

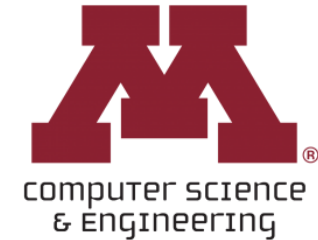
Announcement (0210)

- ❑ HW2 is now due 02/16
- ❑ Slack channels will be created for you in the coming days with your project information – if you have not been assigned to a slack channel for the project by Wednesday (which will have a name like prj-[TA1]-[TA2]-[TeamName]), then reach out to myself in slack
- ❑ Looking for note taker (Please reach out to me over slack – will be added to participation score)



CSCI 5541: Natural Language Processing

Lecture 5: Distributional Semantics and Word Embeddings



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

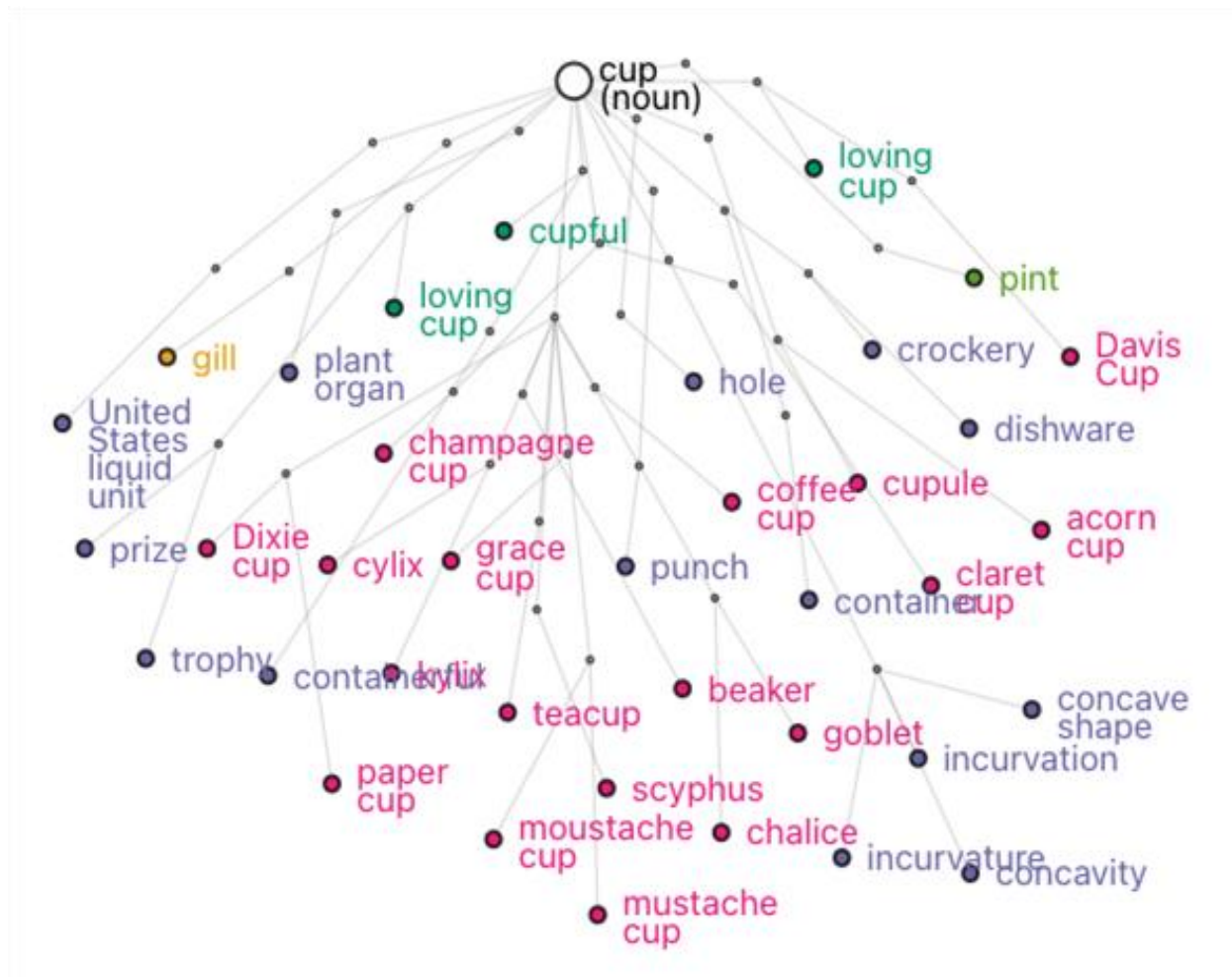
Three ways of looking at word meaning

- ❑ 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
- hyponym
- attribute
- antonym
- holonym
- entailment
- hypernym
- meronym
- cause

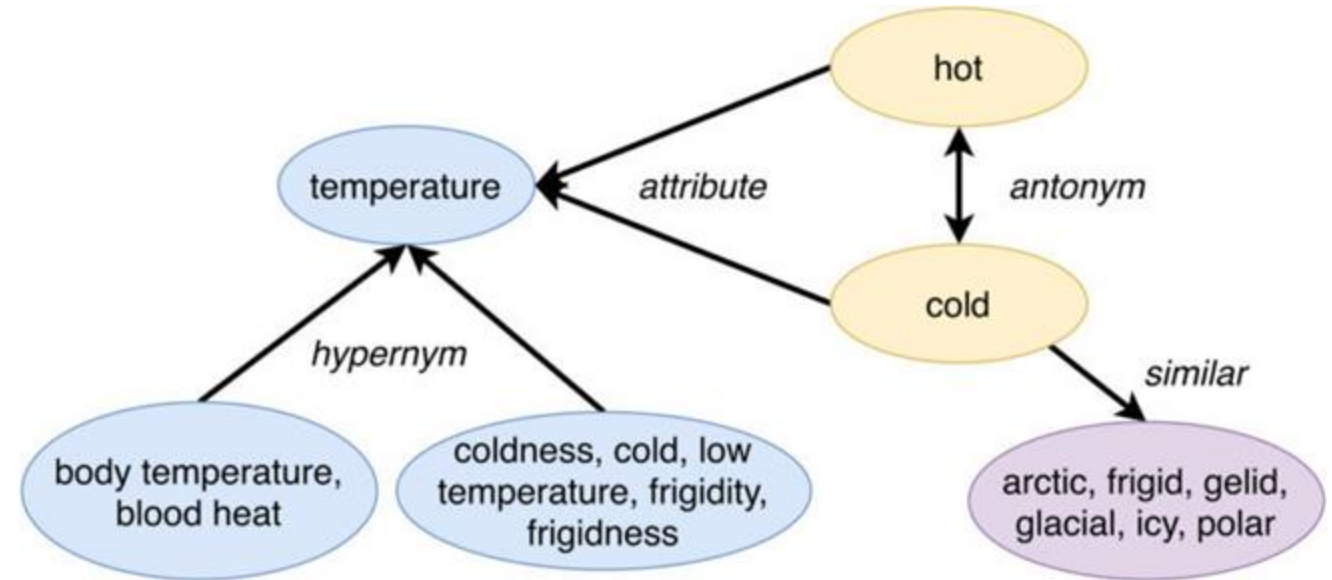


<https://lexical-graph.herokuapp.com/>



Semantic relations

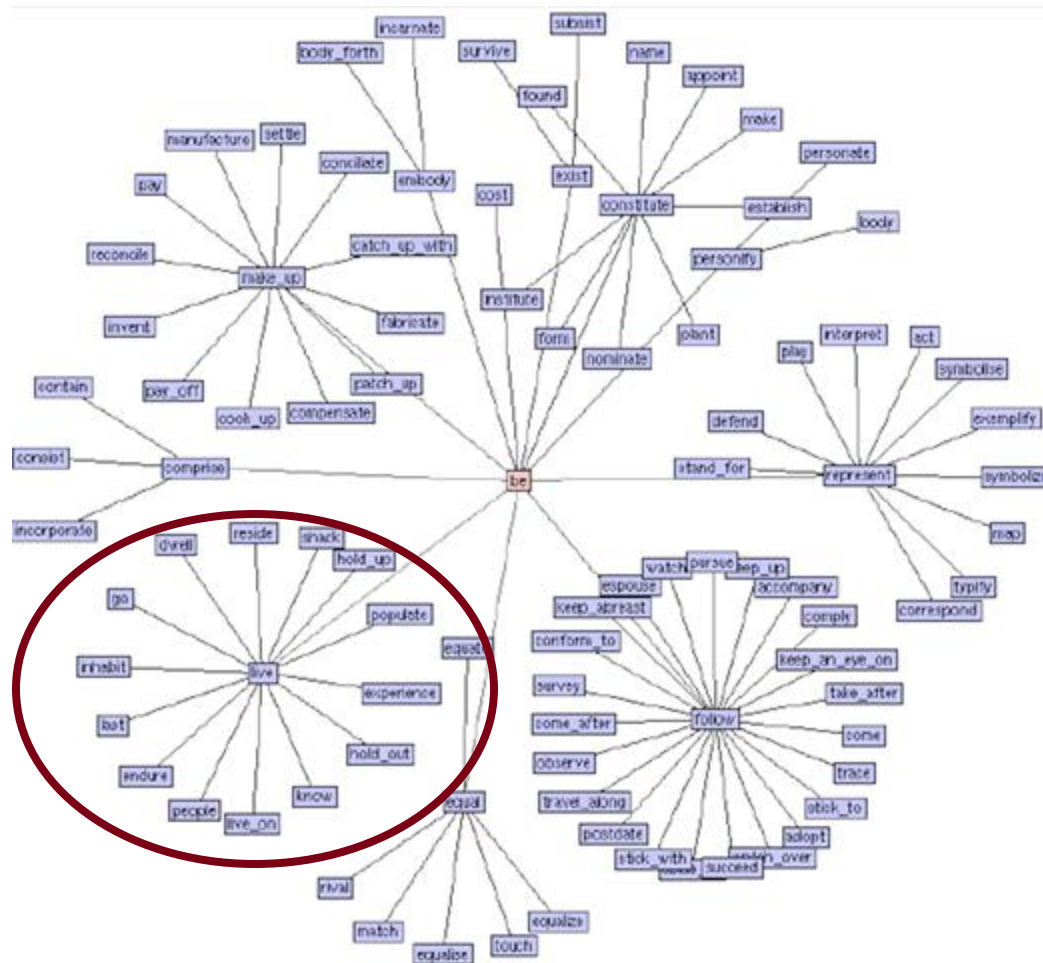
- Synonymy — equivalence
 - <small, little>
- Antonymy — opposition
 - <small, large>
- Meronymy — part-of relation
 - <liver, body>
- Holonymy — has-a relation
 - <body, liver>
- **Hyponymy** — subset; is-a relation
 - <dog, mammal>
- **Hypernymy** — superset
 - <mammal, dog>



WordNet

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

Synset



Tools for WordNet

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

WordNet Search - 3.1
[- WordNet home page](#) - [Glossary](#) - [Help](#)

Word to search for:

Display Options:

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

Noun

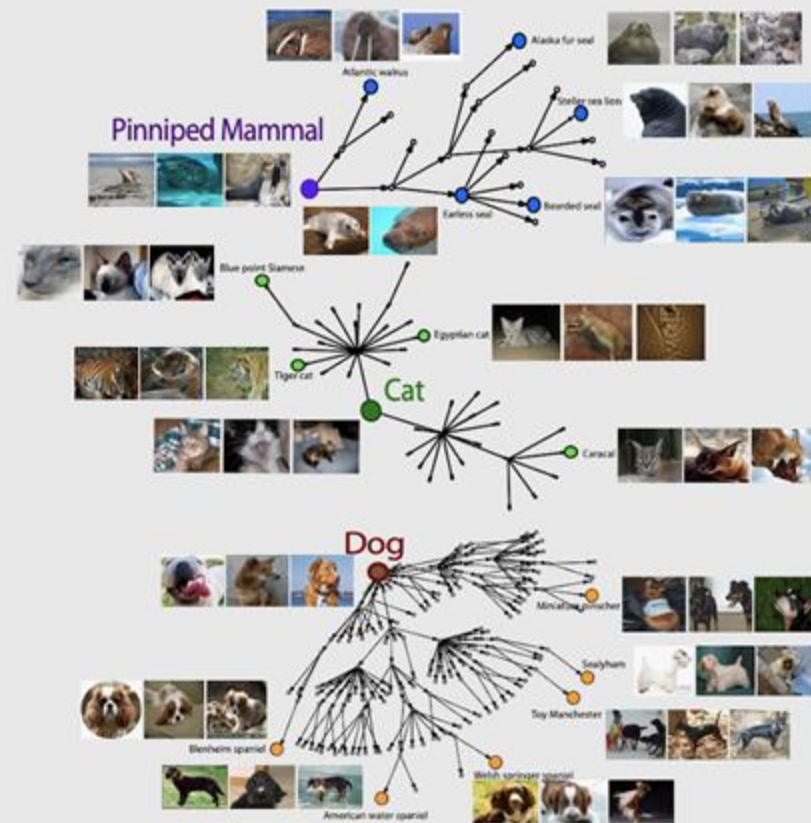
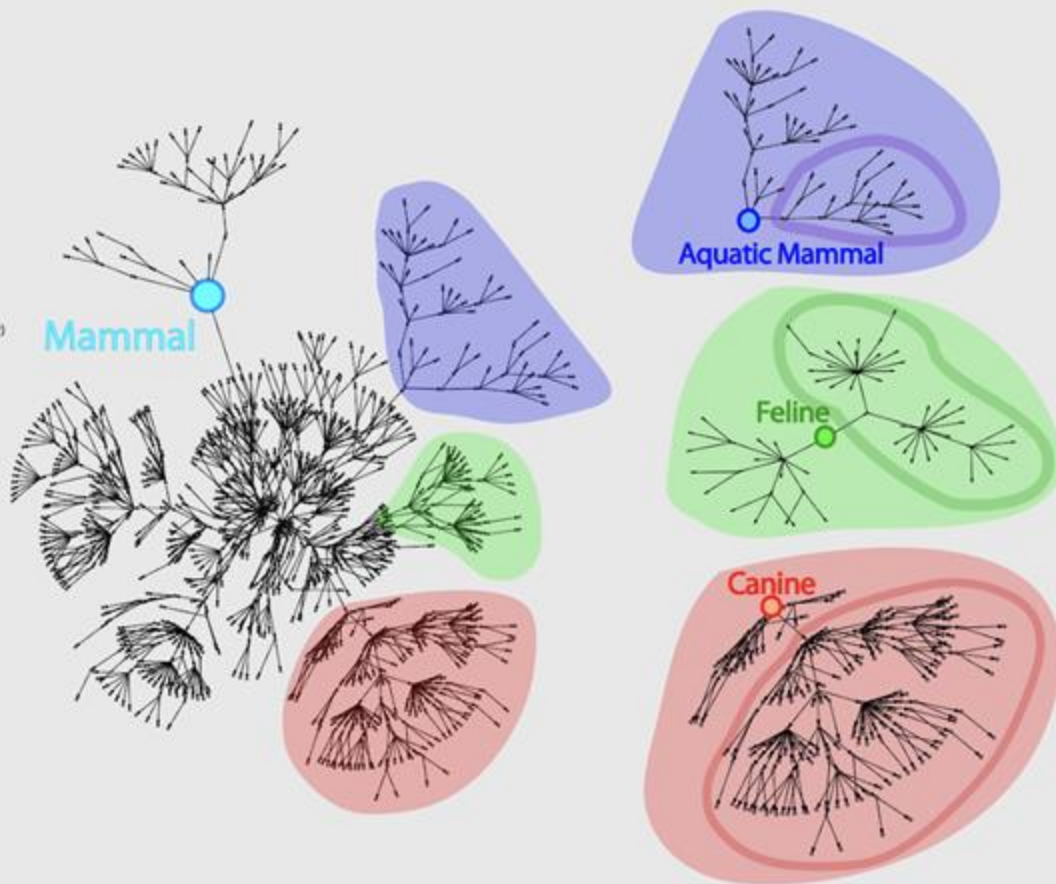
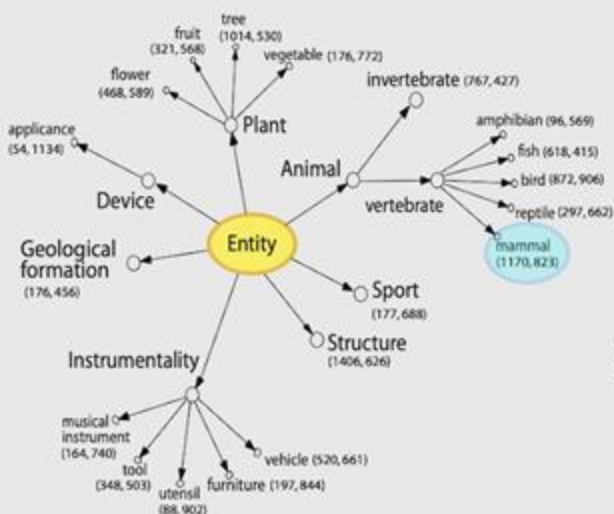
- [S:](#) (n) [jump](#), [leap](#) (a sudden and decisive increase) "a jump in attendance"
- [S:](#) (n) [leap](#), [jump](#), [saltation](#) (an abrupt transition) "a successful leap from college to the major leagues"
- [S:](#) (n) [jump](#) ((film) an abrupt transition from one scene to another)
- [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



ImageNet based on WordNet

ImageNet Trees



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*



  /kʌp/  /kʌp/ 

1   [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





2   [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   [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** 

+ Oxford Collocations Dictionary

— Extra Examples

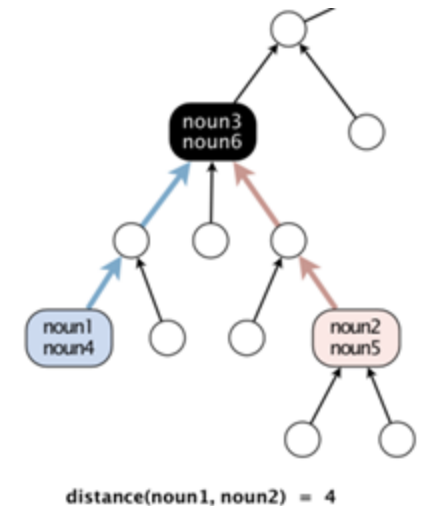
- Customers don't like drinking out of plastic cups.
- She raised her cup to her lips.
- She was so thirsty that she drained her cup.



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>)
 - 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



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



001. t but of different schools. Cook had discovered a BEEF in his possession a few days earlier and, who
002. ily to available canned pork products. Tests with BEEF have been largely unsuccessful because of the
003. ernaster Corps program is to find the reasons for BEEF's low palatability and means of overcoming it.
004. rade 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. ogresses and coat-cutting measures are planned to BEEF up performance. Analysts at Fairbas are looki
011. ration into animal feed, produce 600,000 tonnes of BEEF, which alone is worth 11,000 million at whol
012. pay on the French wards. We served them up corned BEEF, cheese, pickles and hot cocoa, and they thro
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 ¼ 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. Molgaat 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. blouse, 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 ground 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 RATIONS)_ - To increase rate
026. in the prevention of liver abscesses in feed-lot BEEF cattle. Prevention of bacterial pneumonia, sh
027. Feeder, and in controlling scours. _HOW TO FEED: BEEF AND DAIRY CALVES_ - 0.2 gram Dymefan per head
028. gain and improve 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
1.3
-4.5

e.g., shape
e.g., color
e.g., smell

Learned latent feature/component:

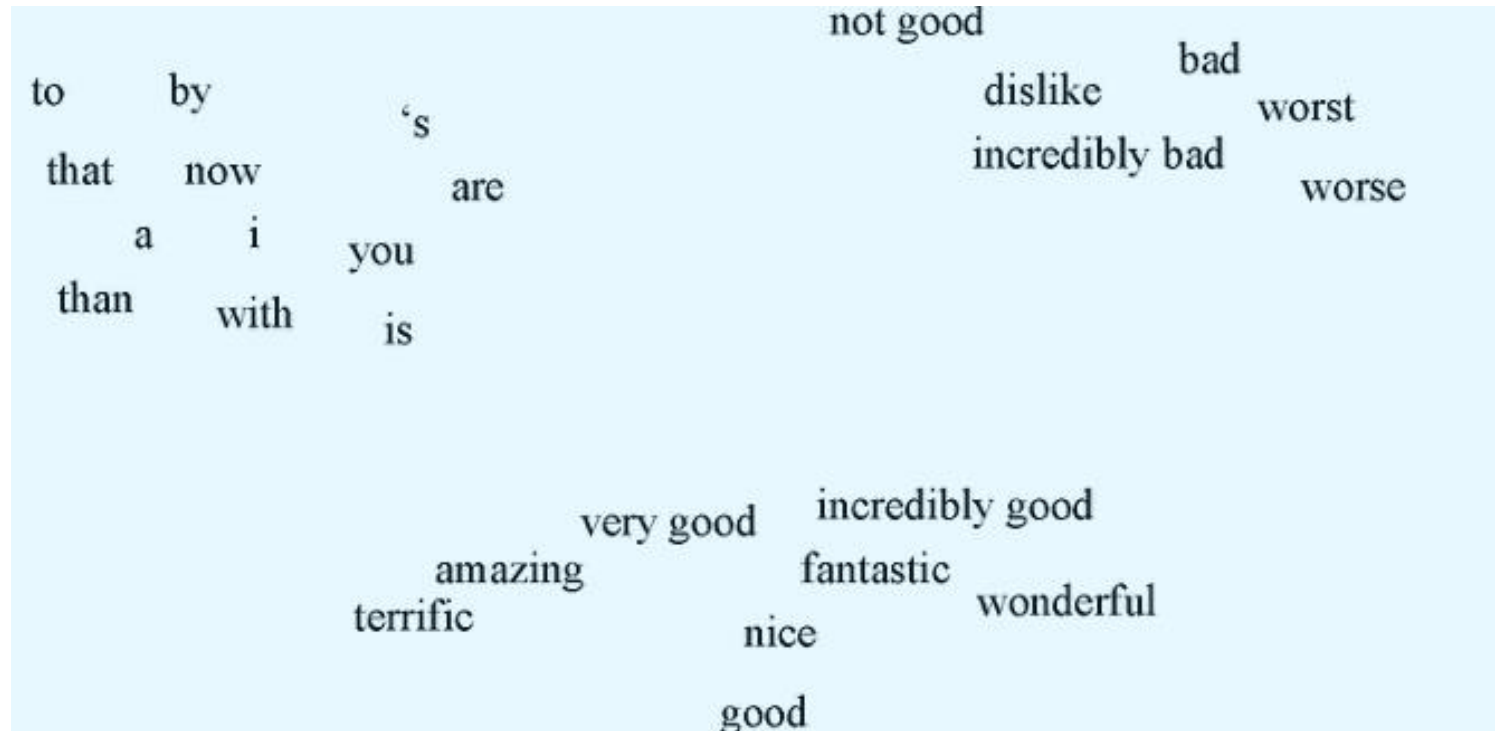
Learning a vector representation of "Beef" from context



Model of meaning focusing on similarity

- Each word = a vector
- Not just “word”; Similar words are nearby in space
- The standard way to represent meaning in “NLP” these days

0.7
1.3
-4.5



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**."



