "you lucky dog"
- S: (n) cad, bounder, blackguard, dog, hound, heel (someone who is morally reprehensible) "you dirty dog"
- S: (n) frank, frankfurter, hotdog, hot dog, dog, wiener, wienerwurst, weenie (a smooth-textured sausage of minced beef or pork usually smoked; often served on a bread roll)
- S: (n) pawl, detent, click, dog (a hinged catch that fits into a notch of a ratchet to move a wheel forward or prevent it from moving backward)
- S: (n) andiron, firedog, dog, dog-iron (metal supports for logs in a fireplace) "the andirons were too hot to touch"
- …

# Tools for WordNet

- Original English WordNet project:
  - [https://wordnet.princeton.edu/related-projects](https://wordnet.princeton.edu/related-projects)
- WordNets in the World (many languages available):
  - [http://globalwordnet.org/resources/wordnets-in-the-world/](http://globalwordnet.org/resources/wordnets-in-the-world/)



WordNet Search - 3.1
- WordNet home page - Glossary - Help

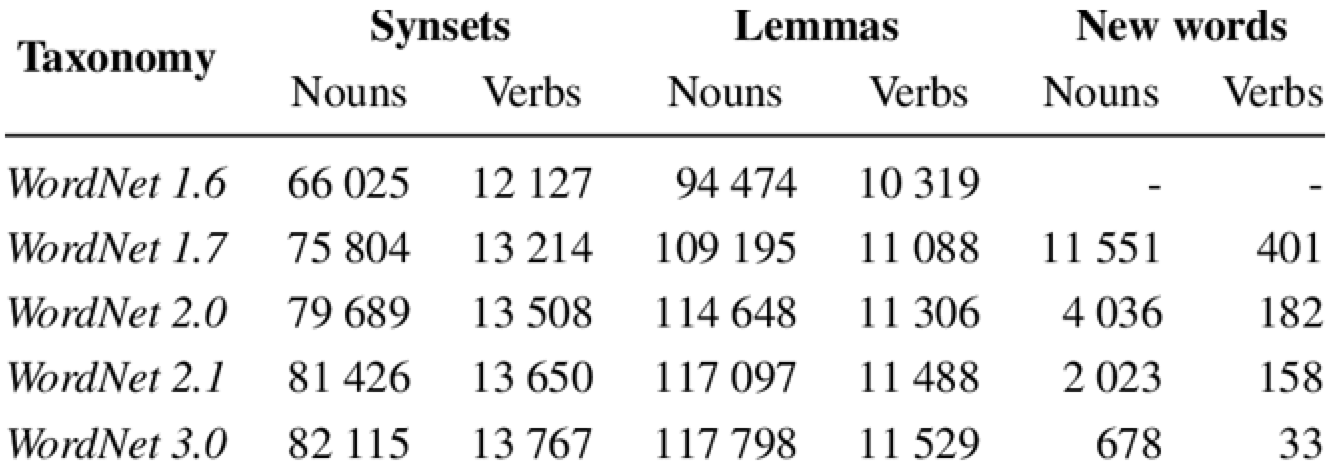
Word to search for: jump

Display Options: (Select option to change)

Key: "S:" = Show Synset (semantic) relations, "W:" = Show Word (lexical) relations
Display options for sense: (gloss) "an example sentence"

**Noun**

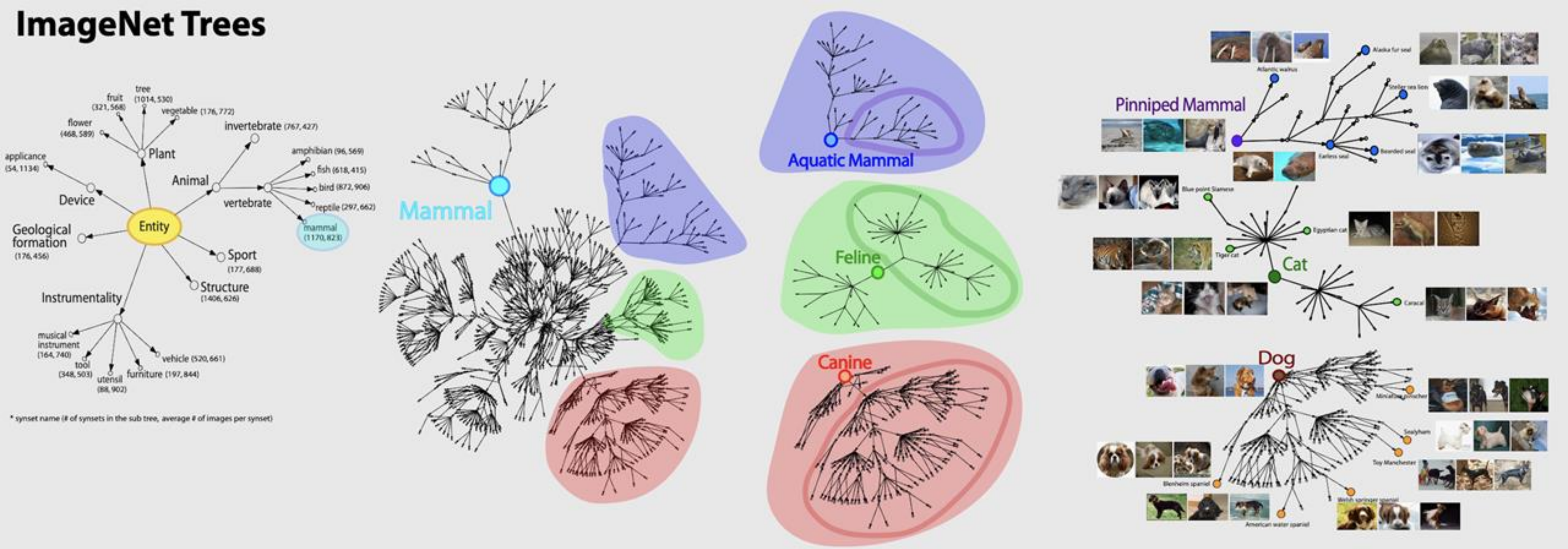
- S: (n) **jump**, leap (a sudden and decisive increase) *"a jump in attendance"*
- S: (n) leap, **jump**, saltation (an abrupt transition) *"a successful leap from college to the major leagues"*
- S: (n) **jump** ((film) an abrupt transition from one scene to another)
- S: (n) startle, **jump**, start (a sudden involuntary movement) *"he awoke with a start"*
- S: (n) **jump**, parachuting (descent with a parachute) *"he had done a lot of parachuting in the army"*
- S: (n) **jump**, jumping (the act of jumping; propelling yourself off the ground) *"he advanced in a series of jumps"; "the jumping was unexpected"*



| Taxonomy | Synsets |  | Lemmas |  | New words |  |
|---|---|---|---|---|---|---|
|  | Nouns | Verbs | Nouns | Verbs | Nouns | Verbs |
| *WordNet 1.6* | 66 025 | 12 127 | 94 474 | 10 319 | - | - |
| *WordNet 1.7* | 75 804 | 13 214 | 109 195 | 11 088 | 11 551 | 401 |
| *WordNet 2.0* | 79 689 | 13 508 | 114 648 | 11 306 | 4 036 | 182 |
| *WordNet 2.1* | 81 426 | 13 650 | 117 097 | 11 488 | 2 023 | 158 |
| *WordNet 3.0* | 82 115 | 13 767 | 117 798 | 11 529 | 678 | 33 |

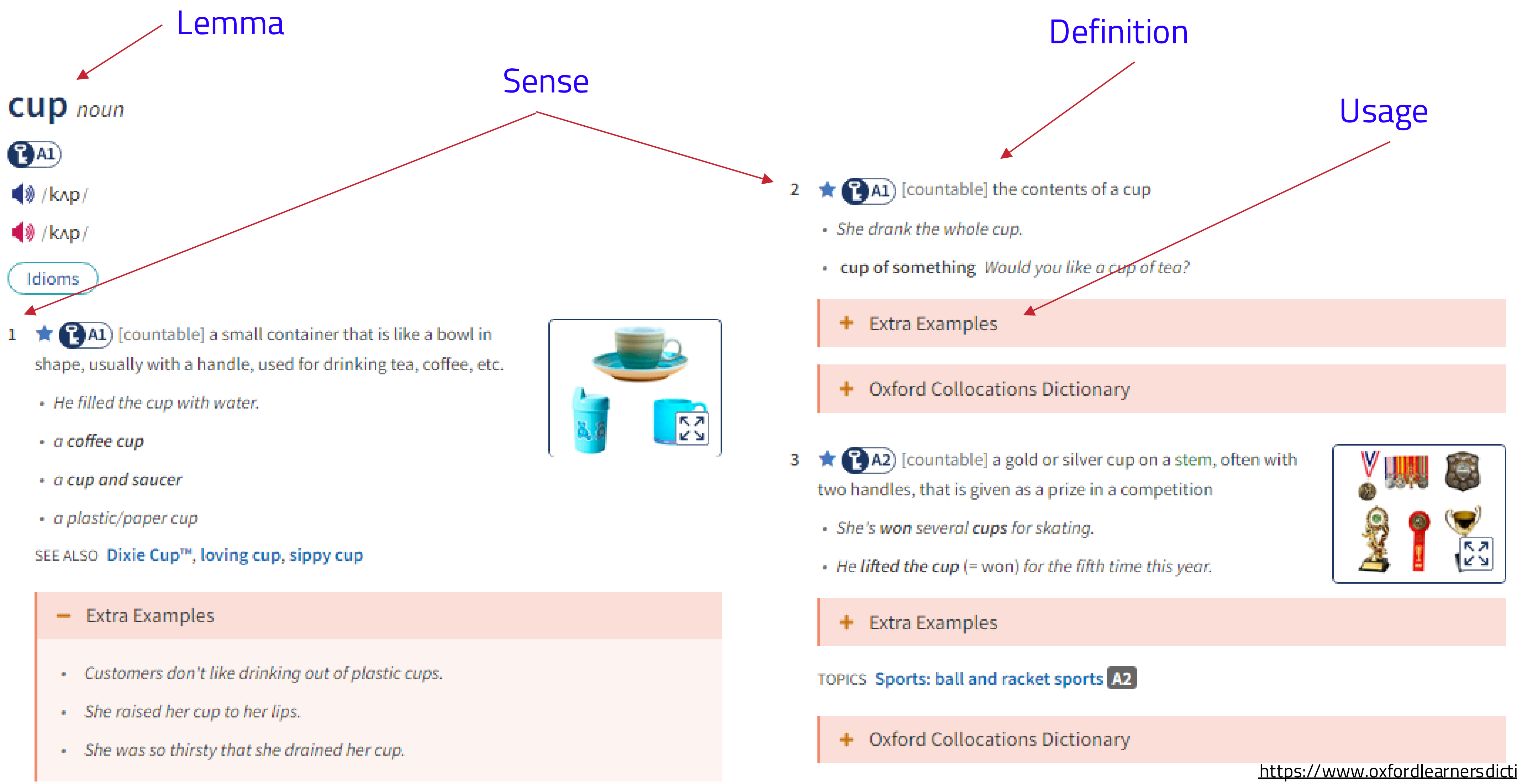
Taxonomy Enrichment with Text and Graph Vector Representations

# ImageNet based on WordNet



L. Fei-Fei and J. Deng. **ImageNet: Where have we been? Where are we going?**, *CVPR Beyond ImageNet Large Scale Visual Recognition Challenge workshop, 2017*, pdf

# Dictionary: A Database for Lexical Semantics



Lemma

Sense

Definition

Usage

**cup** *noun*

A1

/kʌp/

/kʌp/

Idioms

1 ★ A1 [countable] a small container that is like a bowl in shape, usually with a handle, used for drinking tea, coffee, etc.

- *He filled the cup with water.*
- *a* ***coffee cup***
- *a* ***cup and saucer***
- *a plastic/paper cup*

SEE ALSO **Dixie Cup™**, **loving cup**, **sippy cup**

− Extra Examples

- *Customers don't like drinking out of plastic cups.*
- *She raised her cup to her lips.*
- *She was so thirsty that she drained her cup.*

2 ★ A1 [countable] the contents of a cup

- *She drank the whole cup.*
- **cup of something** *Would you like a cup of tea?*

+ Extra Examples

+ Oxford Collocations Dictionary

3 ★ A2 [countable] a gold or silver cup on a stem, often with two handles, that is given as a prize in a competition

- *She's* ***won*** *several* ***cups*** *for skating.*
- *He* ***lifted the cup*** (= won) *for the fifth time this year.*

+ Extra Examples

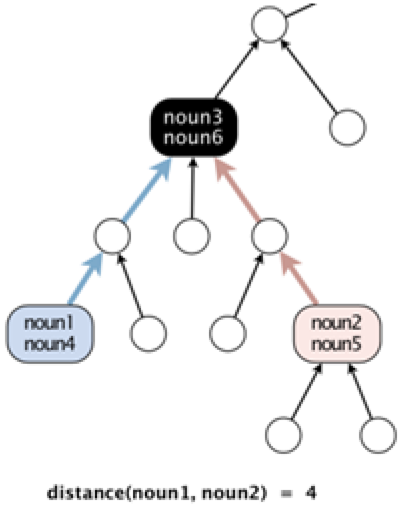
TOPICS **Sports: ball and racket sports** A2

+ Oxford Collocations Dictionary

https://www.oxfordlearnersdictionaries.com/

# Limitations of WordNet and ontological semantics

- WordNet is a useful resource, but there are intrinsic limits
  - It requires many years of manual effort by experts like skilled lexicographers
  - Some lexicographers are not skilled, and this has led to inconsistencies ([https://en.wikipedia.org/wiki/WordNet](https://en.wikipedia.org/wiki/WordNet))
  - Ontology is only as good as ontologists; not driven by data
- Sparse
  - wicked, sick, badass, ninja..
- Hard to compute word relationships



Let's look into lexical semantics driven by data and not rely on lexicographers

# Three ways of looking at word meaning

- Decompositional
    - What characteristics/components of what the word represents
- Ontological
    - How the meaning of the word relates to the meanings of other words
- **Distributional**
    - What contexts the word is found in, relative to other words

# Assumptions in distributional semantics

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

Wittgenstein PI 43

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

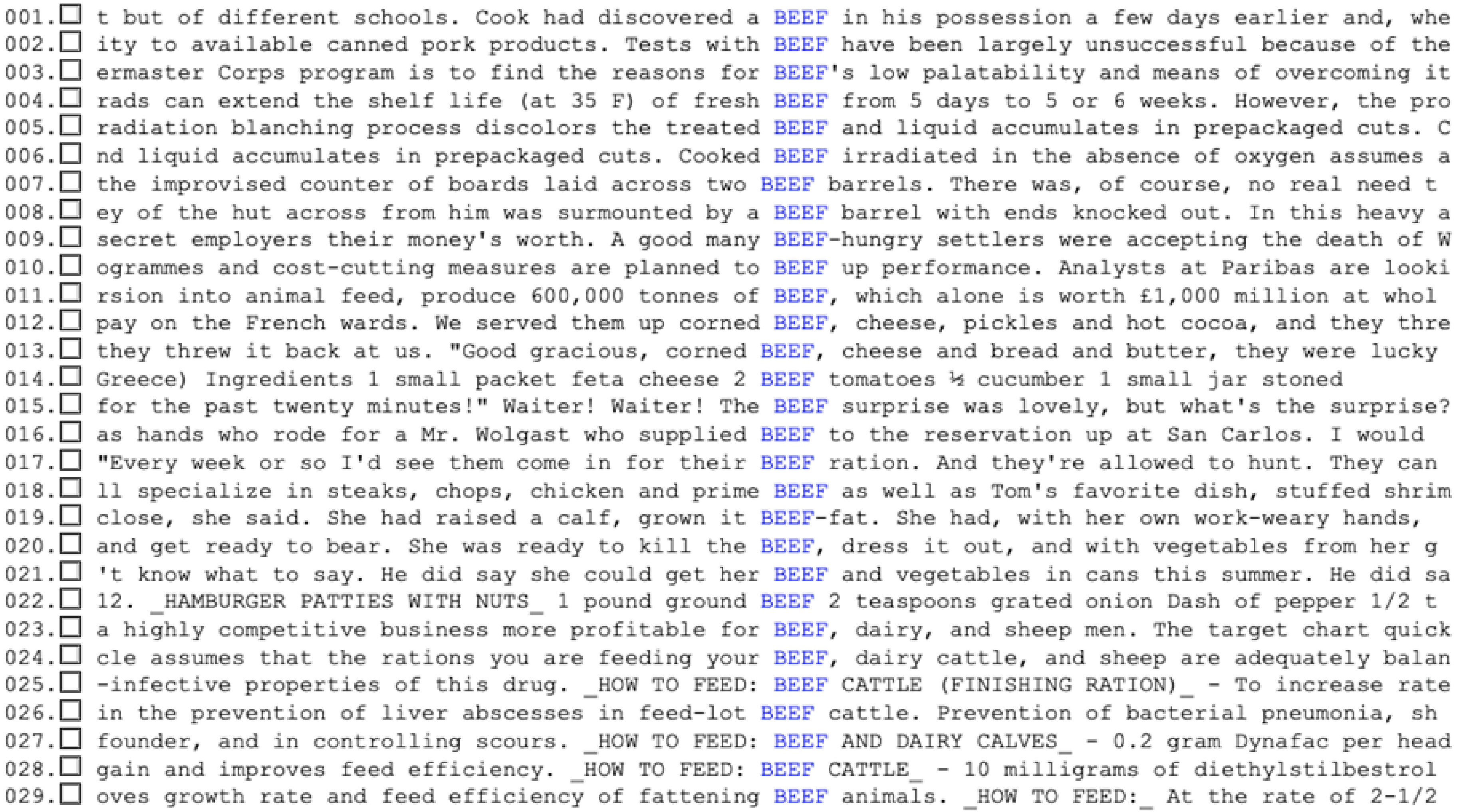
Firth, J. R. 1957:11

"If A and B have almost identical environments we say that they are synonyms."

Harris 1954

What does “beef” mean?

# Beef



```
001.☐ t but of different schools. Cook had discovered a BEEF in his possession a few days earlier and, whe
002.☐ ity to available canned pork products. Tests with BEEF have been largely unsuccessful because of the
003.☐ ermaster Corps program is to find the reasons for BEEF's low palatability and means of overcoming it
004.☐ rads can extend the shelf life (at 35 F) of fresh BEEF from 5 days to 5 or 6 weeks. However, the pro
005.☐ radiation blanching process discolors the treated BEEF and liquid accumulates in prepackaged cuts. C
006.☐ nd liquid accumulates in prepackaged cuts. Cooked BEEF irradiated in the absence of oxygen assumes a
007.☐ the improvised counter of boards laid across two BEEF barrels. There was, of course, no real need t
008.☐ ey of the hut across from him was surmounted by a BEEF barrel with ends knocked out. In this heavy a
009.☐ secret employers their money's worth. A good many BEEF-hungry settlers were accepting the death of W
010.☐ ogrammes and cost-cutting measures are planned to BEEF up performance. Analysts at Paribas are looki
011.☐ rsion into animal feed, produce 600,000 tonnes of BEEF, which alone is worth £1,000 million at whol
012.☐ pay on the French wards. We served them up corned BEEF, cheese, pickles and hot cocoa, and they thre
013.☐ they threw it back at us. "Good gracious, corned BEEF, cheese and bread and butter, they were lucky
014.☐ Greece) Ingredients 1 small packet feta cheese 2 BEEF tomatoes ½ cucumber 1 small jar stoned
015.☐ for the past twenty minutes!" Waiter! Waiter! The BEEF surprise was lovely, but what's the surprise?
016.☐ as hands who rode for a Mr. Wolgast who supplied BEEF to the reservation up at San Carlos. I would
017.☐ "Every week or so I'd see them come in for their BEEF ration. And they're allowed to hunt. They can
018.☐ ll specialize in steaks, chops, chicken and prime BEEF as well as Tom's favorite dish, stuffed shrim
019.☐ close, she said. She had raised a calf, grown it BEEF-fat. She had, with her own work-weary hands,
020.☐ and get ready to bear. She was ready to kill the BEEF, dress it out, and with vegetables from her g
021.☐ 't know what to say. He did say she could get her BEEF and vegetables in cans this summer. He did sa
022.☐ 12. _HAMBURGER PATTIES WITH NUTS_ 1 pound ground BEEF 2 teaspoons grated onion Dash of pepper 1/2 t
023.☐ a highly competitive business more profitable for BEEF, dairy, and sheep men. The target chart quick
024.☐ cle assumes that the rations you are feeding your BEEF, dairy cattle, and sheep are adequately balan
025.☐ -infective properties of this drug. _HOW TO FEED: BEEF CATTLE (FINISHING RATION)_ - To increase rate
026.☐ in the prevention of liver abscesses in feed-lot BEEF cattle. Prevention of bacterial pneumonia, sh
027.☐ founder, and in controlling scours. _HOW TO FEED: BEEF AND DAIRY CALVES_ - 0.2 gram Dynafac per head
028.☐ gain and improves feed efficiency. _HOW TO FEED: BEEF CATTLE_ - 10 milligrams of diethylstilbestrol
029.☐ oves growth rate and feed efficiency of fattening BEEF animals. _HOW TO FEED:_ At the rate of 2-1/2
```

Sentences from the brown corpus. Extracted from the concordancer in The Compleat Lexical Tutor, http://www.lextutor.ca/

# Beef

- Suppose you see these sentences:
  - "there was the smell of **roast** beef"
  - "I smell **seared** beef, and my stomach twists from hunger."

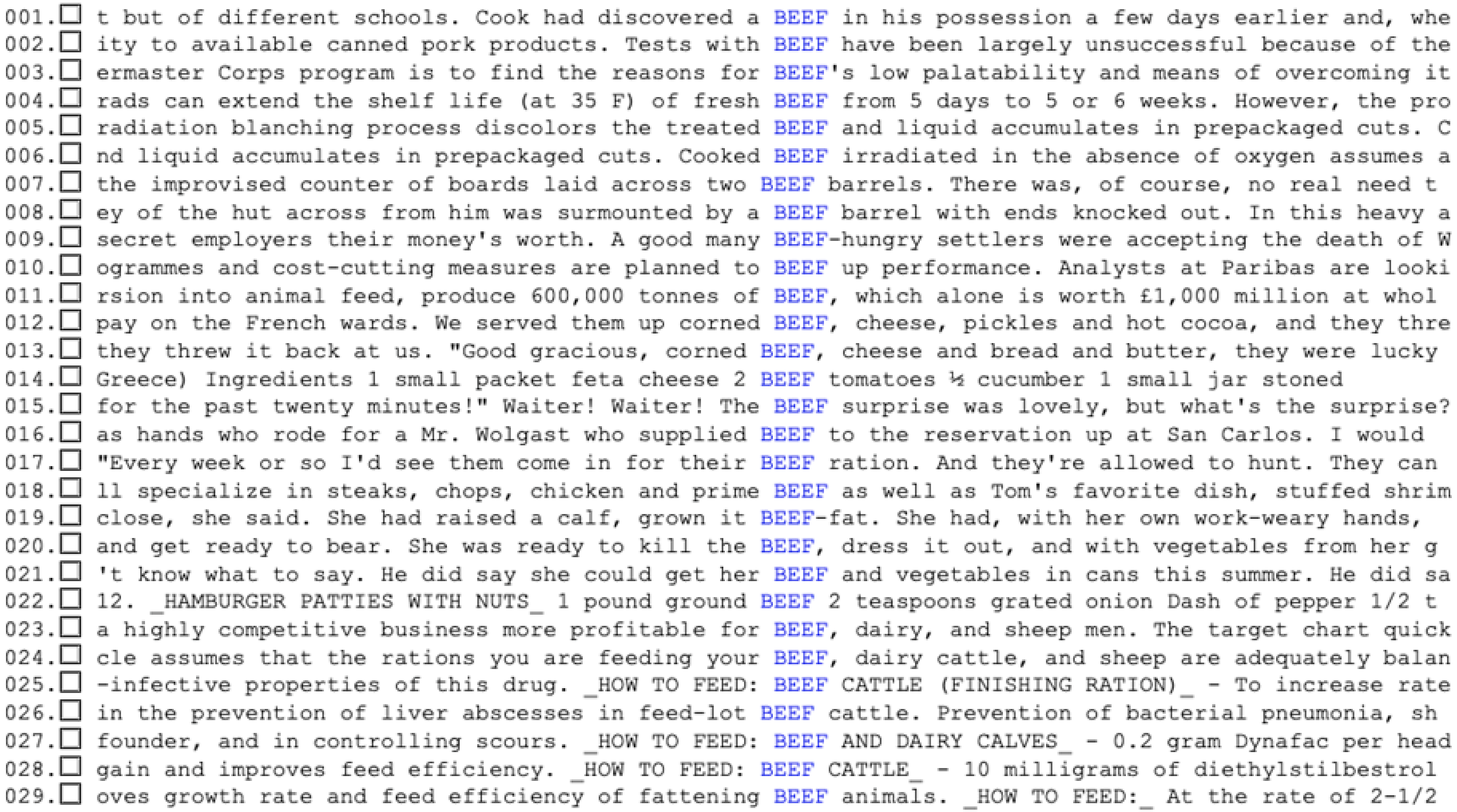
- Suppose you've also seen these:
  - ..a piece of meat **roasted** or for roasting..
  - ..So long as the meat was **seared**, they gulped down several times their own weight every day, ...

Beef is a meat **roasted** or **seared** with smell.

Beef is the culinary name for meat from cattle.

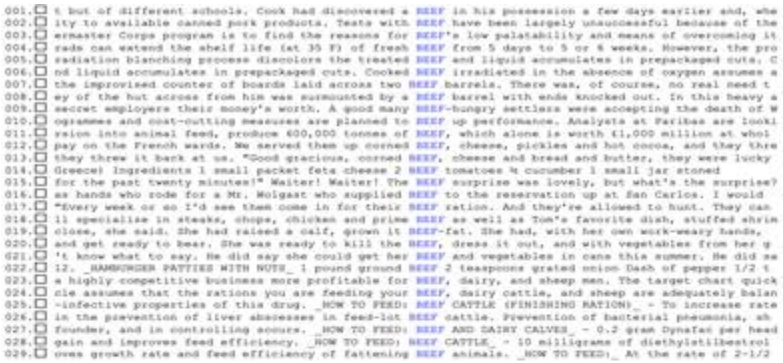
Wikipedia, "beef"

# Beef

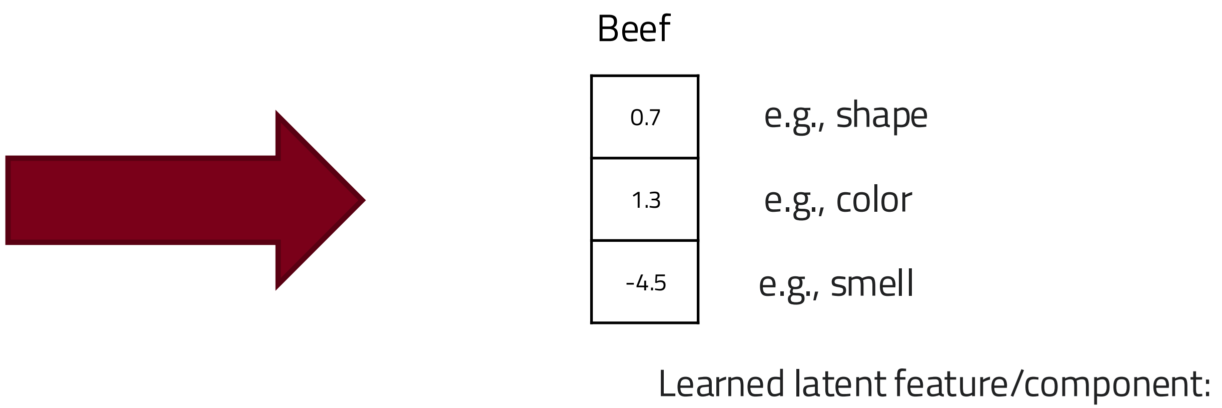


```
001.☐ t but of different schools. Cook had discovered a BEEF in his possession a few days earlier and, whe
002.☐ ity to available canned pork products. Tests with BEEF have been largely unsuccessful because of the
003.☐ ermaster Corps program is to find the reasons for BEEF's low palatability and means of overcoming it
004.☐ rads can extend the shelf life (at 35 F) of fresh BEEF from 5 days to 5 or 6 weeks. However, the pro
005.☐ radiation blanching process discolors the treated BEEF and liquid accumulates in prepackaged cuts. C
006.☐ nd liquid accumulates in prepackaged cuts. Cooked BEEF irradiated in the absence of oxygen assumes a
007.☐ the improvised counter of boards laid across two BEEF barrels. There was, of course, no real need t
008.☐ ey of the hut across from him was surmounted by a BEEF barrel with ends knocked out. In this heavy a
009.☐ secret employers their money's worth. A good many BEEF-hungry settlers were accepting the death of W
010.☐ ogrammes and cost-cutting measures are planned to BEEF up performance. Analysts at Paribas are looki
011.☐ rsion into animal feed, produce 600,000 tonnes of BEEF, which alone is worth £1,000 million at whol
012.☐ pay on the French wards. We served them up corned BEEF, cheese, pickles and hot cocoa, and they thre
013.☐ they threw it back at us. "Good gracious, corned BEEF, cheese and bread and butter, they were lucky
014.☐ Greece) Ingredients 1 small packet feta cheese 2 BEEF tomatoes ½ cucumber 1 small jar stoned
015.☐ for the past twenty minutes!" Waiter! Waiter! The BEEF surprise was lovely, but what's the surprise?
016.☐ as hands who rode for a Mr. Wolgast who supplied BEEF to the reservation up at San Carlos. I would
017.☐ "Every week or so I'd see them come in for their BEEF ration. And they're allowed to hunt. They can
018.☐ ll specialize in steaks, chops, chicken and prime BEEF as well as Tom's favorite dish, stuffed shrim
019.☐ close, she said. She had raised a calf, grown it BEEF-fat. She had, with her own work-weary hands,
020.☐ and get ready to bear. She was ready to kill the BEEF, dress it out, and with vegetables from her g
021.☐ 't know what to say. He did say she could get her BEEF and vegetables in cans this summer. He did sa
022.☐ 12. _HAMBURGER PATTIES WITH NUTS_ 1 pound ground BEEF 2 teaspoons grated onion Dash of pepper 1/2 t
023.☐ a highly competitive business more profitable for BEEF, dairy, and sheep men. The target chart quick
024.☐ cle assumes that the rations you are feeding your BEEF, dairy cattle, and sheep are adequately balan
025.☐ -infective properties of this drug. _HOW TO FEED: BEEF CATTLE (FINISHING RATION)_ - To increase rate
026.☐ in the prevention of liver abscesses in feed-lot BEEF cattle. Prevention of bacterial pneumonia, sh
027.☐ founder, and in controlling scours. _HOW TO FEED: BEEF AND DAIRY CALVES_ - 0.2 gram Dynafac per head
028.☐ gain and improves feed efficiency. _HOW TO FEED: BEEF CATTLE_ - 10 milligrams of diethylstilbestrol
029.☐ oves growth rate and feed efficiency of fattening BEEF animals. _HOW TO FEED:_ At the rate of 2-1/2
```

Sentences from the brown corpus. Extracted from the concordancer in The Compleat Lexical Tutor, http://www.lextutor.ca/



```
001.☐ t but of different schools. Cook had discovered a BEEF in his possession a few days earlier and, whe
002.☐ ity to available canned pork products. Tests with BEEF have been largely unsuccessful because of the
003.☐ ermaster Corps program is to find the reasons for BEEF's low palatability and means of overcoming it
004.☐ rads can extend the shelf life (at 35 F) of fresh BEEF from 5 days to 5 or 6 weeks. However, the pro
005.☐ radiation blanching process discolors the treated BEEF and liquid accumulates in prepackaged cuts. C
006.☐ nd liquid accumulates in prepackaged cuts. Cooked BEEF irradiated in the absence of oxygen assumes a
007.☐ the improvised counter of boards laid across two BEEF barrels. There was, of course, no real need t
008.☐ ey of the hut across from him was surmounted by a BEEF barrel with ends knocked out. In this heavy a
009.☐ secret employers their money's worth. A good many BEEF-hungry settlers were accepting the death of W
010.☐ ogrammes and cost-cutting measures are planned to BEEF up performance. Analysts at Paribas are looki
011.☐ rsion into animal feed, produce 600,000 tonnes of BEEF, which alone is worth £1,000 million at whol
012.☐ pay on the French wards. We served them up corned BEEF, cheese, pickles and hot cocoa, and they thre
013.☐ they threw it back at us. "Good gracious, corned BEEF, cheese and bread and butter, they were lucky
014.☐ Greece) Ingredients 1 small packet feta cheese 2 BEEF tomatoes ½ cucumber 1 small jar stoned
015.☐ for the past twenty minutes!" Waiter! Waiter! The BEEF surprise was lovely, but what's the surprise?
016.☐ as hands who rode for a Mr. Wolgast who supplied BEEF to the reservation up at San Carlos. I would
017.☐ "Every week or so I'd see them come in for their BEEF ration. And they're allowed to hunt. They can
018.☐ ll specialise in steaks, chops, chicken and prime BEEF as well as Tom's favorite dish, stuffed shrim
019.☐ close, she said. She had raised a calf, grown it BEEF-fat. She had, with her own work-weary hands,
020.☐ and get ready to bear. She was ready to kill the BEEF, dress it out, and with vegetables from her g
021.☐ 't know what to say. He did say she could get her BEEF and vegetables in cans this summer. He did se
022.☐ 12. _HAMBURGER PATTIES WITH NUTS_ 1 pound ground BEEF 2 teaspoons grated onion Dash of pepper 1/2 t
023.☐ a highly competitive business more profitable for BEEF, dairy, and sheep men. The target chart quick
024.☐ cle assumes that the rations you are feeding your BEEF, dairy cattle, and sheep are adequately balan
025.☐ -infective properties of this drug. _HOW TO FEED: BEEF CATTLE (FINISHING RATION)_ - To increase rate
026.☐ in the prevention of liver abscesses in feed-lot BEEF cattle. Prevention of bacterial pneumonia, sh
027.☐ founder, and in controlling scours. _HOW TO FEED: BEEF AND DAIRY CALVES_ - 0.2 gram Dynafac per head
028.☐ gain and improves feed efficiency. _HOW TO FEED: BEEF CATTLE_ - 10 milligrams of diethylstilbestrol
029.☐ oves growth rate and feed efficiency of fattening BEEF animals. _HOW TO FEED:_ At the rate of 2-1/2
```



Beef

| | |
|---|---|
| 0.7 | e.g., shape |
| 1.3 | e.g., color |
| -4.5 | e.g., smell |

Learned latent feature/component:

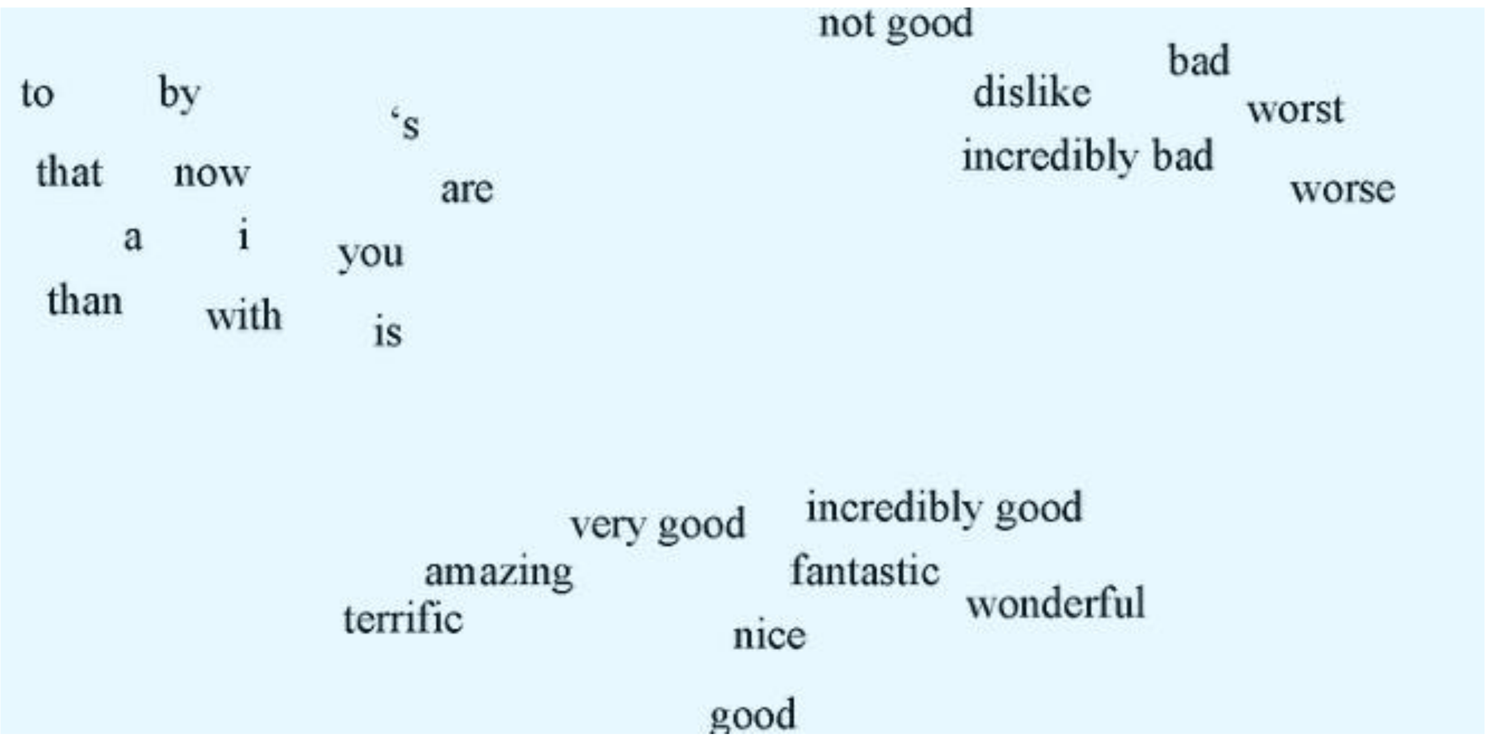
Learning a vector representation of "Beef" from context

# Model of meaning focusing on similarity

- Each word = a vector

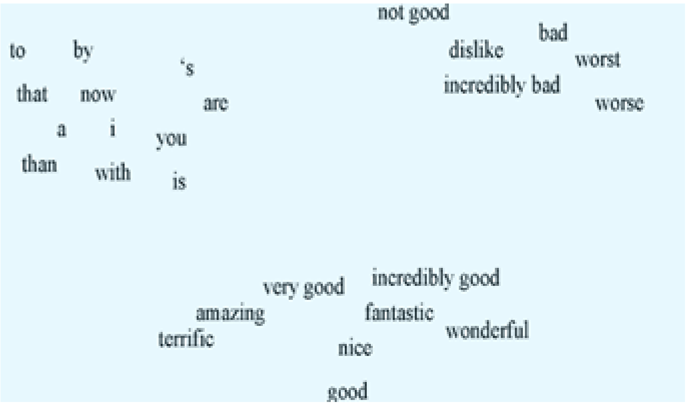
| |
|---|
| 0.7 |
| 1.3 |
| -4.5 |

- Not just "word"; Similar words are nearby in space
- The standard way to represent meaning in "NLP" these days

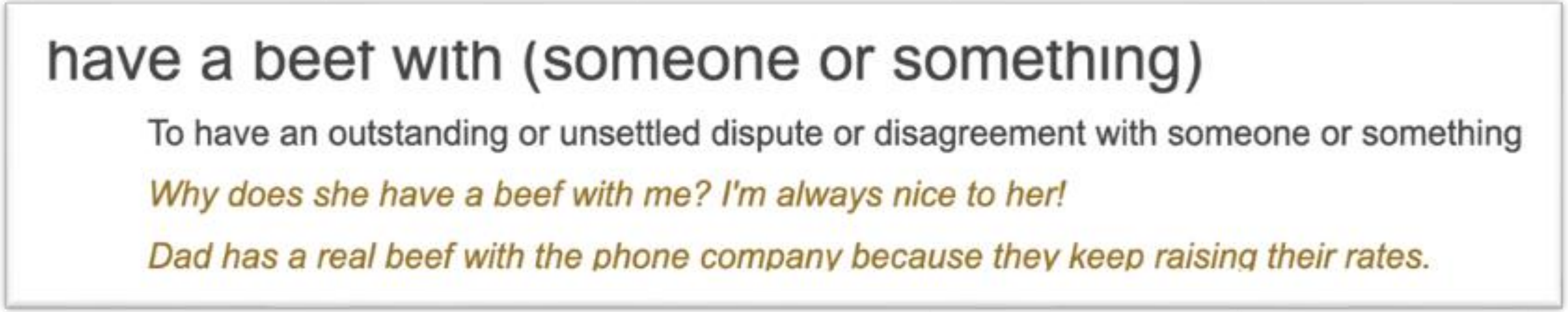


# Distributed representation

- Vector representation that encodes information about the distribution of contexts a word appears in
- Words that appear in similar contexts have similar representations (and similar meanings, by the distributional hypothesis).
- We have several different ways we can encode the notion of "context."



# Another example of polysemy

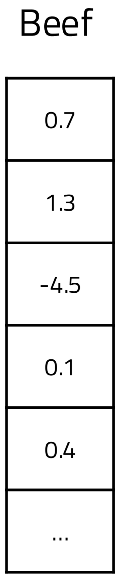


**have a beef with (someone or something)**

To have an outstanding or unsettled dispute or disagreement with someone or something

*Why does she have a beef with me? I'm always nice to her!*

*Dad has a real beef with the phone company because they keep raising their rates.*



Beef

0.7

1.3

-4.5

0.1

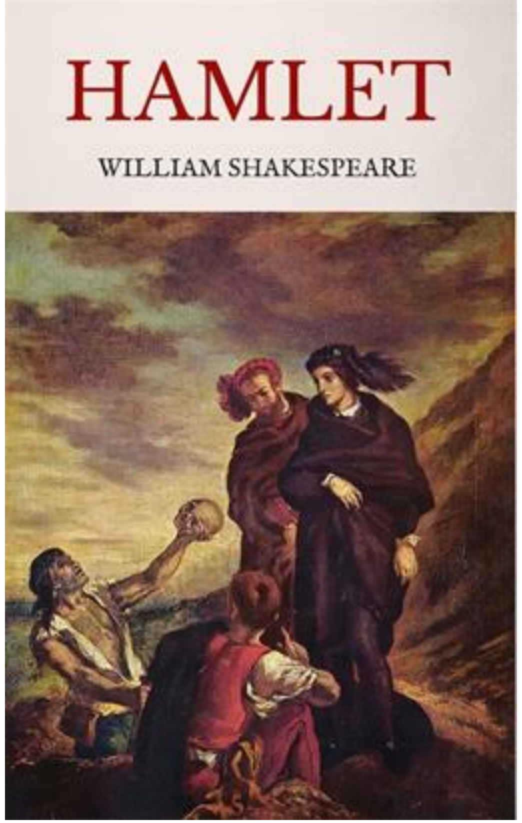
0.4

…

Polysemy should be learned to some extent, assuming that the word vector is trained using ***diverse textual contexts*** and the ***dimensions of the vector*** is large enough

# Different kinds of encoding “context”

- **Count-based**
  - PMI, TF-IDF
- Distributed prediction-based (type) embeddings
  - Word2vec, GloVe, Fasttext
- Distributed contextual (token) embeddings from language models
  - ELMo, BERT, GPT
- Many more variants
  - Multilingual / multi-sense / syntactic embeddings, etc


HAMLET
WILLIAM SHAKESPEARE

MICHAEL FASSBENDER MARION COTILLARD
MACBETH
A FILM BY JUSTIN KURZEL

Original Motion Picture Soundtrack
A Film By Franco Zeffirelli
Romeo & Juliet
Music by Nino Rota

# Term-document matrix

| | Hamlet | Macbeth | Romeo & Juliet | Richard III | Julius Caesar | Tempest |
|---|---|---|---|---|---|---|
| knife | 1 | 1 | 4 | 2 | | 2 |
| dog | | | | 6 | 12 | 2 |
| sword | 2 | 2 | 7 | 5 | | 5 |
| love | 64 | | 135 | 63 | | 12 |
| like | 75 | 38 | 34 | 36 | 34 | 41 |
| … | | | | | | |

Context = appearing in the same document.

# Word vectors

| | | | | | | |
|---|---|---|---|---|---|---|
| knife | 1 | 1 | 4 | 2 | | 2 |
| sword | 2 | 2 | 7 | 5 | | 5 |

Vector representation of the term;
vector size = number of documents

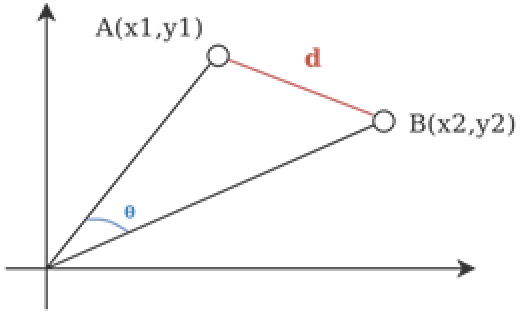
# Cosine Similarity

- ❑ Calculate the cosine similarity between the two word vectors, to judge the degree of their similarity [Salton 1971]

$$cos\,(x,y) = \frac{\sum_{i=1}^{F} x_i y_i}{\sqrt{\sum_{i=1}^{F} x_i^2}\,\sqrt{\sum_{i=1}^{F} y_i^2}}$$

Note:

- ❑ Euclidean distance measures the magnitude of distance between two points
- ❑ Cosine similarity measures their orientation



https://cmry.github.io/notes/euclidean-v-cosine

|  | Hamlet | Macbeth | Romeo & Juliet | Richard III | Julius Caesar | Tempest |
|---|---|---|---|---|---|---|
| **knife** | 1 | 1 | 4 | 2 |  | 2 |
| dog |  |  |  | 6 | 12 | 2 |
| **sword** | 2 | 2 | 7 | 5 |  | 5 |
| love | 64 |  | 135 | 63 |  | 12 |
| like | 75 | 38 | 34 | 36 | 34 | 41 |
| … |  |  |  |  |  |  |

|  |  |
|---|---|
| cos (knife, knife) | 1.0 |
| cos (knife, dog) | 0.11 |
| cos (knife, sword) | 0.99 |
| cos (knife, love) | 0.65 |
| cos (knife, like) | 0.61 |

Not all dimensions are equally informative.
Let's weight dimensions!

# TF-IDF

- Term frequency ($TF_{t,d}$) = the number of times terms $t$ occurs in document $d$
  - Several variants: e.g., passing through log function

- Inverse document frequency ($IDF_d$) = inverse function of number of documents containing ($D_t$) among total number of documents $N$.

$$tfidf\ (t, d) = tf_{t,d} \ \ \times log\frac{N}{D_t}$$

| | Hamlet | Macbeth | Romeo & Juliet | Richard III | Julius Caesar | Tempest |
|---|---|---|---|---|---|---|
| knife | 1 | 1 | 4 | 2 | | 2 |
| dog | | | | 6 | 12 | 2 |
| sword | 2 | 2 | 7 | 5 | | 5 |
| love | 64 | | 135 | 63 | | 12 |
| like | 75 | 38 | 34 | 36 | 34 | 41 |
| … | | | | | | |

| IDF |
|---|
| 0.07 |
| 0.30 |
| 0.07 |
| 0.20 |
| 0.00 |
| |

$$tfidf\ (t,d) = tf_{t,d} \ \times log\frac{N}{D_t}$$

IDF indicates the informativeness of the terms when comparing documents.

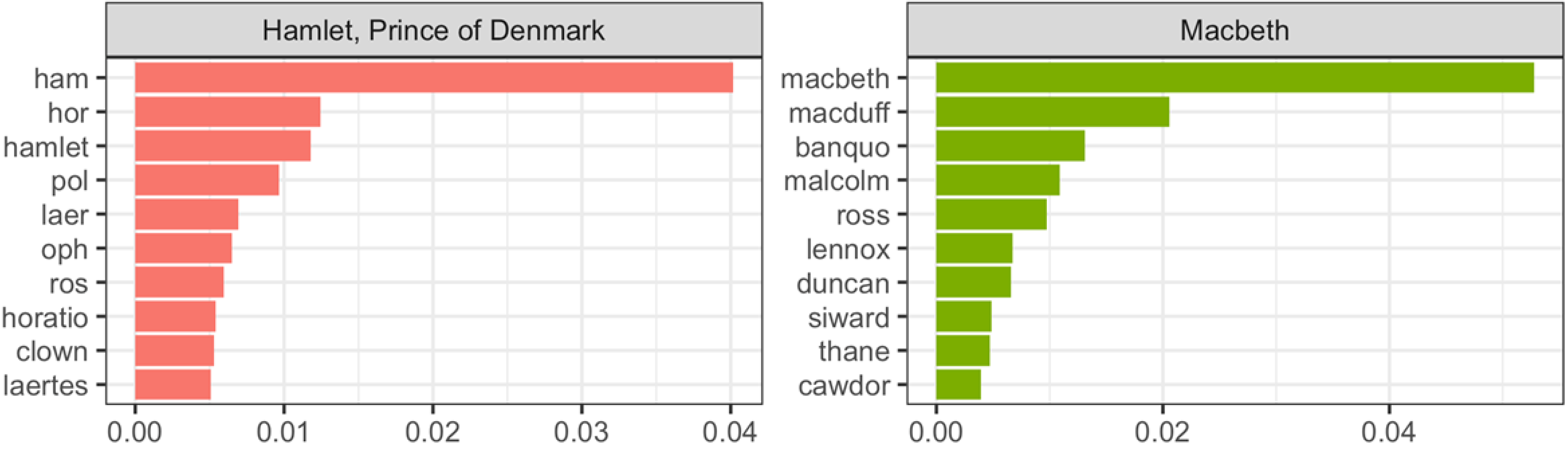
| | | | | | | |
|---|---|---|---|---|---|---|
| knife | 0.07 | 0.07 | 0.28 | 0.14 | 0 | 0.14 |
| dog | 0 | 0 | 0 | 1.8 | 3.6 | 0.6 |

| | Hamlet | Macbeth | Romeo & Juliet | Richard III | Julius Caesar | Tempest |
|---|---|---|---|---|---|---|
| knife | 1 | 1 | 4 | 2 | | 2 |
| dog | | | | 6 | 12 | 2 |
| sword | 2 | 2 | 7 | 5 | | 5 |
| love | 64 | | 135 | 63 | | 12 |
| like | 75 | 38 | 34 | 36 | 34 | 41 |
| … | | | | | | |

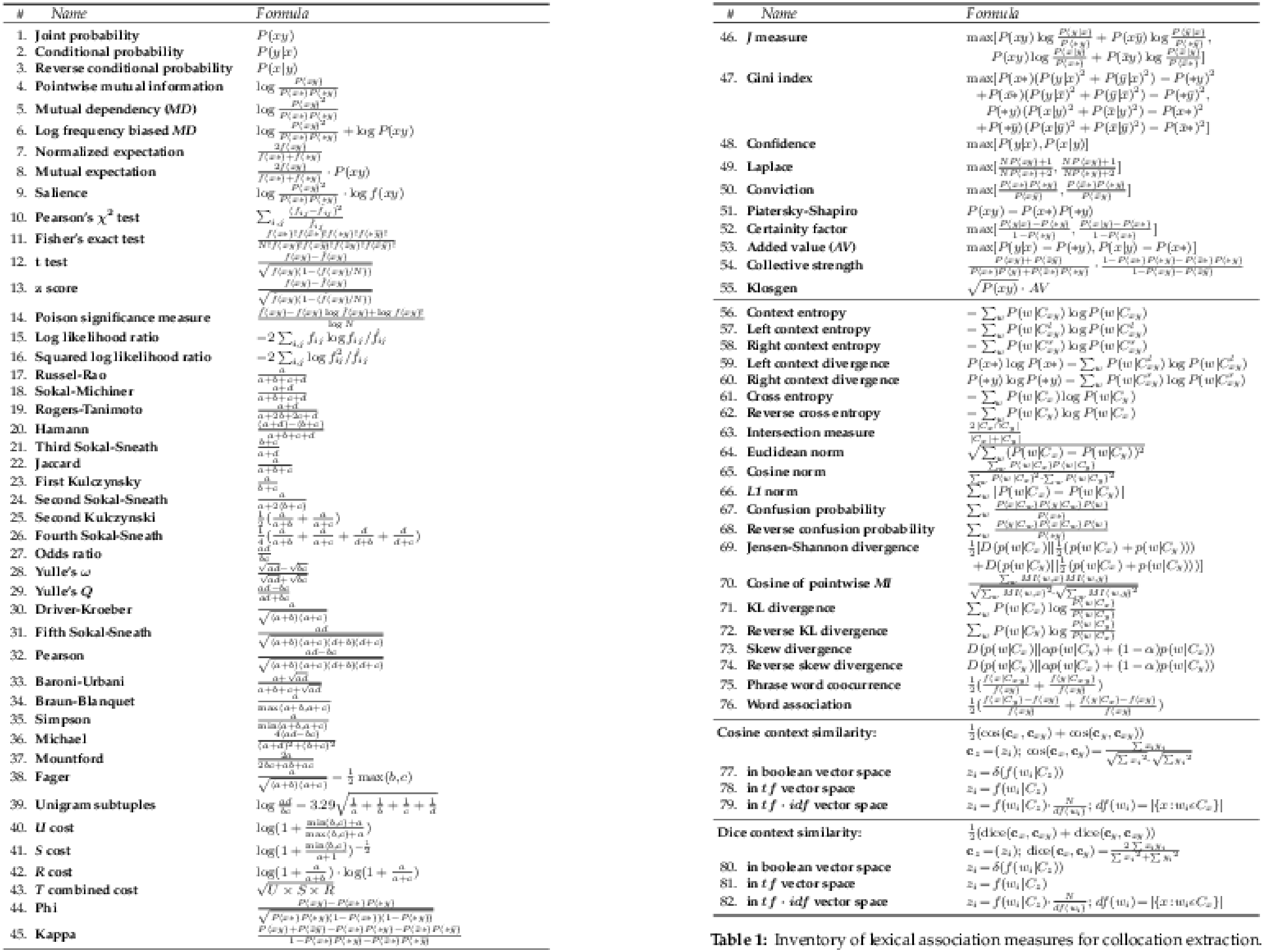
| IDF |
|---|
| 0.07 |
| 0.30 |
| 0.07 |
| 0.20 |
| 0.00 |
| |

$$tfidf\,(t,d) = tf_{t,d} \times log\frac{N}{D_t}$$

IDF indicates the informativeness of the terms when comparing documents.