601. □ t but of different schools. Cook had discovered a BEEF in his possession a few days earlier and, who
602. □ ily to available named pork products. Tests with BEEF have been largely unsuccessful because of the
603. □ ewaster Corps program is to find the reasons for BEEF's low palatability and means of overcoming it
604. □ sale can extend the shelf life (at 35 F) of fresh BEEF from 3 days to 5 or 6 weeks. However, the pro
605. □ radiation blanching process dissolves the treated BEEF and liquid accumulates in wrapped cuts. C
606. □ nd liquid accumulates in wrapped cuts. Cooked BEEF irradiated in the absence of oxygen assumes a
607. □ the improved counter of boric acid about two BEEF barrels. There was, of course, no real need t
608. □ ay of the but across from him was mounted by a BEEF barrel with ends knocked out. In this heavy a
609. □ several employees their money's worth. A good many BEEF-hungry settlers were accepting the death of W
610. □ grammer and constructing measures are planned to BEEF up performance. Analysts at Fairfax are looki
611. □ xplain into animal feed, produce 405,550 tonnes of BEEF, which alone is worth 1,000 million at whol
612. □ pay on the French wards. We served them up second BEEF, cheese, pickles and hot sauce, and they thro
613. □ they threw it back at us. "Good gracious, burned BEEF, cheese and bread and butter, they were lucky
614. □ Greeney Ingredients 1 small packet feta cheese 2 BEEF tomatoes 4 cucumber 1 small jar ground
615. □ for the past twenty minutes!" Walter? Walter! The BEEF surprise was lovely. Not what's the surprise?
616. □ as heads who code for a Mr. Wilgeat who supplied BEEF to the reservation up at San Carlos. I would
617. □ "Every week or so I'd see them come in for their BEEF ration. And they're allowed to hunt. They can
618. □ it specializes in steaks, chops, chicken and prime BEEF as well as Tom's favorite dish, stuffed chick
619. □ close, she said, she had raised a calf, grown it BEEF-fat. She had, with her own work-wear hands,
620. □ and get ready to bear. She was ready to kill the BEEF, dress it out, and with vegetables from her g
621. □ "I know what to say. He did say she could get her BEEF and vegetables in case this summer. He did sa
622. □ 12. _HAMBURGER PATTIES WITH BUTE_ 1 pound ground BEEF 2 teaspoons grated onion Dash of pepper 1/2 t
623. □ a highly competitive business more profitable for BEEF, dairy, and sheep men. The target chart quick
624. □ cle assumes that the rations you are feeding your BEEF, dairy cattle, and sheep are adequately balanc
625. □ -infective properties of this drug. _HOW TO FEED_ BEEF CATTLE (FINISHING RATIONS). - To increase rate
626. □ in the prevention of liver abscesses in feed-lot BEEF cattle. Prevention of bacterial pneumonia, an
627. □ founder, and in controlling scours. _HOW TO FEED_ BEEF AND DAIRY CALVES. - 6.2 gram Dymafar per head
628. □ gain and improve feed efficiency. _HOW TO FEED_ BEEF CATTLE. - 18 milligrams of diethylstilbestrol
629. □ over growth rate and feed efficiency of fattening BEEF animals. _HOW TO FEED_ At the rate of 2-1/2



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
...

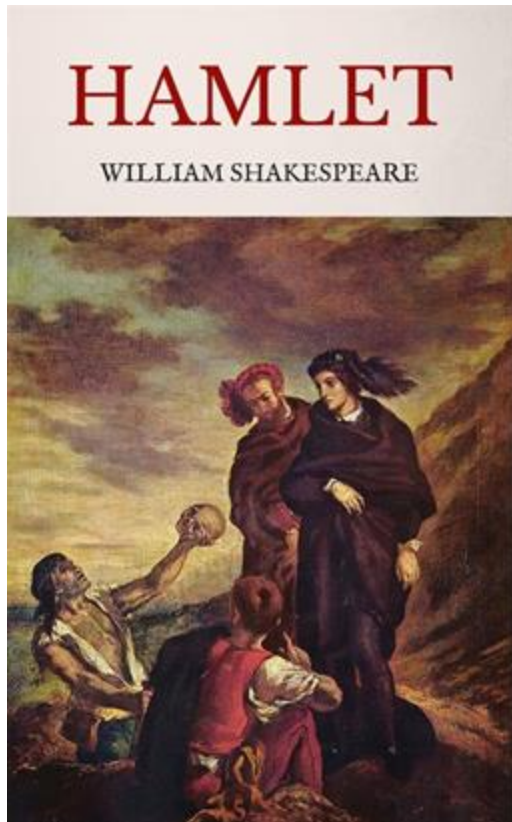
Polysemy should be learned to some extent, assuming that the word vector is trained using ***diverse textual contexts*** and the ***dimensions of the vector*** is large enough



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





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

		Macbeth				
knife	1	1	4	2		2
sword	2	2	7	5		5
		38				

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

Vector representation of the **document**;
vector size = **number of vocabulary**



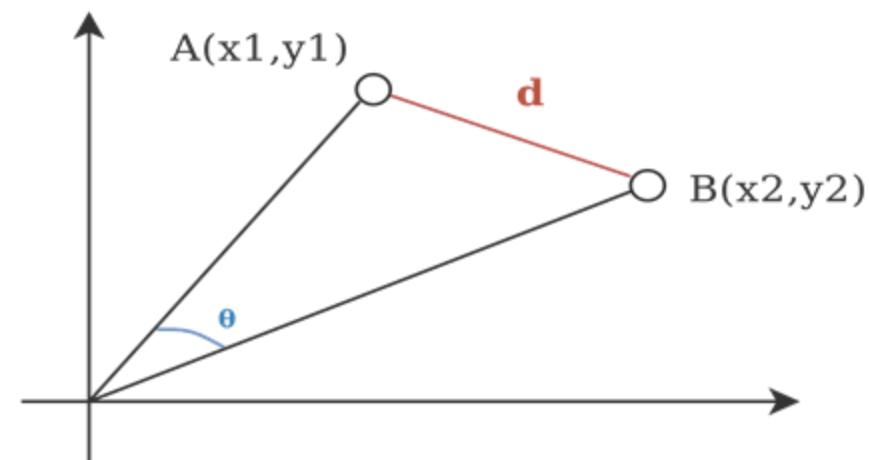
Cosine Similarity

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

Note:

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

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



<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
...						

$\cos(\text{knife}, \text{knife})$ 1.0
 $\cos(\text{knife}, \text{dog})$ 0.11
 $\cos(\text{knife}, \text{sword})$ **0.99**
 $\cos(\text{knife}, \text{love})$ 0.65
 $\cos(\text{knife}, \text{like})$ 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

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

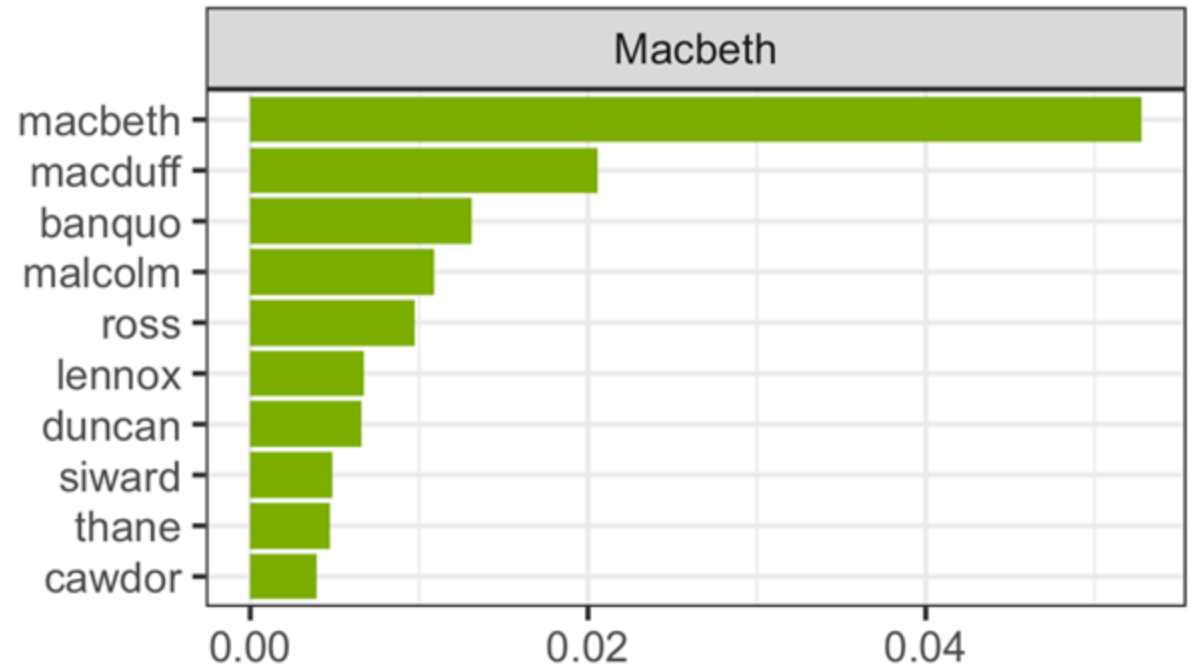
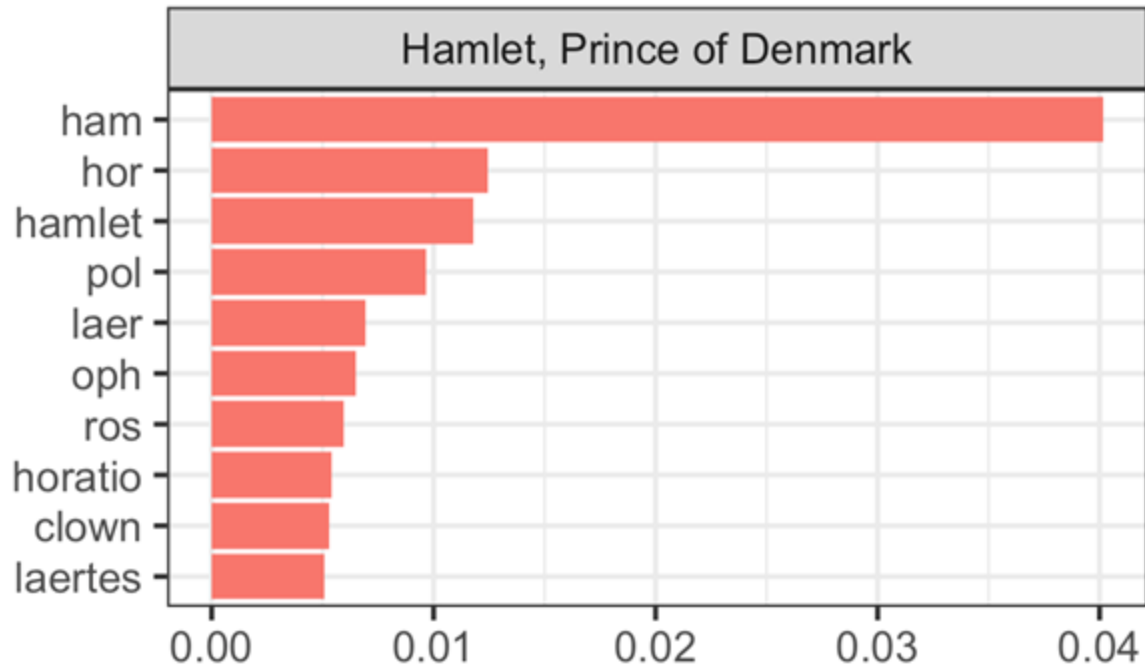
IDF indicates the **informativeness** of the terms when comparing documents.



	Hamlet	Macbeth	Romeo & Juliet	Richard III	Julius Caesar	Tempest
knife	0.07	0.07	0.28	0.14	0	0.14
dog	0	0	0	1.8	3.6	0.6
sword						
love						
like						
...						

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





<https://dataviz21.classes.andrewheiss.com/example/13-example/>



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

#	Name	Formula
1.	Joint probability	$P(xy)$
2.	Conditional probability	$P(y x)$
3.	Reverse conditional probability	$P(x y)$
4.	Pointwise mutual information	$\log \frac{P(xy)}{P(x)P(y)}$
5.	Mutual dependency (MD)	$\log \frac{P(x,y)^2}{P(x)P(y)}$
6.	Log frequency biased MD	$\log \frac{P(x,y)^2}{P(x)P(y)} + \log P(xy)$
7.	Normalized expectation	$\frac{2f(xy)}{f(x)+f(y)}$
8.	Mutual expectation	$\frac{2f(xy)}{f(x)+f(y)} \cdot P(xy)$
9.	Saliency	$\log \frac{P(xy)}{P(x)P(y)} \cdot \log f(xy)$
10.	Pearson's χ^2 test	$\sum_{i,j} \frac{(f_{ij} - f_{ij}^e)^2}{f_{ij}^e}$
11.	Fisher's exact test	$\frac{f(x+1)f(y+1)f(x,y)}{Nf(x)f(y)f(x,y)}$
12.	t test	$\frac{f(x,y) - f(x)f(y)}{\sqrt{f(x)(1-f(x)/N)}}$
13.	z score	$\frac{f(x,y) - f(x)f(y)}{\sqrt{f(x)(1-f(x)/N)}}$
14.	Poisson significance measure	$\frac{f(x,y) - f(x)f(y) \log f(x,y) + \log f(x,y)}{\log N}$
15.	Log likelihood ratio	$-2 \sum_{i,j} f_{ij} \log f_{ij} / f_{ij}^e$
16.	Squared log likelihood ratio	$-2 \sum_{i,j} \log f_{ij}^2 / f_{ij}^e$
17.	Russel-Rao	$\frac{a+b+c+d}{a+d}$
18.	Sokal-Michiner	$\frac{a+d}{a+b+c+d}$
19.	Rogers-Tanimoto	$\frac{a+2b+d}{a+b+c+d}$
20.	Hamann	$\frac{a+d}{a+b+c+d}$
21.	Third Sokal-Sneath	$\frac{a+d}{a+b+c}$
22.	Jaccard	$\frac{a}{a+b+c}$
23.	First Kulczynski	$\frac{a+b}{a+c}$
24.	Second Sokal-Sneath	$\frac{a+2(b+c)}{a+b+c}$
25.	Second Kulczynski	$\frac{1}{2} (\frac{a+b}{a+c} + \frac{a+b}{a+d})$
26.	Fourth Sokal-Sneath	$\frac{1}{4} (\frac{a}{a+b} + \frac{a}{a+c} + \frac{d}{a+b} + \frac{d}{a+c})$
27.	Odds ratio	$\frac{ad}{bc}$
28.	Yulle's ω	$\frac{\sqrt{ad} - \sqrt{bc}}{\sqrt{ad} + \sqrt{bc}}$
29.	Yulle's Q	$\frac{ad-bc}{ad+bc}$
30.	Driver-Kroeber	$\frac{\sqrt{(a+b)(a+c)}}{ad}$
31.	Fifth Sokal-Sneath	$\frac{\sqrt{(a+b)(a+c)(d+b)(d+c)}}{ad-bc}$
32.	Pearson	$\frac{\sqrt{(a+b)(a+c)(d+b)(d+c)}}{ad-bc}$
33.	Baroni-Urbani	$\frac{a + \sqrt{ad}}{a+b+c + \sqrt{ad}}$
34.	Braun-Blanquet	$\frac{\max(a+b, a+c)}{a+b+c + \sqrt{ad}}$
35.	Simpson	$\frac{a}{\min(a+b, a+c)}$
36.	Michael	$\frac{a}{4(ad-bc)}$
37.	Mountford	$\frac{2bc+ab+ac}{2bc+ab+ac}$
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	$\frac{P(x,y) - P(x)P(y)}{\sqrt{P(x)P(y)(1-P(x))(1-P(y))}}$
45.	Kappa	$\frac{P(x,y) + P(x)P(y) - P(x)P(y) - P(x)P(y)}{1 - P(x)P(y) - P(x)P(y)}$

#	Name	Formula
46.	J measure	$\max[P(xy) \log \frac{P(x,y)}{P(x)P(y)} + P(xy) \log \frac{P(y x)}{P(y)}]$
47.	Gini index	$\max[P(x+)(P(y x)^2 + P(y x)^2) - P(x+y)^2]$
48.	Confidence	$\max[P(y x), P(x y)]$
49.	Laplace	$\max[\frac{N P(x,y)+1}{N P(x)+2}, \frac{N P(x,y)+1}{N P(y)+2}]$
50.	Conviction	$\max[\frac{P(x,y)P(x)}{P(x)P(y)}, \frac{P(x,y)P(y)}{P(x)P(y)}]$
51.	Piatersky-Shapiro	$P(xy) - P(x+)P(+y)$
52.	Certainty factor	$\max[\frac{P(x,y) - P(x)P(y)}{1 - P(x)}, \frac{P(x,y) - P(x)P(y)}{1 - P(y)}]$
53.	Added value (AV)	$\max[P(y x) - P(+y), P(x y) - P(x+)]$
54.	Collective strength	$\frac{P(x,y) + P(x)}{P(x) + P(y) + P(x)P(y)} \cdot \frac{1 - P(x)P(y) - P(x)P(y)}{1 - P(x) - P(y)}$
55.	Klosgen	$\sqrt{P(xy) \cdot AV}$
56.	Context entropy	$-\sum_w P(w C_x) \log P(w C_x)$
57.	Left context entropy	$-\sum_w P(w C_x^l) \log P(w C_x^l)$
58.	Right context entropy	$-\sum_w P(w C_x^r) \log P(w C_x^r)$
59.	Left context divergence	$P(x+) \log P(x+) - \sum_w P(w C_x^l) \log P(w C_x^l)$
60.	Right context divergence	$P(+y) \log P(+y) - \sum_w P(w C_x^r) \log P(w C_x^r)$
61.	Cross entropy	$-\sum_w P(w C_x) \log P(w C_y)$
62.	Reverse cross entropy	$-\sum_w P(w C_y) \log P(w C_x)$
63.	Intersection measure	$\frac{2 C_x \cap C_y }{ C_x + C_y }$
64.	Euclidean norm	$\sqrt{\sum_w (P(w C_x) - P(w C_y))^2}$
65.	Cosine norm	$\frac{\sum_w P(w C_x)P(w C_y)}{\sqrt{\sum_w P(w C_x)^2} \sqrt{\sum_w P(w C_y)^2}}$
66.	LI norm	$\frac{\sum_w P(w C_x) - P(w C_y) }{\sum_w P(w C_x) + P(w C_y)}$
67.	Confusion probability	$\sum_w \frac{P(w C_x)P(w C_y)}{P(w C_x) + P(w C_y)}$
68.	Reverse confusion probability	$\sum_w \frac{P(w C_y)P(w C_x)}{P(w C_x) + P(w C_y)}$
69.	Jensen-Shannon divergence	$\frac{1}{2} [D(p(w C_x) \frac{1}{2}(p(w C_x) + p(w C_y))) + D(p(w C_y) \frac{1}{2}(p(w C_x) + p(w C_y)))]$
70.	Cosine of pointwise MI	$\frac{\sum_w MI(w,x)MI(w,y)}{\sqrt{\sum_w MI(w,x)^2} \sqrt{\sum_w MI(w,y)^2}}$
71.	KL divergence	$\sum_w P(w C_x) \log \frac{P(w C_x)}{P(w C_y)}$
72.	Reverse KL divergence	$\sum_w P(w C_y) \log \frac{P(w C_y)}{P(w C_x)}$
73.	Skew divergence	$D(p(w C_x) \alpha p(w C_x) + (1-\alpha)p(w C_y))$
74.	Reverse skew divergence	$D(p(w C_y) \alpha p(w C_x) + (1-\alpha)p(w C_y))$
75.	Phrase word cooccurrence	$\frac{1}{2} (\frac{f(x C_x)}{f(x)} + \frac{f(y C_x)}{f(x)})$
76.	Word association	$\frac{1}{2} (\frac{f(x C_x) - f(x)}{f(x)} + \frac{f(y C_x) - f(x)}{f(x)})$
Cosine context similarity:		$\frac{1}{2} (\cos(\mathbf{e}_x, \mathbf{e}_y) + \cos(\mathbf{e}_y, \mathbf{e}_x))$
77. in boolean vector space		$\mathbf{e}_x = \{z_i\}; \cos(\mathbf{e}_x, \mathbf{e}_y) = \frac{\sum z_i x_i z_i y_i}{\sqrt{\sum z_i^2} \sqrt{\sum y_i^2}}$
78. in tf vector space		$z_i = \delta(f(w_i C_x))$
79. in tf · idf vector space		$z_i = f(w_i C_x) \cdot \frac{N}{df(w_i)}; df(w_i) = \{x: w_i \in C_x\} $
Dice context similarity:		$\frac{1}{2} (\text{dico}(\mathbf{e}_x, \mathbf{e}_y) + \text{dico}(\mathbf{e}_y, \mathbf{e}_x))$
80. in boolean vector space		$\mathbf{e}_x = \{z_i\}; \text{dico}(\mathbf{e}_x, \mathbf{e}_y) = \frac{2 \sum z_i x_i z_i y_i}{\sum z_i^2 + \sum y_i^2}$
81. in tf vector space		$z_i = \delta(f(w_i C_x))$
82. in tf · idf vector space		$z_i = f(w_i C_x) \cdot \frac{N}{df(w_i)}; df(w_i) = \{x: w_i \in C_x\} $

Table 1: Inventory of lexical association measures for collocation extraction. Pecina, P. (2008). Lexical Association Measures: Collocation Extraction.