https://datavizs21.classes.andrewheiss.com/example/13-example/



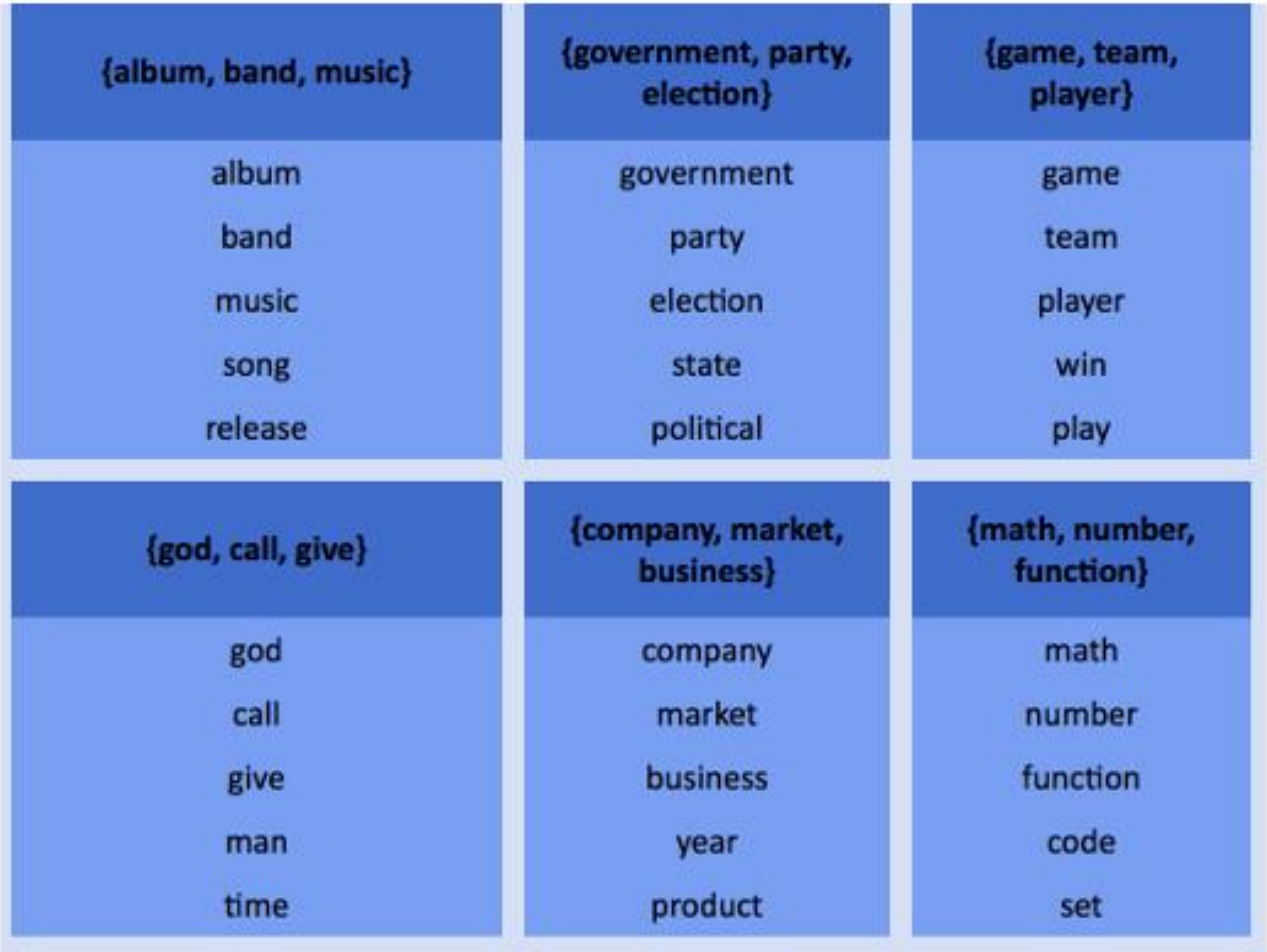
| # | *Name* | *Formula* |
|---|---|---|
| 1. | **Joint probability** | $P(xy)$ |
| 2. | **Conditional probability** | $P(y \mid x)$ |
| 3. | **Reverse conditional probability** | $P(x \mid y)$ |
| 4. | **Pointwise mutual information** | $\log \frac{P(xy)}{P(x*)P(*y)}$ |
| 5. | **Mutual dependency (*MD*)** | $\log \frac{P(xy)^2}{P(x*)P(*y)}$ |
| 6. | **Log frequency biased *MD*** | $\log \frac{P(xy)^2}{P(x*)P(*y)} + \log P(xy)$ |
| 7. | **Normalized expectation** | $\frac{2f(xy)}{f(x*)+f(*y)}$ |
| 8. | **Mutual expectation** | $\frac{2f(xy)}{f(x*)+f(*y)} \cdot P(xy)$ |
| 9. | **Salience** | $\log \frac{P(xy)^2}{P(x*)P(*y)} \cdot \log f(xy)$ |
| 10. | **Pearson's $\chi^2$ test** | $\sum_{i,j} \frac{(f_{ij}-\hat{f}_{ij})^2}{\hat{f}_{ij}}$ |
| 11. | **Fisher's exact test** | $\frac{f(x*)!f(\bar{x}*)!f(*y)!f(*\bar{y})!}{N!f(xy)!f(x\bar{y})!f(\bar{x}y)!f(\bar{x}\bar{y})!}$ |
| 12. | **t test** | $\frac{f(xy)-\hat{f}(xy)}{\sqrt{f(xy)(1-(f(xy)/N))}}$ |
| 13. | **z score** | $\frac{f(xy)-\hat{f}(xy)}{\sqrt{\hat{f}(xy)(1-(\hat{f}(xy)/N))}}$ |
| 14. | **Poison significance measure** | $\frac{\hat{f}(xy)-f(xy)\log \hat{f}(xy)+\log f(xy)!}{\log N}$ |
| 15. | **Log likelihood ratio** | $-2\sum_{i,j} f_{ij} \log f_{ij}/\hat{f}_{ij}$ |
| 16. | **Squared log likelihood ratio** | $-2\sum_{i,j} \log f_{ij}^2/\hat{f}_{ij}$ |
| 17. | **Russel-Rao** | $\frac{a}{a+b+c+d}$ |
| 18. | **Sokal-Michiner** | $\frac{a+d}{a+b+c+d}$ |
| 19. | **Rogers-Tanimoto** | $\frac{a+d}{a+2b+2c+d}$ |
| 20. | **Hamann** | $\frac{(a+d)-(b+c)}{a+b+c+d}$ |
| 21. | **Third Sokal-Sneath** | $\frac{b+c}{a+d}$ |
| 22. | **Jaccard** | $\frac{a}{a+b+c}$ |
| 23. | **First Kulczynsky** | $\frac{a}{b+c}$ |
| 24. | **Second Sokal-Sneath** | $\frac{a}{a+2(b+c)}$ |
| 25. | **Second Kulczynski** | $\frac{1}{2}(\frac{a}{a+b}+\frac{a}{a+c})$ |
| 26. | **Fourth Sokal-Sneath** | $\frac{1}{4}(\frac{a}{a+b}+\frac{a}{a+c}+\frac{d}{d+b}+\frac{d}{d+c})$ |
| 27. | **Odds ratio** | $\frac{ad}{bc}$ |
| 28. | **Yulle's $\omega$** | $\frac{\sqrt{ad}-\sqrt{bc}}{\sqrt{ad}+\sqrt{bc}}$ |
| 29. | **Yulle's *Q*** | $\frac{ad-bc}{ad+bc}$ |
| 30. | **Driver-Kroeber** | $\frac{a}{\sqrt{(a+b)(a+c)}}$ |
| 31. | **Fifth Sokal-Sneath** | $\frac{ad}{\sqrt{(a+b)(a+c)(d+b)(d+c)}}$ |
| 32. | **Pearson** | $\frac{ad-bc}{\sqrt{(a+b)(a+c)(d+b)(d+c)}}$ |
| 33. | **Baroni-Urbani** | $\frac{a+\sqrt{ad}}{a+b+c+\sqrt{ad}}$ |
| 34. | **Braun-Blanquet** | $\frac{a}{\max(a+b,a+c)}$ |
| 35. | **Simpson** | $\frac{a}{\min(a+b,a+c)}$ |
| 36. | **Michael** | $\frac{4(ad-bc)}{(a+d)^2+(b+c)^2}$ |
| 37. | **Mountford** | $\frac{2a}{2bc+ab+ac}$ |
| 38. | **Fager** | $\frac{a}{\sqrt{(a+b)(a+c)}}-\frac{1}{2}\max(b,c)$ |
| 39. | **Unigram subtuples** | $\log\frac{ad}{bc}-3.29\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}$ |
| 40. | ***U* cost** | $\log(1+\frac{\min(b,c)+a}{\max(b,c)+a})$ |
| 41. | ***S* cost** | $\log(1+\frac{\min(b,c)}{a+1})^{-\frac{1}{2}}$ |
| 42. | ***R* cost** | $\log(1+\frac{a}{a+b})\cdot\log(1+\frac{a}{a+c})$ |
| 43. | ***T* combined cost** | $\sqrt{U \times S \times R}$ |
| 44. | **Phi** | $\frac{P(xy)-P(x*)P(*y)}{\sqrt{P(x*)P(*y)(1-P(x*))(1-P(*y))}}$ |
| 45. | **Kappa** | $\frac{P(xy)+P(\bar{x}\bar{y})-P(x*)P(*y)-P(\bar{x}*)P(*\bar{y})}{1-P(x*)P(*y)-P(\bar{x}*)P(*\bar{y})}$ |
| 46. | ***J* measure** | $\max[P(xy)\log\frac{P(y \mid x)}{P(*y)}+P(x\bar{y})\log\frac{P(\bar{y} \mid x)}{P(*\bar{y})},$<br>$P(xy)\log\frac{P(x \mid y)}{P(x*)}+P(\bar{x}y)\log\frac{P(\bar{x} \mid y)}{P(\bar{x}*)}]$ |
| 47. | **Gini index** | $\max[P(x*)(P(y \mid x)^2+P(\bar{y} \mid x)^2)-P(*y)^2$<br>$+P(\bar{x}*)(P(y \mid \bar{x})^2+P(\bar{y} \mid \bar{x})^2)-P(*\bar{y})^2,$<br>$P(*y)(P(x \mid y)^2+P(\bar{x} \mid y)^2)-P(x*)^2$<br>$+P(*\bar{y})(P(x \mid \bar{y})^2+P(\bar{x} \mid \bar{y})^2)-P(\bar{x}*)^2]$ |
| 48. | **Confidence** | $\max[P(y \mid x),P(x \mid y)]$ |
| 49. | **Laplace** | $\max[\frac{NP(xy)+1}{NP(x*)+2},\frac{NP(xy)+1}{NP(*y)+2}]$ |
| 50. | **Conviction** | $\max[\frac{P(x*)P(*\bar{y})}{P(x\bar{y})},\frac{P(\bar{x}*)P(*y)}{P(\bar{x}y)}]$ |
| 51. | **Piatersky-Shapiro** | $P(xy)-P(x*)P(*y)$ |
| 52. | **Certainity factor** | $\max[\frac{P(y \mid x)-P(*y)}{1-P(*y)},\frac{P(x \mid y)-P(x*)}{1-P(x*)}]$ |
| 53. | **Added value (*AV*)** | $\max[P(y \mid x)-P(*y),P(x \mid y)-P(x*)]$ |
| 54. | **Collective strength** | $\frac{P(xy)+P(\bar{x}\bar{y})}{P(x*)P(y)+P(\bar{x}*)P(*\bar{y})}\cdot\frac{1-P(x*)P(*y)-P(\bar{x}*)P(*y)}{1-P(xy)-P(\bar{x}\bar{y})}$ |
| 55. | **Klosgen** | $\sqrt{P(xy)}\cdot AV$ |
| 56. | **Context entropy** | $-\sum_w P(w \mid C_{xy})\log P(w \mid C_{xy})$ |
| 57. | **Left context entropy** | $-\sum_w P(w \mid C^l_{xy})\log P(w \mid C^l_{xy})$ |
| 58. | **Right context entropy** | $-\sum_w P(w \mid C^r_{xy})\log P(w \mid C^r_{xy})$ |
| 59. | **Left context divergence** | $P(x*)\log P(x*)-\sum_w P(w \mid C^l_{xy})\log P(w \mid C^l_{xy})$ |
| 60. | **Right context divergence** | $P(*y)\log P(*y)-\sum_w P(w \mid C^r_{xy})\log P(w \mid C^r_{xy})$ |
| 61. | **Cross entropy** | $-\sum_w P(w \mid C_x)\log P(w \mid C_y)$ |
| 62. | **Reverse cross entropy** | $-\sum_w P(w \mid C_y)\log P(w \mid C_x)$ |
| 63. | **Intersection measure** | $\frac{2\lvert C_x \cap C_y\rvert}{\lvert C_x\rvert+\lvert C_y\rvert}$ |
| 64. | **Euclidean norm** | $\sqrt{\sum_w(P(w \mid C_x)-P(w \mid C_y))^2}$ |
| 65. | **Cosine norm** | $\frac{\sum_w P(w \mid C_x)P(w \mid C_y)}{\sum_w P(w \mid C_x)^2\cdot\sum_w P(w \mid C_y)^2}$ |
| 66. | ***L1* norm** | $\sum_w \lvert P(w \mid C_x)-P(w \mid C_y)\rvert$ |
| 67. | **Confusion probability** | $\sum_w \frac{P(x \mid C_w)P(y \mid C_w)P(w)}{P(x*)}$ |
| 68. | **Reverse confusion probability** | $\sum_w \frac{P(y \mid C_w)P(x \mid C_w)P(w)}{P(*y)}$ |
| 69. | **Jensen-Shannon divergence** | $\frac{1}{2}[D(p(w \mid C_x)\|\frac{1}{2}(p(w \mid C_x)+p(w \mid C_y)))$<br>$+D(p(w \mid C_y)\|\frac{1}{2}(p(w \mid C_x)+p(w \mid C_y)))]$ |
| 70. | **Cosine of pointwise *MI*** | $\frac{\sum_w MI(w,x)MI(w,y)}{\sqrt{\sum_w MI(w,x)^2}\cdot\sqrt{\sum_w MI(w,y)^2}}$ |
| 71. | **KL divergence** | $\sum_w P(w \mid C_x)\log\frac{P(w \mid C_x)}{P(w \mid C_y)}$ |
| 72. | **Reverse KL divergence** | $\sum_w P(w \mid C_y)\log\frac{P(w \mid C_y)}{P(w \mid C_x)}$ |
| 73. | **Skew divergence** | $D(p(w \mid C_x)\|\alpha p(w \mid C_y)+(1-\alpha)p(w \mid C_x))$ |
| 74. | **Reverse skew divergence** | $D(p(w \mid C_y)\|\alpha p(w \mid C_x)+(1-\alpha)p(w \mid C_y))$ |
| 75. | **Phrase word cooccurrence** | $\frac{1}{2}(\frac{f(x \mid C_{xy})}{f(xy)}+\frac{f(y \mid C_{xy})}{f(xy)})$ |
| 76. | **Word association** | $\frac{1}{2}(\frac{f(x \mid C_y)-f(xy)}{f(xy)}+\frac{f(y \mid C_x)-f(xy)}{f(xy)})$ |
| | **Cosine context similarity:** | $\frac{1}{2}(\cos(\mathbf{c}_x,\mathbf{c}_{xy})+\cos(\mathbf{c}_y,\mathbf{c}_{xy}))$<br>$\mathbf{c}_z=(z_i);\ \cos(\mathbf{c}_x,\mathbf{c}_y)=\frac{\sum x_i y_i}{\sqrt{\sum x_i^2}\cdot\sqrt{\sum y_i^2}}$ |
| 77. | **in boolean vector space** | $z_i=\delta(f(w_i \mid C_z))$ |
| 78. | **in *t f* vector space** | $z_i=f(w_i \mid C_z)$ |
| 79. | **in *t f* · *idf* vector space** | $z_i=f(w_i \mid C_z)\cdot\frac{N}{df(w_i)};\ df(w_i)=\lvert\{x:w_i \in C_x\}\rvert$ |
| | **Dice context similarity:** | $\frac{1}{2}(\mathrm{dice}(\mathbf{c}_x,\mathbf{c}_{xy})+\mathrm{dice}(\mathbf{c}_y,\mathbf{c}_{xy}))$<br>$\mathbf{c}_z=(z_i);\ \mathrm{dice}(\mathbf{c}_x,\mathbf{c}_y)=\frac{2\sum x_i y_i}{\sum x_i^2+\sum y_i^2}$ |
| 80. | **in boolean vector space** | $z_i=\delta(f(w_i \mid C_z))$ |
| 81. | **in *t f* vector space** | $z_i=f(w_i \mid C_z)$ |
| 82. | **in *t f* · *idf* vector space** | $z_i=f(w_i \mid C_z)\cdot\frac{N}{df(w_i)};\ df(w_i)=\lvert\{x:w_i \in C_x\}\rvert$ |

**Table 1:** Inventory of lexical association measures for collocation extraction.

Pecina, P. (2008). Lexical Association Measures: Collocation Extraction.

# (Optional) Topic Modeling

- A probabilistic model for discovering hidden “topics” (groups of terms that tend to occur together) in documents.
- Unsupervised (find *interesting structure* in the data)
- Clustering tokens into topics



| {album, band, music} | {government, party, election} | {game, team, player} |
|---|---|---|
| album | government | game |
| band | party | team |
| music | election | player |
| song | state | win |
| release | political | play |

| {god, call, give} | {company, market, business} | {math, number, function} |
|---|---|---|
| god | company | math |
| call | market | number |
| give | business | function |
| man | year | code |
| time | product | set |

# Topic Modeling

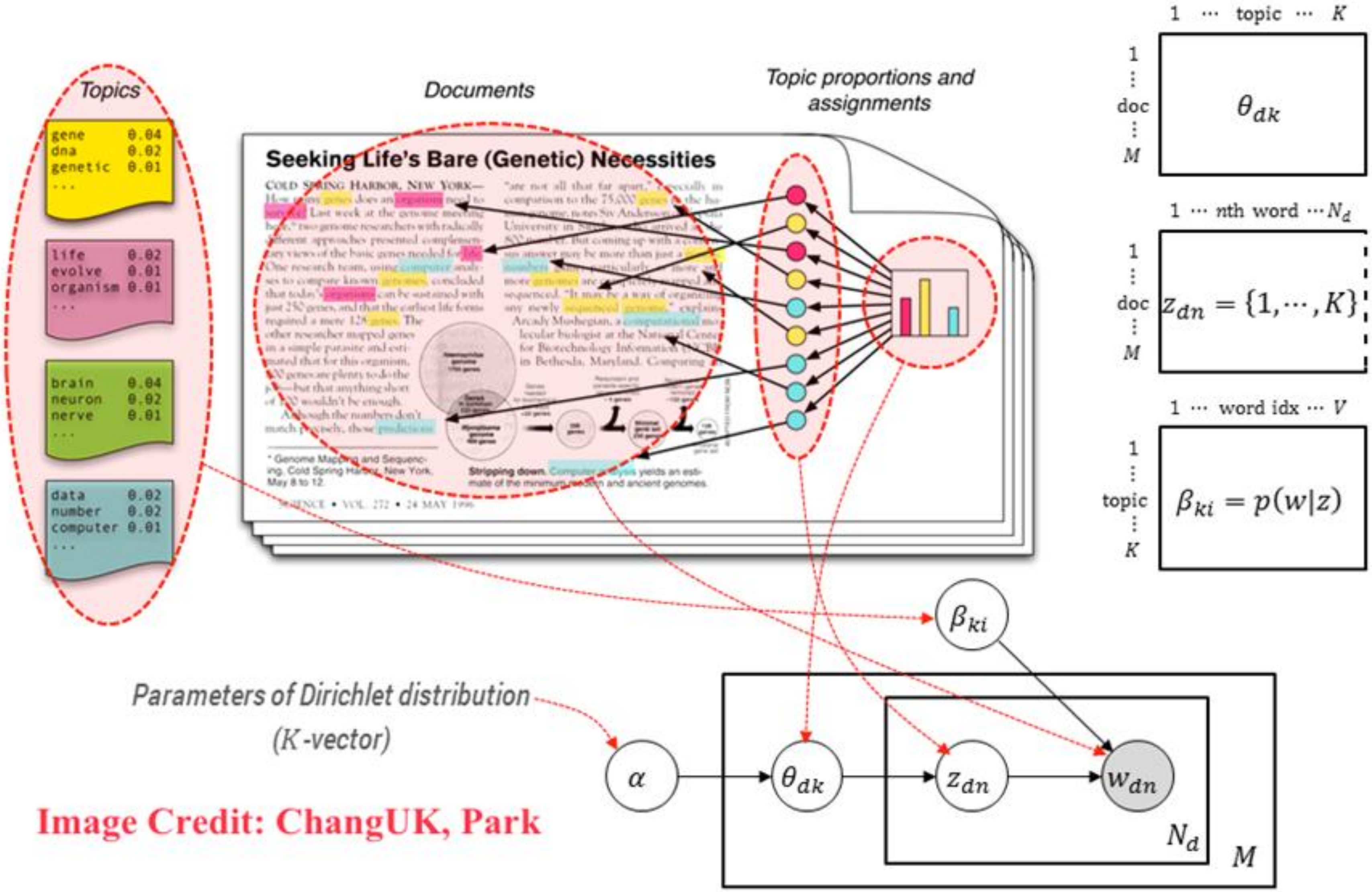


**Image Credit: ChangUK, Park**

# Different kinds of encoding “context”

- ❑ ~~Count-based~~
  - ○ PMI, TF-IDF
- ❑ **Distributed prediction-based (type) embeddings**
  - ○ Word2vec, GloVe, Fasttext
- ❑ **Distributed contextual (token) embeddings from language models**
  - ○ ELMo, BERT, GPT
- ❑ Many more variants
  - ○ Multilingual / multi-sense / syntactic embeddings, etc

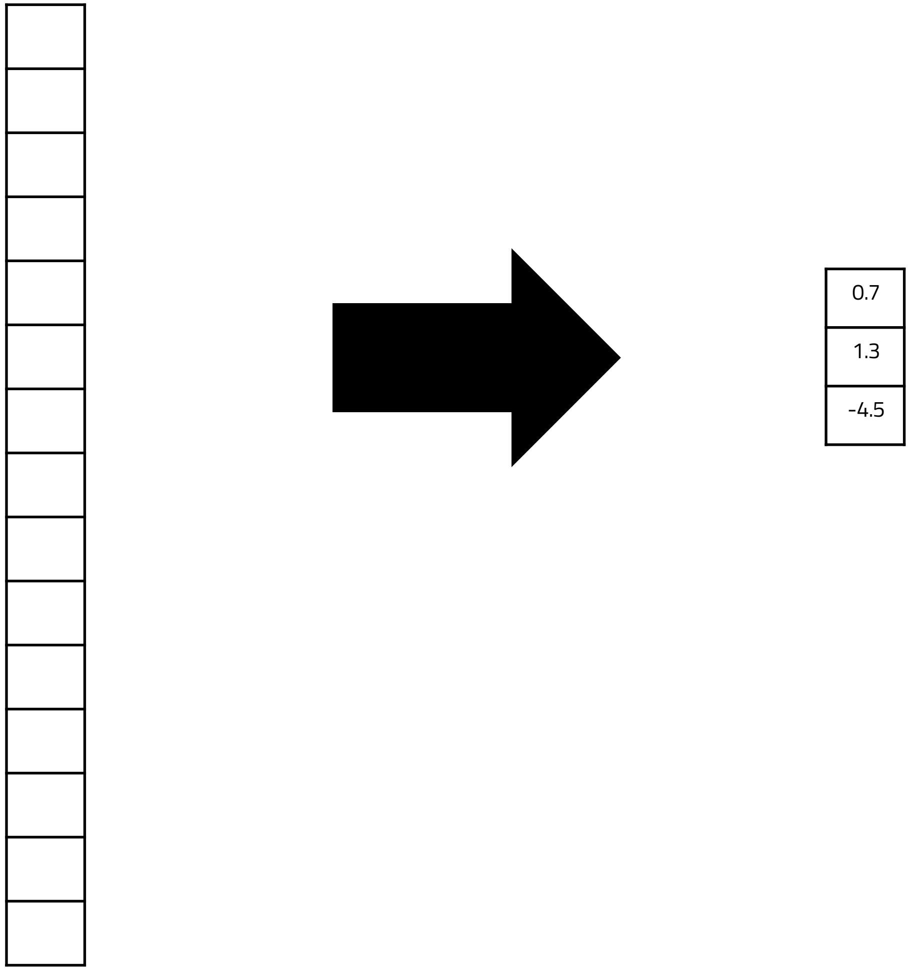
# Sparse vectors

“aardvark”

V-dimensional vector, single 1 for the identity of the element

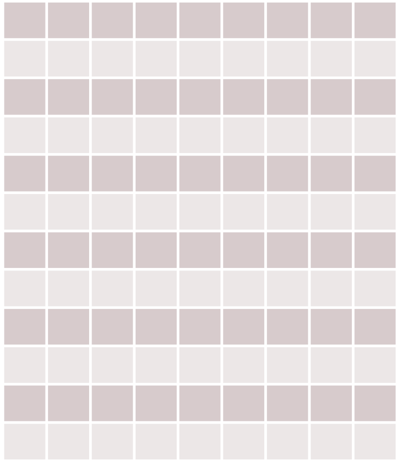
| | |
|---|---|
| a | 0 |
| a | 0 |
| aa | 0 |
| aal | 0 |
| aalii | 0 |
| aam | 0 |
| Aani | 0 |
| aardvark | 1 |
| aardwolf | 0 |
| ... | |
| zythem | 0 |
| Zythia | 0 |
| zythum | 0 |
| Zyzomys | 0 |
| Zyzzogeton | 0 |

# Sparse vectors -> Dense vectors



|  | Hamlet | Macbeth | Romeo & Juliet | Richard III | Julius Caesar | Tempest |
|---|---|---|---|---|---|---|
| **knife** | 1 | 1 | 4 | 2 |  | 2 |
| dog |  |  |  | 6 | 12 | 2 |
| **sword** | 2 | 2 | 7 | 5 |  | 5 |
| love | 64 |  | 135 | 63 |  | 12 |
| like | 75 | 38 | 34 | 36 | 34 | 41 |
| … |  |  |  |  |  |  |

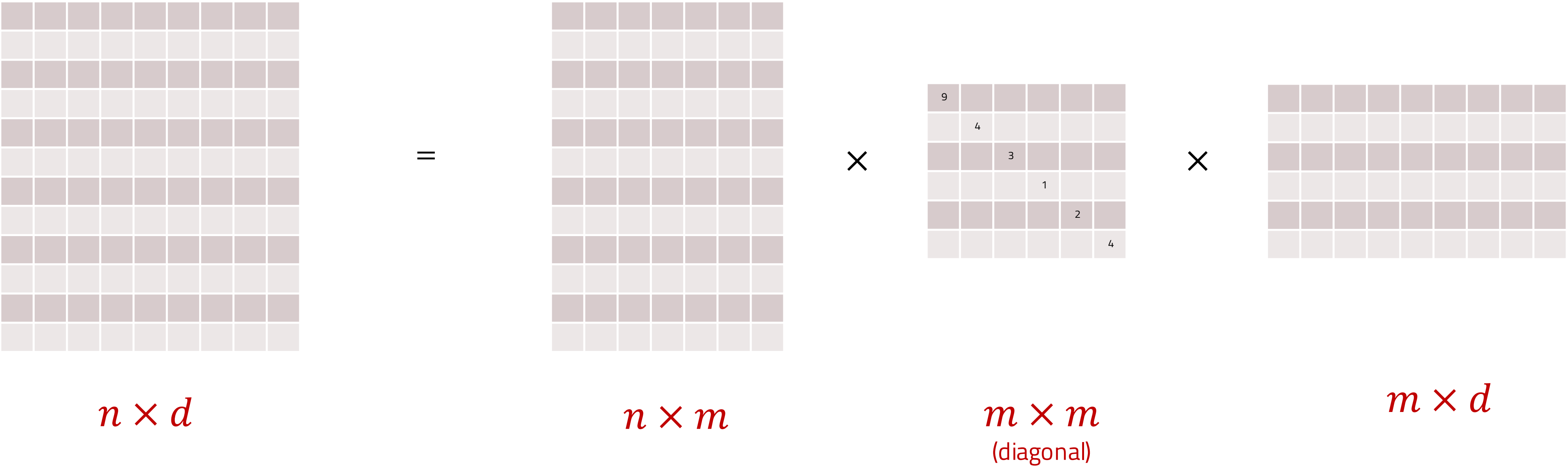
$=$



$n \times d$

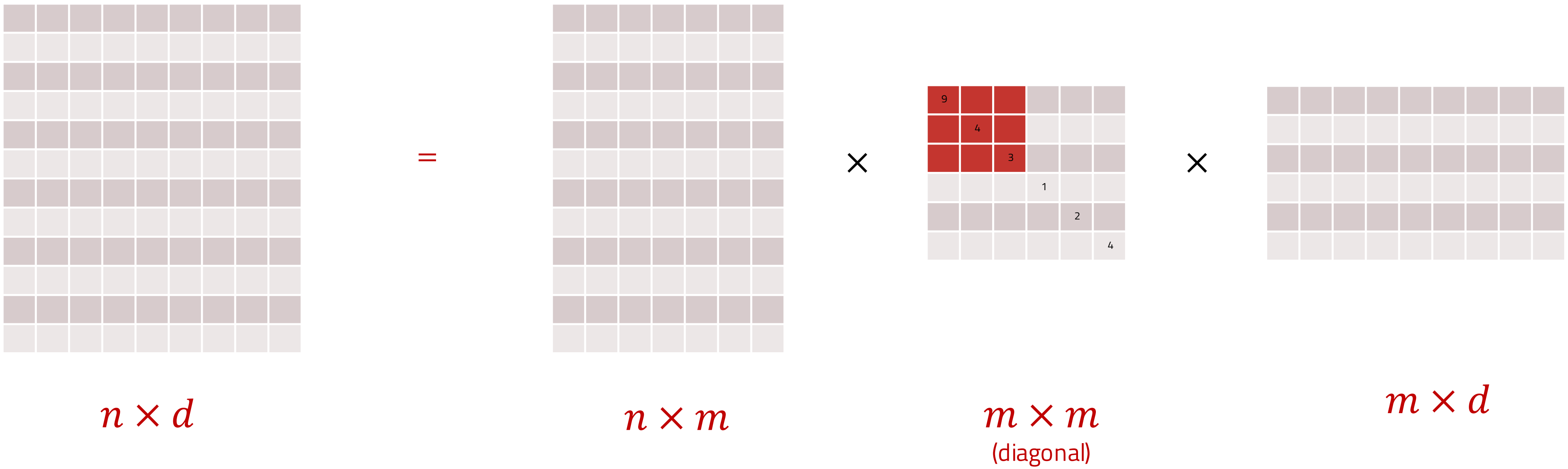
# Singular value decomposition (SVD)

- Any $n \times d$ matrix $X$ can be decomposed into the product of three matrices
  - where $m$ is the number of linearly independent rows



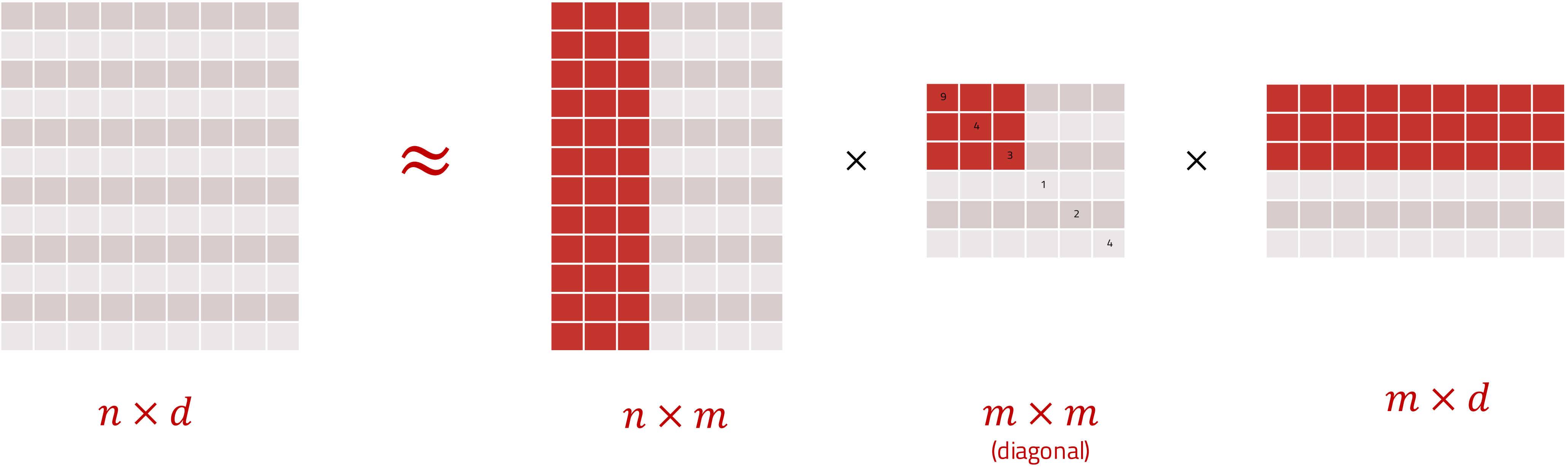
# Singular value decomposition (SVD)

- We can approximate the full matrix by only considering the leftmost k terms in the diagonal matrix



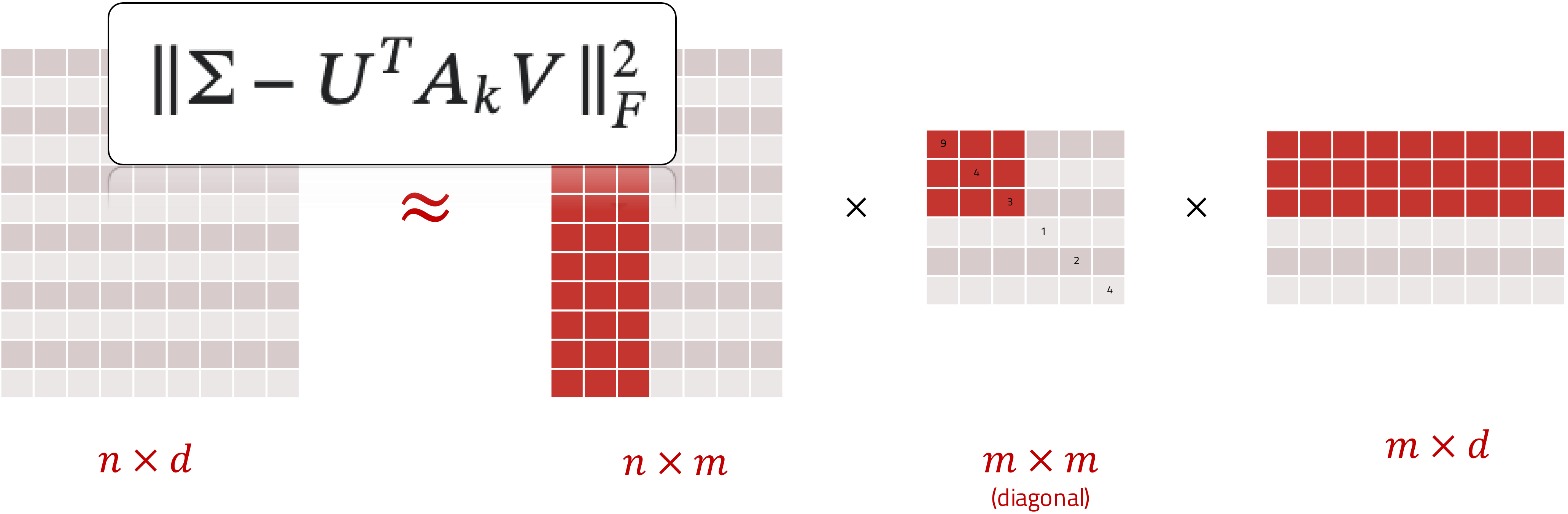
# Singular value decomposition (SVD)

- We can approximate the full matrix by only considering the leftmost k terms in the diagonal matrix



# Singular value decomposition (SVD)

- ❑ We can approximate the full matrix by only considering the leftmost k terms in the diagonal matrix



| | Hamlet | Macbeth | Romeo & Juliet | Richard III | Julius Caesar | Tempest |
|---|---|---|---|---|---|---|
| knife | 1 | 1 | 4 | 2 | | 2 |
| dog | | | | 6 | 12 | 2 |
| sword | 2 | 2 | 7 | 5 | | 5 |
| love | 64 | | 135 | 63 | | 12 |
| like | 75 | 38 | 34 | 36 | 34 | 41 |

| | | | |
|---|---|---|---|
| knife | 0.2 | 0.42 | 0.22 |
| dog | 0.5 | 1.2 | 8.6 |
| sword | -0.2 | 0.7 | -2.2 |
| love | 9.3 | -0.5 | 0.5 |
| like | 0.2 | 4.3 | 0.9 |

$n \times m$

×

| | | |
|---|---|---|
| 0.5 | | |
| | 0.3 | |
| | | 2.5 |

$m \times m$

×

| Hamlet | Macbeth | Romeo & Juliet | Richard III | Julius Caesar | Tempest |
|---|---|---|---|---|---|
| -0.2 | 0.7 | -2.2 | -0.2 | 0.7 | -2.2 |
| -0.2 | 0.7 | -2.2 | 9.3 | -0.5 | 0.5 |
| 9.3 | -0.5 | 0.5 | -0.5 | 0.5 | 9.3 |

$m \times d$

Low-dimensional representation for terms (here 3 dimensions)

Low-dimensional representation for documents (here 3 dimensions)

| | | | |
|---|---|---|---|
| knife | 0.2 | 0.42 | 0.22 |
| dog | 0.5 | 1.2 | 8.6 |
| sword | -0.2 | 0.7 | -2.2 |
| love | 9.3 | -0.5 | 0.5 |
| like | 0.2 | 4.3 | 0.9 |

| | | |
|---|---|---|
| 0.5 | | |
| | 0.3 | |
| | | 2.5 |

| Hamlet | Macbeth | Romeo & Juliet | Richard III | Julius Caesar | Tempest |
|---|---|---|---|---|---|
| -0.2 | 0.7 | -2.2 | -0.2 | 0.7 | -2.2 |
| -0.2 | 0.7 | -2.2 | 9.3 | -0.5 | 0.5 |
| 9.3 | -0.5 | 0.5 | -0.5 | 0.5 | 9.3 |

# Latent semantic analysis

- Latent Semantic Analysis/Indexing is this process of applying SVD to the term-document co-occurrence matrix
  - Terms typically weighted by tf-idf
- This is a form of dimensionality reduction
  - for terms, from a D-dimensional sparse vector to a K-dimensional dense one where K << D.
- Similar kinds:
  - Probabilistic Latent Semantic Indexing (pLSI) (Hofmann, 1999)
  - Nonnegative Matrix Factorization (NMF) (Lee & Seung, 1999)
  - Latent Dirichlet Allocation (LDA) (Blei et al., 2003)

| | #1 | #2 | #3 |
|---|---|---|---|
| knife | 0.2 | 0.42 | 0.22 |
| dog | 0.5 | 1.2 | 8.6 |
| sword | -0.2 | 0.7 | -2.2 |
| love | 9.3 | -0.5 | 0.5 |
| like | 0.2 | 4.3 | 0.9 |

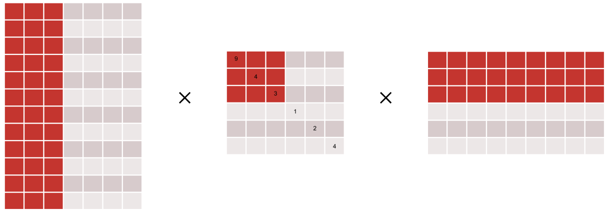
| #1 | #2 | #3 | #4 |
|---|---|---|---|
| music | how | program | 10 |
| film | what | project | 30 |
| theater | about | russian | 11 |
| mr | their | space | 12 |
| this | or | russia | 15 |

(Deerwester et al. 1998)

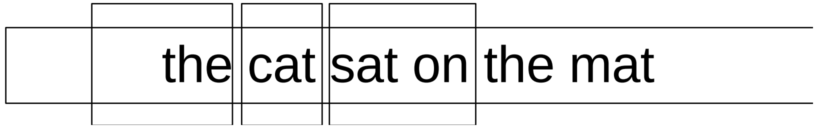
# Count-based vs Prediction-based Methods

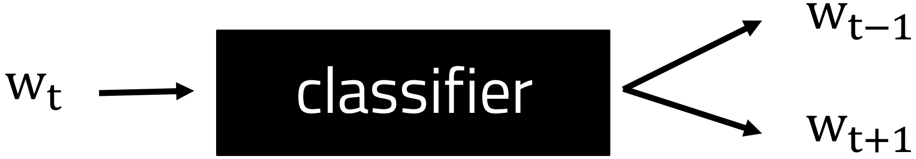
**LSA**, **HAL** (Lund & Burgess)
**Hellinger-PCA** (Rohde et al, Lebret & Collobert)

| | Hamlet | Macbeth |
|---|---|---|
| knife | 1 | 1 |
| dog | | |
| sword | 2 | 2 |
| love | 64 | |
| like | 75 | 38 |



**Skip-gram/CBOW** (Mikolovet al)
**NLM, HLBL, RNN** (Bengioet al; Collobert & Weston; Huang et al; Mnih & Hinton)





# Recap: Text Classification

x = "Today's weather is great"

$$P(y \mid x)$$

y = {positive, negative}

$\hat{y}$ = positive

|Y| = **2**

$x_{<t}$ = "Today's weather is"

$$P(x_t \mid x_{<t})$$

$x_t$ = {a, aa .. apple .. banana .. great .. good .. zebra ..}

$\hat{x}$ = great

|X| = **V (vocabulary size)**

$x_{<t}$ = "Today 's [ ] is great"

$$P(x_t \mid x_{t-2,t-1,\ t+1,\ t+2})$$

$x_t$ = {a, aa .. apple .. banana .. great .. good .. zebra ..}

$\hat{x}$ = weather

|X| = **V (vocabulary size)**

# Recap: Text Classification

$x_{t-2}$ = [ ] .. weather .. ..

$$P(x_{t-2} \mid x_t)$$

$x_{t-1}$ = .. [ ] weather .. ..

$$P(x_{t-1} \mid x_t)$$

$$P(x_{t+1} \mid x_t)$$

$x_{t+1}$ = .. .. weather [ ] ..

$$P(x_{t+2} \mid x_t)$$

$x_{t+2}$ = .. .. weather .. [ ]

$x_{<t}$ = “Today ’s [ ] is great”

$$P(x_t \mid x_{t-2,t-1,\, t+1,\, t+2})$$

$x_t$ = {a, aa .. apple .. banana .. great .. good .. zebra ..}

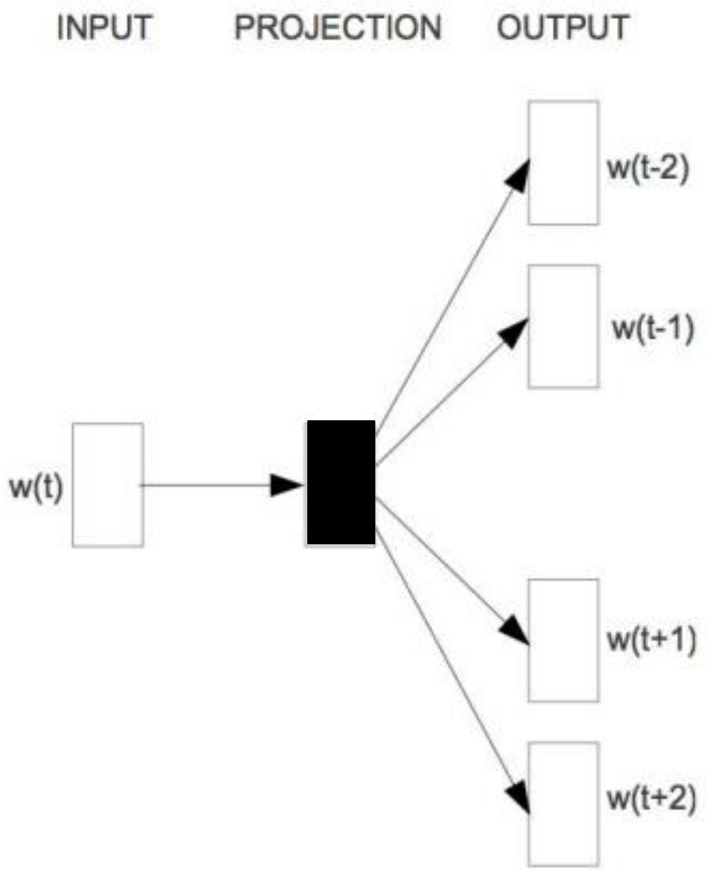
$\hat{x}$ = weather

|X| = **V (vocabulary size)**

# Dense vectors from prediction (not count)

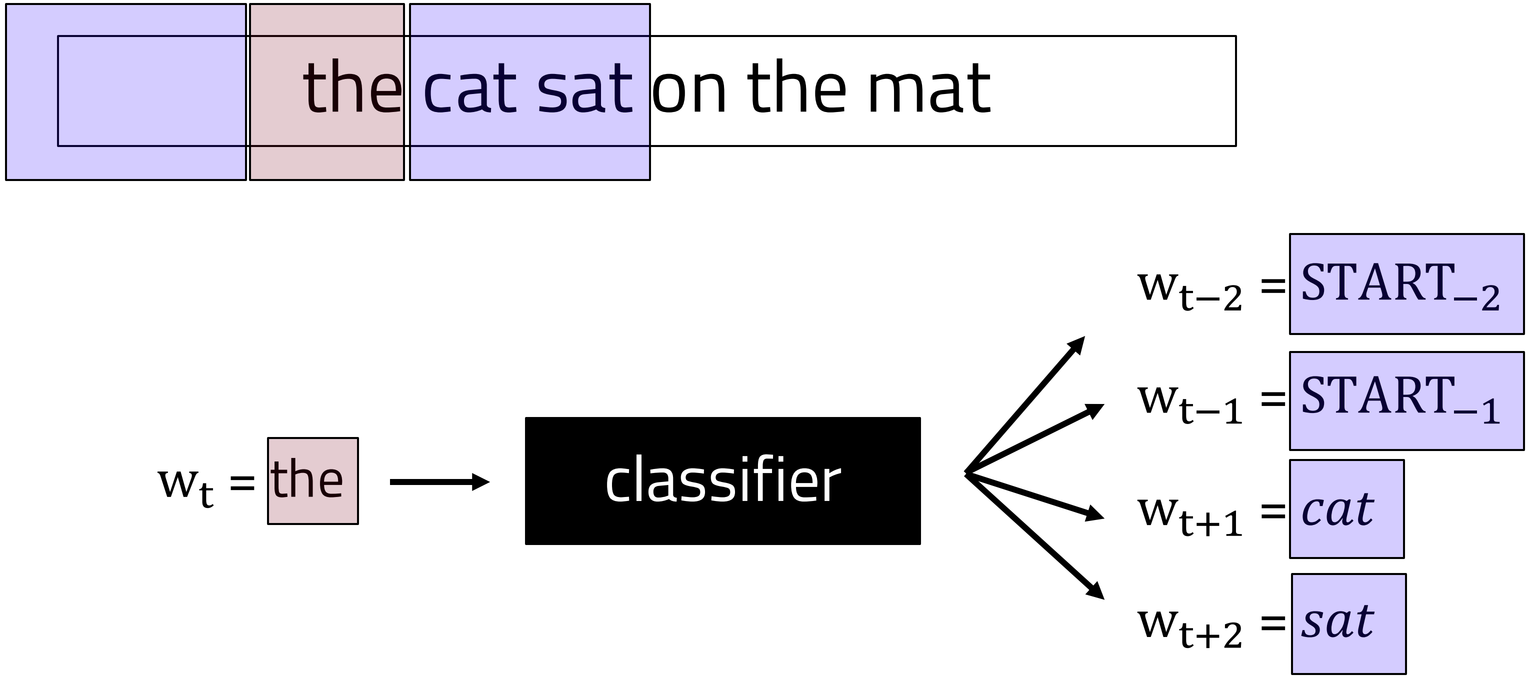
the cat sat on the mat

**Skipgram model**: given a single word in a sentence, predict the words in a context window around it.



(Mikolove et al., 14)