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



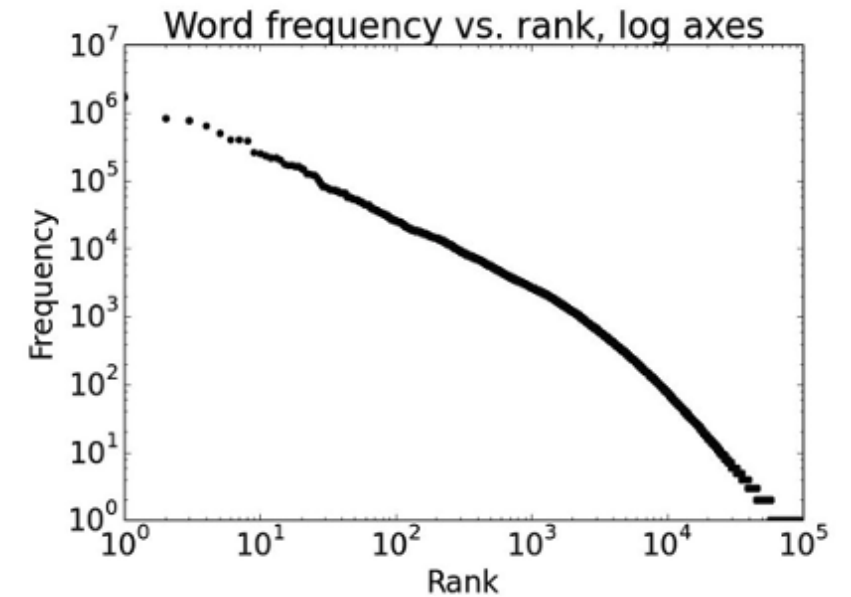
Sparse vectors



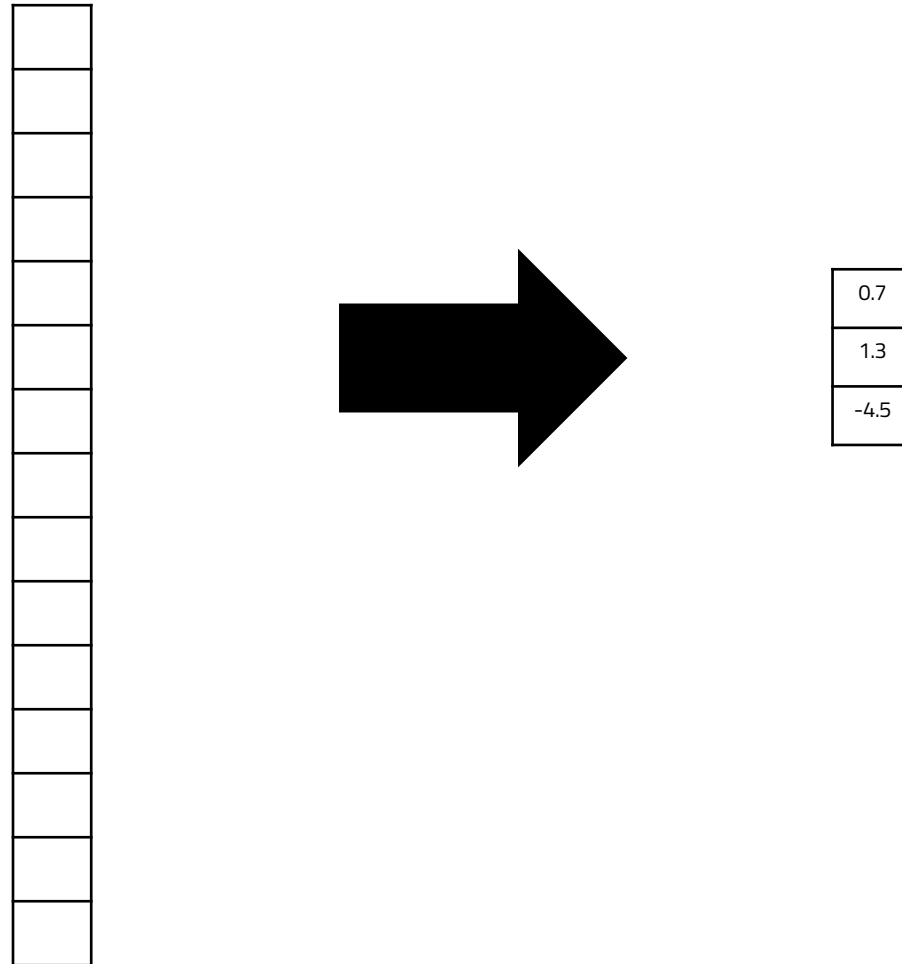
"aardvark"

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

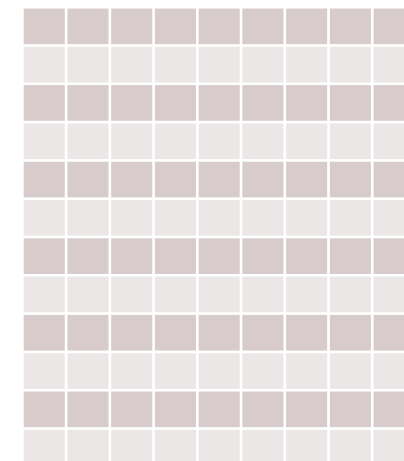


Sparse vectors -> Dense vectors



	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
...						

=

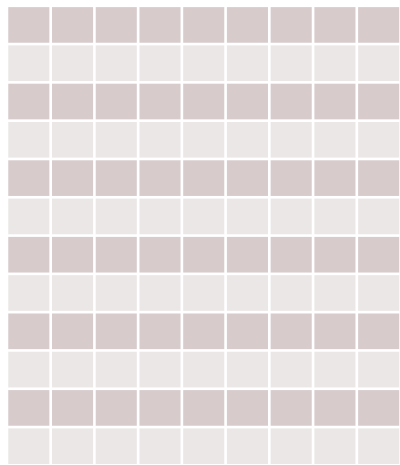


$n \times d$



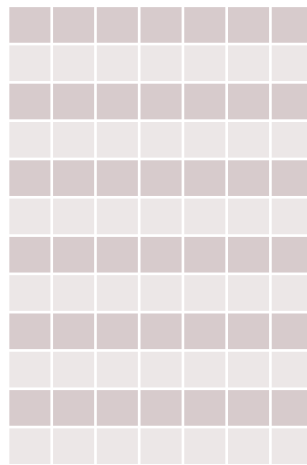
Singular value decomposition (SVD)

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



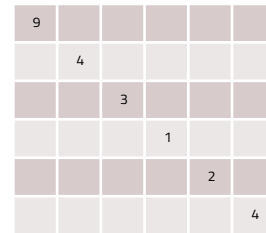
$n \times d$

=



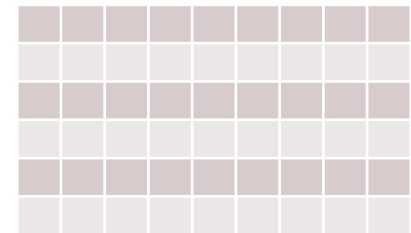
$n \times m$

×



$m \times m$
(diagonal)

×

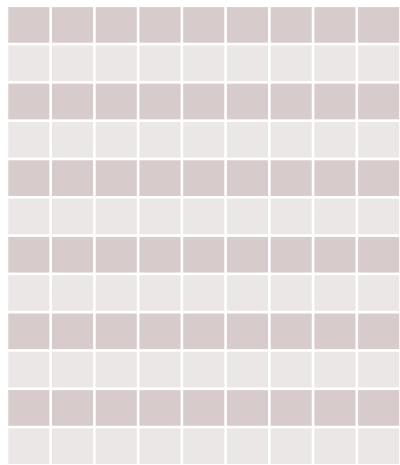


$m \times d$



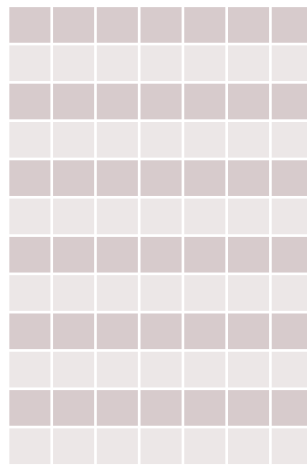
Singular value decomposition (SVD)

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



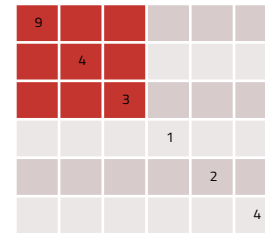
$n \times d$

=



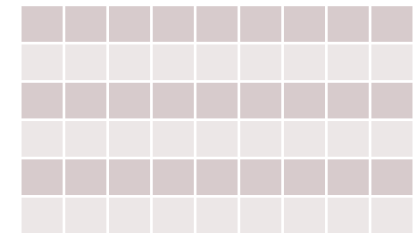
$n \times m$

×



$m \times m$
(diagonal)

×

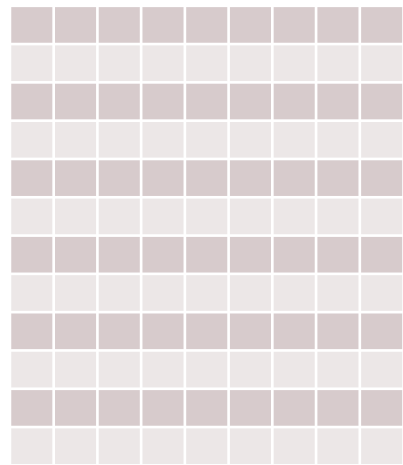


$m \times d$



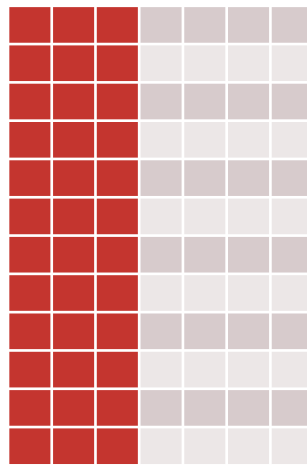
Singular value decomposition (SVD)

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



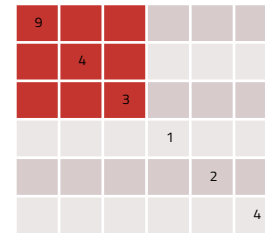
$n \times d$

\approx



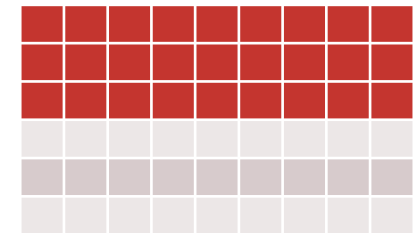
$n \times m$

\times



$m \times m$
(diagonal)

\times

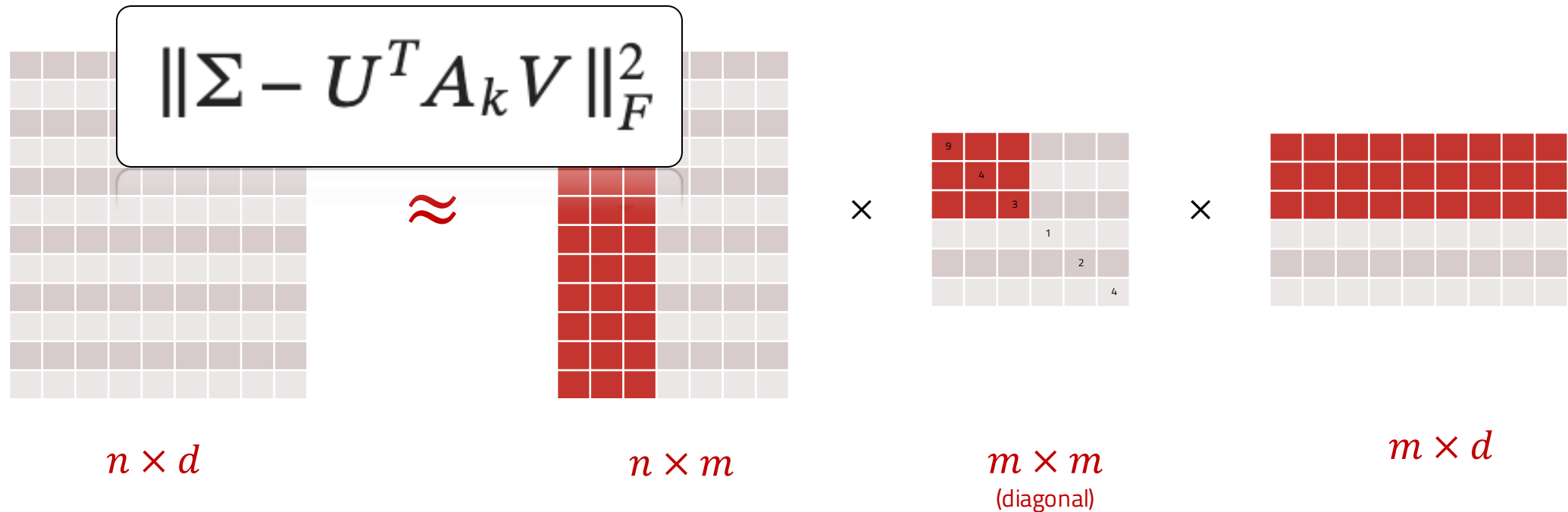


$m \times d$



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
knife	-0.2	0.7	-2.2	-0.2	0.7	-2.2
dog	-0.2	0.7	-2.2	9.3	-0.5	0.5
sword	9.3	-0.5	0.5	-0.5	0.5	9.3

$m \times d$



Low-dimensional representation
for **terms** (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

Low-dimensional representation
for **documents** (here 3 dimensions)

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 $K \ll D$.
- 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

Sort by probabilities

#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)



Latent semantic analysis



001. a lot of different schools. Cook had discovered a beef in his possession a few days earlier and, the
 002. try to available common pork products. Taste with beef have been largely unaccounted because of the
 003. whether those groups is to find the reason for meat's low palatability and means of processing it.
 004. same can extend the shelf life set 20% of fresh beef from 5 days to 3 or 4 weeks. However, the pro
 005. antibiotic bleaching process discolors the treated beef and liquid accumulated in packaged cuts. C
 006. of liquid accumulated in packaged cuts. Cooked beef irradiated in the absence of oxygen assumes a
 007. the important source of food. Each source has beef bacteria. These are, of course, no real need a
 008. up of the hot source from his was contaminated by a beef barrel with white knocked out. In this heavy a
 009. second emphasize their meat's with. A good meat beef-hungry activities were accepting the death of a
 010. expenses and cost-cutting measures are planned to beef up performance. Multiple at parties are looki
 011. main into animal feed, produce with the source of beef, which alone to work almost million of what
 012. say on the French table. He served them up green beef, cheese, potatoes and hot coffee, and they like
 013. they live in back of. "Good afternoon, good beef, cheese and bread and butter, they were happy
 014. bread ingredients I read about the source of beef source's source I read. In bread
 015. for the best family source? Wait! Wait! The beef source was lovely, but what's the surprise?
 016. to home who come to a. "But what the surprise? Not to the destination of the destination. I guess
 017. heavy work as it's see them come in for their beef value. And they're allowed to buy. They can
 018. to vegetables to bread, chips, chicken and give beef as well as their favorite drink, which often
 019. since, she said, she had raised a calf, given it beef-fat, the best, with her own work-week habits,
 020. and get ready to heat, she was ready to kill the beef, down to cut, and with vegetables for her a
 021. "I know what to say. He did say she could get her beef and vegetables in case this summer. He did n
 022. a highly competitive business were profitable for beef, dairy, and sheep see. The target short quick
 023. he assumes that the various you saw feeding your beef, dairy cattle, and sheep are obviously been
 024. industrial perspective of this beef. HOW TO FEED BEEF CATTLE (INCLUDING PARTIAL). - To increase rate
 025. in the production of live animals in beefing beef cattle. Possession of technical personnel, an
 026. number, and in controlling source. HOW TO FEED BEEF AND DAIRY CATTLE. - 0.2 gram protein per head
 027. and and improve feed efficiency. HOW TO FEED BEEF CATTLE. - 10 milligrams of 4000000000000000
 028. one growth rate and feed efficiency of fattening beef animals. HOW TO FEED. At the rate of 2-10



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

	#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

Sort by probabilities

	#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)



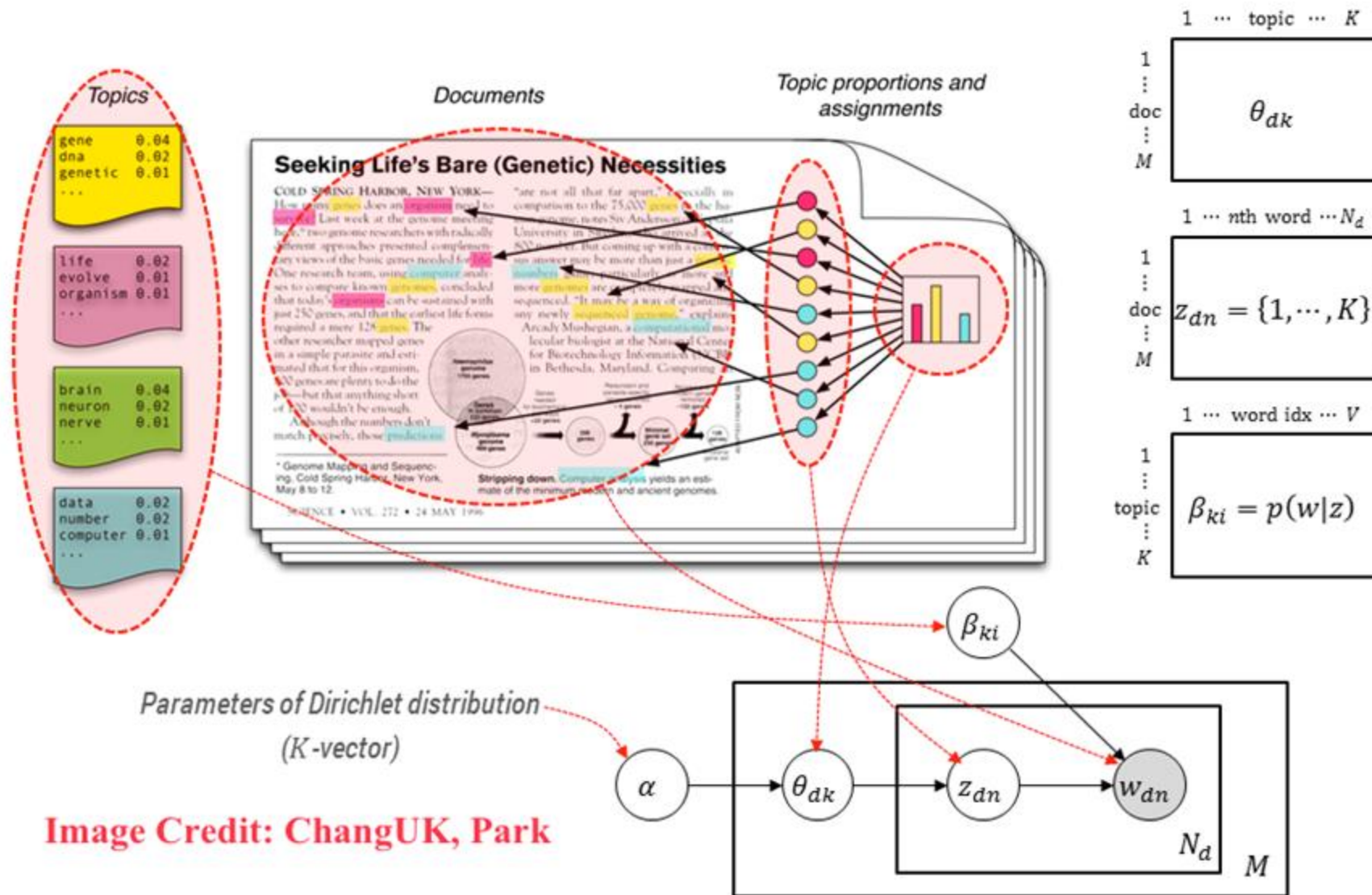
Topic Modeling

- 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

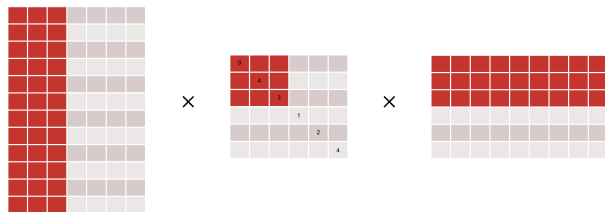


Count-based vs Prediction-based Methods

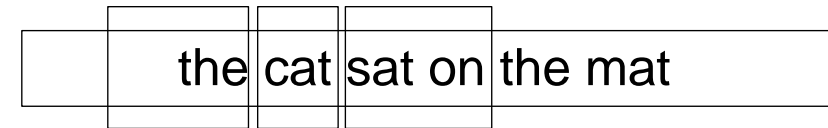
To obtain good dense representations

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

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



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



Text Classification Revisited

$x = \text{"Today's weather is great"}$

$$P(y | x)$$

$y = \{\text{positive, negative}\}$

$\hat{y} = \text{positive}$

$$|Y| = 2$$

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

$$P(x_t | x_{<t})$$

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

$\hat{x} = \text{great}$

$$|X| = V \text{ (vocabulary size)}$$

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

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

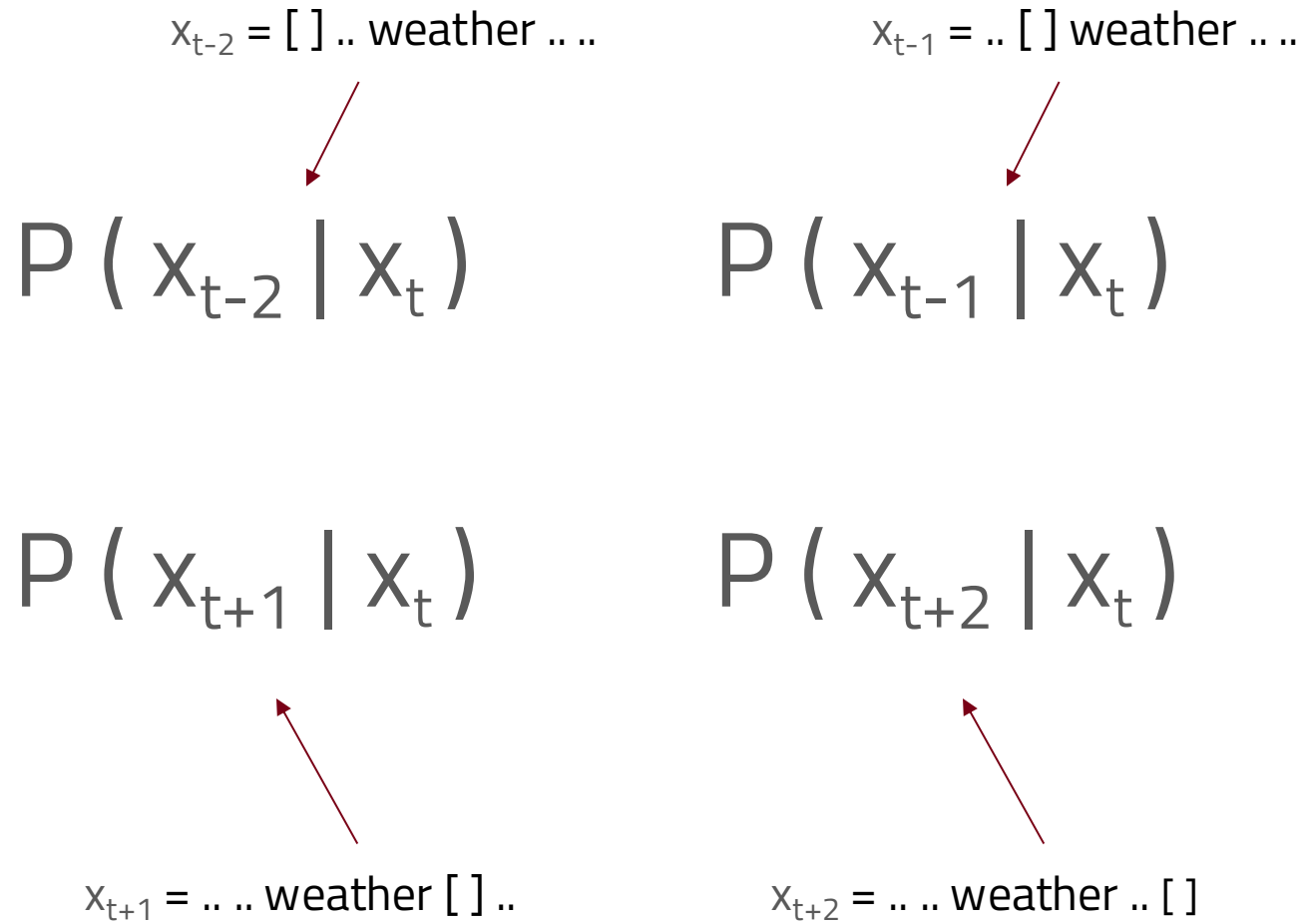
$x_t = \{\text{a, aa .. apple .. banana .. great .. good .. zebra ..}\}$