# Dense vectors from prediction (not count)



the cat sat on the mat

$w_t$ = the

classifier

$w_{t-2}$ = $START_{-2}$

$w_{t-1}$ = $START_{-1}$

$w_{t+1}$ = *cat*

$w_{t+2}$ = *sat*

$P(x_{t-2} \mid x_t)$

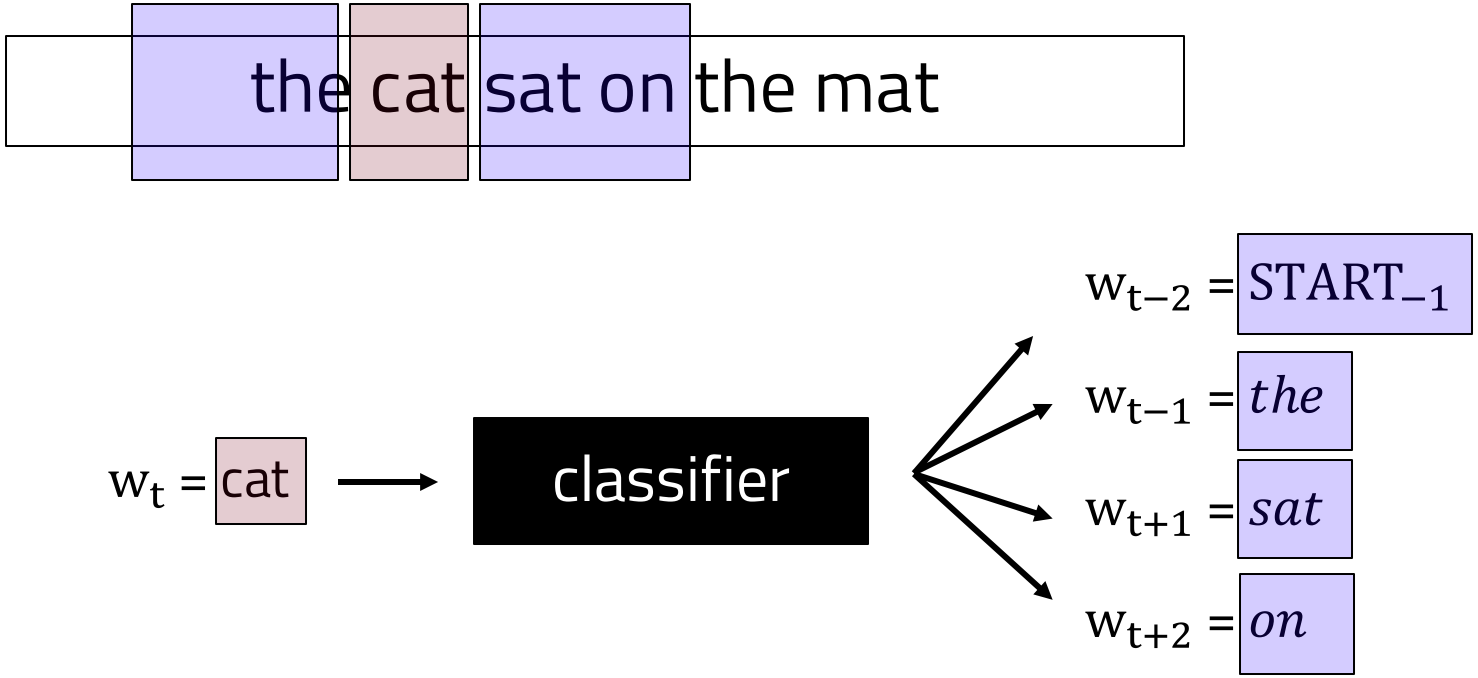
$P(x_{t-1} \mid x_t)$

$P(x_{t+1} \mid x_t)$

$P(x_{t+2} \mid x_t)$

Context window size = 2

# Dense vectors from prediction (not count)



the cat sat on the mat

$w_t$ = cat → classifier

$w_{t-2} = START_{-1}$

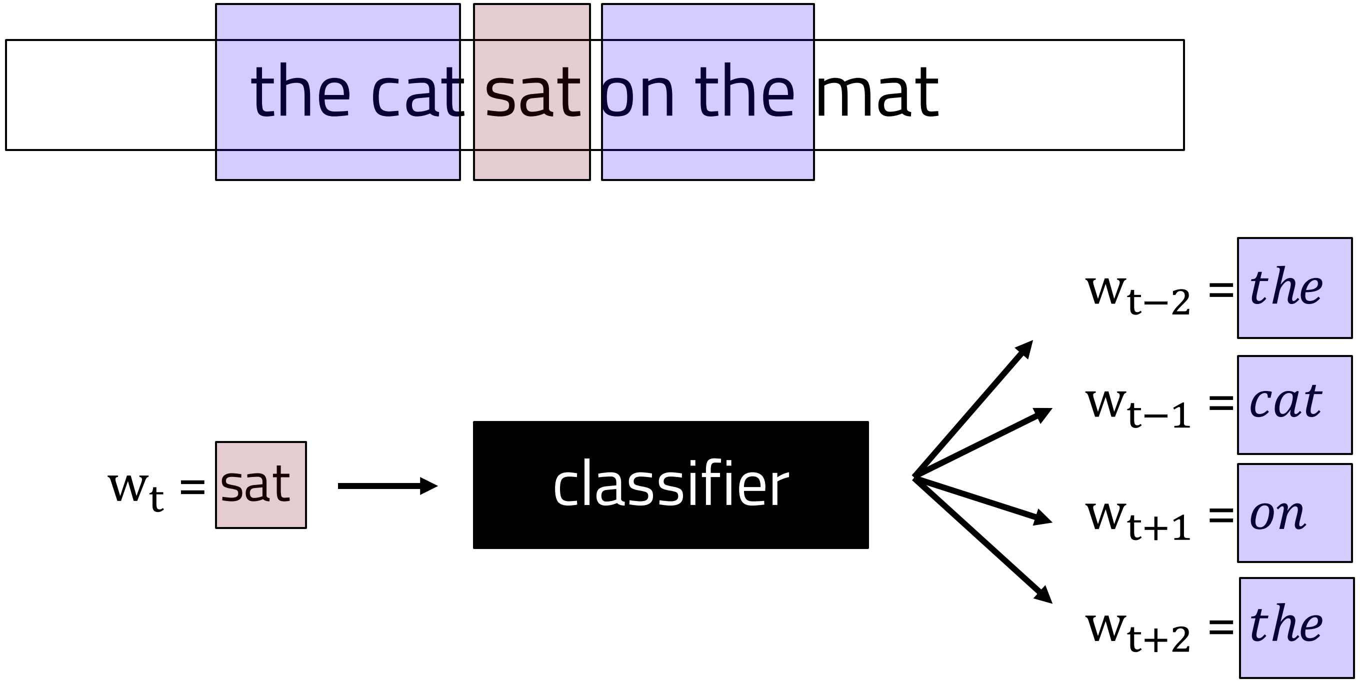
$w_{t-1} = the$

$w_{t+1} = sat$

$w_{t+2} = on$

Context window size = 2

# Dense vectors from prediction (not count)



the cat sat on the mat

$w_t$ = sat → classifier

$w_{t-2} = the$

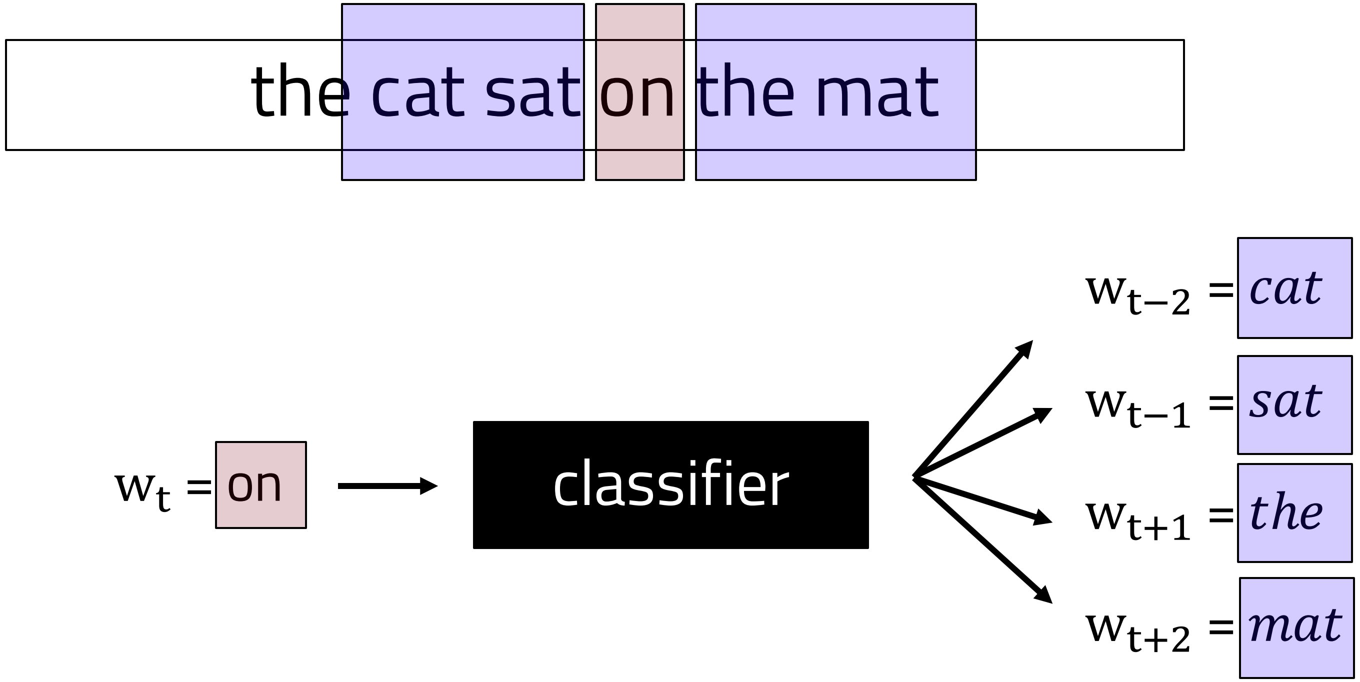
$w_{t-1} = cat$

$w_{t+1} = on$

$w_{t+2} = the$

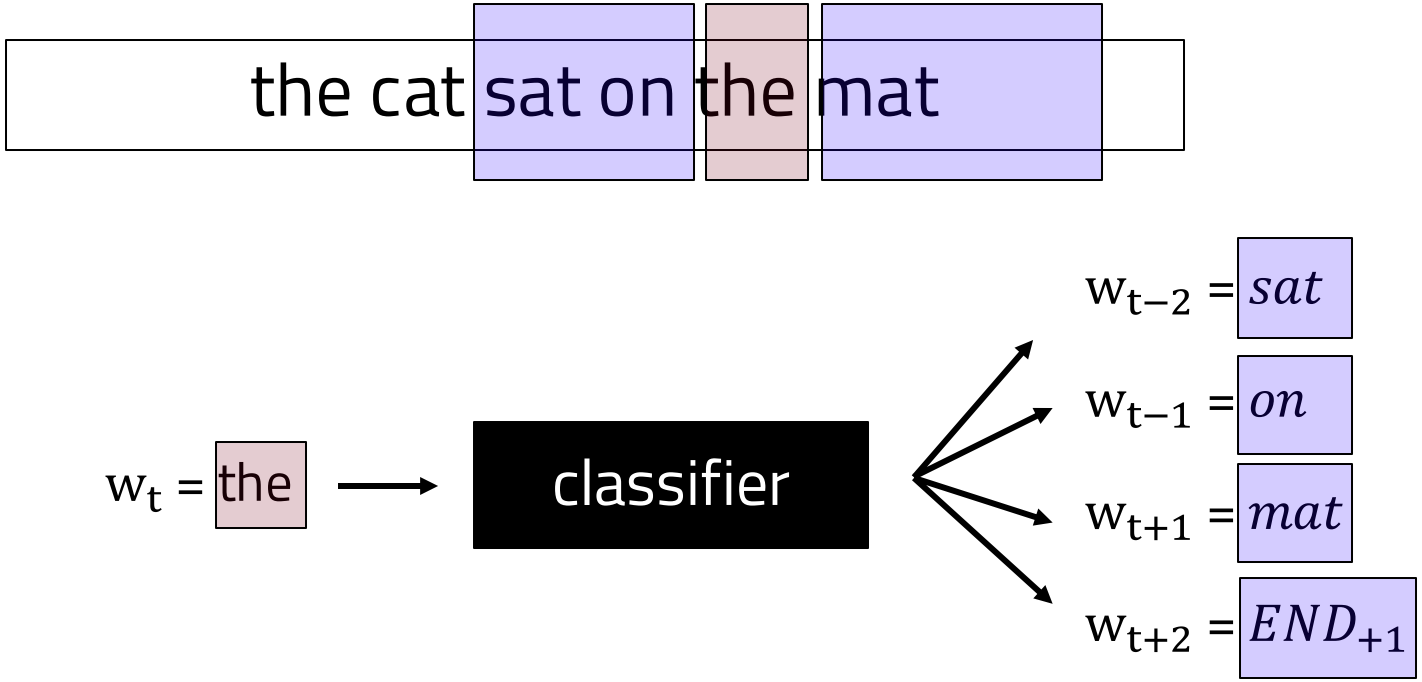
Context window size = 2

# Dense vectors from prediction (not count)



Context window size = 2

# Dense vectors from prediction (not count)



the cat sat on the mat

$w_t =$ the

classifier

$w_{t-2} = sat$

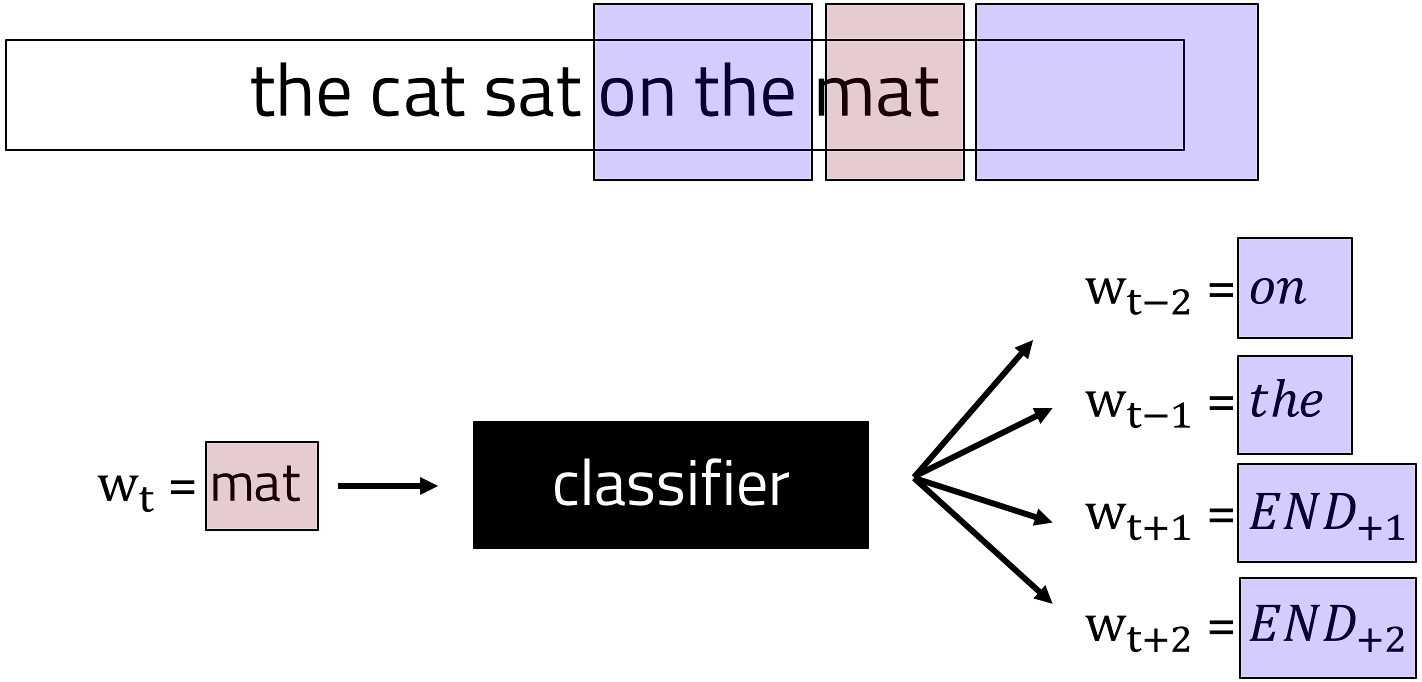
$w_{t-1} = on$

$w_{t+1} = mat$

$w_{t+2} = END_{+1}$

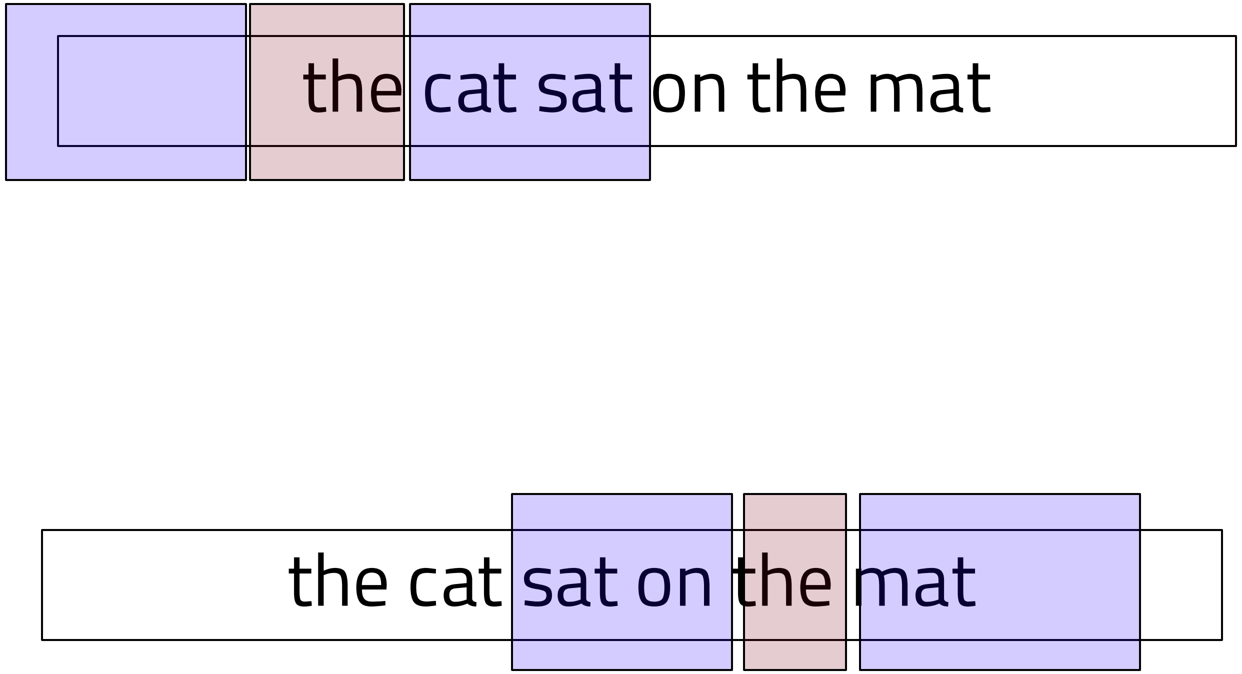
Context window size = 2

# Dense vectors from prediction (not count)



Context window size = 2

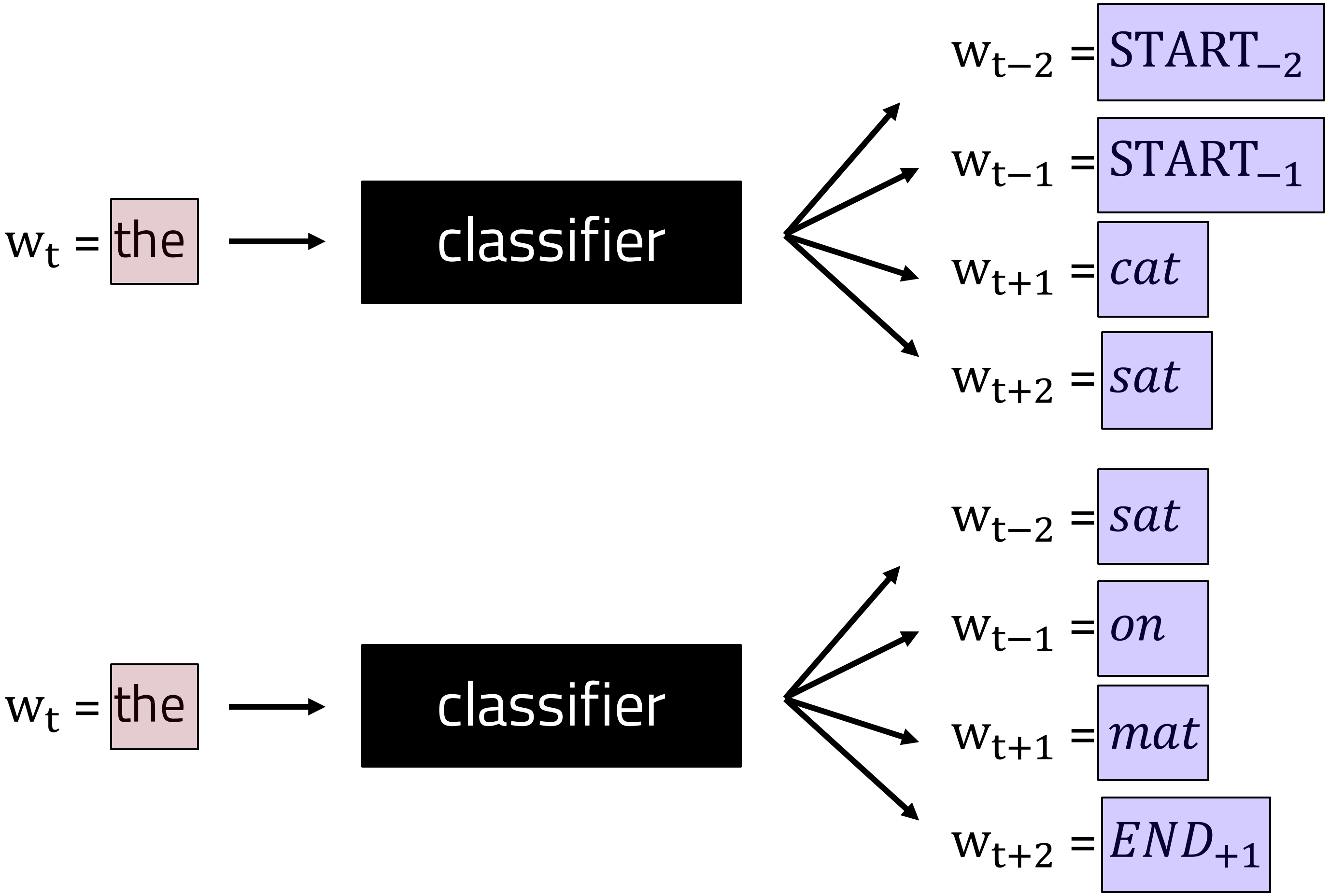
# Dense vectors from prediction (not count)



the cat sat on the mat

the cat sat on the mat

# Dense vectors from prediction (not count)



$w_t =$ the → classifier

$w_{t-2} = START_{-2}$

$w_{t-1} = START_{-1}$

$w_{t+1} = cat$

$w_{t+2} = sat$

$w_t =$ the → classifier

$w_{t-2} = sat$

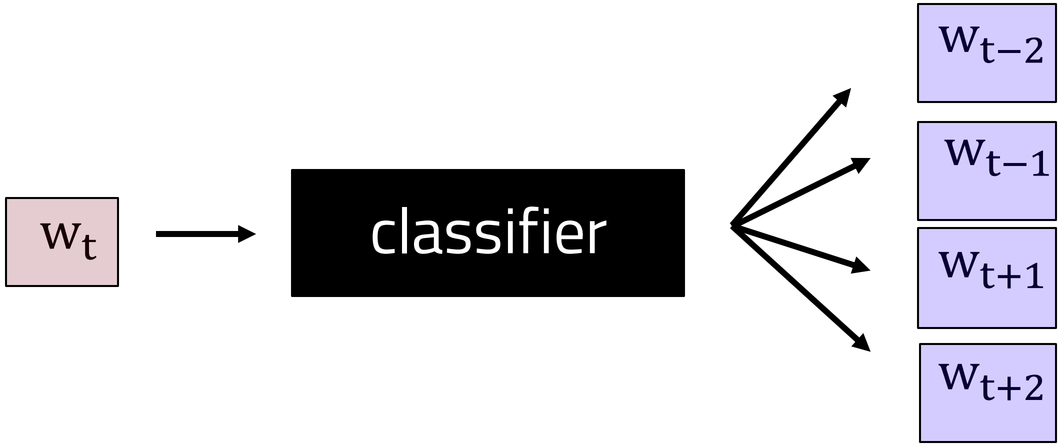
$w_{t-1} = on$

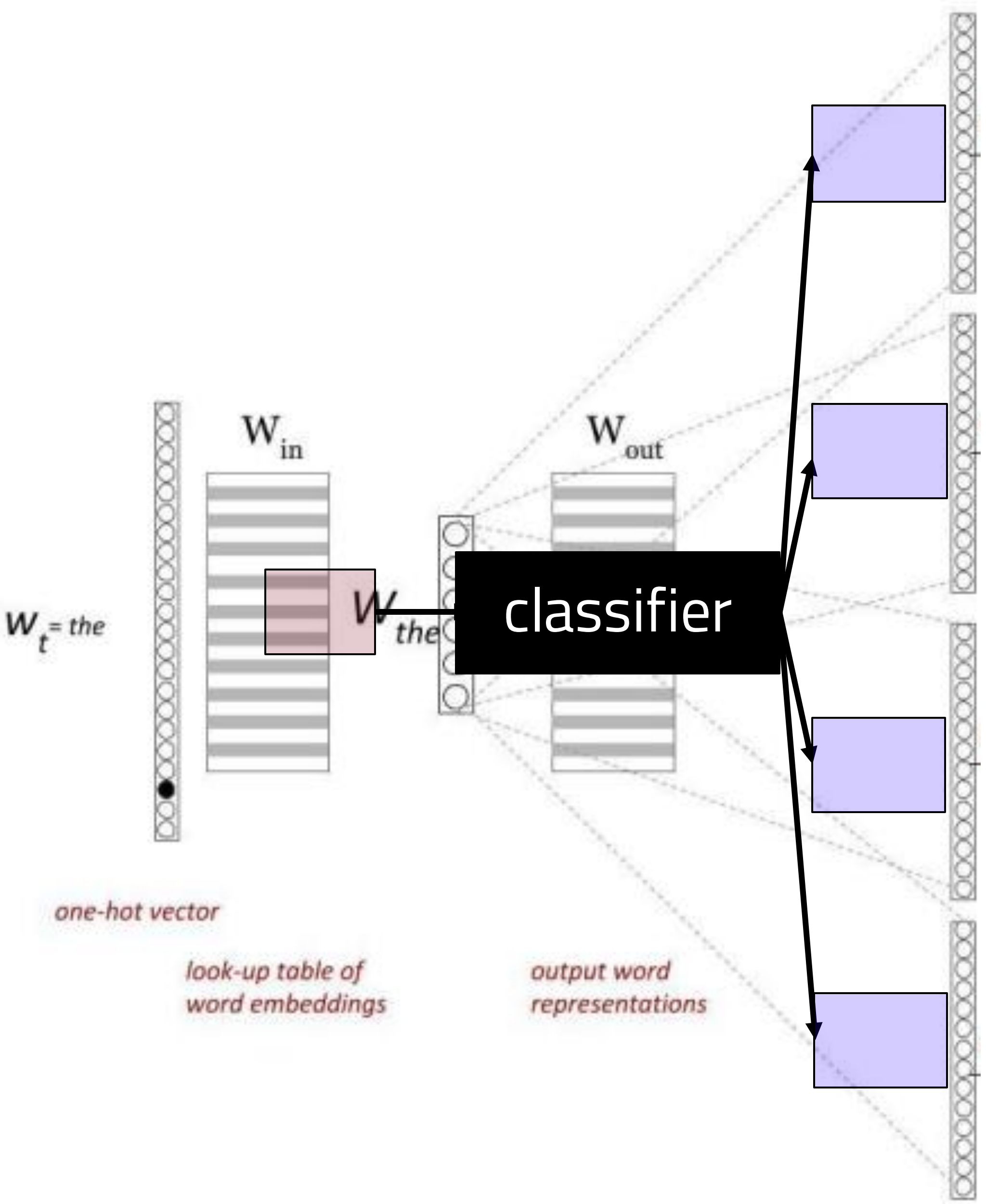
$w_{t+1} = mat$

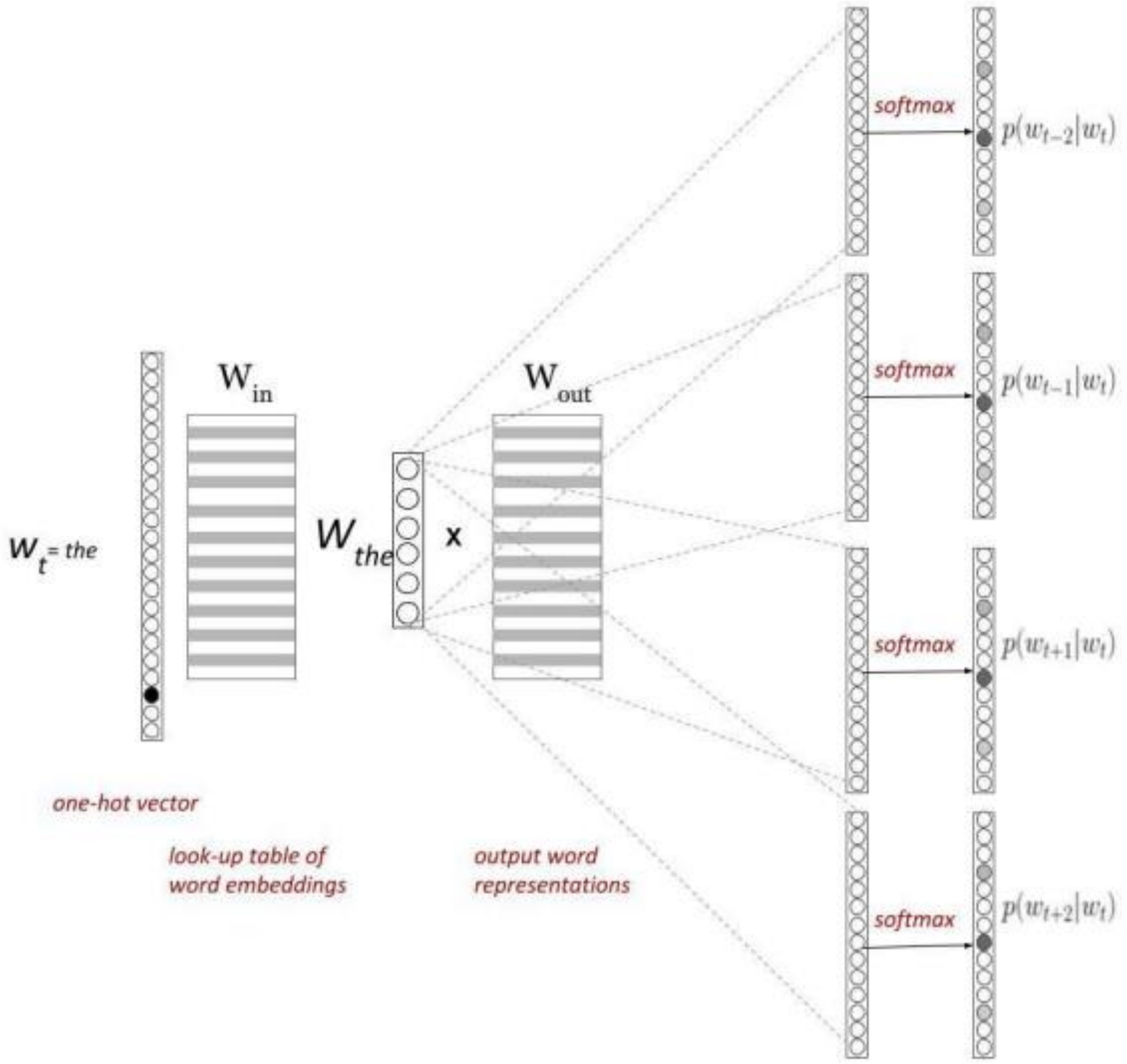
$w_{t+2} = END_{+1}$

Context window size = 2

# Dense vectors from prediction (not count)




$w_t$ = the
$W_{in}$
$W_{the}$
$W_{out}$
classifier
one-hot vector
look-up table of word embeddings
output word representations



$W_{in}$
$W_{out}$
$w_t$ = the
$W_{the}$
x
softmax
softmax
softmax
softmax
$p(w_{t-2}|w_t)$
$p(w_{t-1}|w_t)$
$p(w_{t+1}|w_t)$
$p(w_{t+2}|w_t)$
one-hot vector
look-up table of word embeddings
output word representations

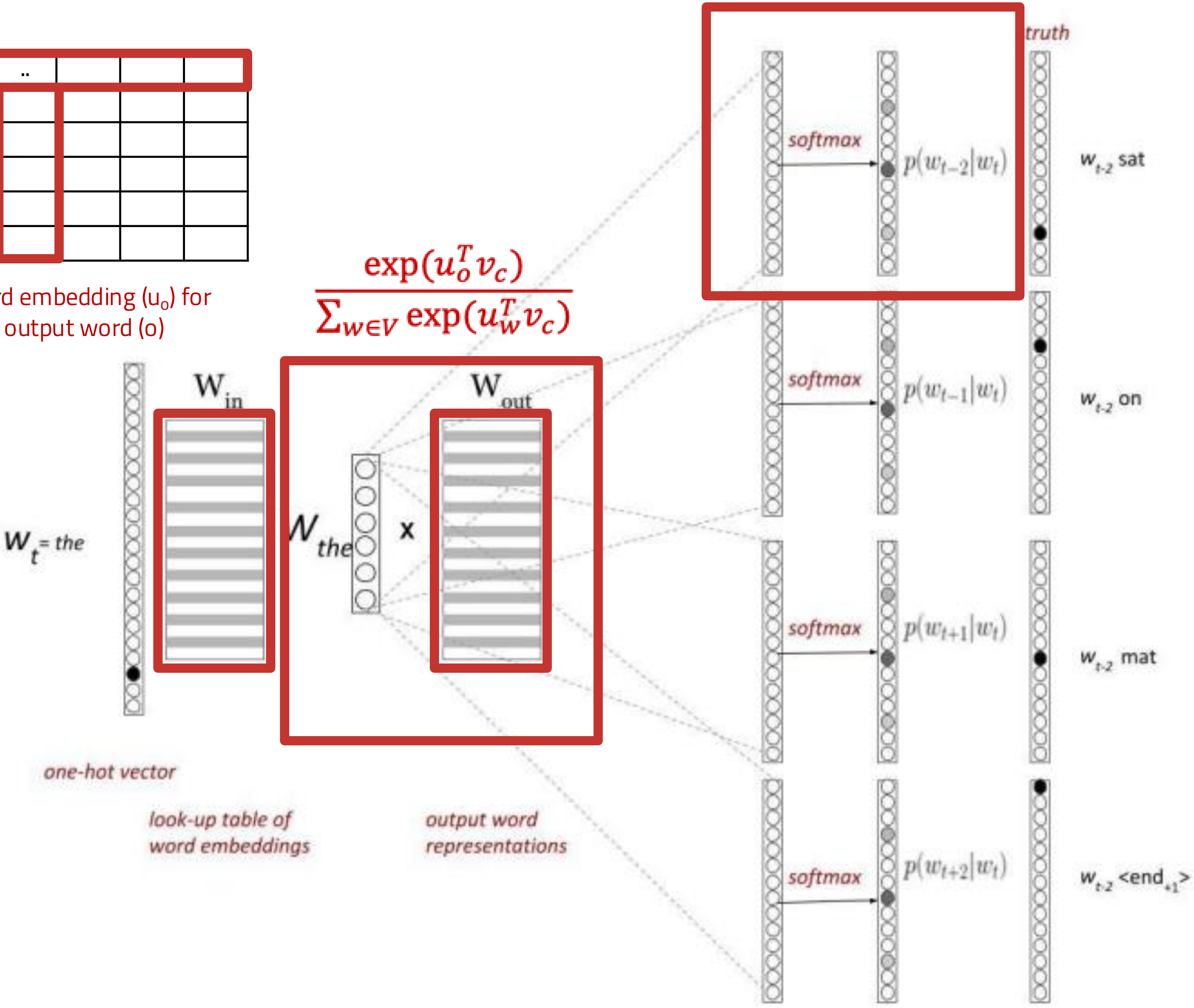
V

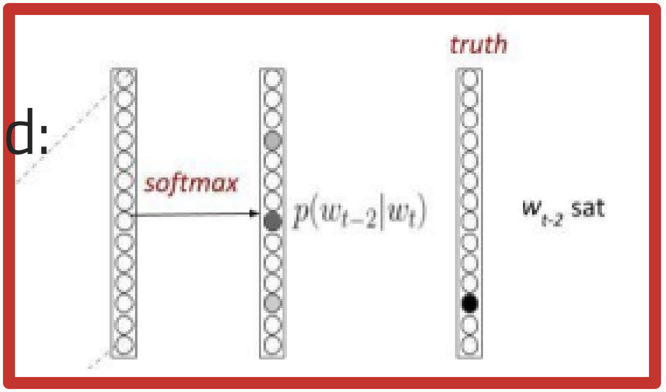
| the | cat | mat | on | sat | .. |  |  |  |
|---|---|---|---|---|---|---|---|---|
| 5.2 | 1.5 | ... |  |  |  |  |  |  |
| 0.5 | 0.4 | ... |  |  |  |  |  |  |
| -6.2 | 0.6 | .. |  |  |  |  |  |  |
| 0.5 | -3.4 | .. |  |  |  |  |  |  |
| ... |  |  |  |  |  |  |  |  |

Word embedding ($v_c$) for center word (c) “the”

Word embedding ($u_o$) for output word (o)

$$\frac{\exp(u_o^T v_c)}{\sum_{w\in V}\exp(u_w^T v_c)}$$





The objective function $J(\theta)$ is the average negative log likelihood:

$$J(\theta) = -\frac{1}{T}\sum_{t=1}^{T}\sum_{-m\le j\le m, j\ne 0} \boxed{\log P(w_{t+j} \mid w_t; \theta)}$$

All word vectors

For a center word $c$ and $a$ context word $o$ :

$$x_i = \quad P(o \mid c) = \frac{\exp\boxed{(u_o^T v_c)}}{\boxed{\sum_{w\in V}\exp(u_w^T v_c)}}$$

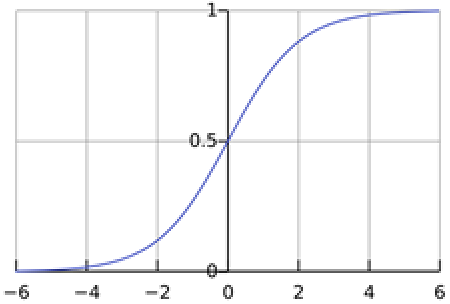
Dor product compares similarity of $o$ and $c$ . $u^T v = u \cdot v = \sum_{i=1}^{n} u_i v_i$

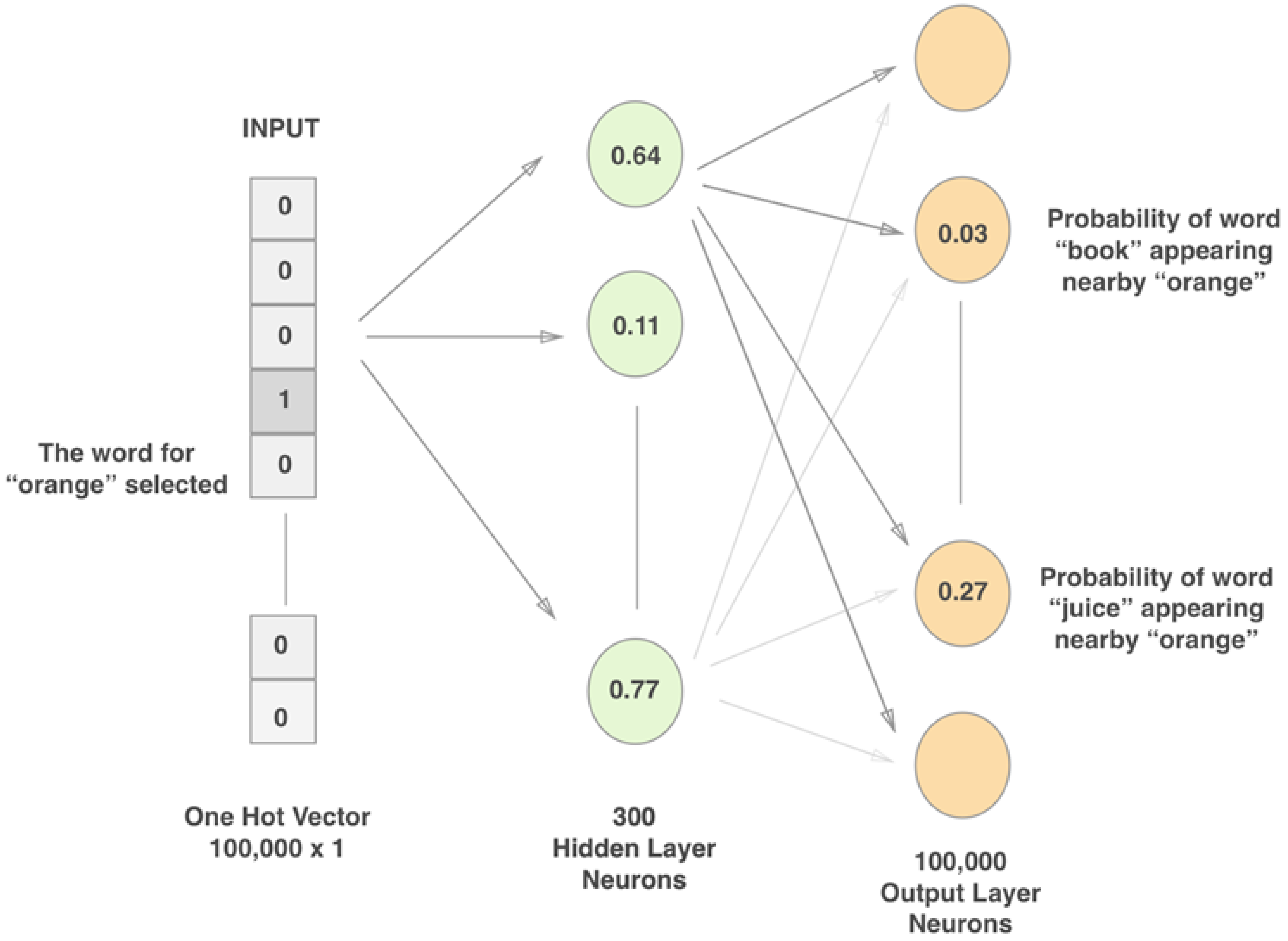
Normalize over entire vocabulary to give probability distribution

“soft” because still assigns some probability to smaller $x_i$

$$\boxed{soft}\boxed{max}(x_i) = \frac{\exp(x_i)}{\sum_{j=1}\exp(x_j)} = p_i$$

“max” because amplifies probability of largest $x_i$





INPUT

The word for “orange” selected

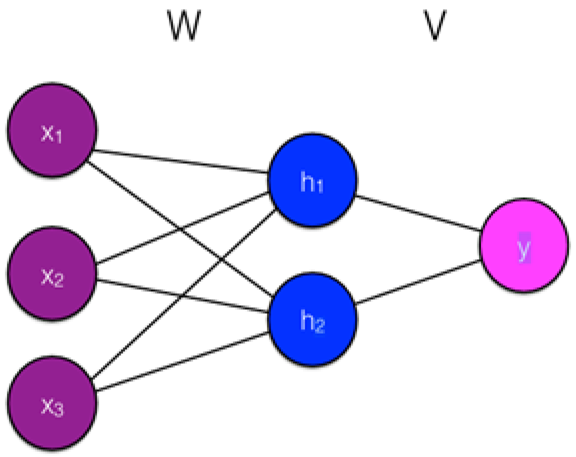
One Hot Vector 100,000 x 1

300 Hidden Layer Neurons

100,000 Output Layer Neurons

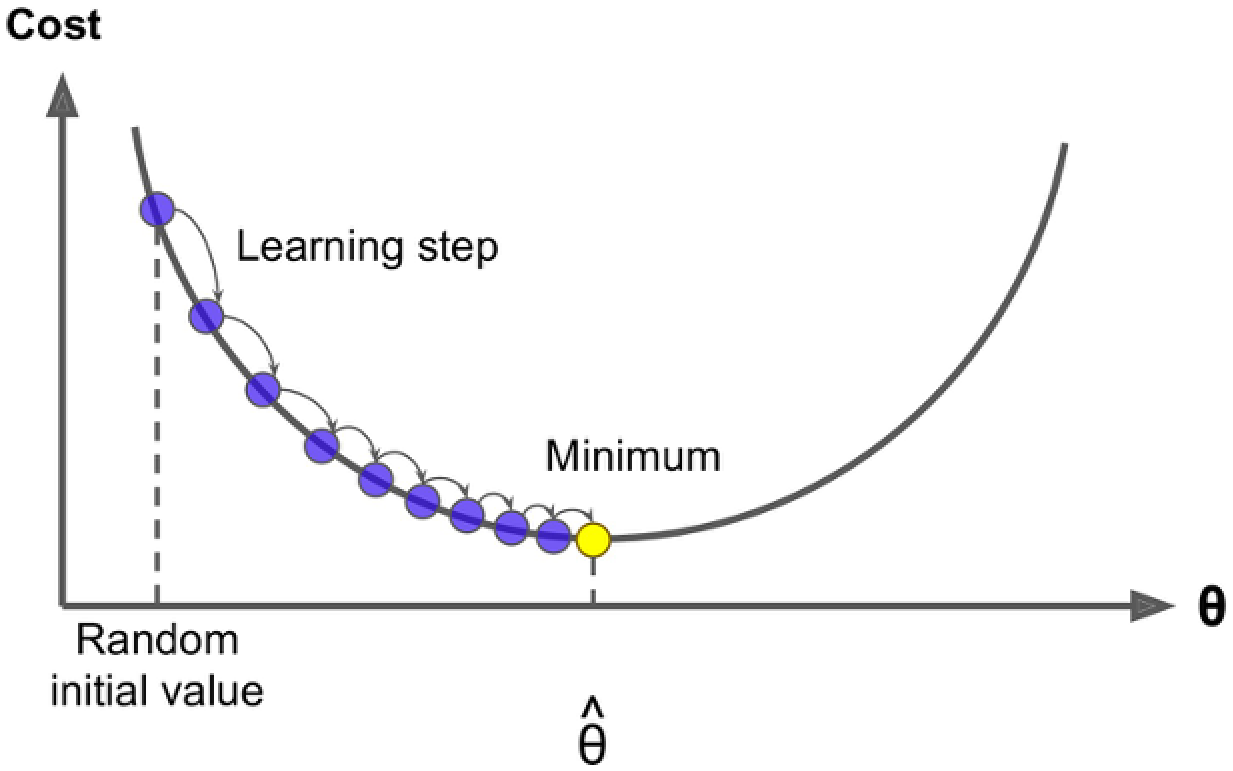
Probability of word “book” appearing nearby “orange”

Probability of word “juice” appearing nearby “orange”



# Minimize the objective function $J(\theta)$ using *gradient descent*

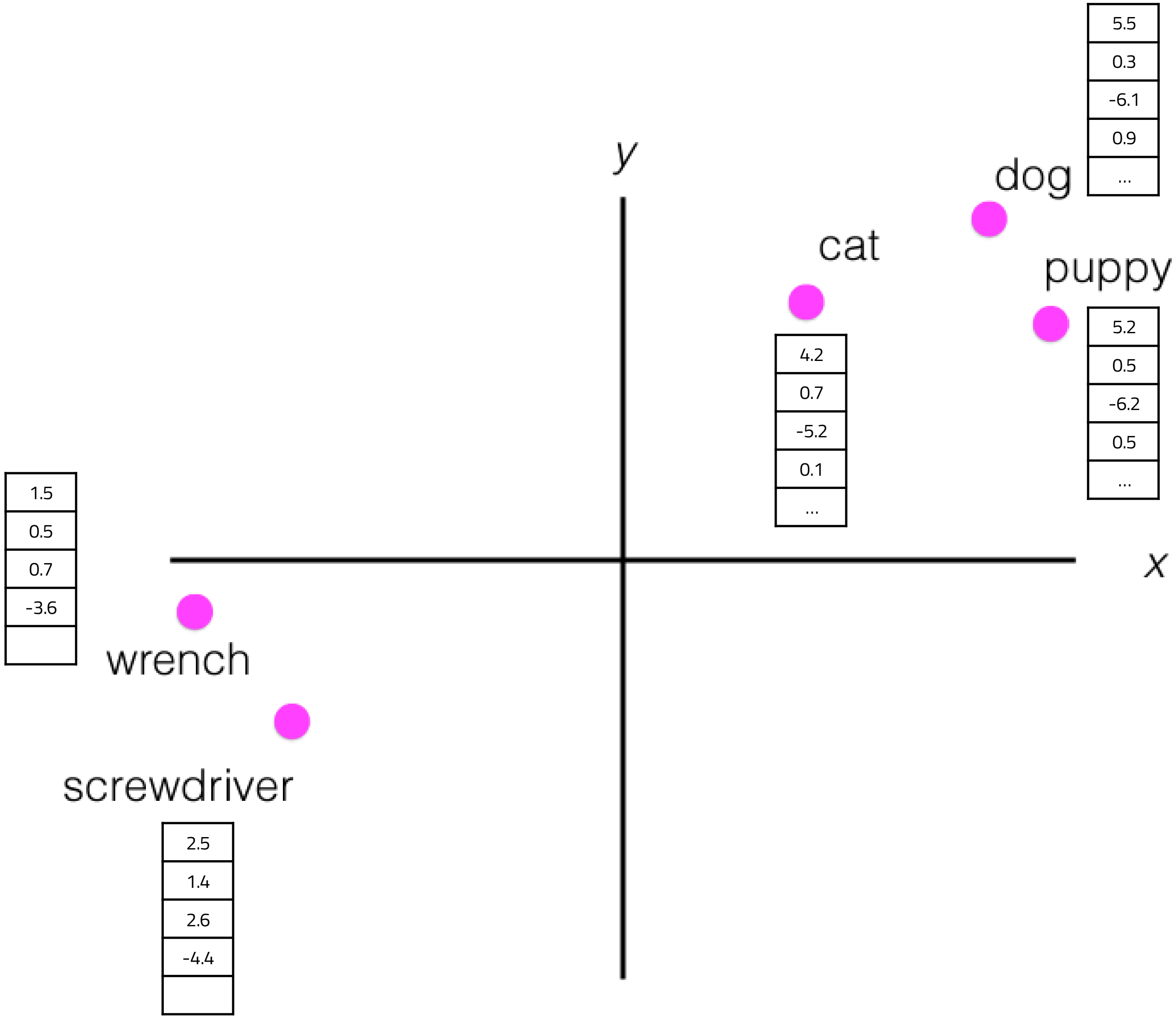
**Idea**: for current value of $\theta$ , calculate gradient of $J(\theta)$ then take small step in direction of negative gradient. Repeat this until convergence



# Two kinds of training data

- The labeled data for specific tasks
  - Labeled sentiment for movie reviews (~2K labels/reviews, ~1.5 words)
  - Used for supervised models

- Unlabeled text for representation learning
  - Trillions of words (Wikipedia, web text, books, etc)
  - Used for word distributed representations

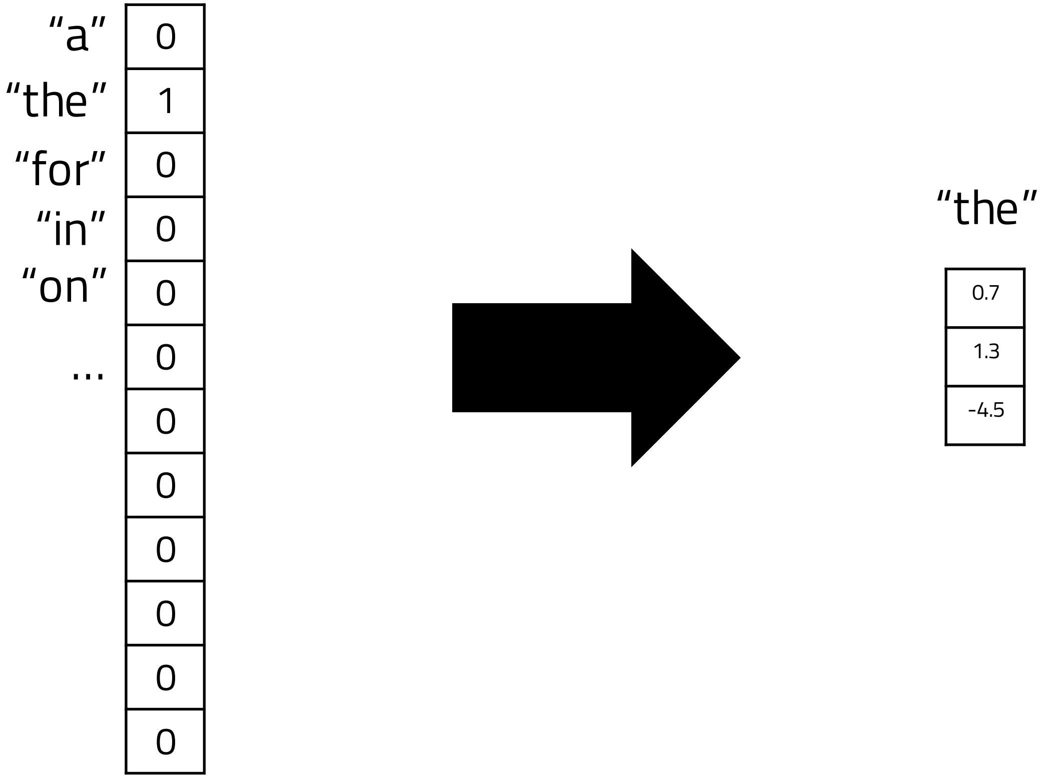

y
x
dog
5.5
0.3
-6.1
0.9
...
cat
4.2
0.7
-5.2
0.1
...
puppy
5.2
0.5
-6.2
0.5
...
1.5
0.5
0.7
-3.6
wrench
screwdriver
2.5
1.4
2.6
-4.4

# Why *dog* and *cat* are in similar positions

| | | | | | | |
|---|---|---|---|---|---|---|
| the | black | **dog** | jumped | on | the | table |
| the | black | **cat** | jumped | on | the | table |
| the | black | **puppy** | jumped | on | the | table |

| | | | | | | |
|---|---|---|---|---|---|---|
| the | black | **wrench** | jumped | on | the | table |
| the | black | **shoe** | jumped | on | the | table |

# Dimensionality reduction

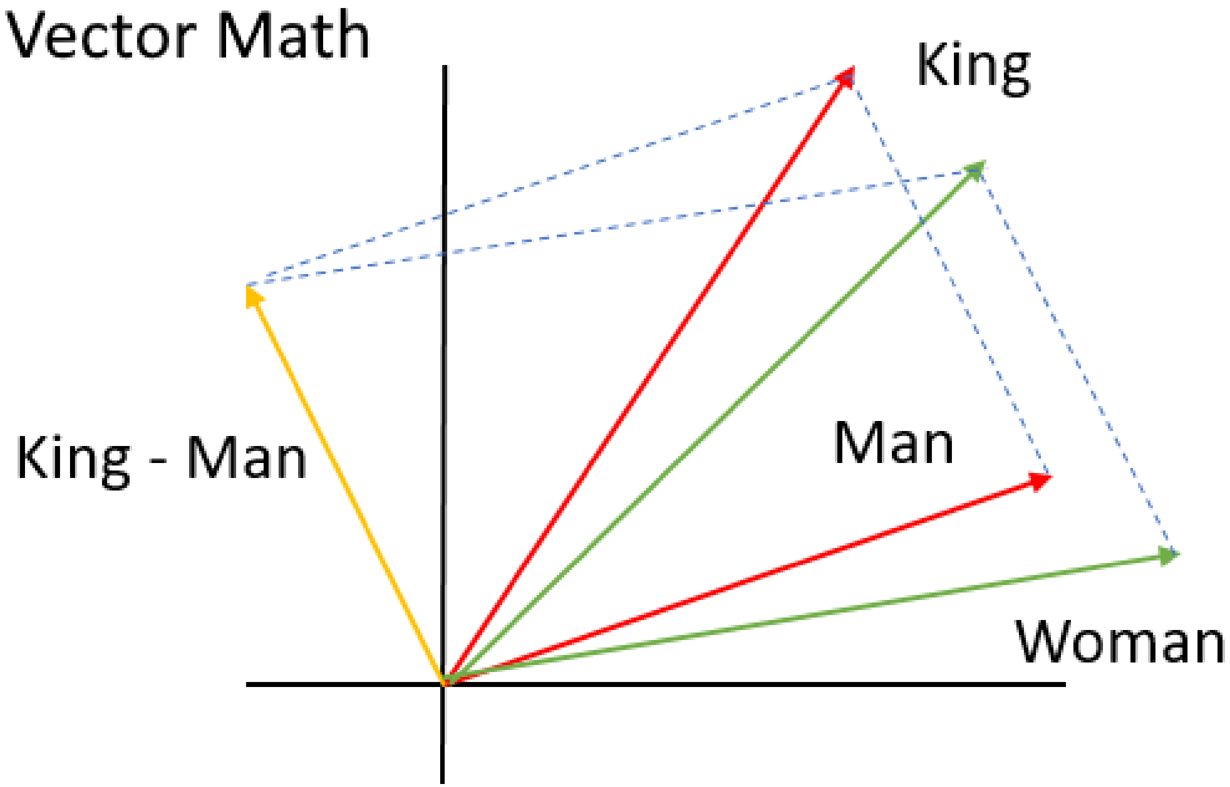


“a” 0
“the” 1
“for” 0
“in” 0
“on” 0
… 0
0
0
0
0
0
0

“the”
0.7
1.3
-4.5

V-dimensional space (1-hot)
Representations for all words are completely independent

3-dimensional space
Representations are not structured



v(“King”) – v(“Man”) + v(“Woman”) =

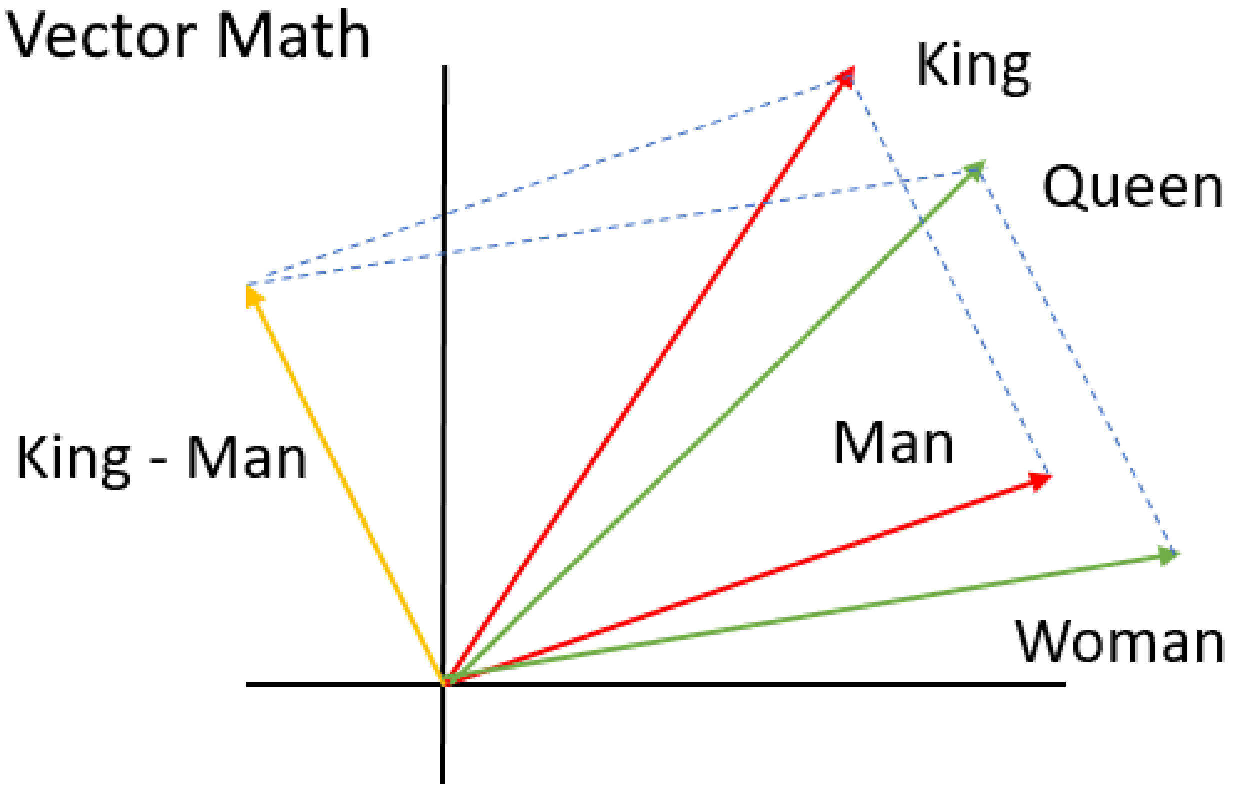
$$\begin{bmatrix}0.7\\1.3\\-4.5\\\ldots\end{bmatrix} - \begin{bmatrix}5.2\\0.5\\-6.2\\0.5\\\ldots\end{bmatrix} + \begin{bmatrix}4.2\\0.7\\-5.2\\0.1\\\ldots\end{bmatrix} = \begin{bmatrix}5.2\\0.5\\-6.2\\0.5\\\ldots\end{bmatrix}$$

Closest vector

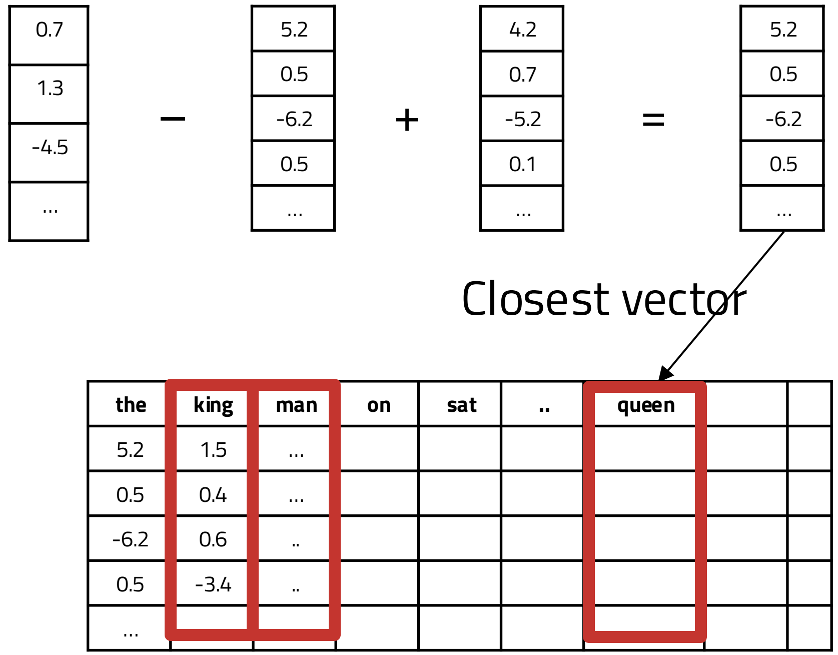
| the | king | man | on | sat | .. | | | |
|---|---|---|---|---|---|---|---|---|
| 5.2 | 1.5 | … | | | | | | |
| 0.5 | 0.4 | … | | | | | | |
| -6.2 | 0.6 | .. | | | | | | |
| 0.5 | -3.4 | .. | | | | | | |
| … | | | | | | | | |

Mikolov et al. 2013 show that vector representations have some potential for analogical reasoning through vector arithmetic.

Mikolov et al., (2013), “Linguistic Regularities in Continuous Space Word Representations” (NAACL)



v(“King”) – v(“Man”) - v(“Woman”) =



$$\begin{bmatrix}0.7\\1.3\\-4.5\\\dots\end{bmatrix} - \begin{bmatrix}5.2\\0.5\\-6.2\\0.5\\\dots\end{bmatrix} + \begin{bmatrix}4.2\\0.7\\-5.2\\0.1\\\dots\end{bmatrix} = \begin{bmatrix}5.2\\0.5\\-6.2\\0.5\\\dots\end{bmatrix}$$

Closest vector

| the | king | man | on | sat | .. | queen | | |
|---|---|---|---|---|---|---|---|---|
| 5.2 | 1.5 | … | | | | | | |
| 0.5 | 0.4 | … | | | | | | |
| -6.2 | 0.6 | .. | | | | | | |
| 0.5 | -3.4 | .. | | | | | | |
| … | | | | | | | | |

Mikolov et al. 2013 show that vector representations have some potential for analogical reasoning through vector arithmetic.

Mikolov et al., (2013), “Linguistic Regularities in Continuous Space Word Representations” (NAACL)

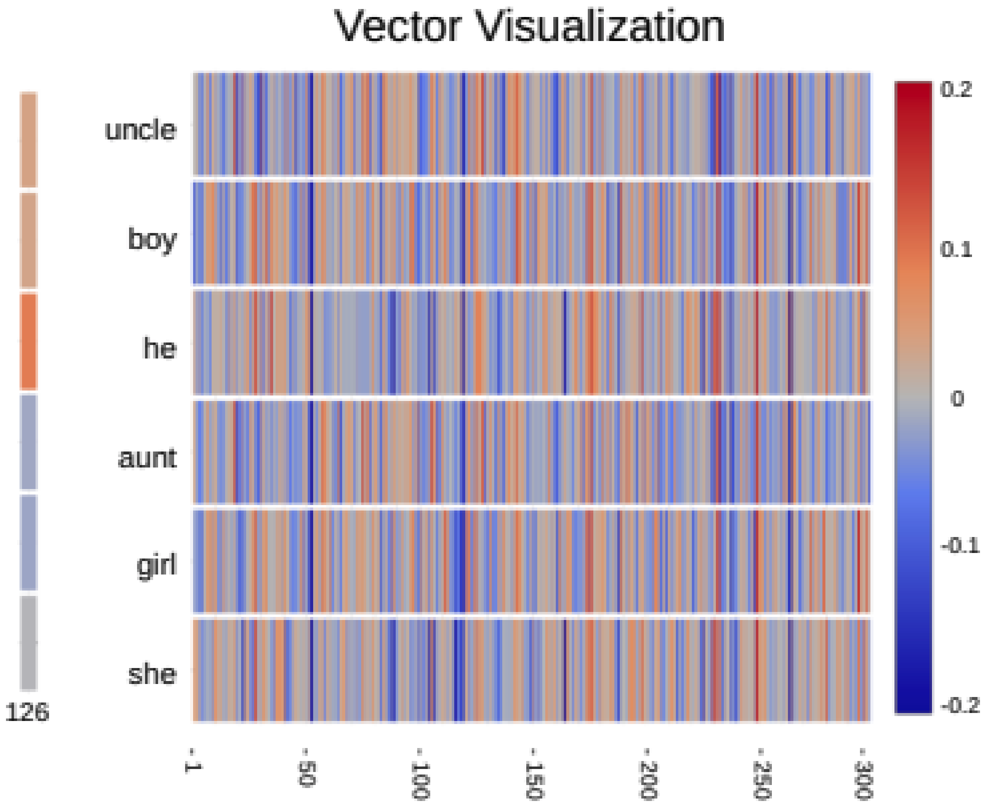


Figure 2: Embedding vectors for three male words (“uncle”, “boy”, “he”) and three female words (“aunt”, “girl”, “she”). Component 126, shown magnified at left, is positive for the male words and negative for the female words.

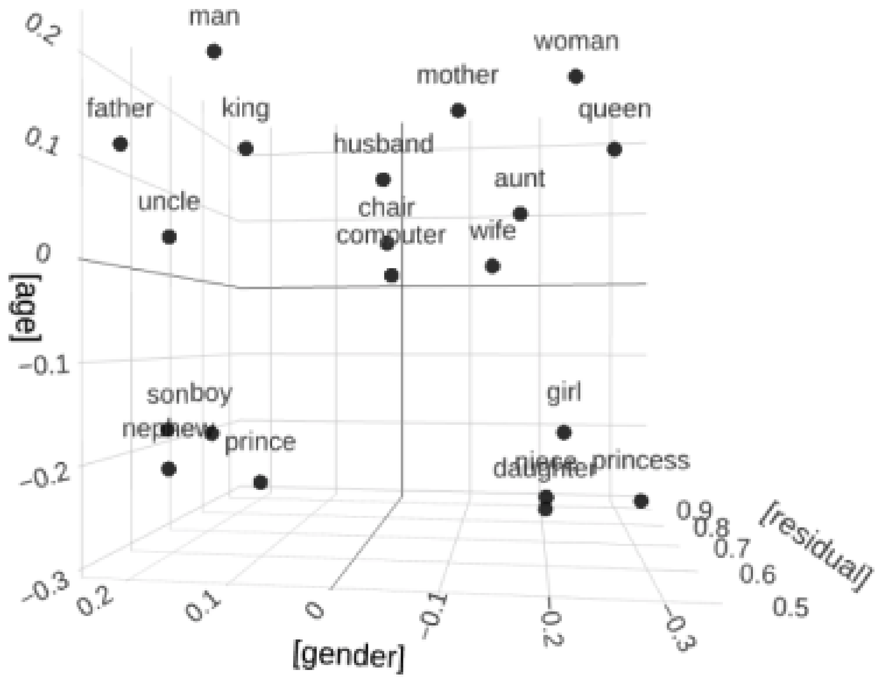


Figure 3: Words plotted in our 3D semantic space. Male words appear in the positive (left) half of the x-axis; female words in the negative (right) half. Adult words are in the positive (top) half of the y-axis; youth words in the negative (bottom) half. The third dimension is the “semantic residual”, explained in the main text.

Interactive Visualizations of Word Embeddings for K-12 Students. AAAI-22

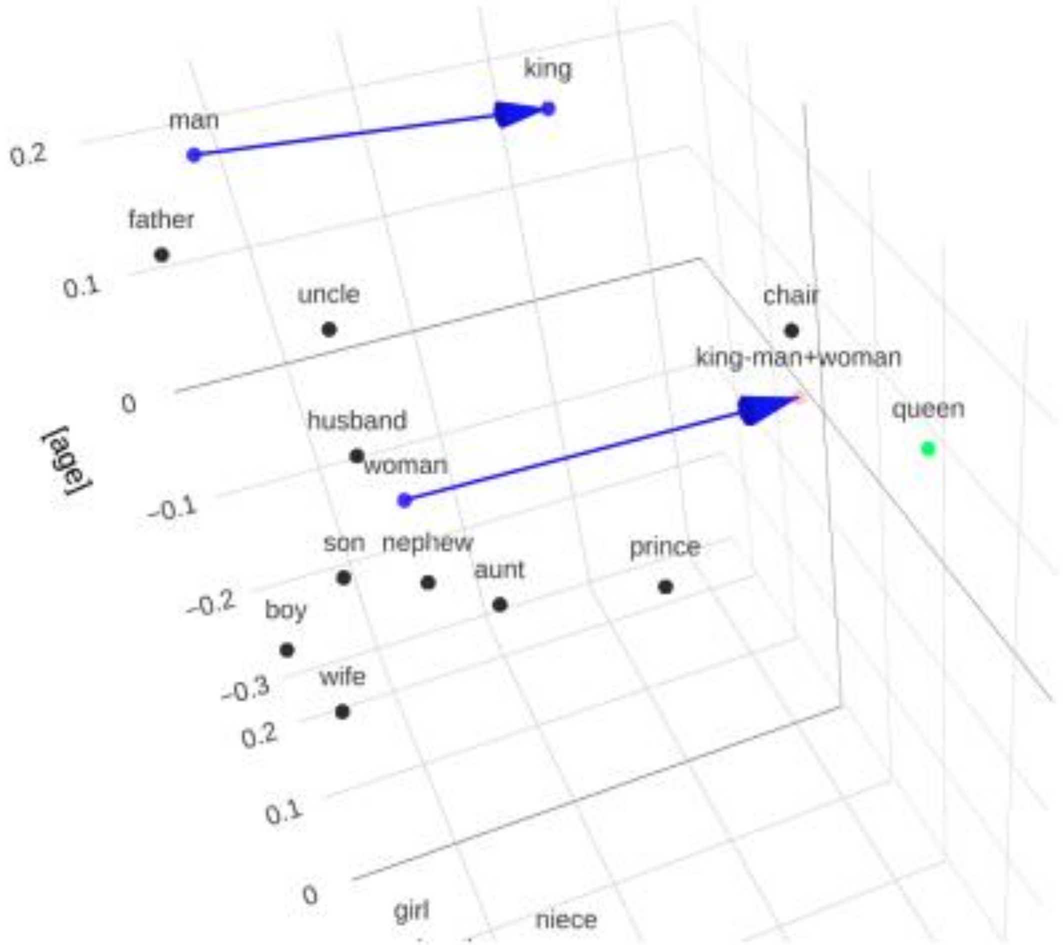


Figure 5: Analogy by vector arithmetic: “man” is to “king” as “woman” is to “king − man + woman” = “queen”.

# Low-dimensional, distributed representations

- Two similar words (e.g., synonyms or words under the same class) have similar distributional properties

- In neural models, replace the initial V-dimensional sparse vector with much smaller k-dimensional dense vectors

- Low-dimensional, dense word representations are extraordinarily powerful and are a large part of why neural network models have been so successful for NLP

# Count-based vs Prediction-based Methods

**LSA, HAL** (Lund & Burgess)

**Hellinger-PCA** (Rohde et al, Lebret & Collobert)

- Fast training
- Efficient usage of statistics
- Primarily used to capture word similarity
- Disproportionate importance given to large counts

**Skip-gram**/b (Mikolovet al)

**NLM**, **HLBL**, **RNN** (Bengioet al; Collobert & Weston; Huang et al; Mnih & Hinton)

- Scales with corpus size
- Inefficient usage of statistics
- Generated improved performance on other tasks
- Can capture complex patterns beyond word similarity

# Count-based and Prediction-based Methods

- ❑ Strong connection between count-based methods and prediction-based methods (Levy and Goldberg 2014)

- ❑ Skip-gram objective is equivalent to matrix factorization with PMI and discount for number of samples k

$$M_{w,c} = \mathrm{PMI}(w, c) - \log(k)$$

Neural Word Embedding as Implicit Matrix Factorization, (Levy & Goldberg, 2014)

# Other techniques and embeddings not covered

- ❑ Contrastive learning with negative samples
- ❑ Other variants
  - ~~○ Word2Vec (Mikolove et al., 14)~~
    - ✓ [https://code.google.com/archive/p/word2vec/](https://code.google.com/archive/p/word2vec/)
  - ○ GloVe (Pennington et al., 14)
    - ✓ [http://nlp.stanford.edu/projects/glove/](http://nlp.stanford.edu/projects/glove/)
  - ○ FastText (Bojanowski et al.’ 17)
    - ✓ [http://www.fasttext.cc/](http://www.fasttext.cc/)

# Word2Vec Demo

- Pre-trained word2vec models:
  - https://code.google.com/archive/p/word2vec/
- Gensim:
  - https://radimrehurek.com/gensim/auto_examples/tutorials/run_word2vec.html
- Online demos:
  - http://nlp.polytechnique.fr/word2vec
  - http://vectors.nlpl.eu/explore/embeddings/en/
  - https://remykarem.github.io/word2vec-demo/

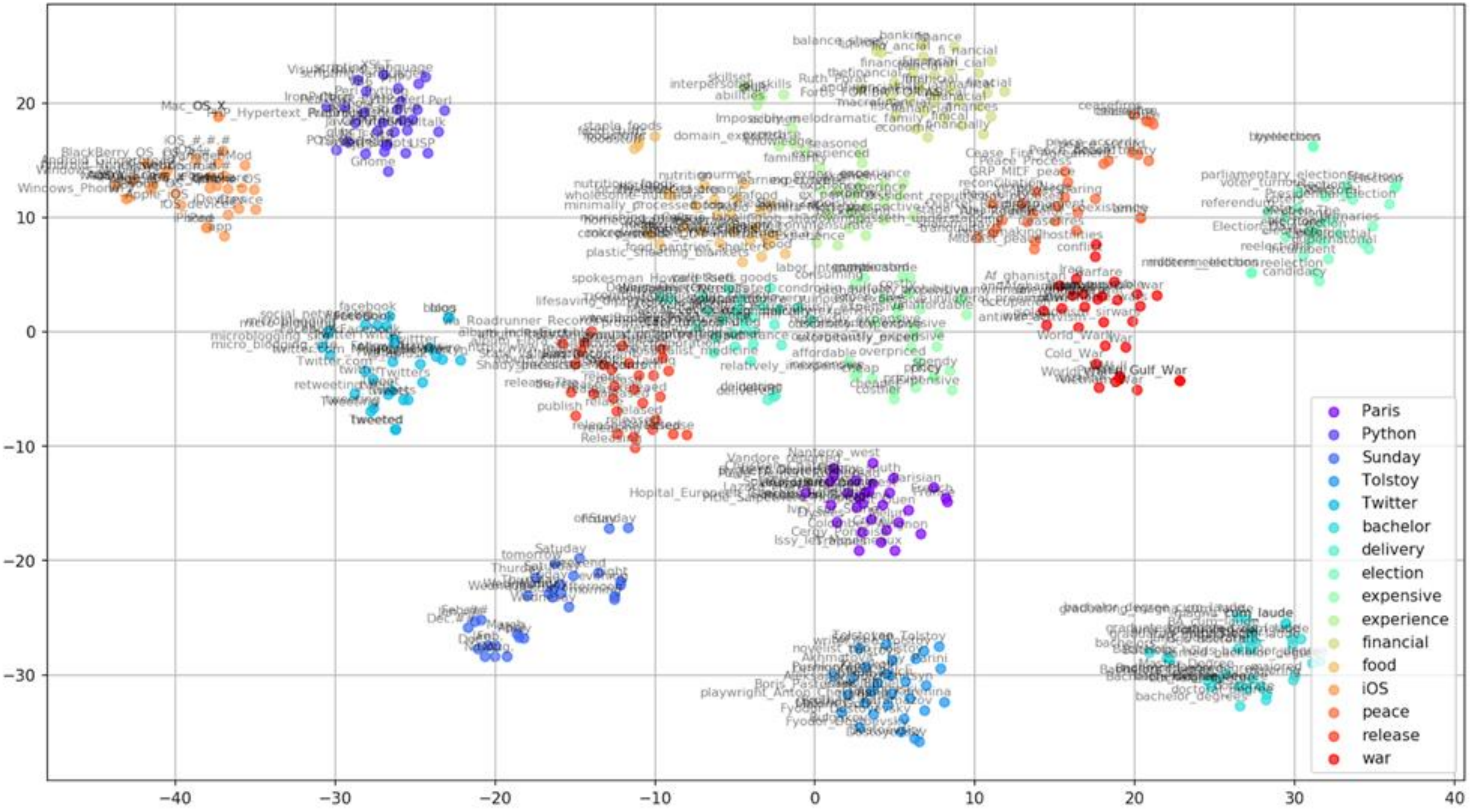
# Types of Evaluation

# Types of Evaluation

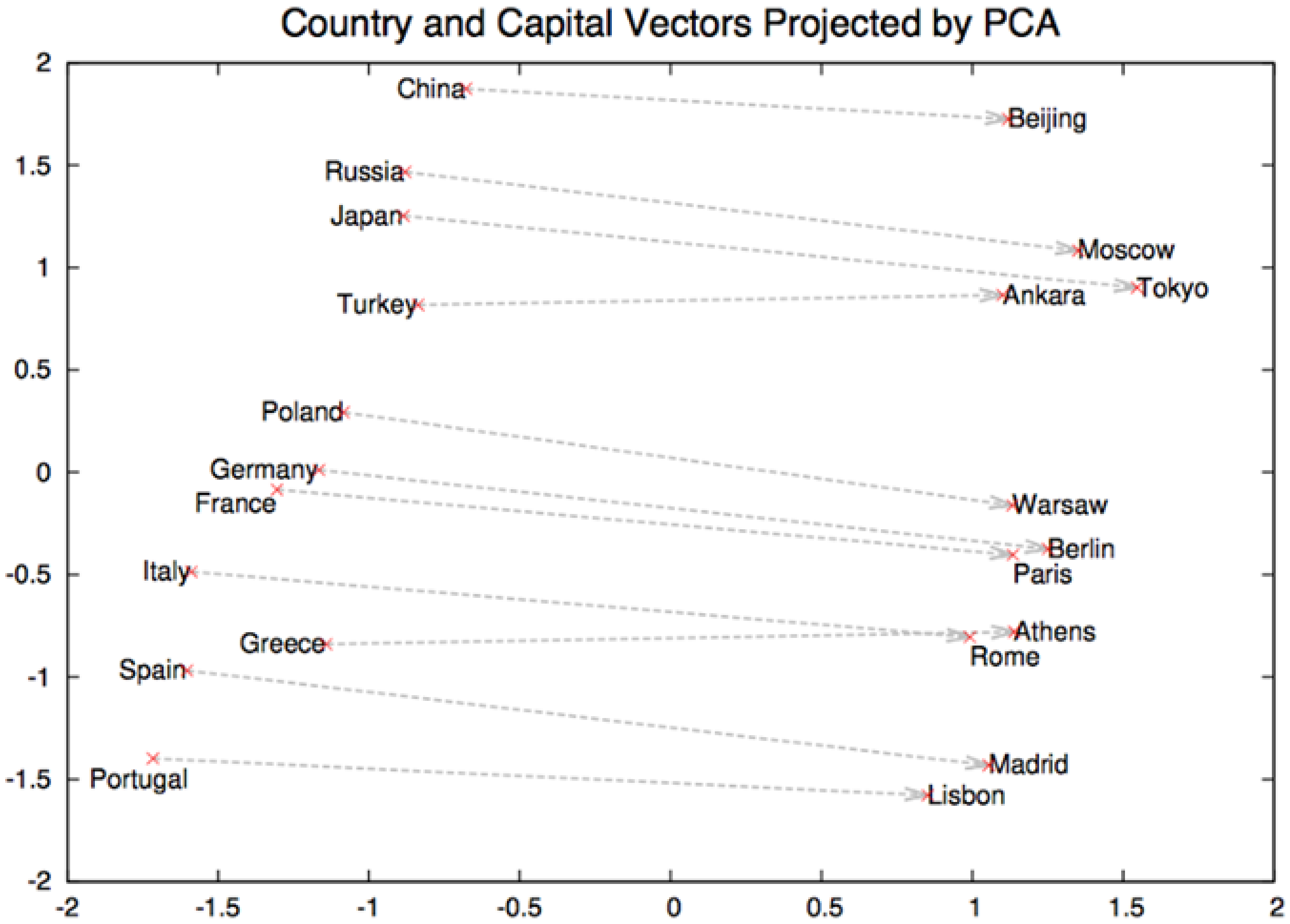
- Intrinsic vs Extrinsic
  - Intrinsic: How good is it based on its features?
  - Extrinsic: How useful is it downstream?