$\hat{x} = \text{weather}$

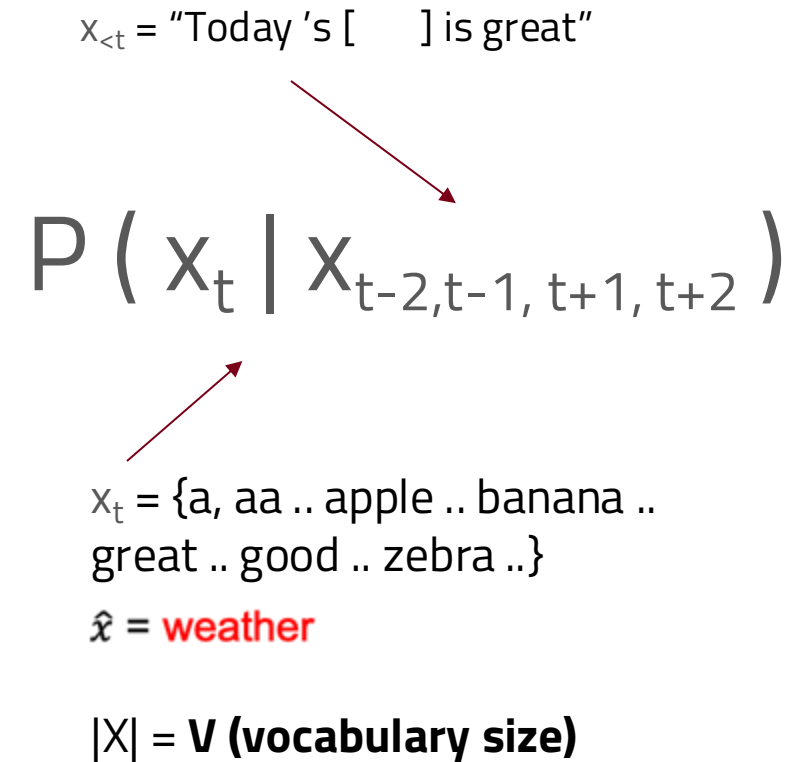
$$|X| = V \text{ (vocabulary size)}$$



Text Classification Revisited



Predict the neighboring word(s) from the middle word



Predict the middle word from neighboring words

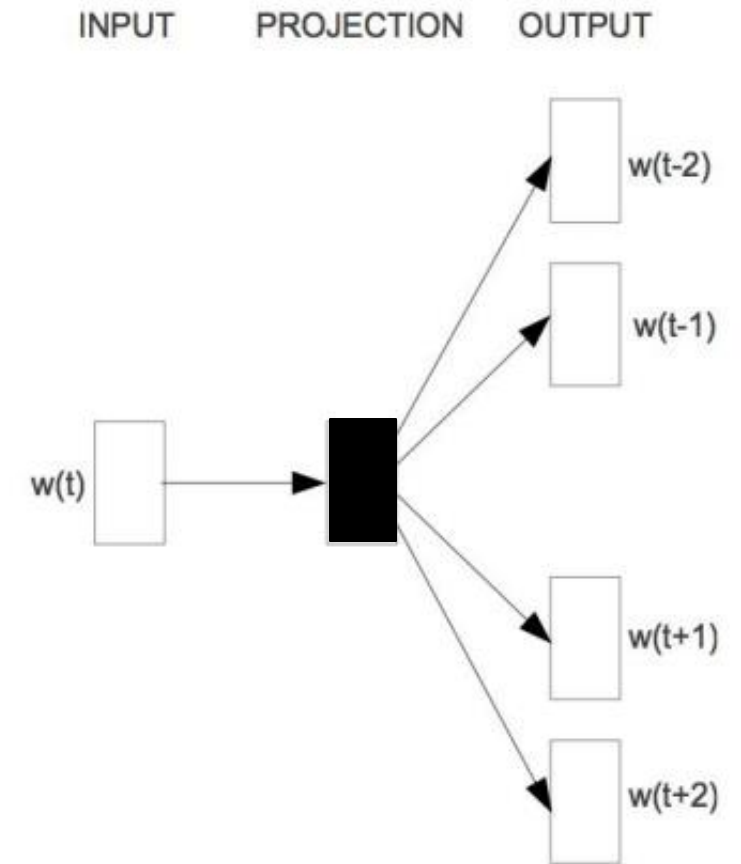


Dense vectors from prediction (not count)

the cat sat on the mat

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

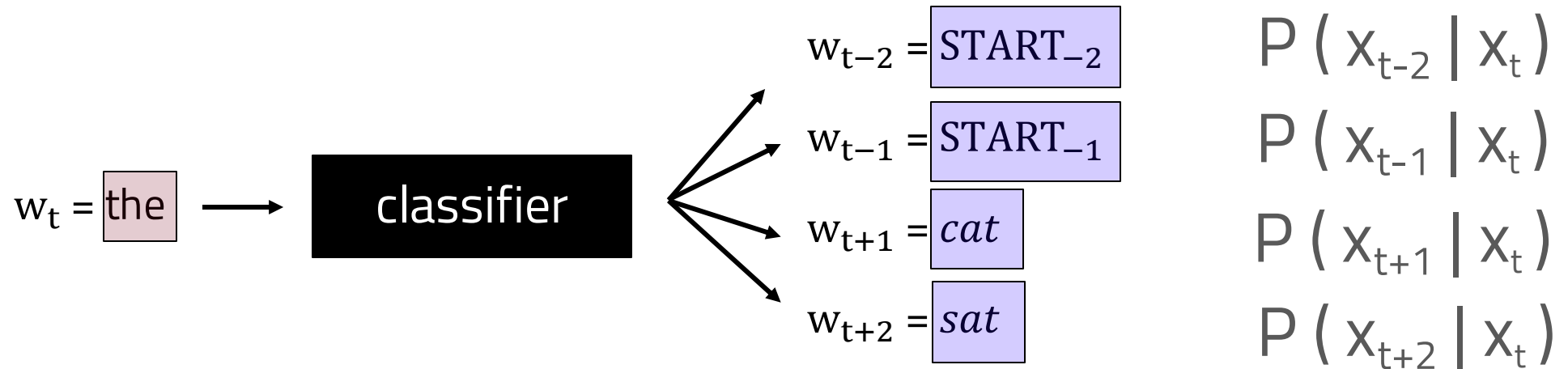
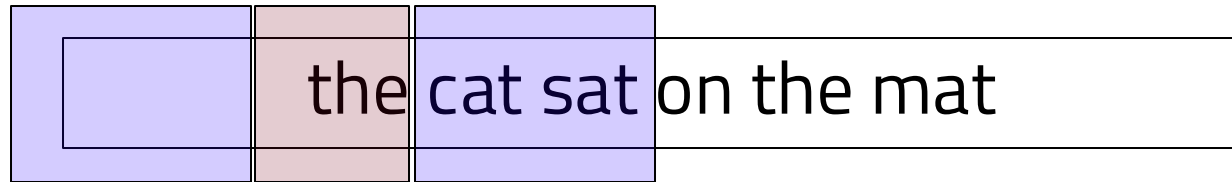
Predict the neighboring word(s) from the middle word



(Mikolove et al., 14)



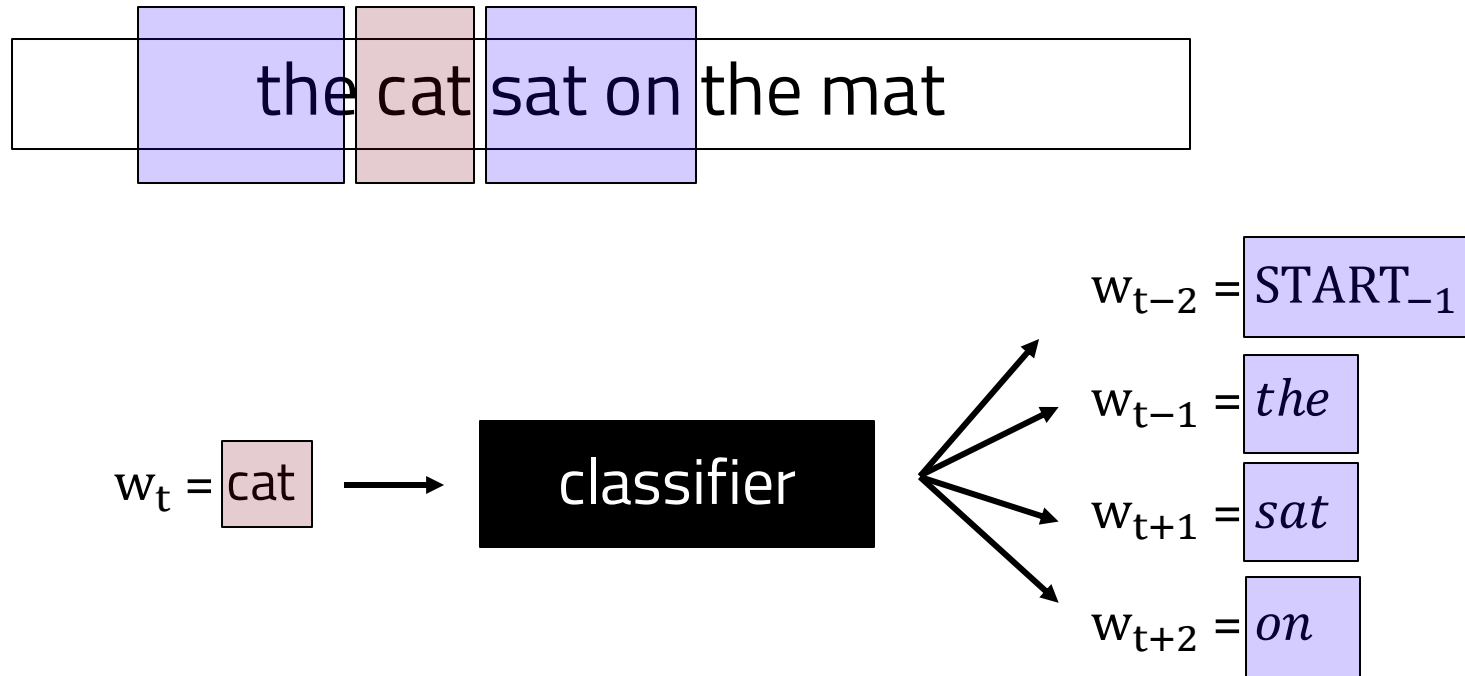
Dense vectors from prediction (not count)



Context window size = 2



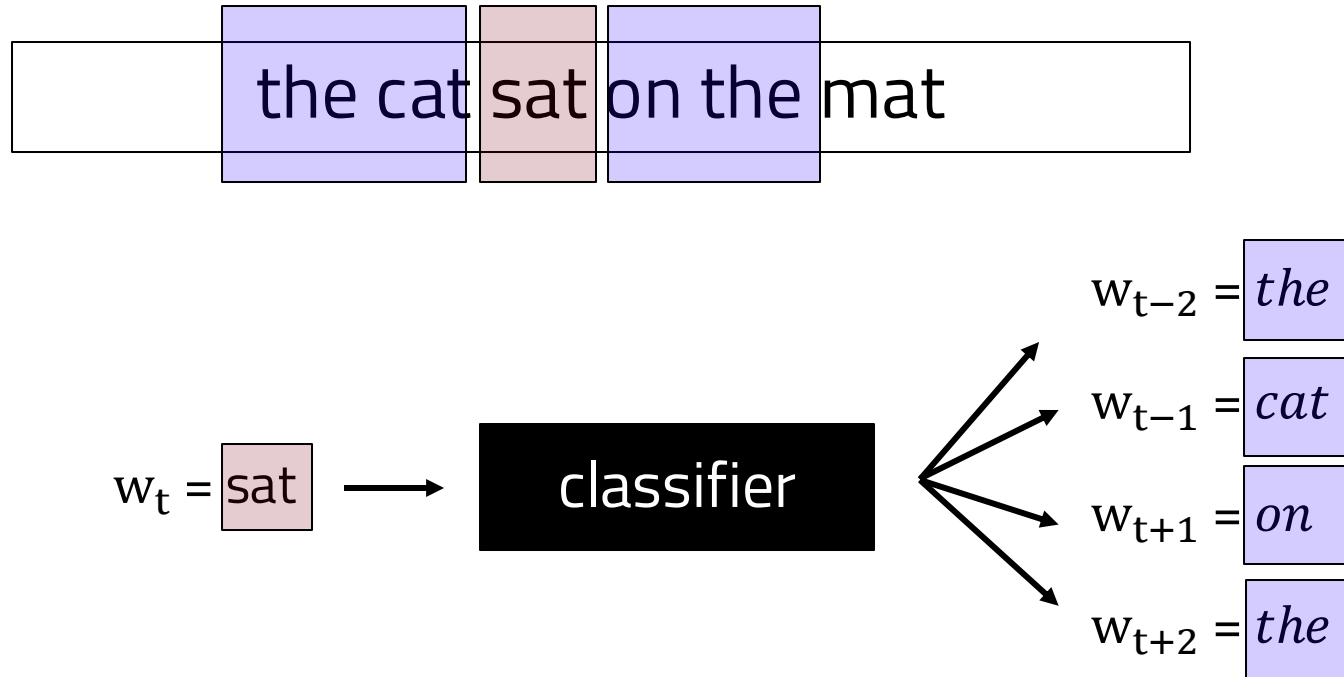
Dense vectors from prediction (not count)



Context window size = 2

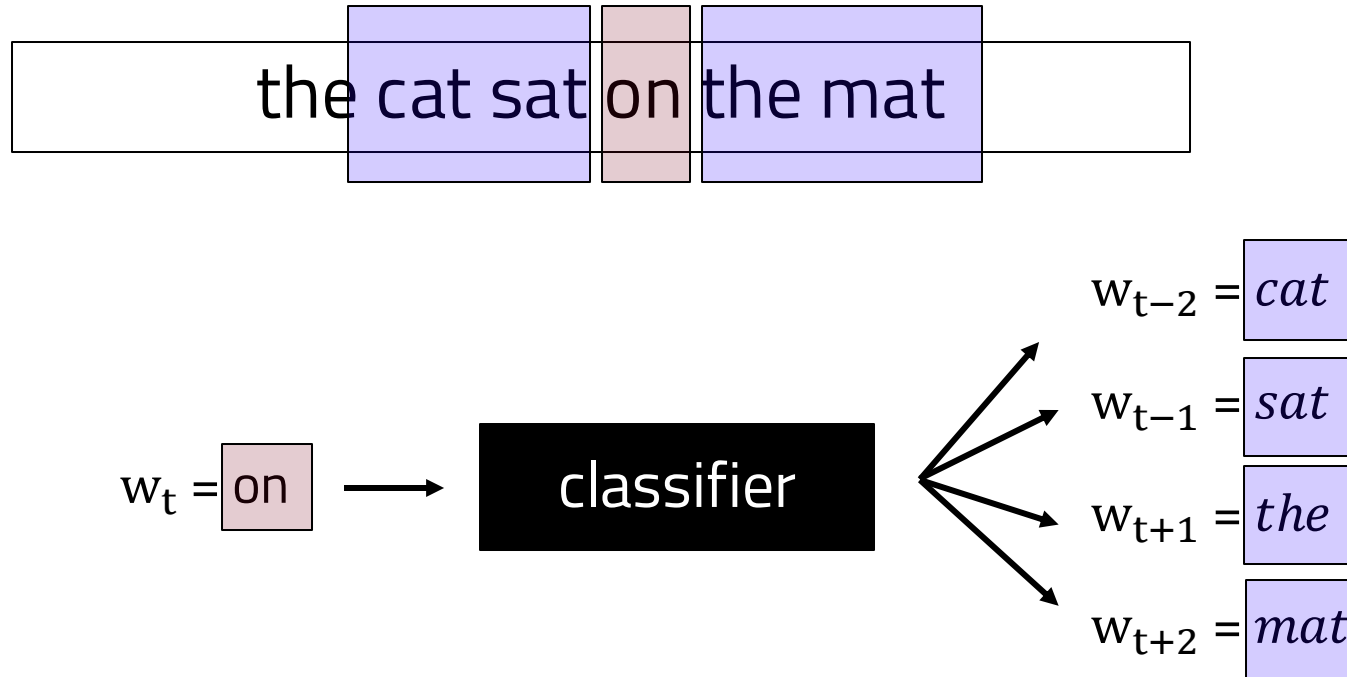


Dense vectors from prediction (not count)



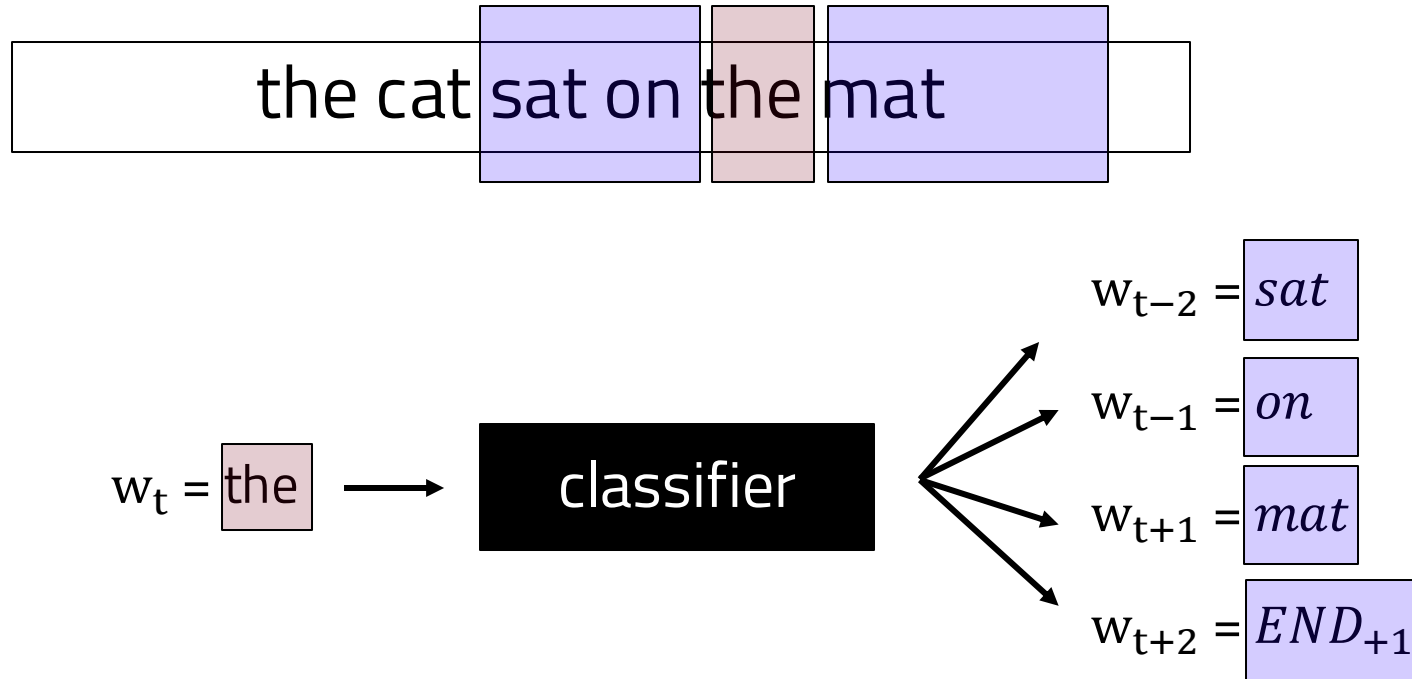
Context window size = 2

Dense vectors from prediction (not count)



Context window size = 2

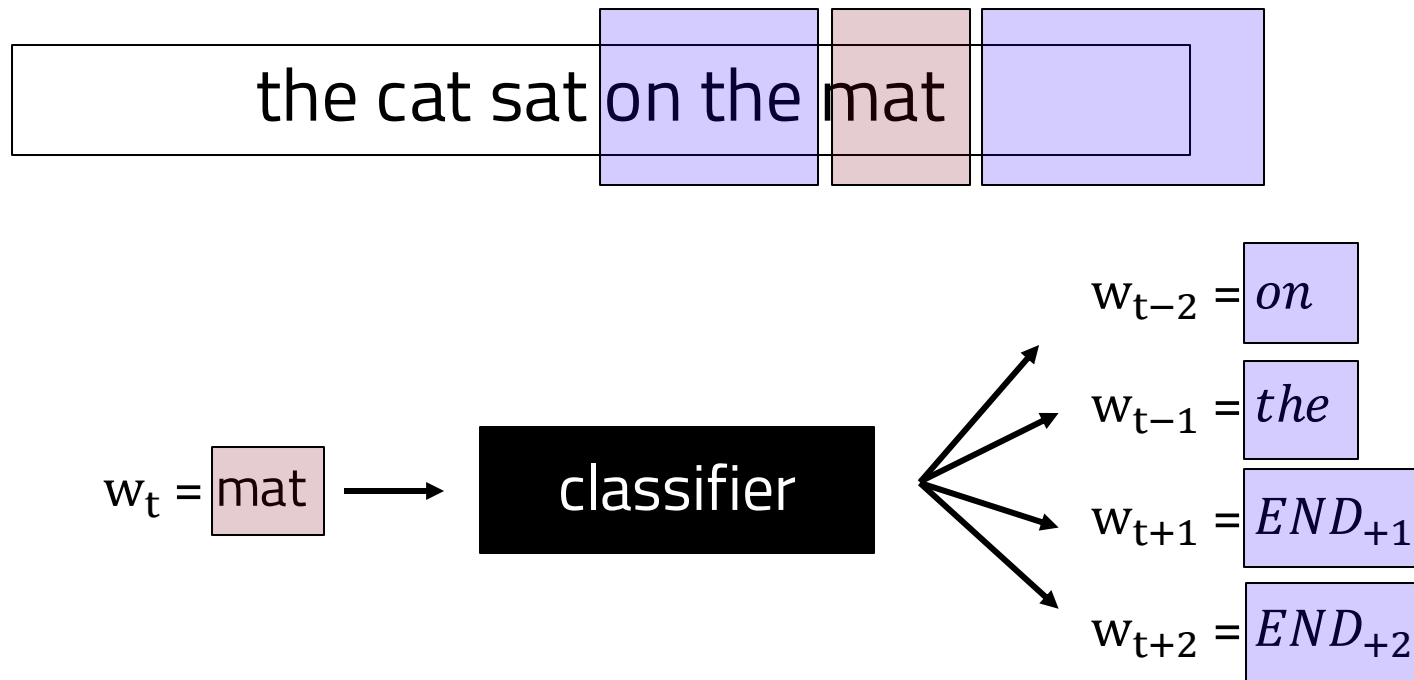
Dense vectors from prediction (not count)