- Qualitative vs. Quantitative
  - Qualitative: Examine the characteristics of examples.
  - Quantitative: Calculate statistics

# Visualization of Embeddings



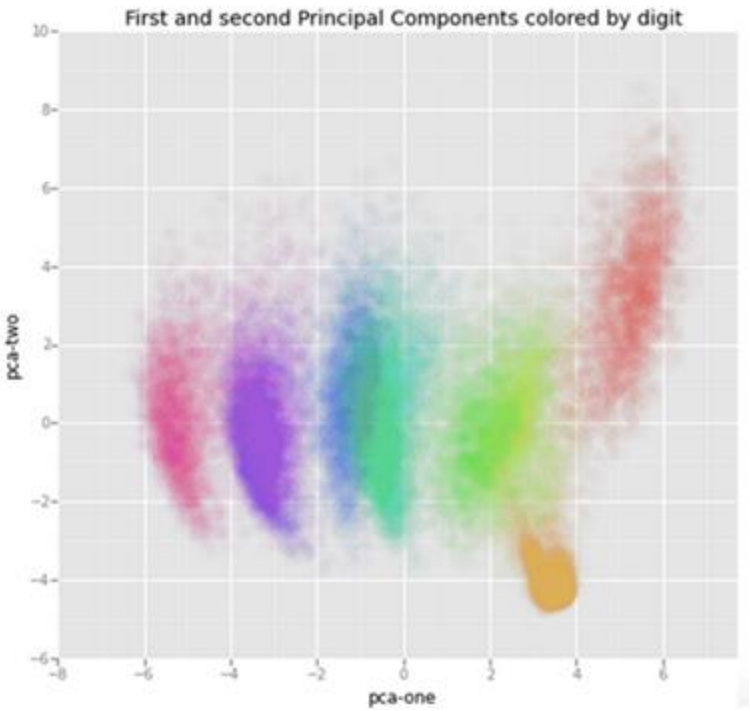
# Visualization of Embeddings

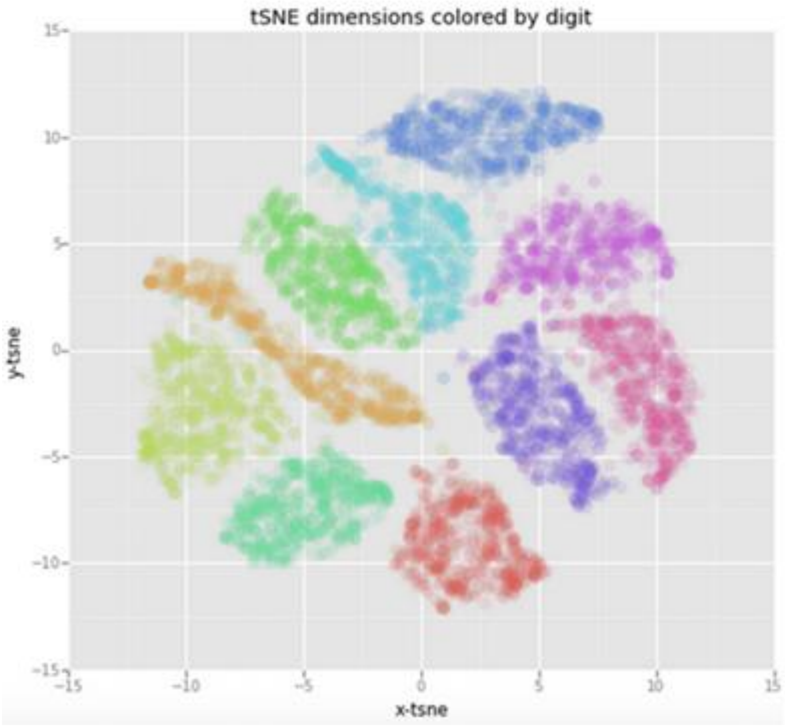


# Linear and Non-linear Projection

- Non-linear projections group things that are close in high-dimensional space
  - e.g. SNE/t-SNE (van der Maaten and Hinton 2008) group things that give each other a high probability according to a Gaussian

PCA



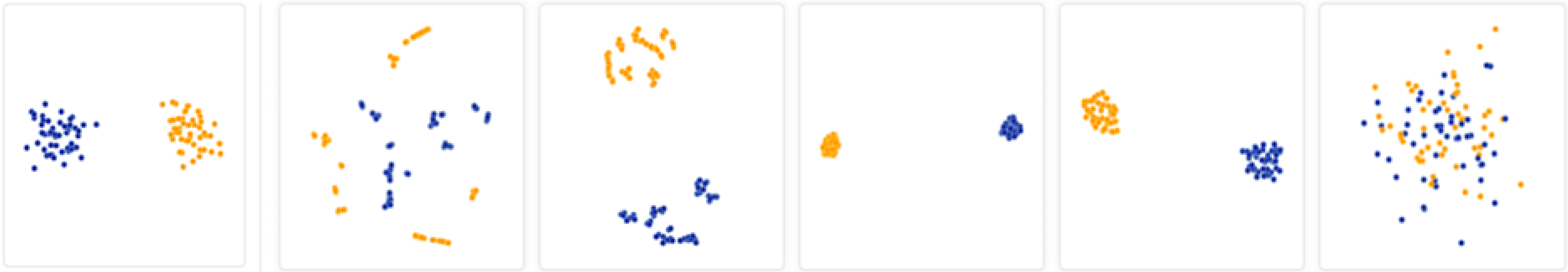


T-SNE

Image from Derksen (2016)

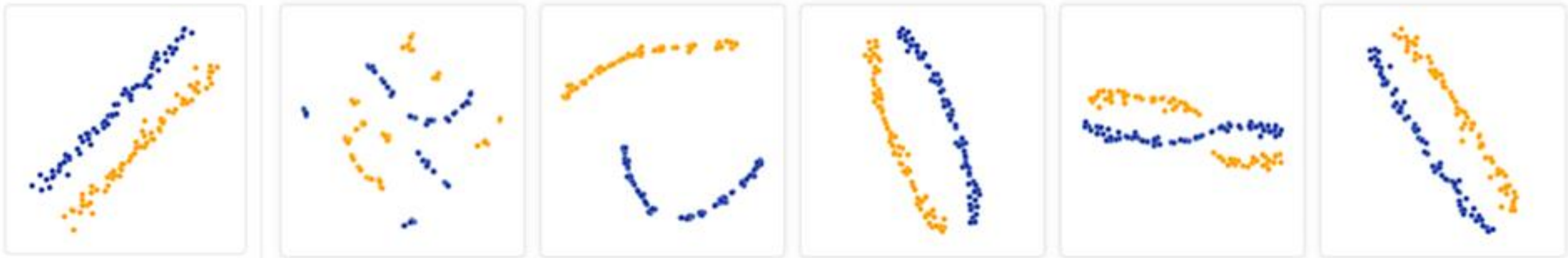
# t-SNE Visualization can be Misleading! (Wattenberg et al. 2016)

## Settings matter



*Original* | Perplexity: 2 Step: 5,000 | Perplexity: 5 Step: 5,000 | Perplexity: 30 Step: 5,000 | Perplexity: 50 Step: 5,000 | Perplexity: 100 Step: 5,000

## Linear correlations cannot be interpreted



*Original* | Perplexity: 2 Step: 5,000 | Perplexity: 5 Step: 5,000 | Perplexity: 30 Step: 5,000 | Perplexity: 50 Step: 5,000 | Perplexity: 100 Step: 5,000

# Intrinsic Evaluation of Embeddings

- ❑ **Relatedness**: The correlation between embedding cosine similarity and human eval of similarity?
- ❑ **Analogy**: Find x for "a is to b, as x is to y".

- ❑ **Categorization**: Create clusters based on the embeddings, and measure purity of clusters.
- ❑ **Selectional Preference**: Determine whether a noun is a typical argument of a verb.

(categorization from Schnabel et al 2015)

# Intrinsic evaluation:

Ask humans how similar two words are

Relatedness:
correlation (Spearman/Pearson) between vector similarity of pair of words and human judgments

| Word 1 | Word 2 | 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)

WordSim-353 dataset (Finkelstein et al., 2002)

# Intrinsic evaluation:

Analogical reasoning (Mikolov et al., 2013).

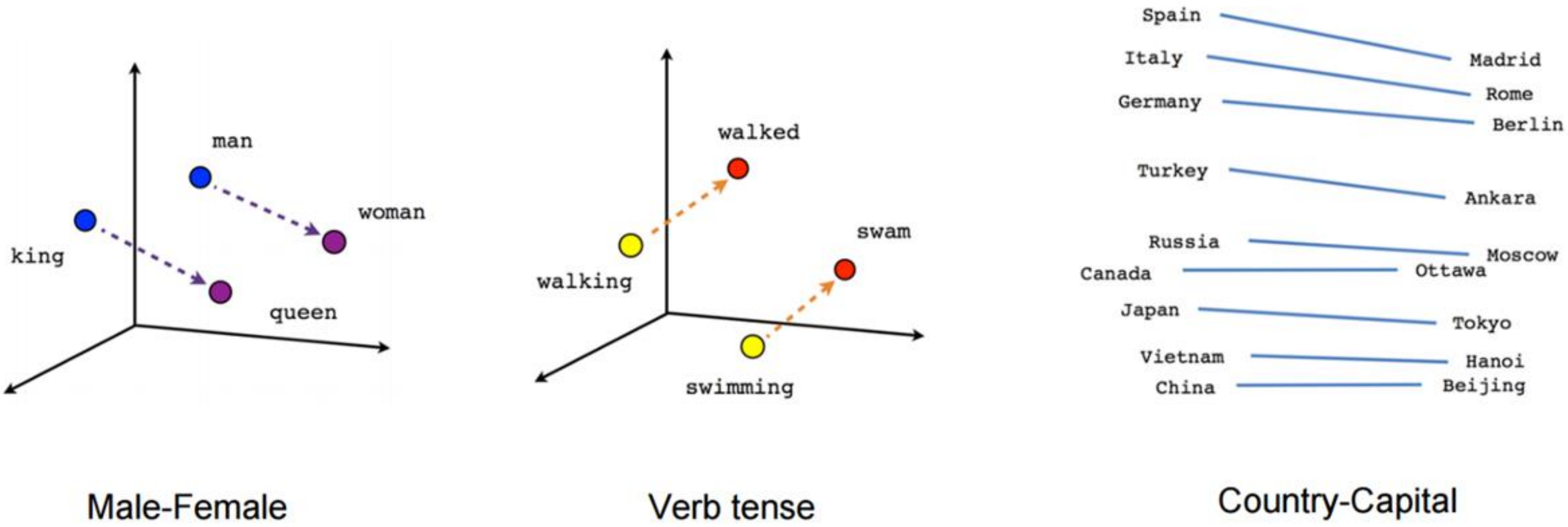
For analogy Germany : Berlin :: France : ?,
find closest vector to v("Berlin") – v("Germany")+v("France")

| | | |
|---|---|---|
| possibly | impossibly | Certain |
| generating | generated | Shrinking |
| think | thinking | Look |
| Baltimore | Maryland | Minneapolis |
| shrinking | shrank | Slowing |
| Rabat | Morocco | Astana |

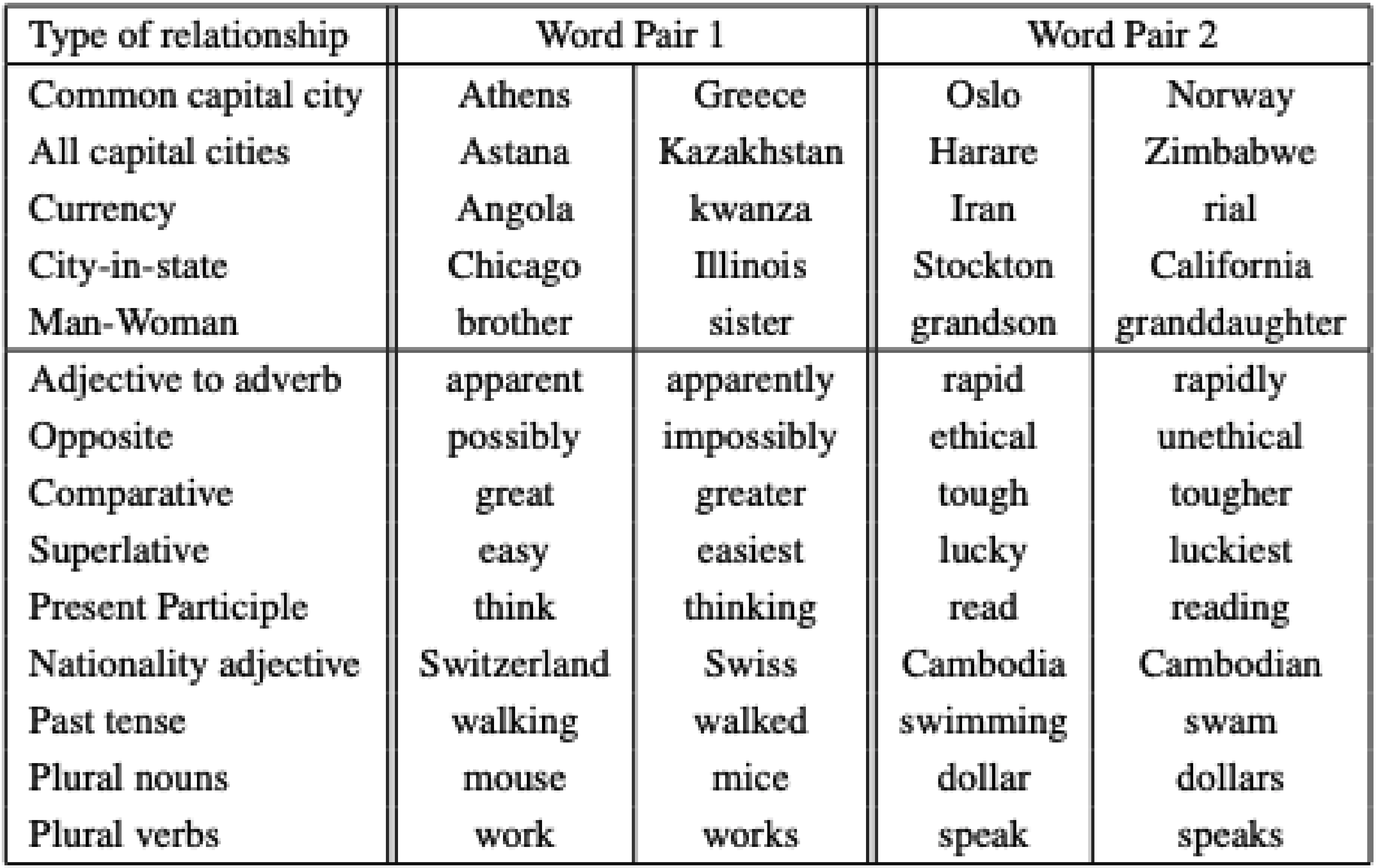
| |
|---|
| Uncertain |
| Shrank |
| Looking |
| Minnesota |
| Slowed |
| Kazakhstan |

# Intrinsic evaluation:

Analogical reasoning (Mikolov et al., 2013).



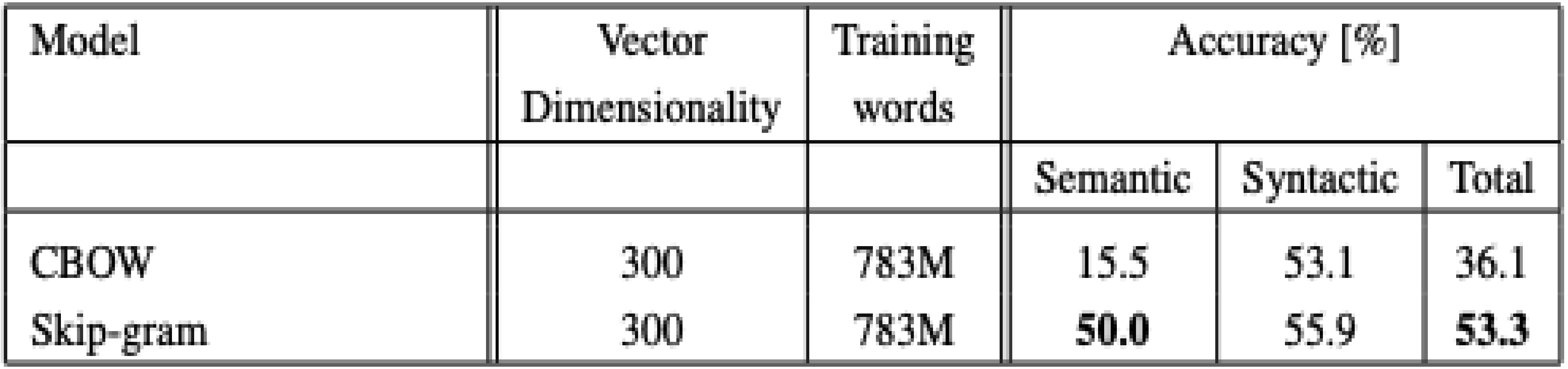
# Analogical reasoning test



| Type of relationship | Word Pair 1 | | Word Pair 2 | |
|---|---|---|---|---|
| Common capital city | Athens | Greece | Oslo | Norway |
| All capital cities | Astana | Kazakhstan | Harare | Zimbabwe |
| Currency | Angola | kwanza | Iran | rial |
| City-in-state | Chicago | Illinois | Stockton | California |
| Man-Woman | brother | sister | grandson | granddaughter |
| Adjective to adverb | apparent | apparently | rapid | rapidly |
| Opposite | possibly | impossibly | ethical | unethical |
| Comparative | great | greater | tough | tougher |
| Superlative | easy | easiest | lucky | luckiest |
| Present Participle | think | thinking | read | reading |
| Nationality adjective | Switzerland | Swiss | Cambodia | Cambodian |
| Past tense | walking | walked | swimming | swam |
| Plural nouns | mouse | mice | dollar | dollars |
| Plural verbs | work | works | speak | speaks |

Mikolov et al. 2013

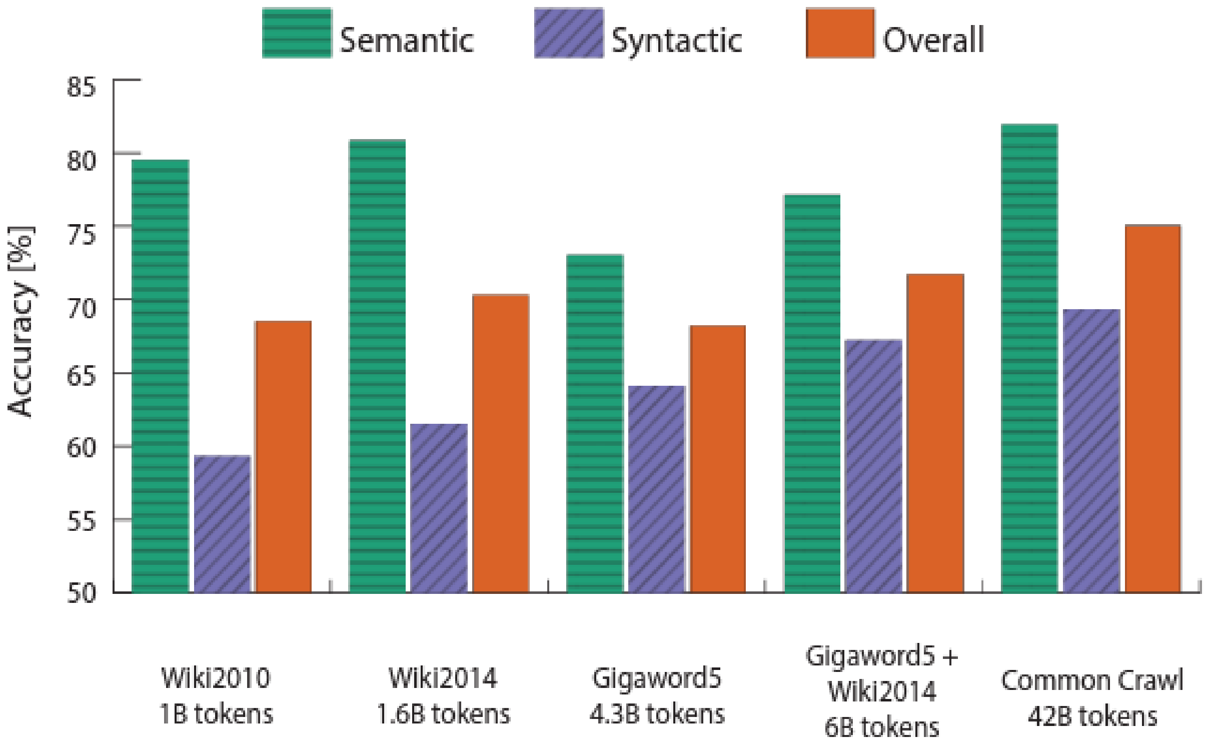
# Analogical reasoning test



| Model | Vector Dimensionality | Training words | Accuracy [%] | | |
|---|---|---|---|---|---|
| | | | Semantic | Syntactic | Total |
| CBOW | 300 | 783M | 15.5 | 53.1 | 36.1 |
| Skip-gram | 300 | 783M | **50.0** | 55.9 | **53.3** |

Mikolov et al. 2013

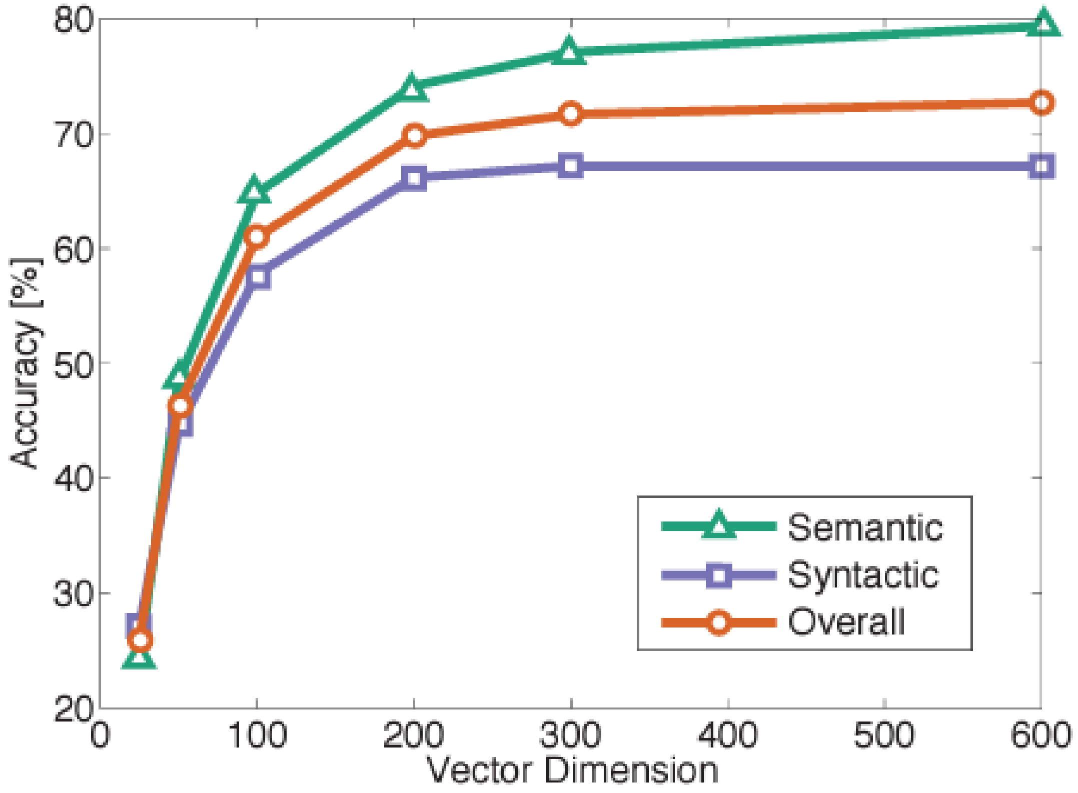
# Analogy evaluation and hyper-parameters



- ❑ More data helps
- ❑ Wikipedia is better than news text

Mikolov et al. 2013

# Analogy evaluation and hyper-parameters

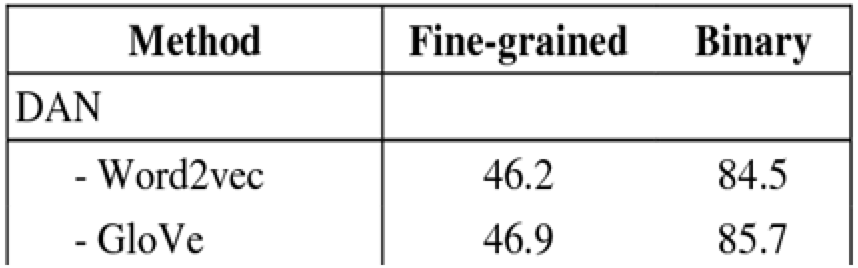


- ❑ Dimensionality
- ❑ Good dimension is ~300

Mikolov et al. 2013

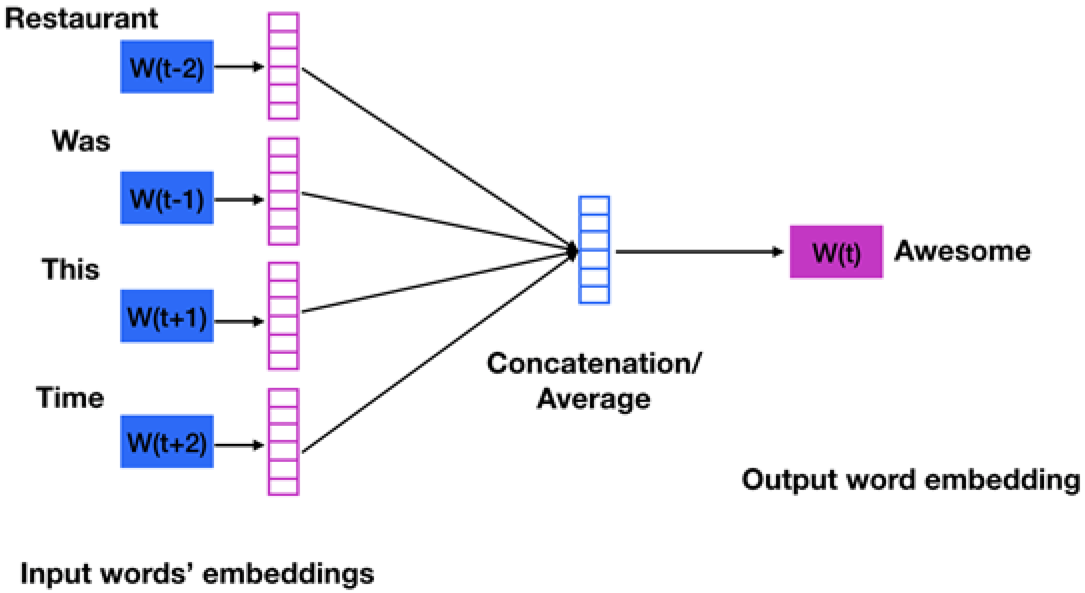
# Extrinsic Evaluation

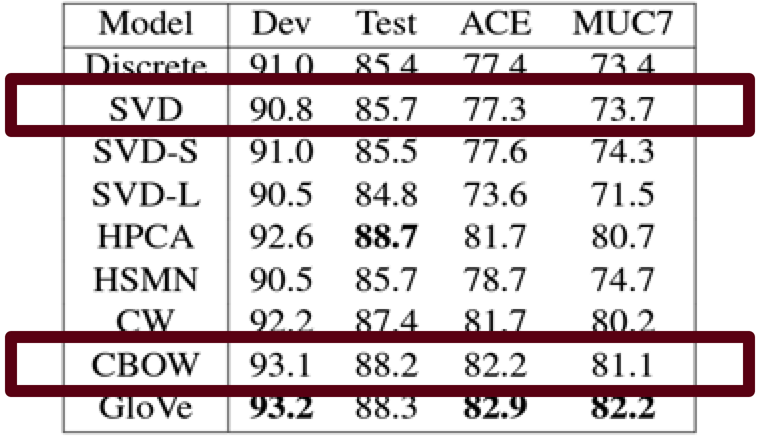
Be aware and use the best one for the task



| Method | Fine-grained | Binary |
|---|---|---|
| DAN | | |
| - Word2vec | 46.2 | 84.5 |
| - GloVe | 46.9 | 85.7 |

Sentiment classification





| Model | Dev | Test | ACE | MUC7 |
|---|---|---|---|---|
| Discrete | 91.0 | 85.4 | 77.4 | 73.4 |
| SVD | 90.8 | 85.7 | 77.3 | 73.7 |
| SVD-S | 91.0 | 85.5 | 77.6 | 74.3 |
| SVD-L | 90.5 | 84.8 | 73.6 | 71.5 |
| HPCA | 92.6 | **88.7** | 81.7 | 80.7 |
| HSMN | 90.5 | 85.7 | 78.7 | 74.7 |
| CW | 92.2 | 87.4 | 81.7 | 80.2 |
| CBOW | 93.1 | 88.2 | 82.2 | 81.1 |
| GloVe | **93.2** | 88.3 | **82.9** | **82.2** |

Named Entity Recognition: identifying references to a person, organization or location:

# When are Pre-trained Embeddings Useful?

- Basically, when training data is insufficient
  - E.g. Low-resource languages

- **Very useful**: tagging, parsing, text classification

- **Less useful:** machine translation

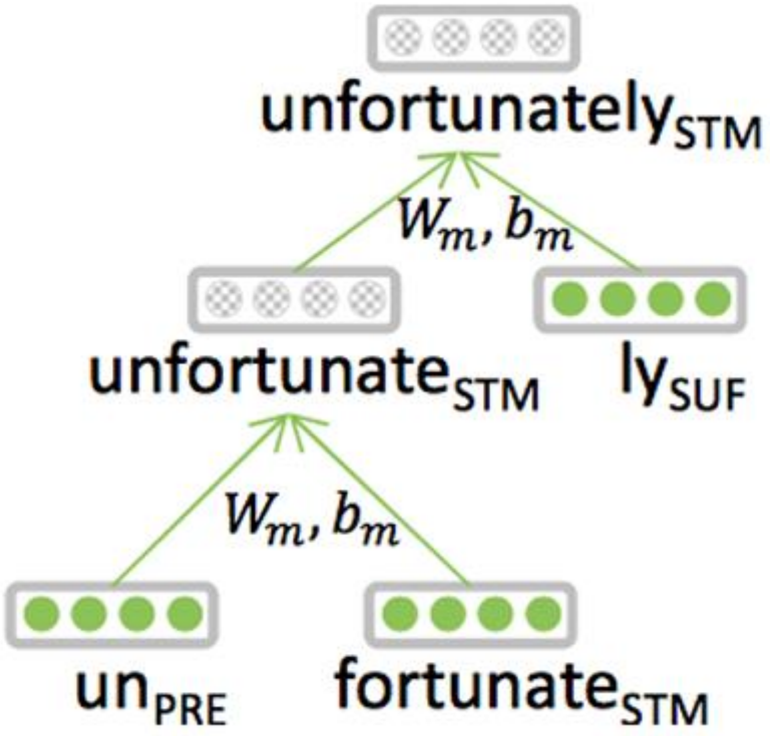
- **Basically not useful:** language modeling

# Limitations of Word Embeddings

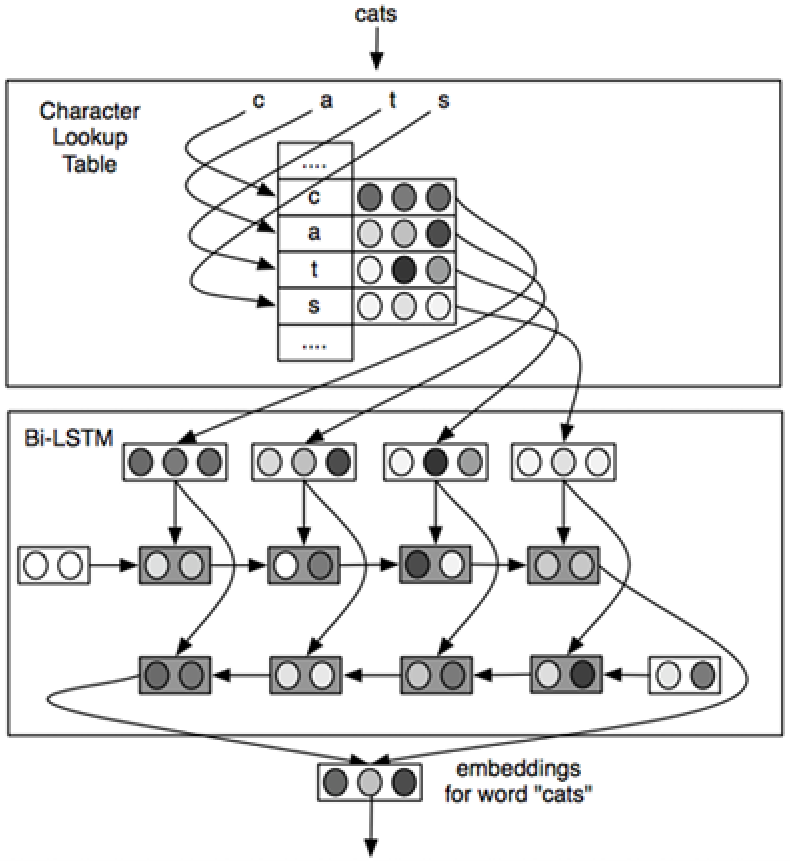
# Limitations of Embeddings

- Sensitive to **superficial differences** (dog / dogs)
  - E.g. misspellings: “minuscule” → “miniscule”
  - E.g. compounded/prefixed/suffixed words split into “wrong” subwords “descheduled” ⇒ [ “des”, “##ched”, “##uled” ]
- **Not necessarily coordinated** with knowledge or across languages
- Can encode **bias** (encode stereotypical gender roles, racial biases)

# Sub-word Embeddings

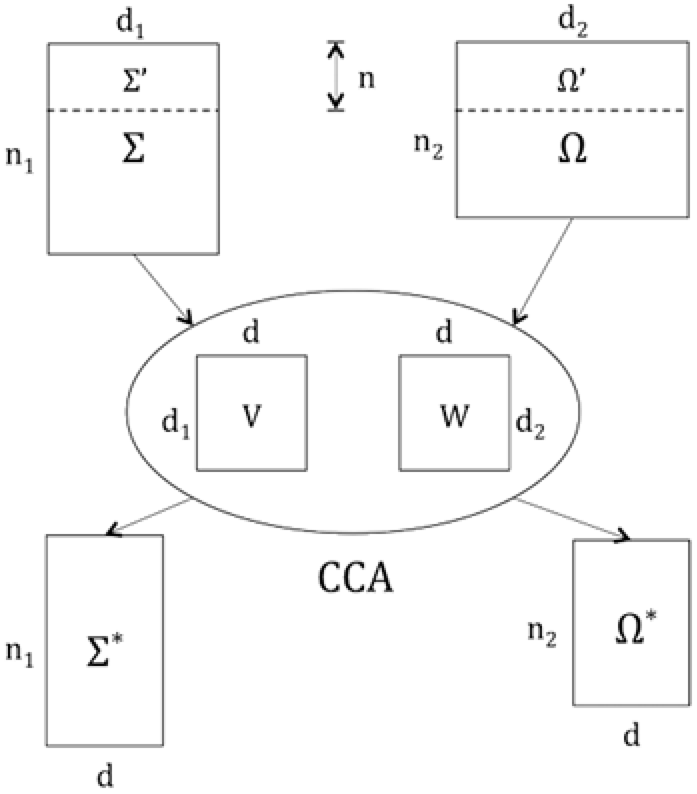


Morpheme-based (Luong et al. 2013)

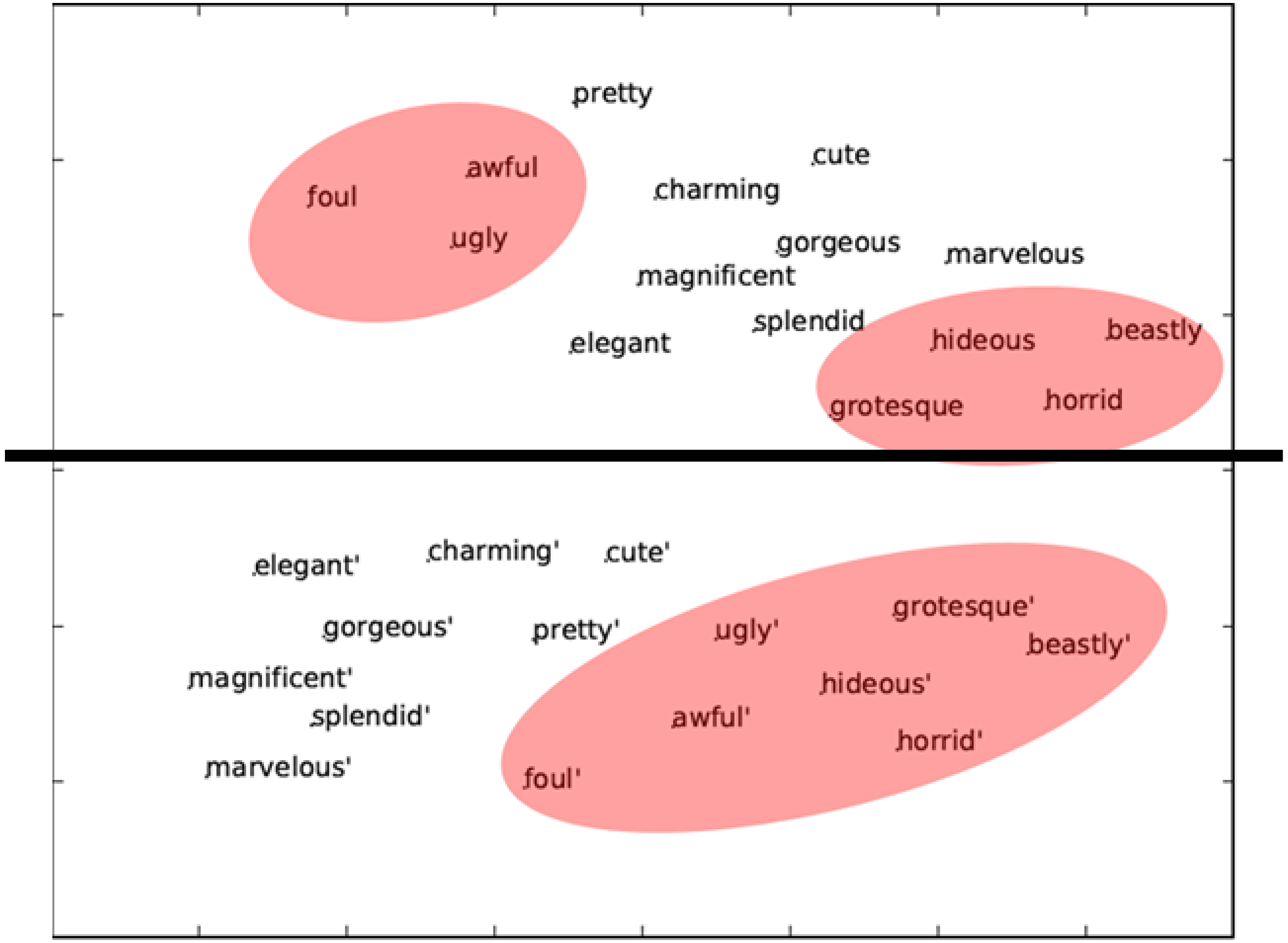


Character-based (Ling et al. 2015)

# Multilingual Coordination of Embeddings using dictionaries



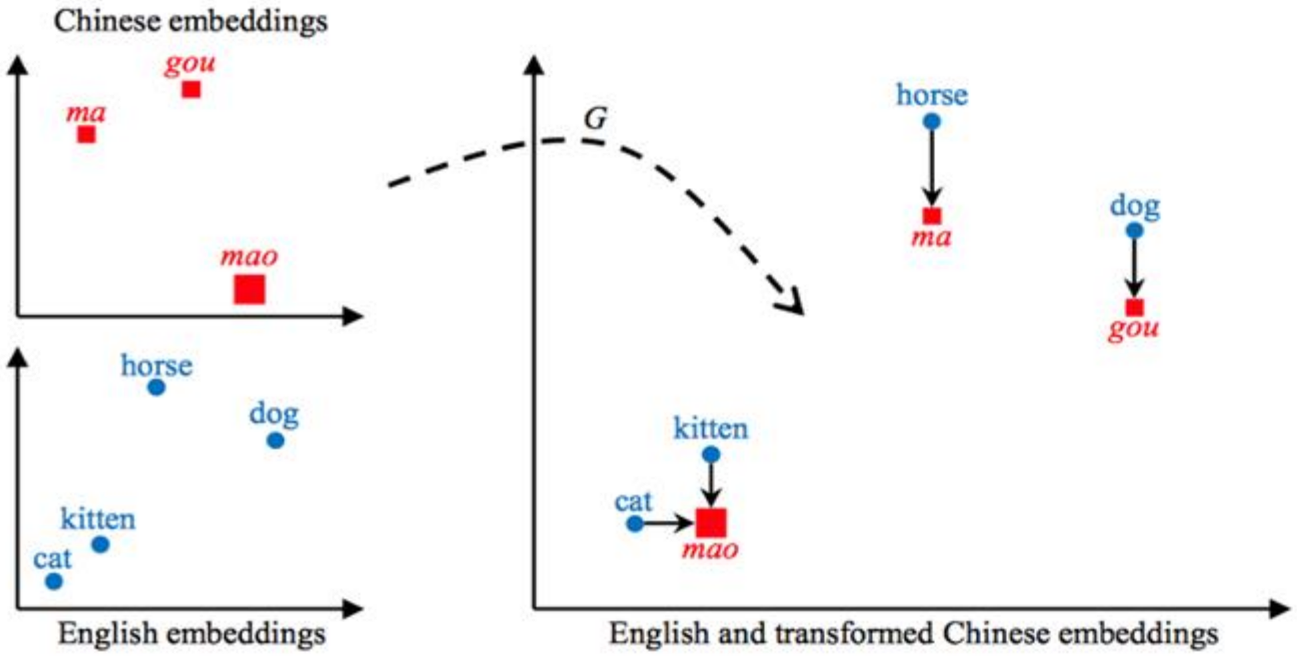
Improving Vector Space Word Representations Using Multilingual Correlation (Faruqui & Dyer, 2014)



Monolingual (top) and multilingual (bottom) word projections of the antonyms (shown in red) and synonyms of "beautiful"

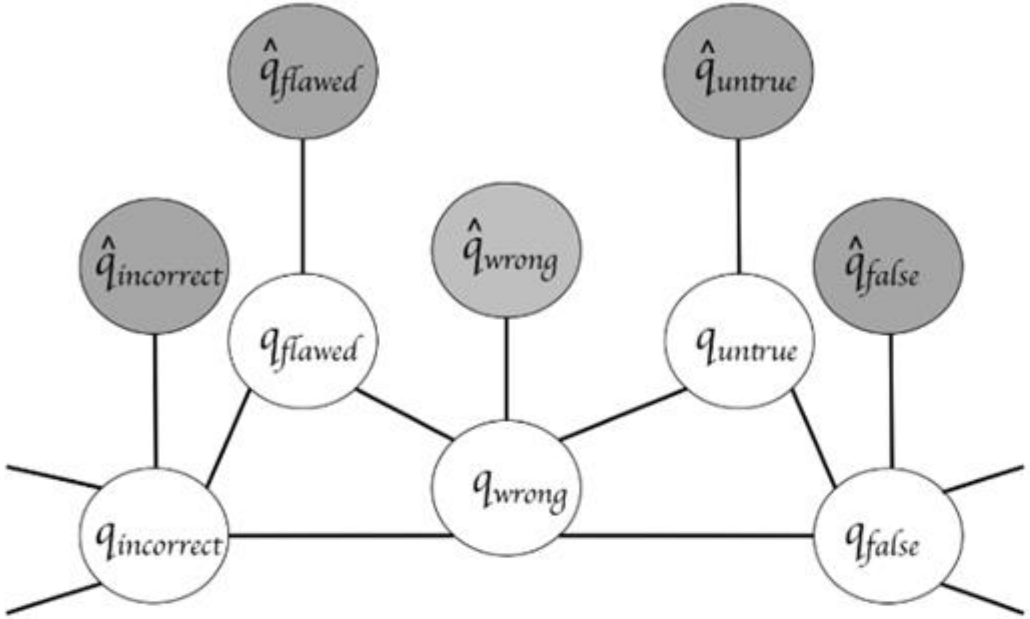
# Unsupervised Coordination of Embeddings

- ❑ In some cases, we can do it with no dictionary at all!
  - ○ Just use identical words, e.g. the digits (Artexte et al. 2017)
  - ○ Or, just match distributions (Zhang et al. 2017)



# Retrofitting of Embeddings to Existing Lexicons

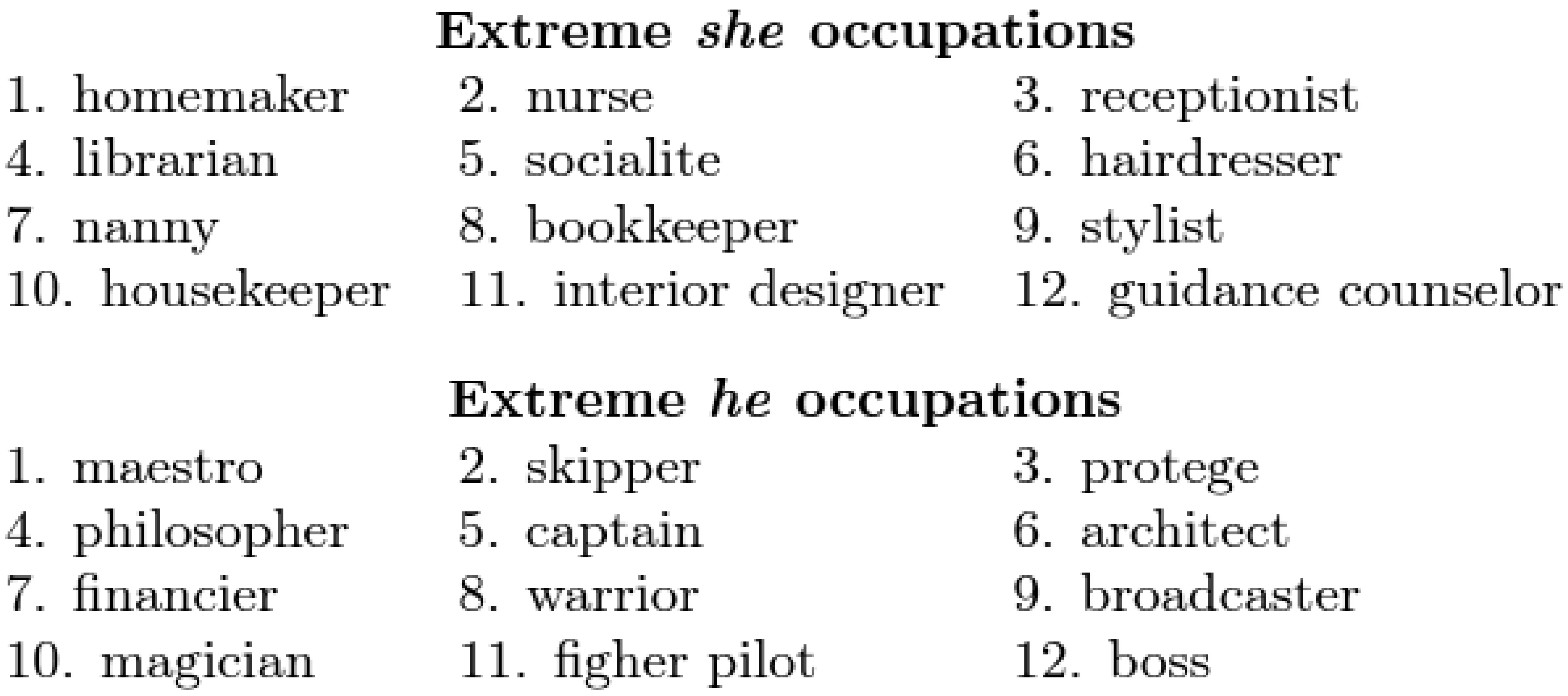
- ❑ Make word vectors to match with existing lexicon like WordNet (Faruqui et al. 2015)



$$\Psi(Q) = \sum_{i=1}^{n} \left[ \alpha_i \|q_i - \hat{q}_i\|^2 + \sum_{(i,j)\in E} \beta_{ij} \|q_i - q_j\|^2 \right]$$

# De-biasing Word Embeddings

Word embeddings reflect bias in statistics



**Extreme *she* occupations**

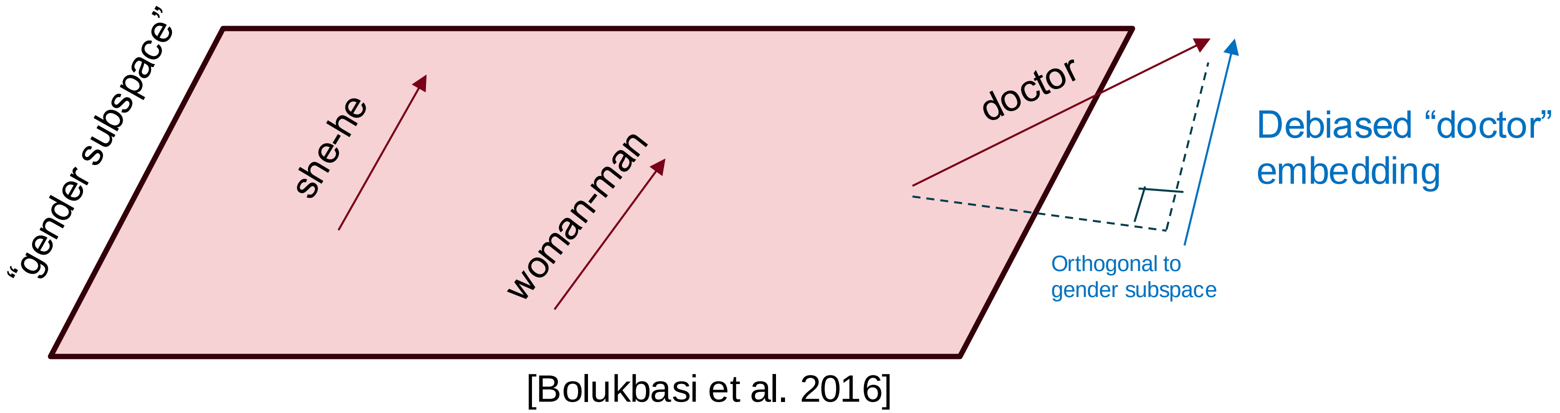
1. homemaker
2. nurse
3. receptionist
4. librarian
5. socialite
6. hairdresser
7. nanny
8. bookkeeper
9. stylist
10. housekeeper
11. interior designer
12. guidance counselor

**Extreme *he* occupations**

1. maestro
2. skipper
3. protege
4. philosopher
5. captain
6. architect
7. financier
8. warrior
9. broadcaster
10. magician
11. figher pilot
12. boss

(Bolukbasi et al. 2016)

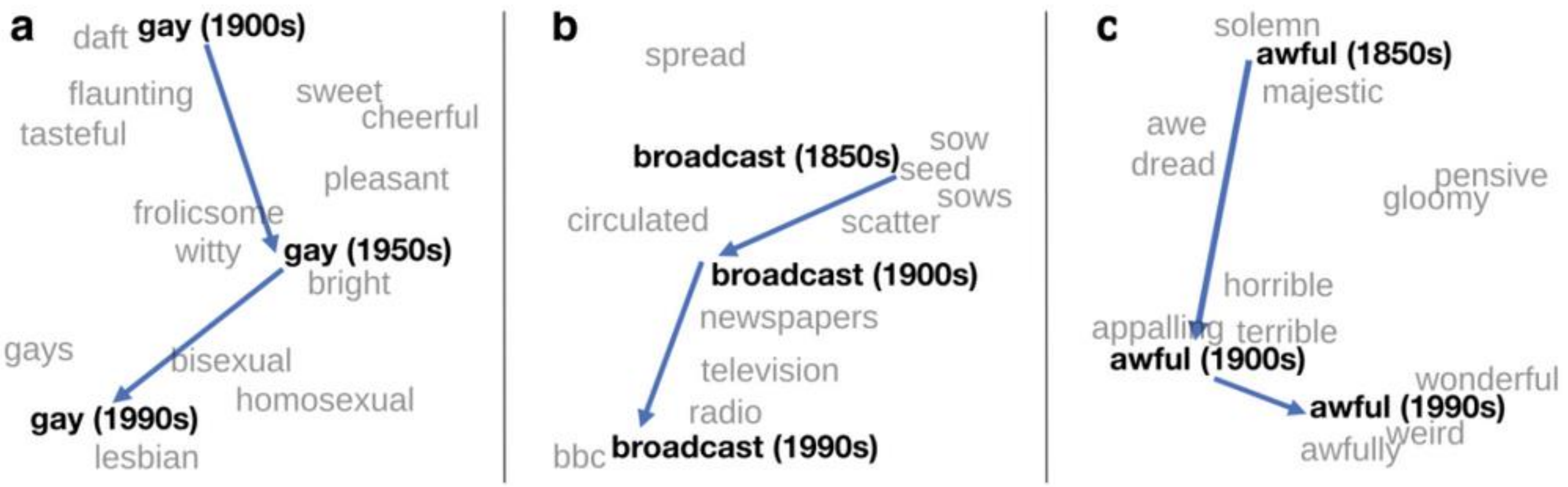
# De-biasing Word Embeddings



[Bolukbasi et al. 2016]

Identify pairs to "neutralize", find the direction of the trait to neutralize, and ensure that they are neutral in that direction

(Bolukbasi et al. 2016)



semantic to model temporal word analogy or relatedness (Szymanski, 2017; Rosin et al., 2017) or to capture the dynamics of semantic relations (Kutuzov et al., 2017)

# Different kinds of encoding “context”

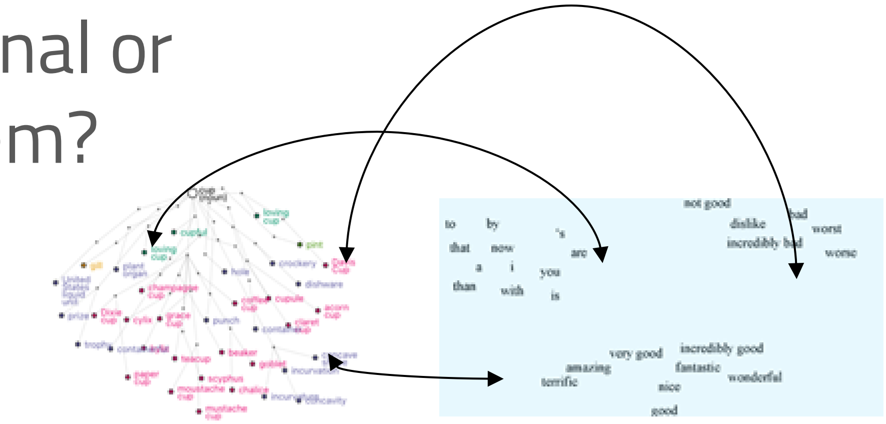
- ❑ ~~Count-based~~
  - ○ PMI, TF-IDF
- ❑ ~~Distributed prediction-based (type) embeddings~~
  - ○ Word2vec, GloVe, Fasttext
- ❑ **Distributed contextual (token) embeddings from language models**
  - ○ ELMo, BERT, GPT
- ❑ ~~Many more variants~~
  - ○ Multilingual / multi-sense / syntactic embeddings, etc

# Questions

- ☐ How do you define "beautiful"? Is it more difficult to define "beautiful" than it is to define "cup"?

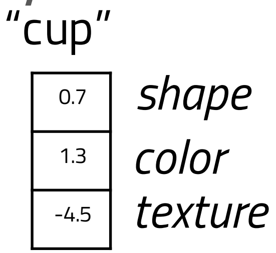
- ☐ How does your representation know "elephant" is bigger than "mouse"?

- ☐ Distributional semantics can be de-compositional or ontologically represented? How to combine them?



# Questions

- We’ve just learned how to learn the meaning of “bank” from data as a dense vector. What if meaning of “bank” can be different by context? Can we learn the vectors dynamically adaptable by context?
- How do you interpret the vector? You only know the “relationship” between words but not meaning of word itself. Does each dimension of the vector in distributional semantics correspond to “component” in the decompositional semantics?
- Some words like “war” include various information. Can we quantify the abstract nature of words in distributed representations?



“cup”

0.7 *shape*

1.3 *color*

-4.5 *texture*