Context window size = 2



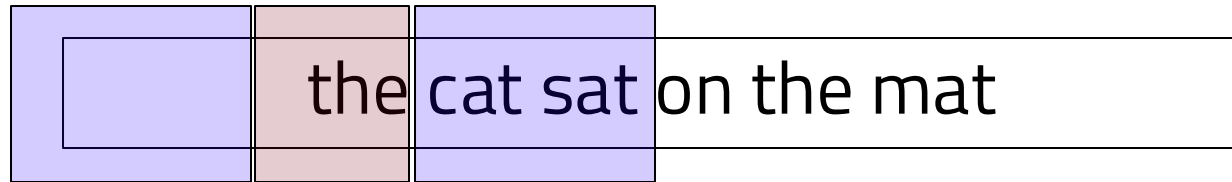
Dense vectors from prediction (not count)



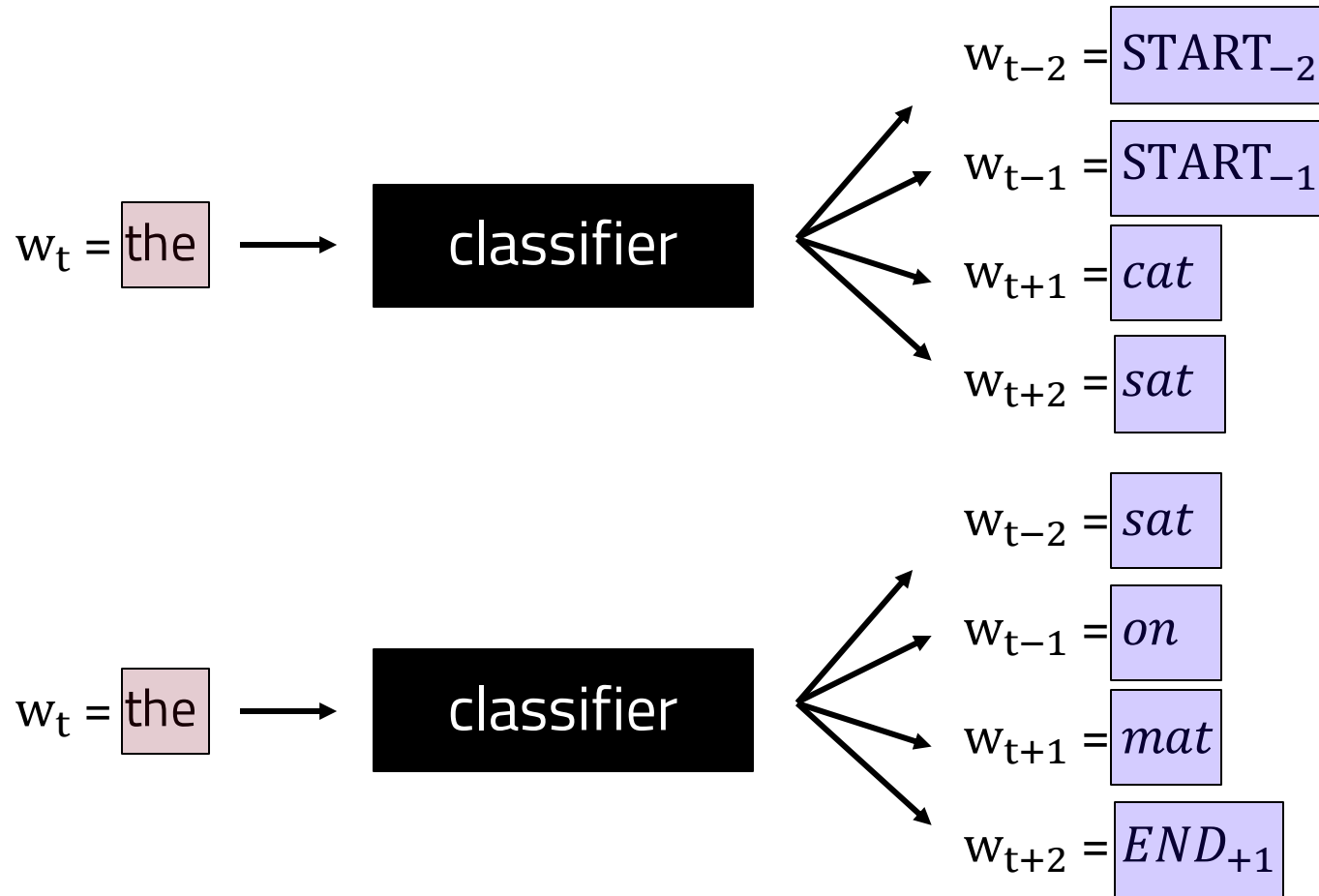
Context window size = 2



Same context but different output?



Same context but different output?



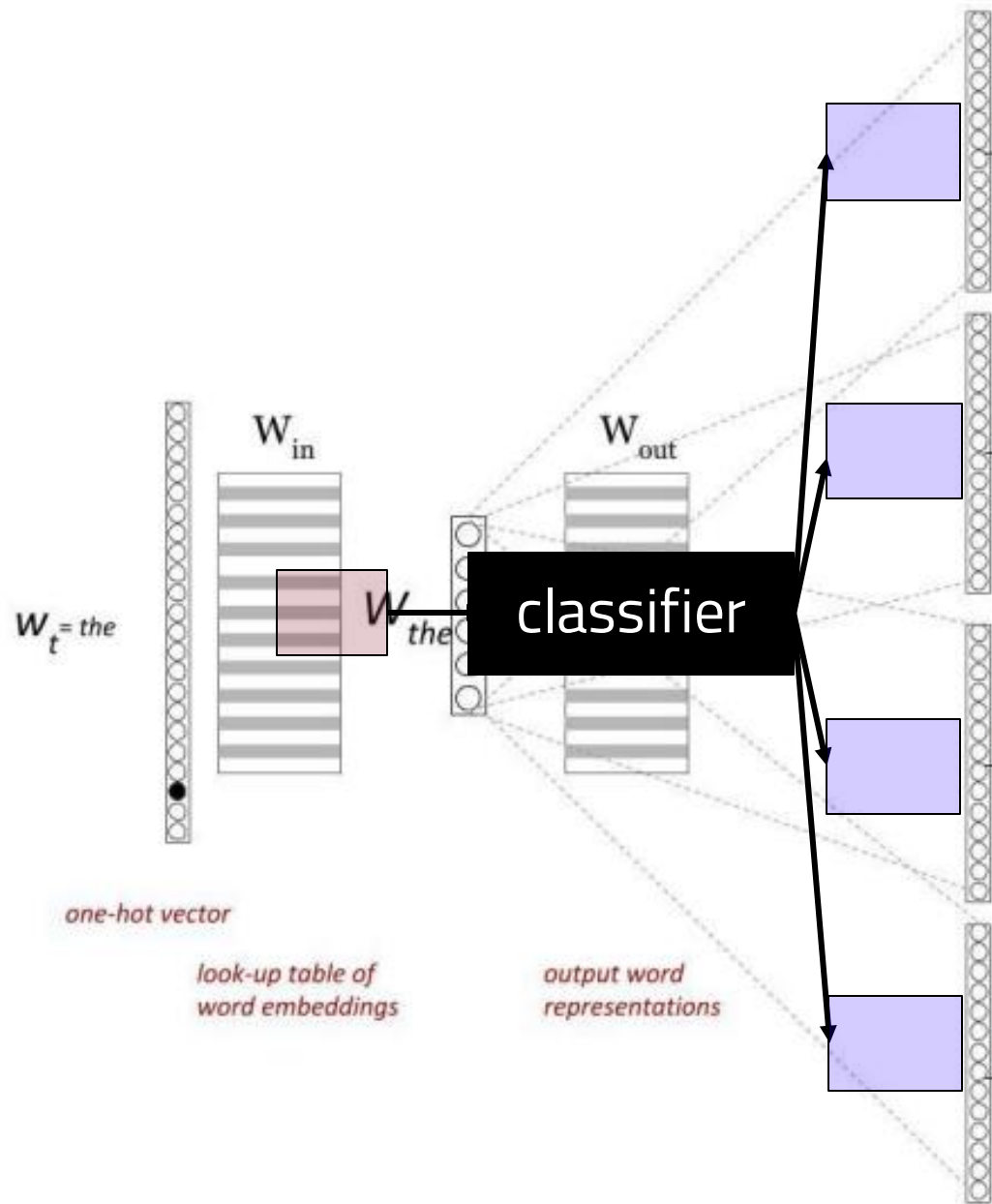
Learning
*multiple
senses/context*

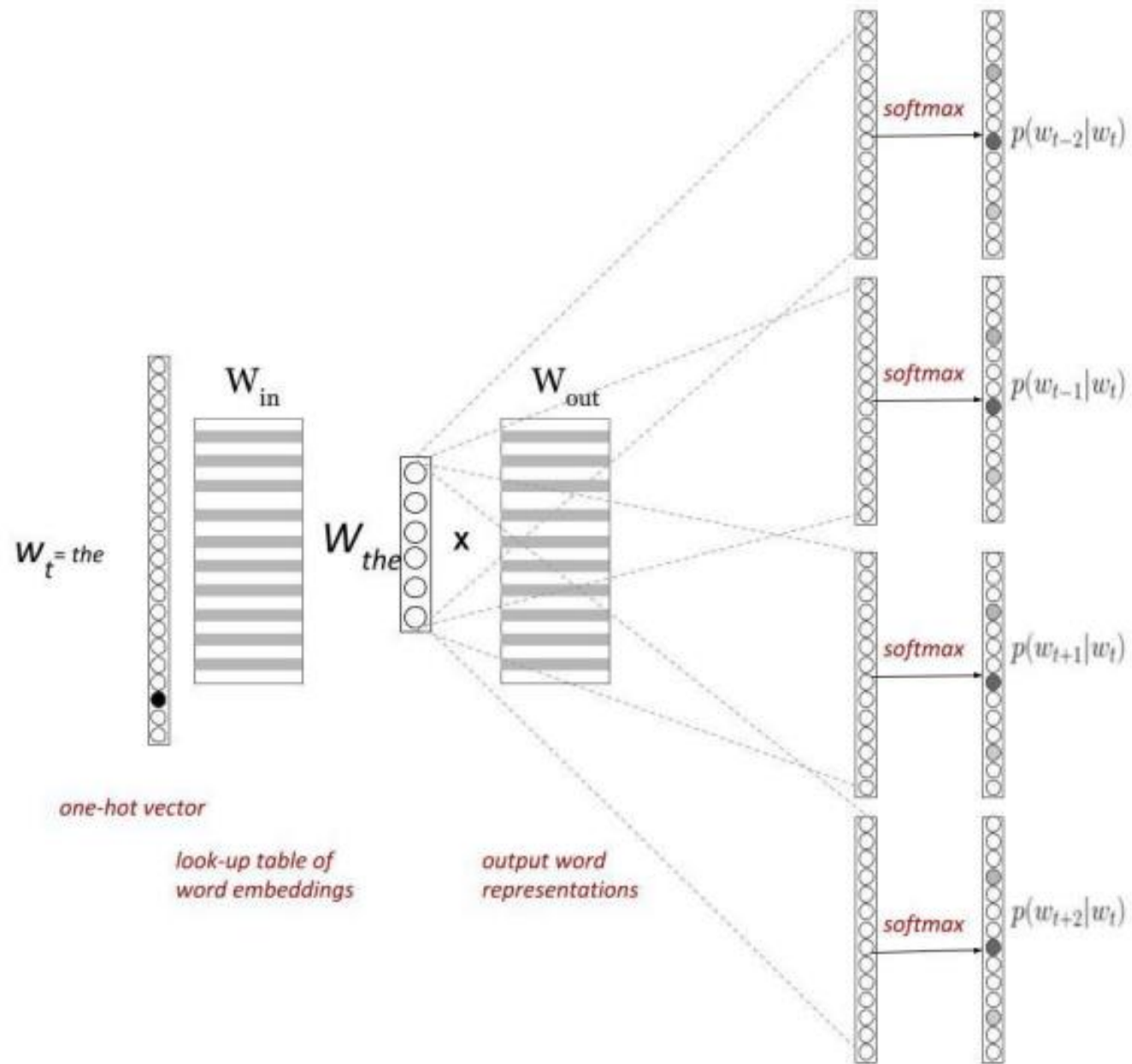
Context window size = 2



Dense vectors from prediction (not count)







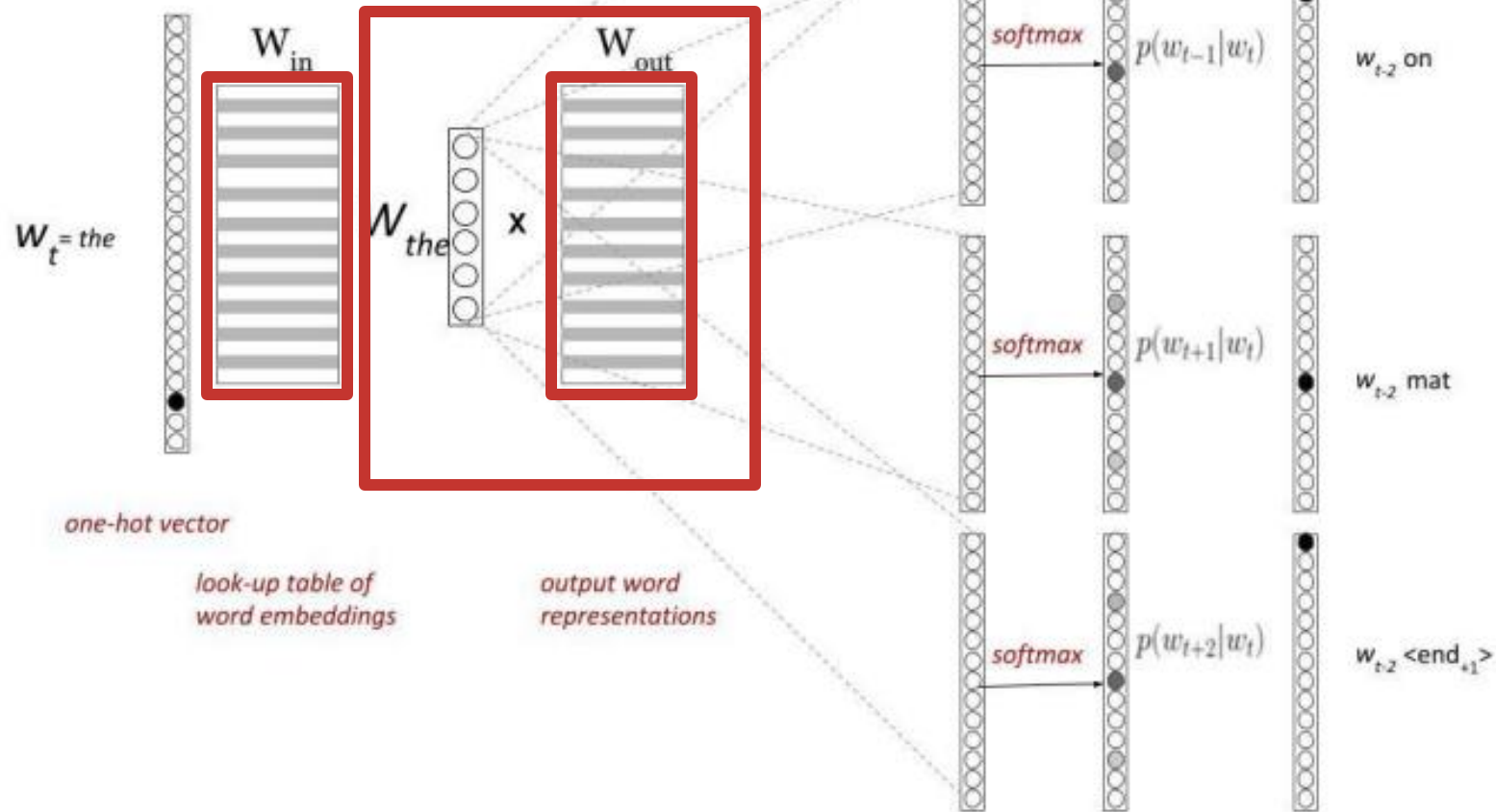
V

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"

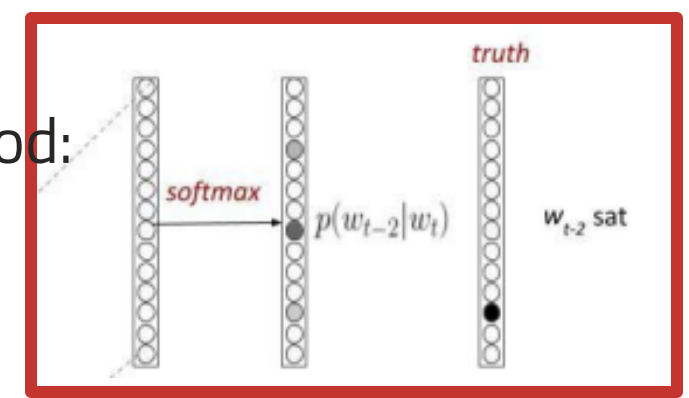
Word embedding (u_o) for output word (o)

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



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

$$J(\theta) = -\frac{1}{T} \sum_{t=1}^T \sum_{-m \leq j \leq m, j \neq 0} \log P(w_{t+j} | w_t; \theta)$$



All word vectors

the	cat	mat	on	sat	-			
5.2	1.5	...						
0.5	0.4	...						
-6.2	0.6	..						
0.5	-3.6	..						
...								

For a center word c and a context word o :

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

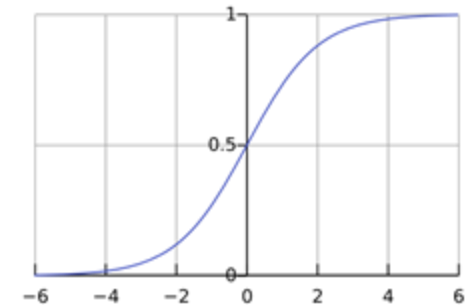
Dot product compares similarity of o and c . $u^T v = u \cdot v = \sum_{i=1}^n u_i v_i$

Normalize over entire vocabulary to give probability distribution

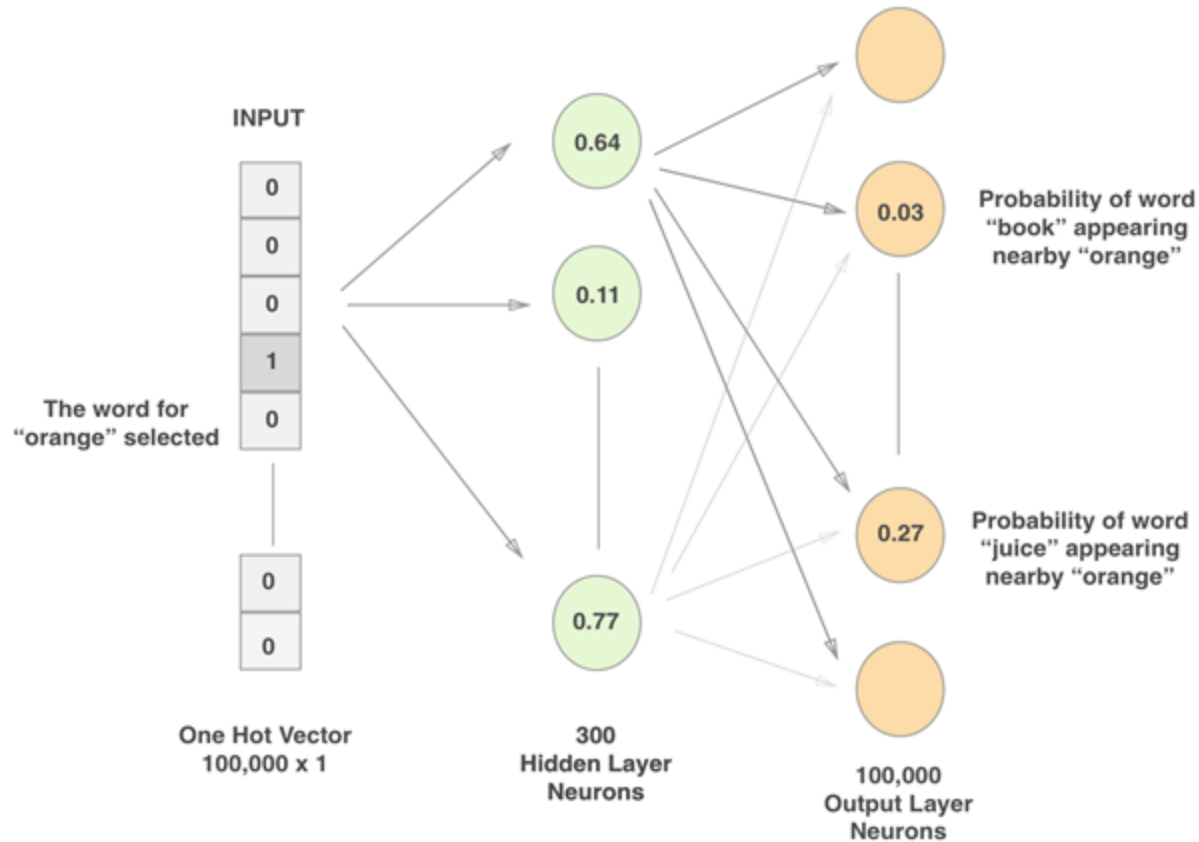
"soft" because still assigns some probability to smaller x_i

"max" because amplifies probability of largest x_i

$$\text{softmax}(x_i) = \frac{\exp(x_i)}{\sum_{j=1} \exp(x_j)} = p_i$$



$$P(x_{t-2} | x_t)$$

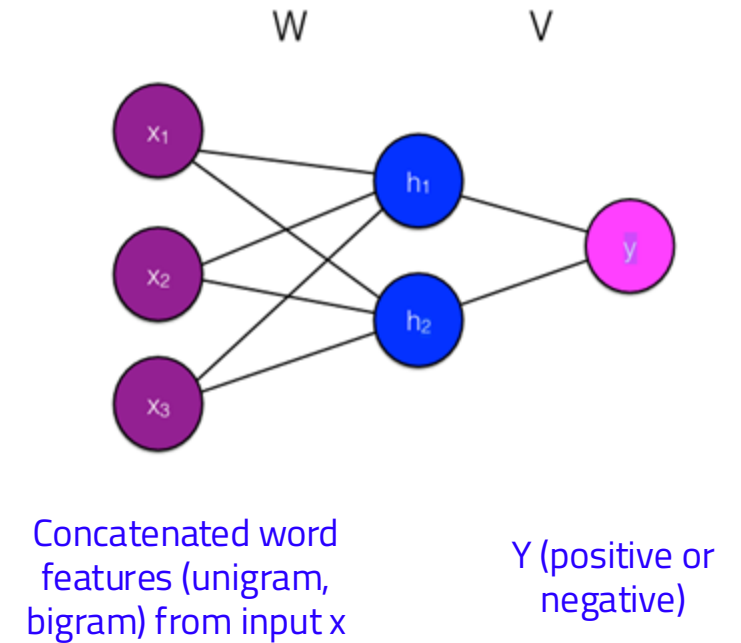


Context word 1-hot vector
(middle word in Skipgram)

Fixed-dimension word
embeddings to learn

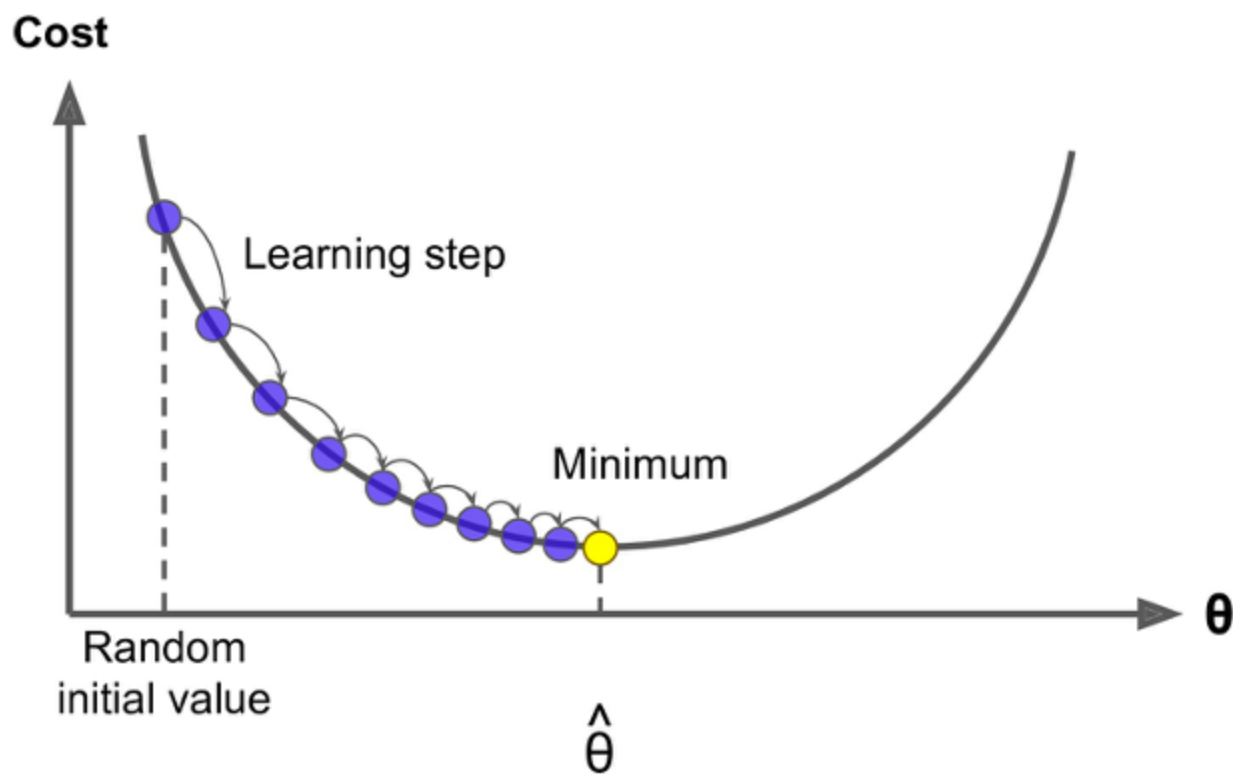
Output word to predict
(neighboring words in Skipgram)

$$P(y | x)$$



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

Idea: for current value of θ , 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

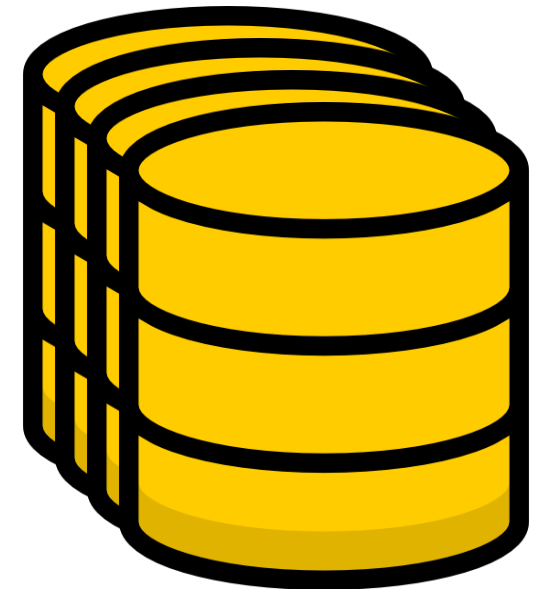
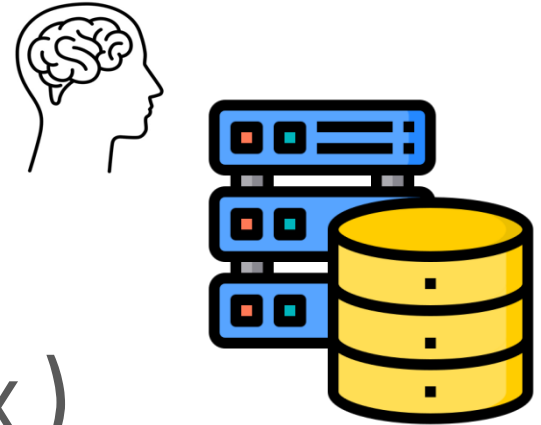
- Labeled sentiment for movie reviews (~2K labels/reviews, ~1.5 words)
- Used for **supervised** models (text classification)

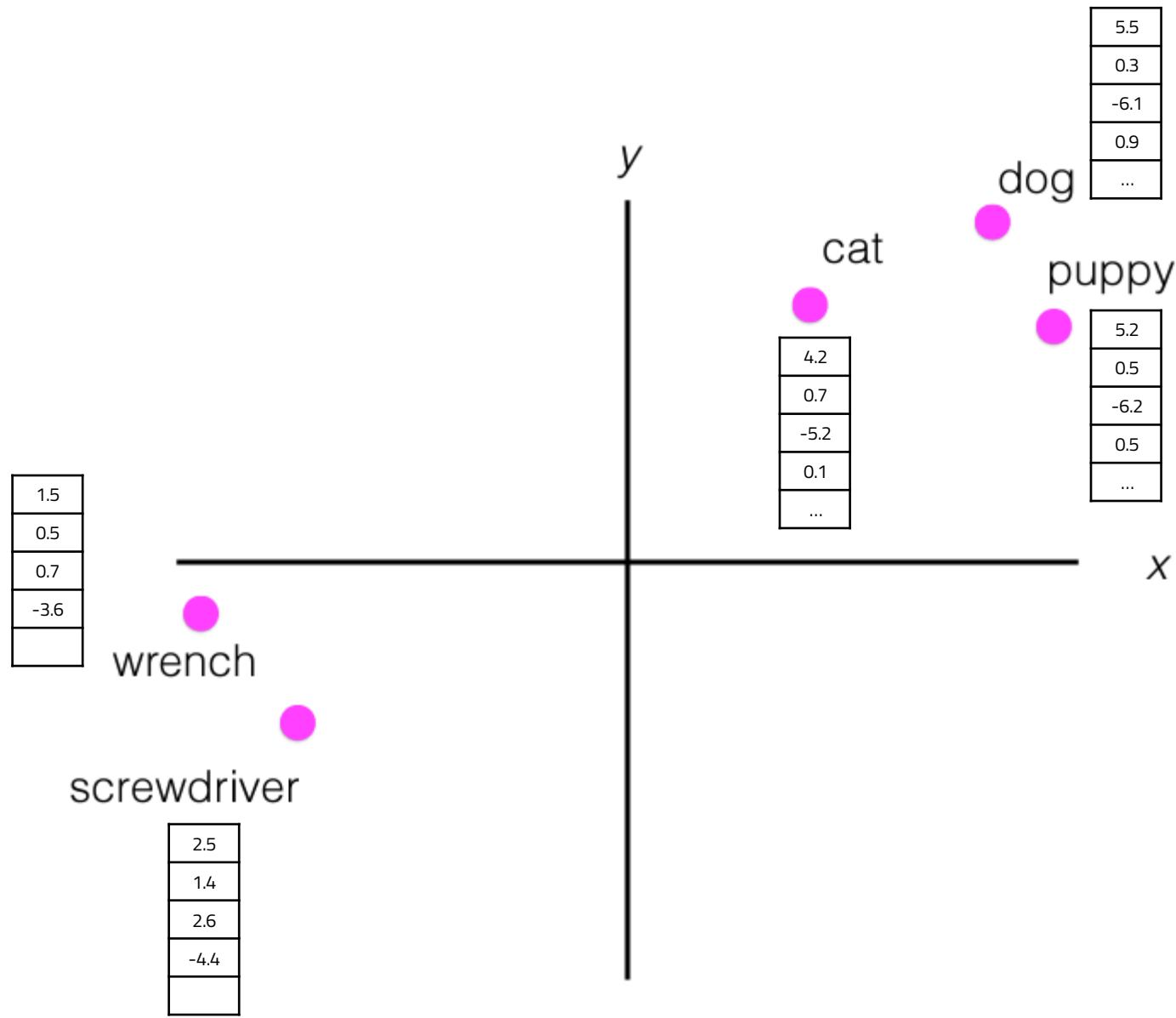
$$P(y | x)$$

□ Unlabeled text for representation learning

- Trillions of words (Wikipedia, web text, books, etc)
- Used for **word distributed representations** (skipgram training)

$$P(x_{t-2} | x_t)$$





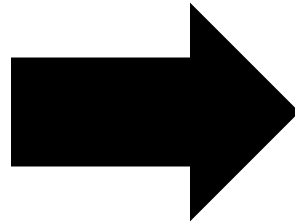
Why *dog* and *cat* are in similar positions

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"

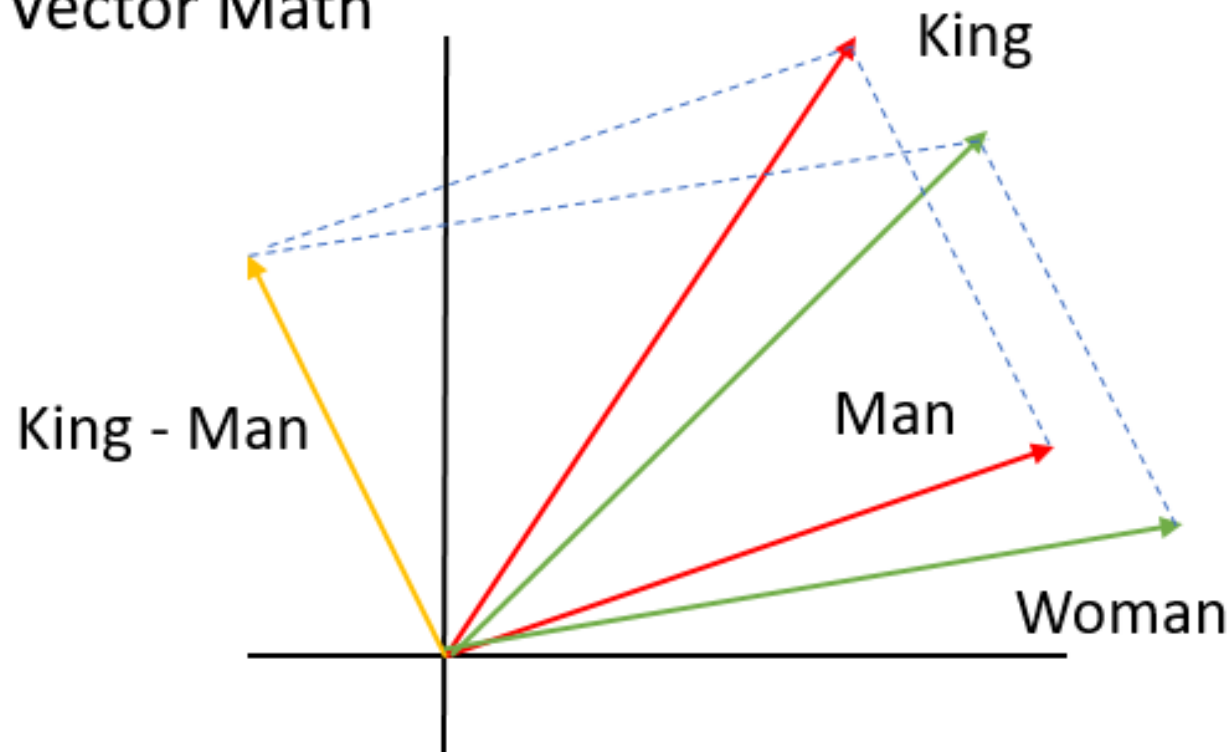
0.7
1.3
-4.5

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

3-dimensional space
Representations are not structured



Vector Math



$$v(\text{"King"}) - v(\text{"Man"}) + v(\text{"Woman"}) =$$

0.7	-	5.2	+	4.2	=	5.2
1.3		0.5		0.7		0.5
-4.5		-6.2		-5.2		-6.2
...		0.5		0.1		0.5
...	

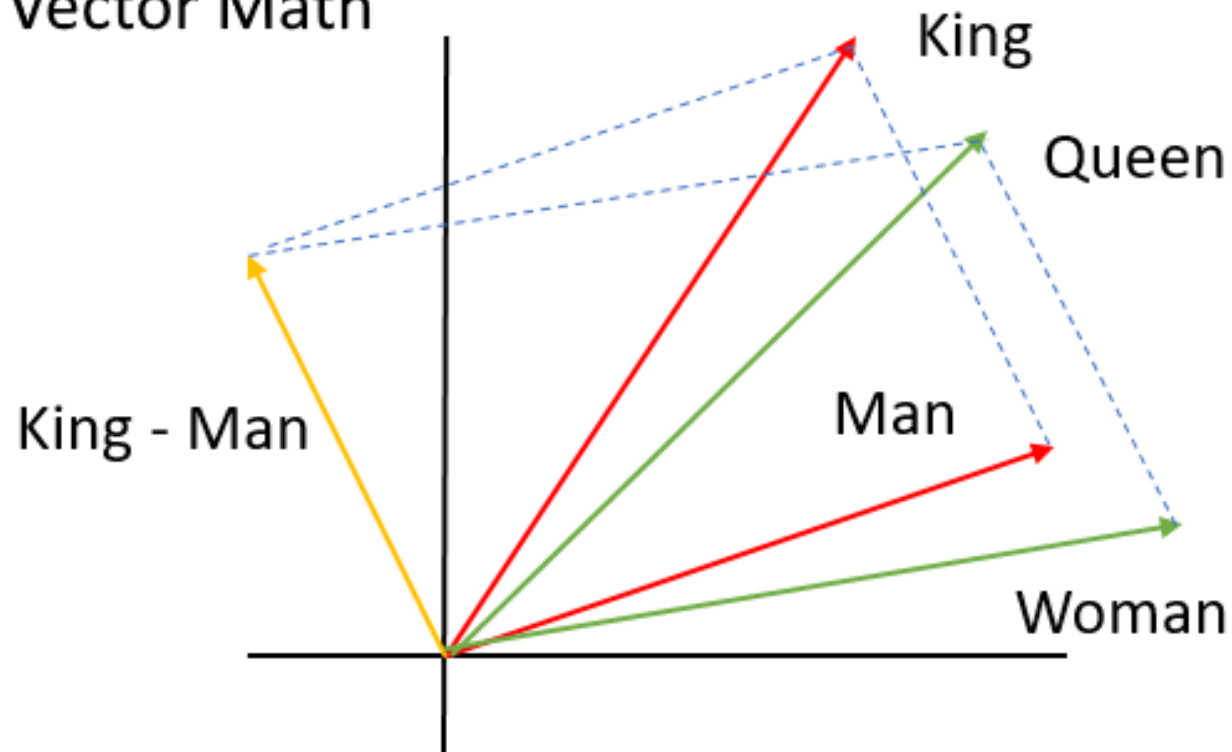
Closest vector

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**.



Vector Math



$$v(\text{"King"}) - v(\text{"Man"}) - v(\text{"Woman"}) =$$

0.7	-	5.2	+	4.2	=	5.2
1.3		0.5		0.7		0.5
-4.5		-6.2		-5.2		-6.2
...		0.5		0.1		0.5
...	

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**.





Boy



Girl



Vector Visualization

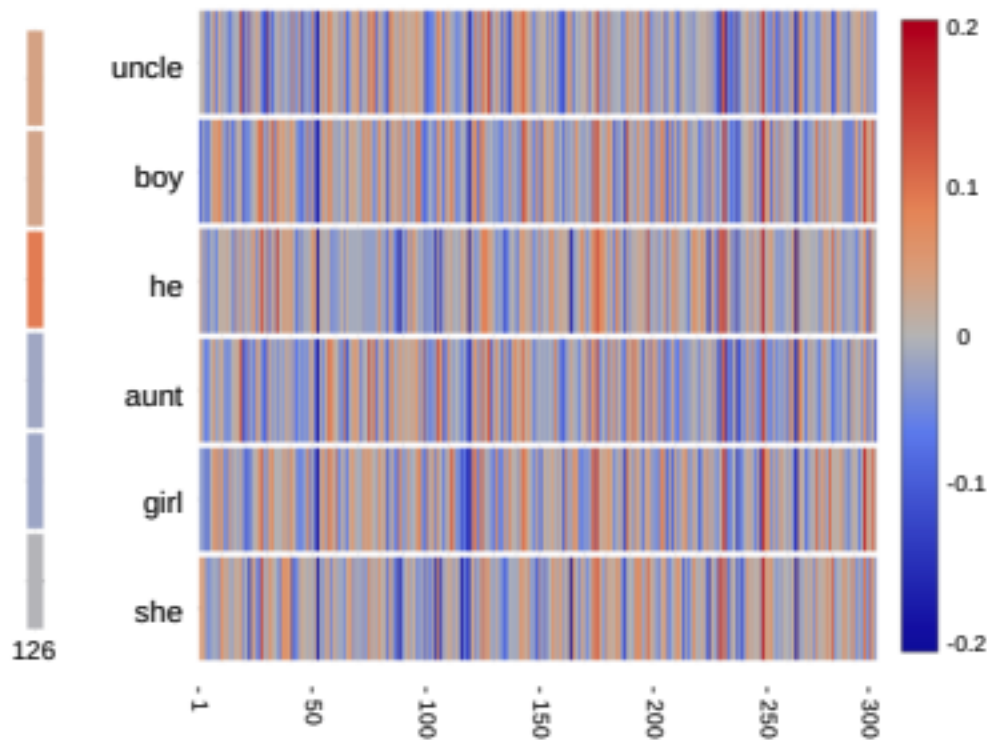


Figure 2: Embedding vectors for three male words (“uncle”, “boy”, “he”) and three female words (“aunt”, “girl”, “she”). Component 126, shown magnified at left, is positive for the male words and negative for the female words.

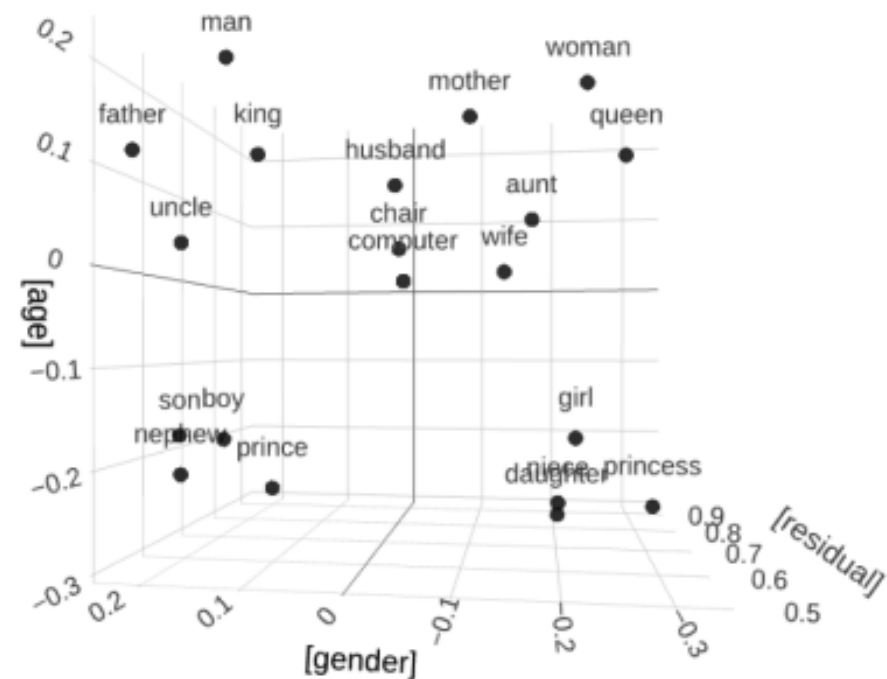


Figure 3: Words plotted in our 3D semantic space. Male words appear in the positive (left) half of the x-axis; female words in the negative (right) half. Adult words are in the positive (top) half of the y-axis; youth words in the negative (bottom) half. The third dimension is the “semantic residual”,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

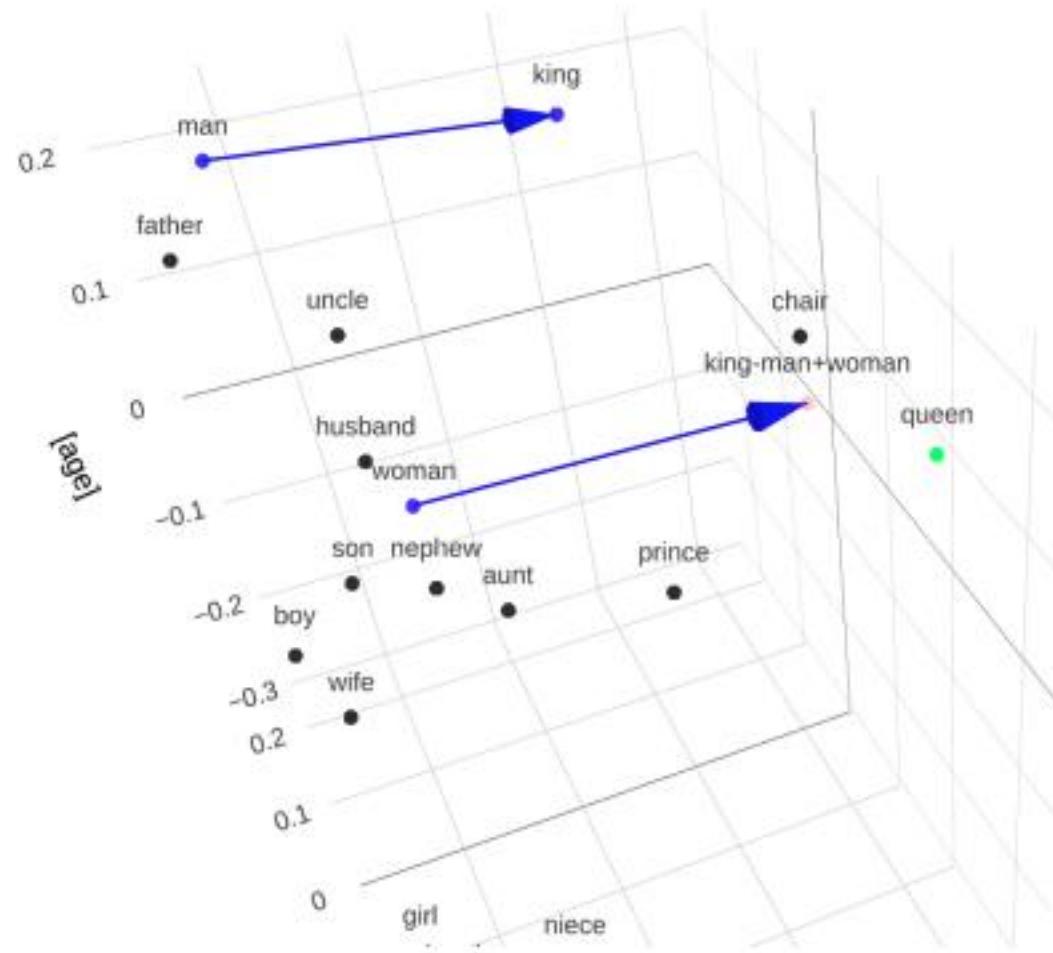


Figure 5: Analogy by vector arithmetic: “man” is to “king” as “woman” is to “king - man + woman” = “queen”.

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

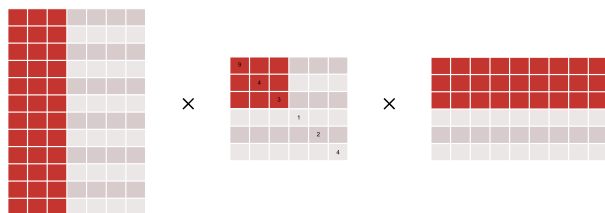


Count-based vs Prediction-based Methods

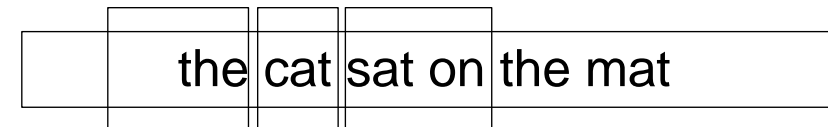
To obtain good dense representations

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

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



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



Count-based vs Prediction-based Methods

LSA, HAL (Lund & Burgess)

Hellinger-PCA (Rohde et al, Lebet & Collobert)

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

Skip-gram/b (Mikolov et al)

NLM, HLBL, RNN (Bengio et 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

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

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/>

- GloVe (Pennington et al., 14)

- ✓ <http://nlp.stanford.edu/projects/glove/>

- FastText (Bojanowski et al.' 17)

- ✓ <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/>



Evaluation:

How representative is your learned vector?



Types of Evaluation

□ Intrinsic vs Extrinsic

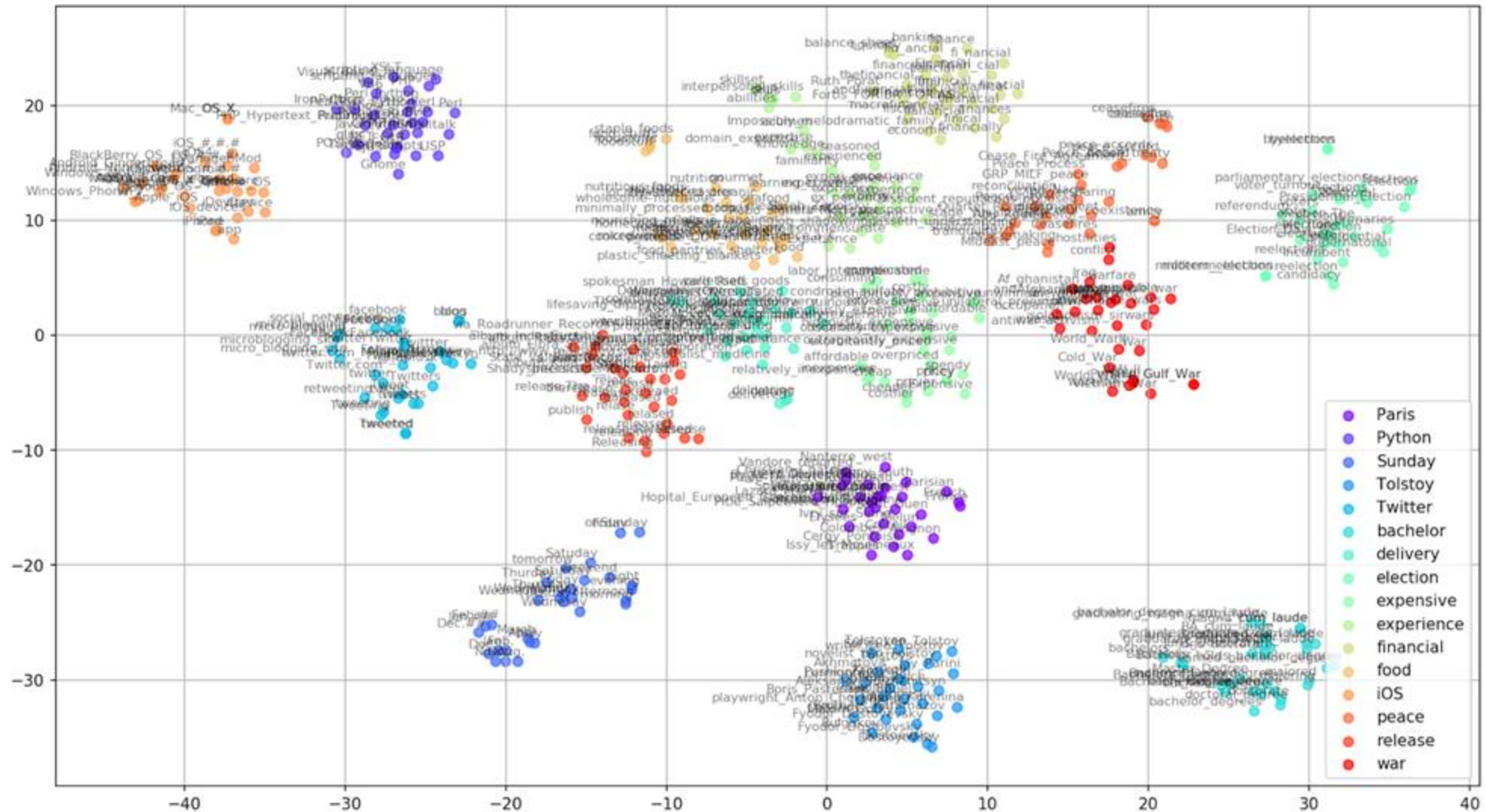
- Intrinsic: How good is it based on its features?
- Extrinsic: How useful is it downstream?

□ Qualitative vs. Quantitative

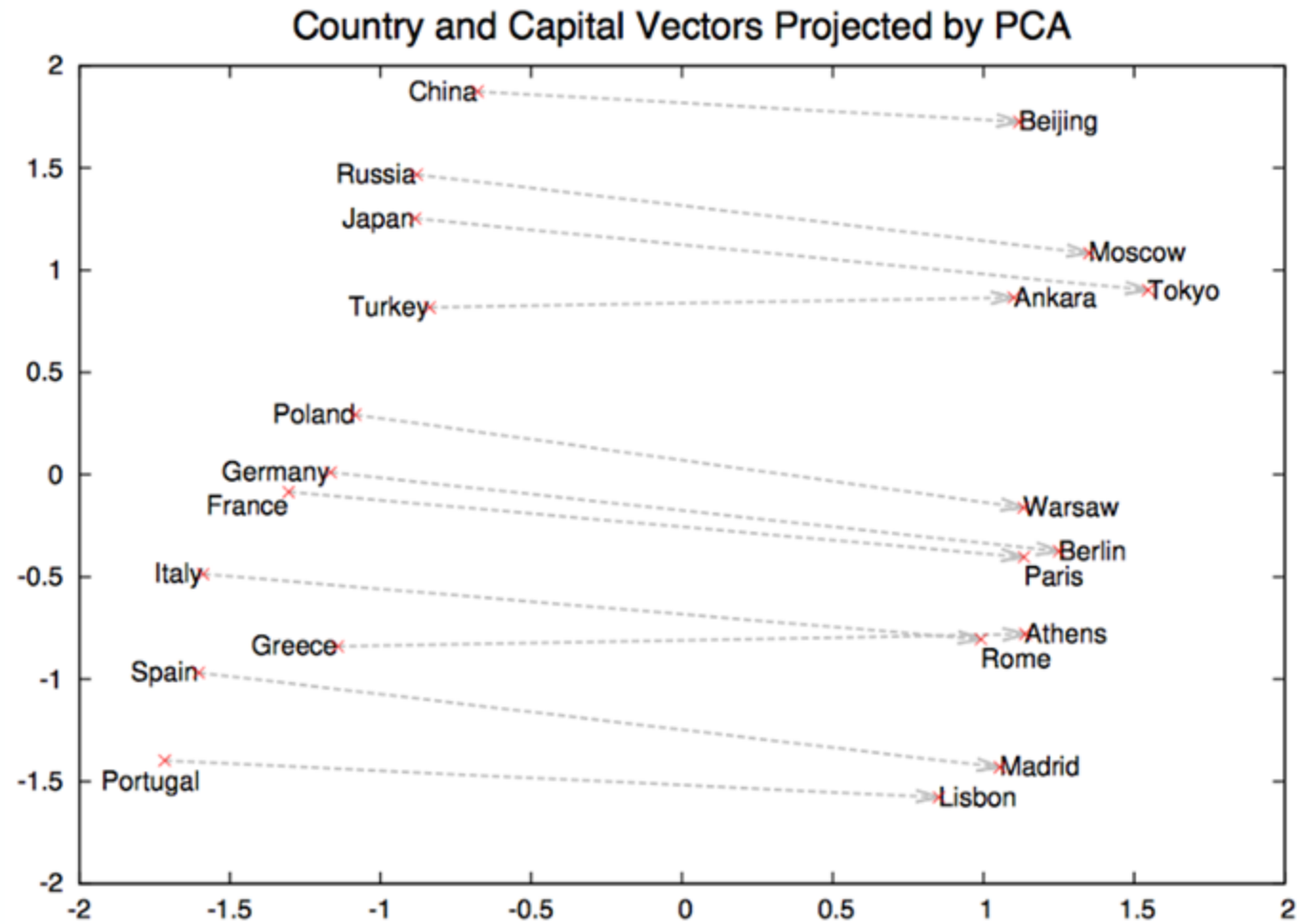
- Qualitative: Examine the characteristics of examples.
- Quantitative: Calculate statistics



Visualization of Embeddings



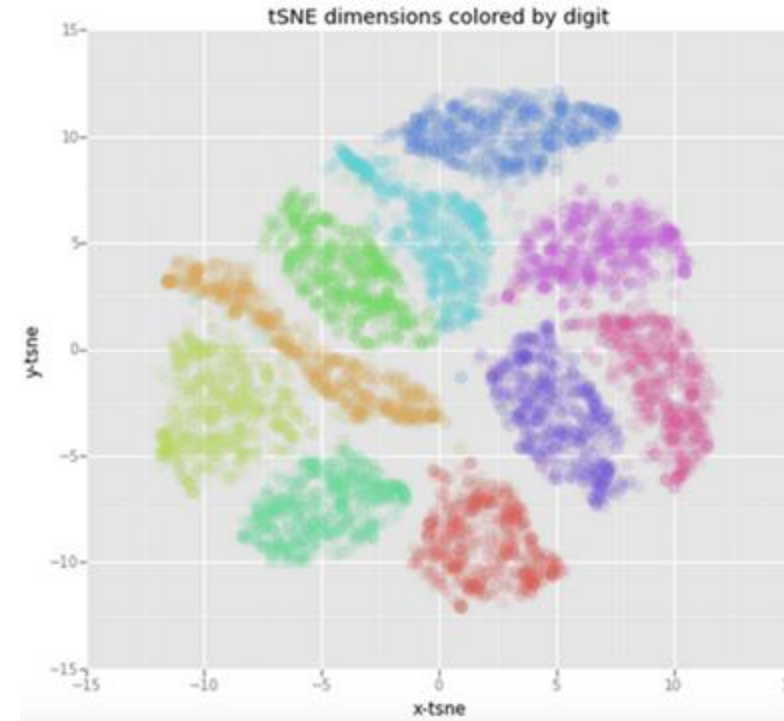
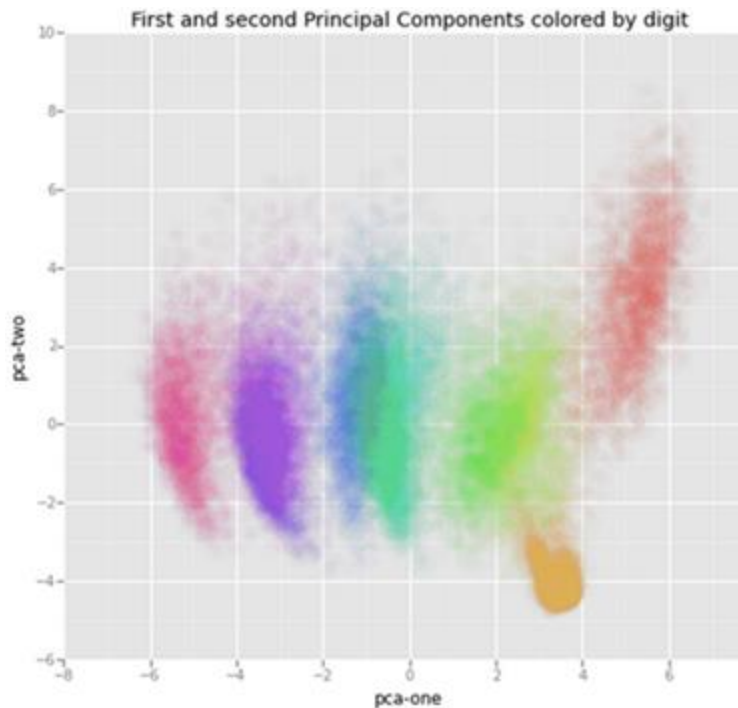
Visualization of Embeddings



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

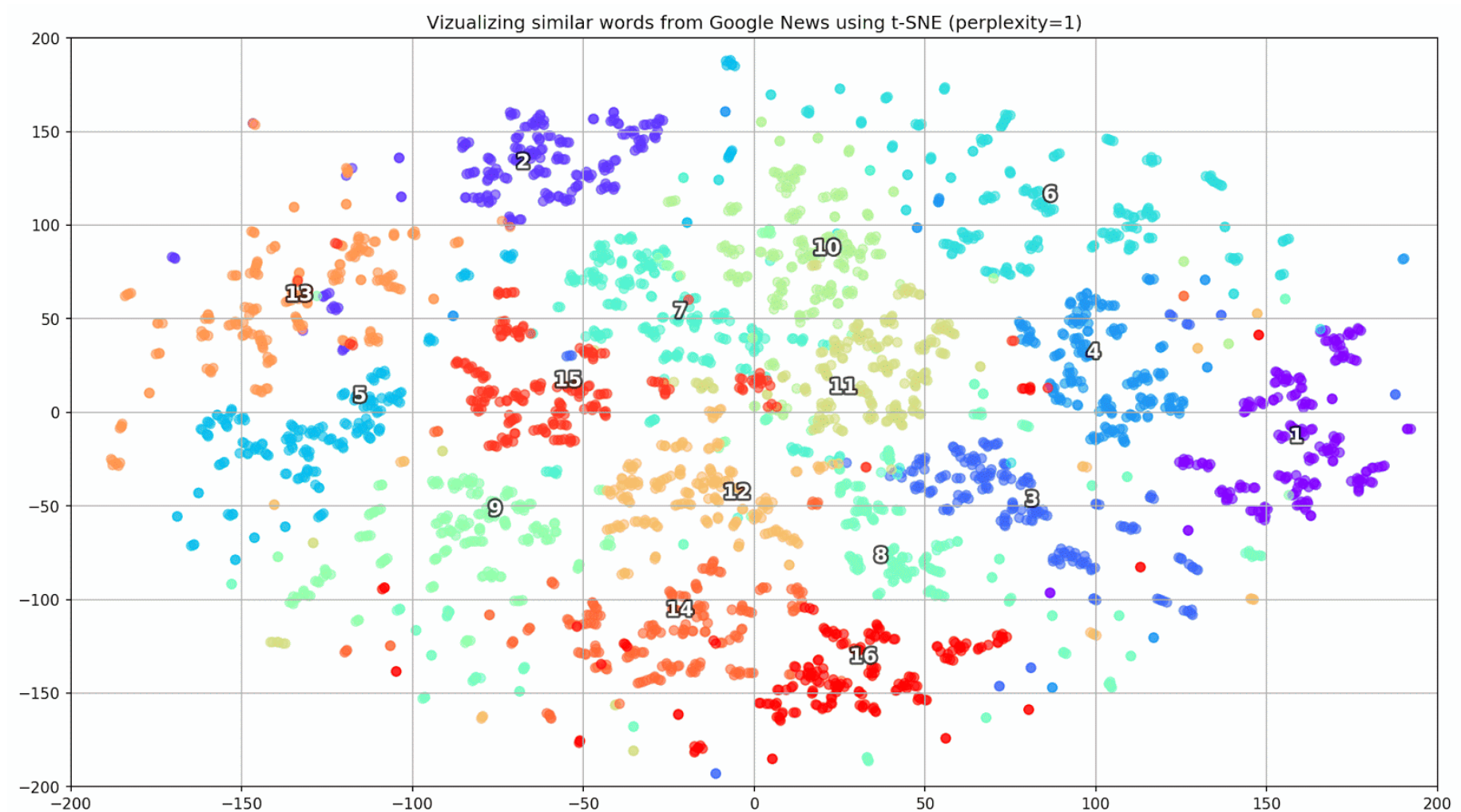


T-SNE

Image from Derksen (2016)



Hyper-parameters in t-SNE (e.g., perplexity)



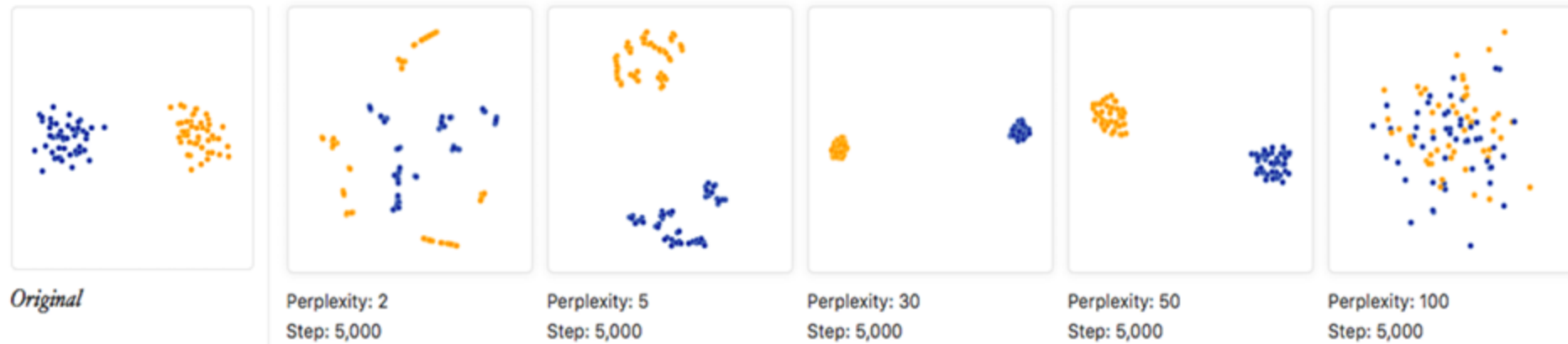
<https://towardsdatascience.com/google-news-and-leo-tolstoy-visualizing-word2vec-word-embeddings-with-t-sne-11558d8bd4d>



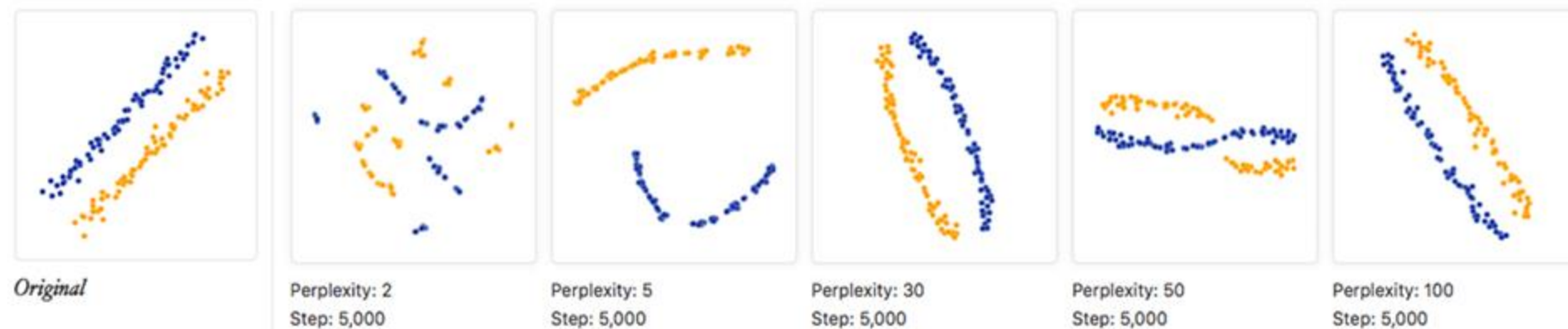
t-SNE Visualization can be Misleading!

(Wattenberg et al. 2016)

Settings matter



Linear correlations cannot be interpreted



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).

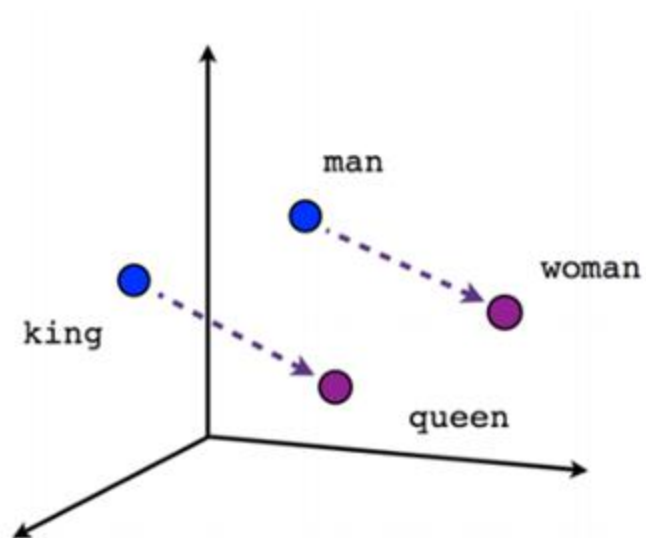
For analogy **Germany : Berlin :: France : ?**,
find closest vector to $v(\text{"Berlin"}) - v(\text{"Germany"}) + v(\text{"France"})$

possibly	impossibly	Certain	Uncertain
generating	generated	Shrinking	Shrank
think	thinking	Look	Looking
Baltimore	Maryland	Minneapolis	Minnesota
shrinking	shrank	Slowing	Slowed
Rabat	Morocco	Astana	Kazakhstan

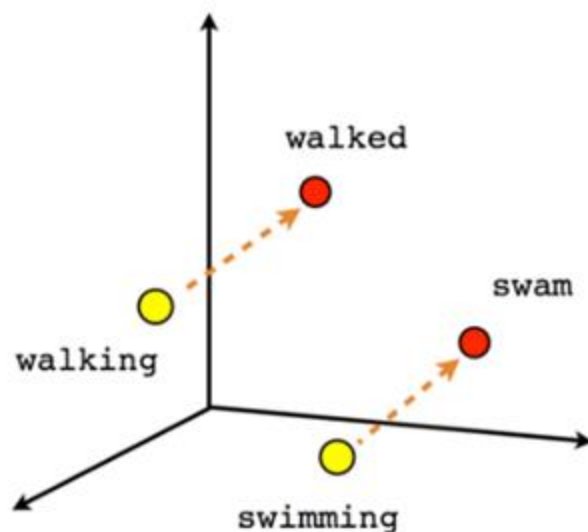


Intrinsic evaluation:

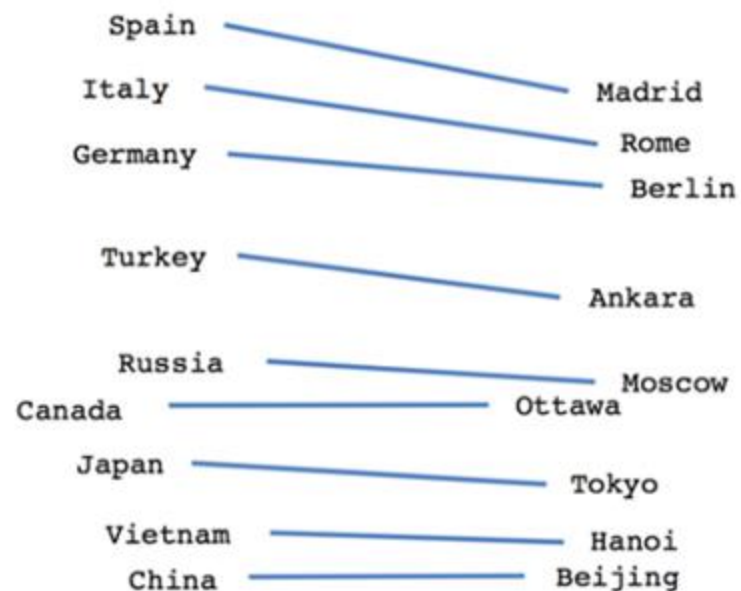
Analogical reasoning (Mikolov et al., 2013).



Male-Female



Verb tense



Country-Capital



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

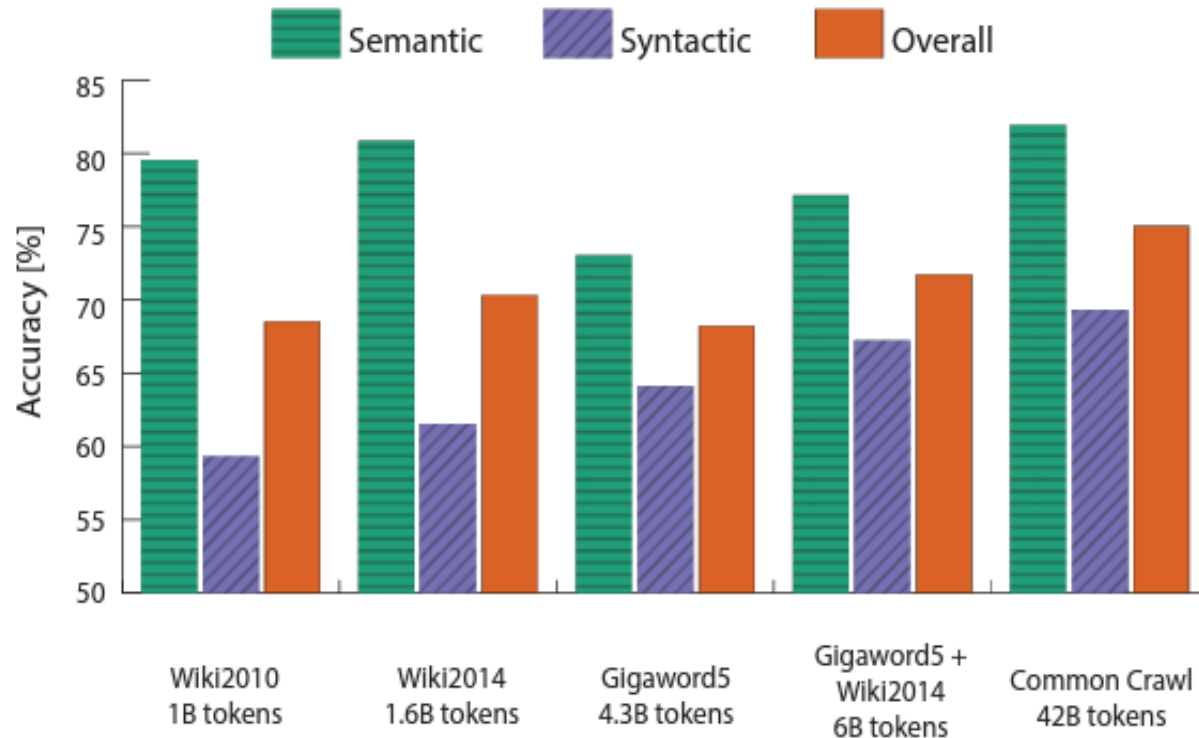


Analogical reasoning test

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



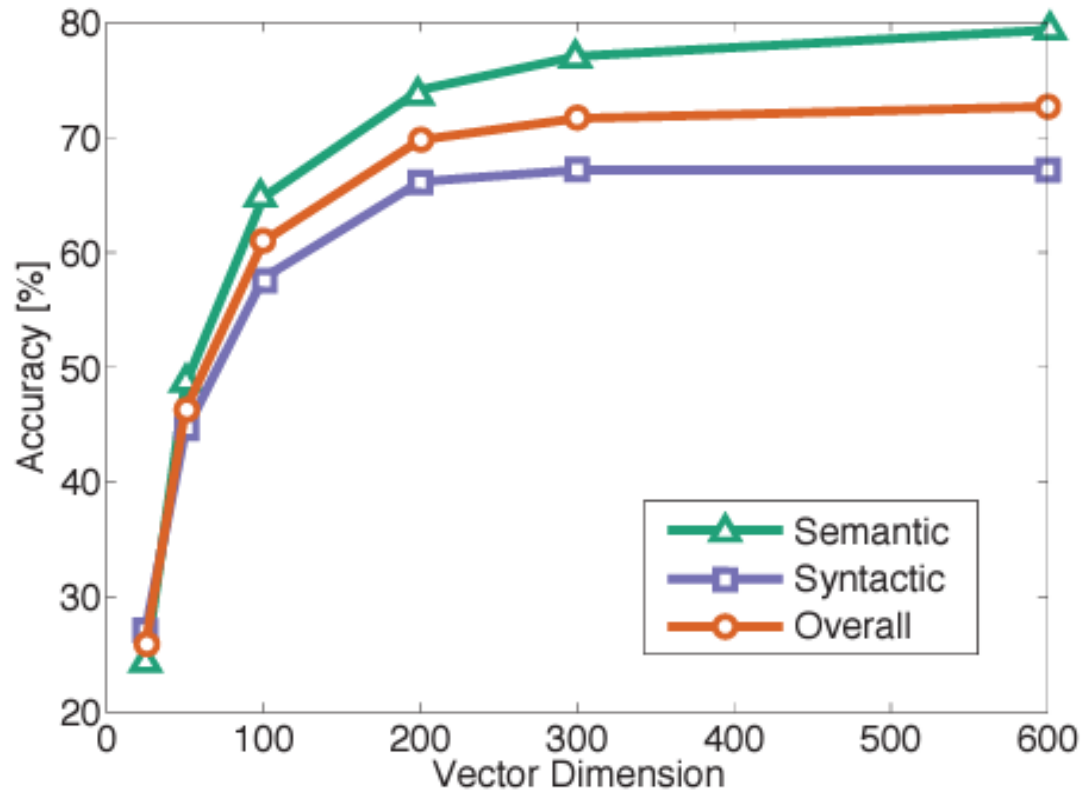
Analogy evaluation and hyper-parameters



- More data helps
- Wikipedia is better than news text



Analogy evaluation and hyper-parameters



- Dimensionality
- Good dimension is ~300

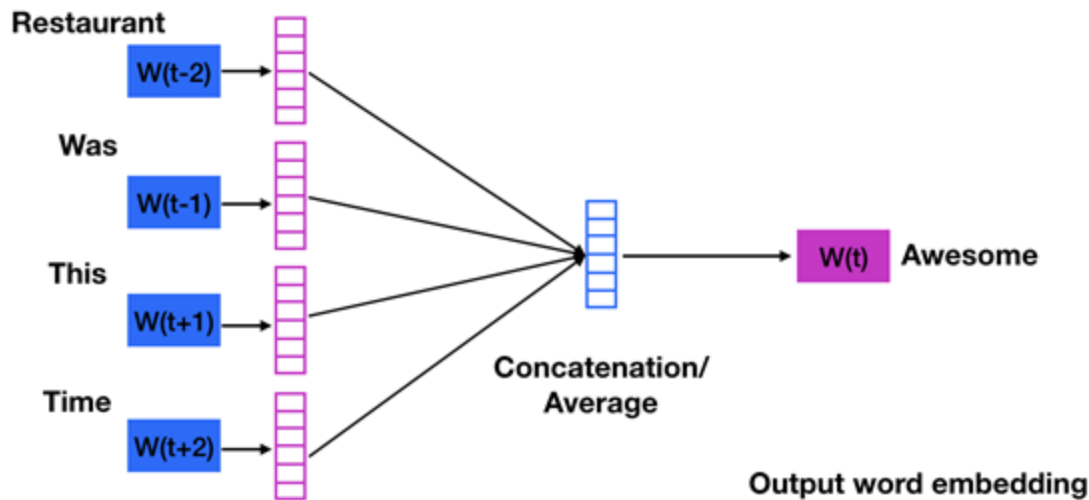


Extrinsic Evaluation

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



Input words' embeddings

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

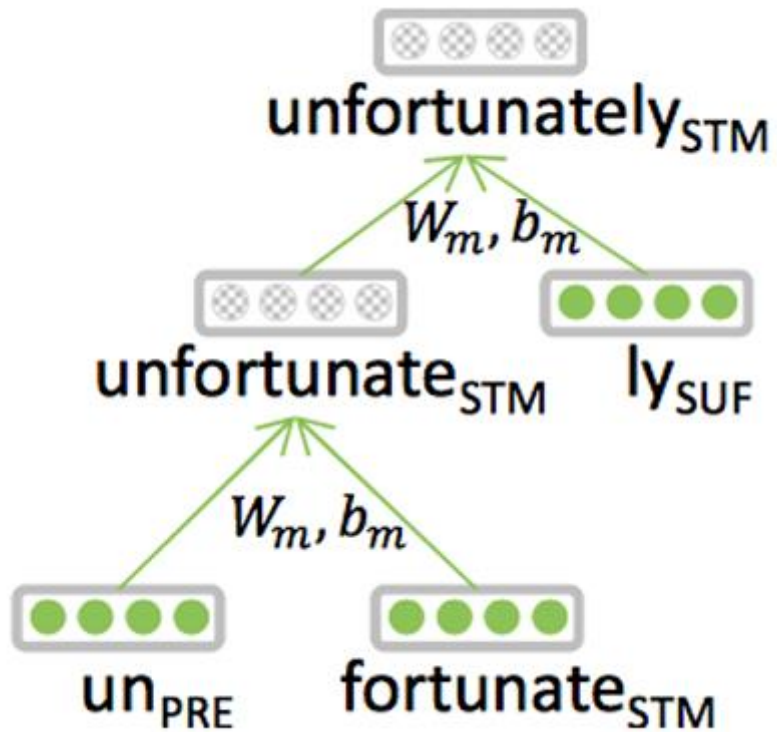


Limitations of Embeddings

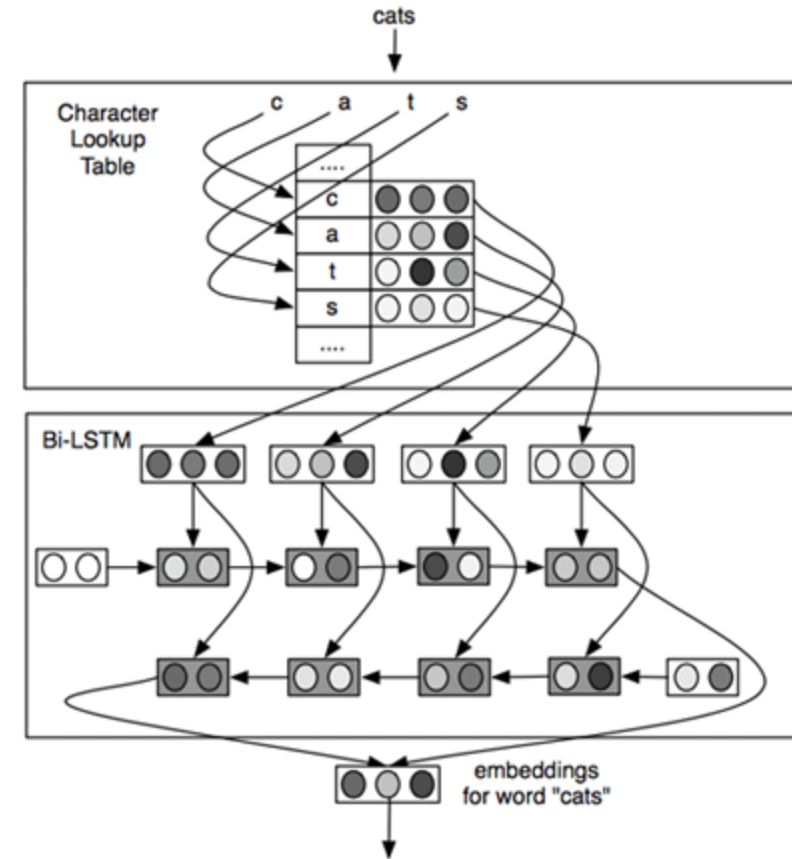
- ❑ Sensitive to **superficial differences** (dog / dogs)
 - E.g. misspellings: “minuscule” → “miniscule”
 - E.g. compounded/prefixed/suffixed words split into “wrong” subwords
“descheduled” ⇒ [“des”, “##ched”, “##uled”]
- ❑ **Not necessarily coordinated** with knowledge or across languages
- ❑ Can encode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



Sub-word Embeddings

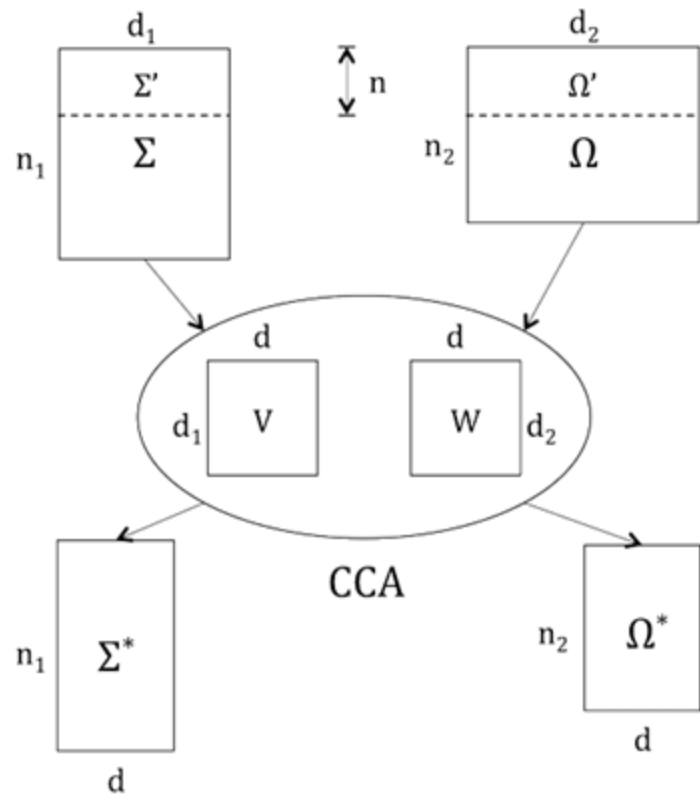


Morpheme-based (Luong et al. 2013)

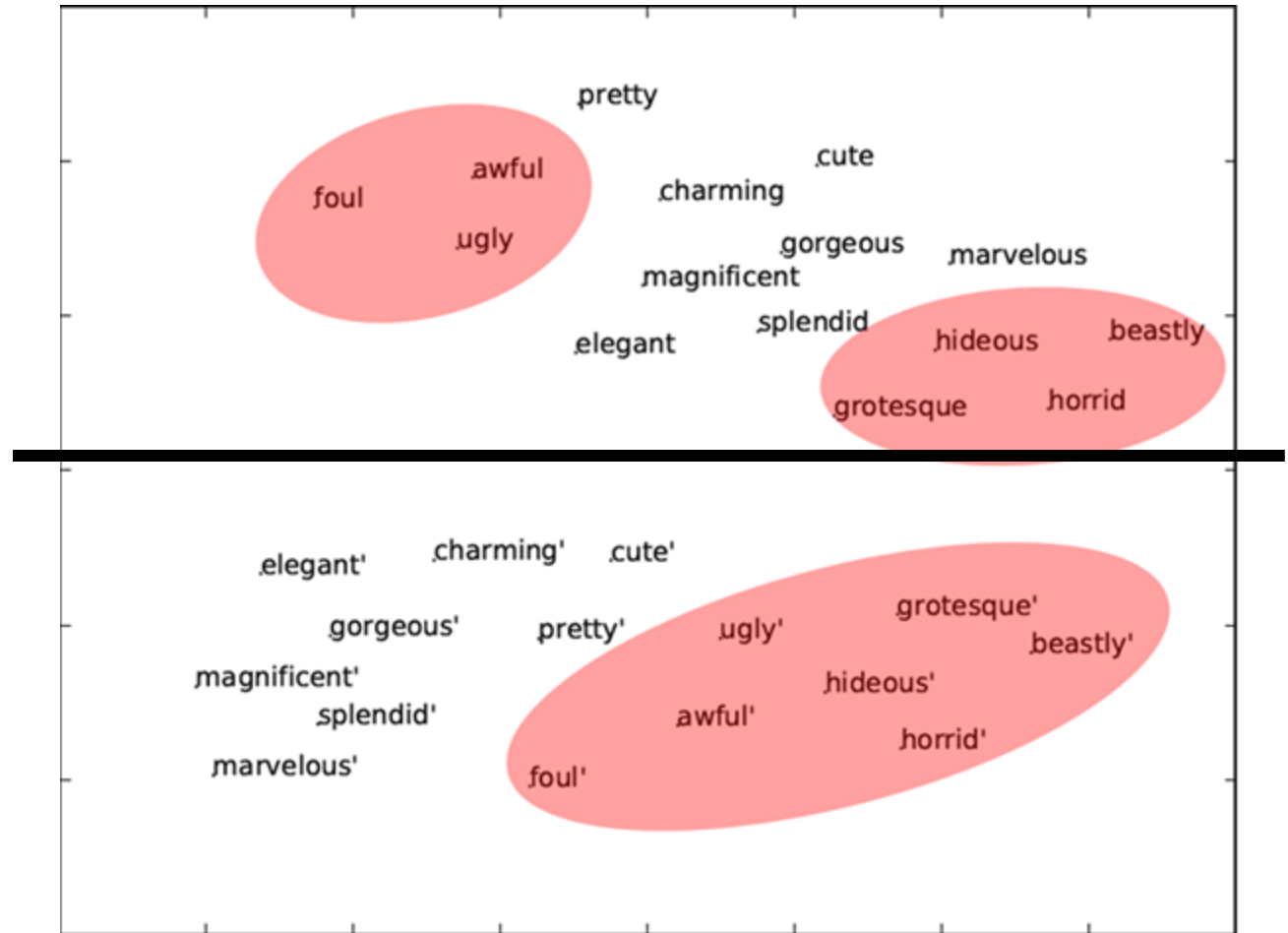


Character-based (Ling et al. 2015)

Multilingual Coordination of Embeddings using dictionaries



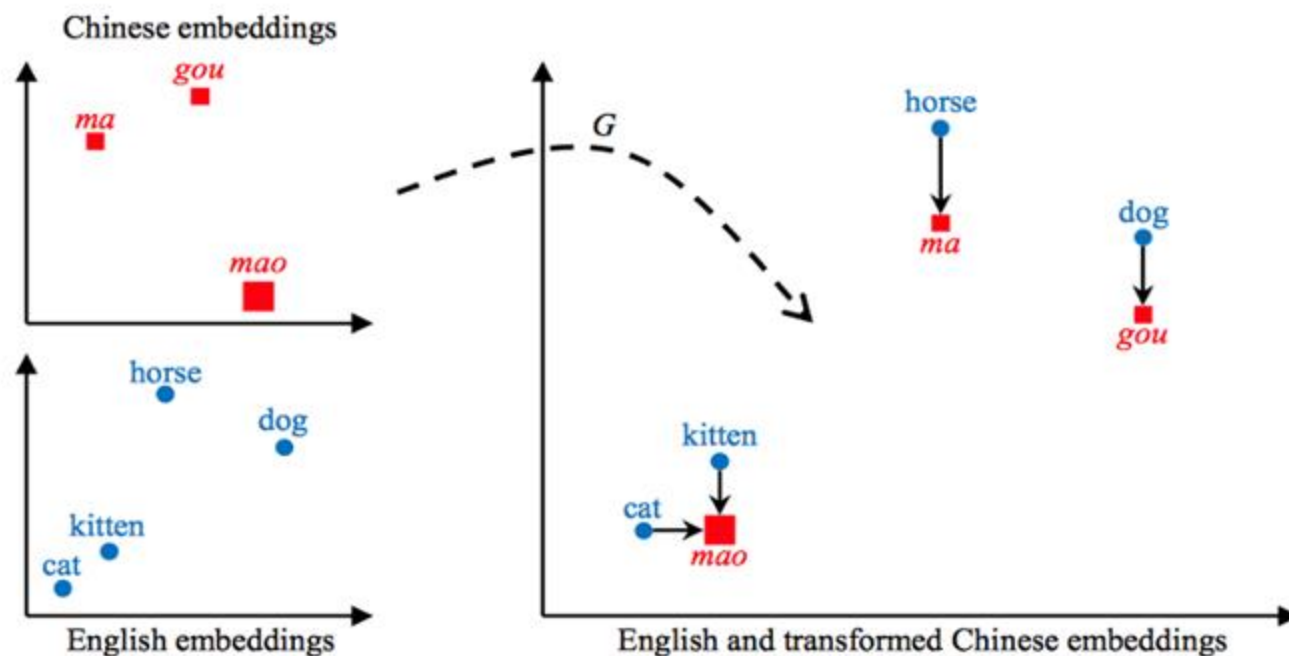
Improving Vector Space Word Representations Using Multilingual Correlation (Faruqui & Dyer, 2014)



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

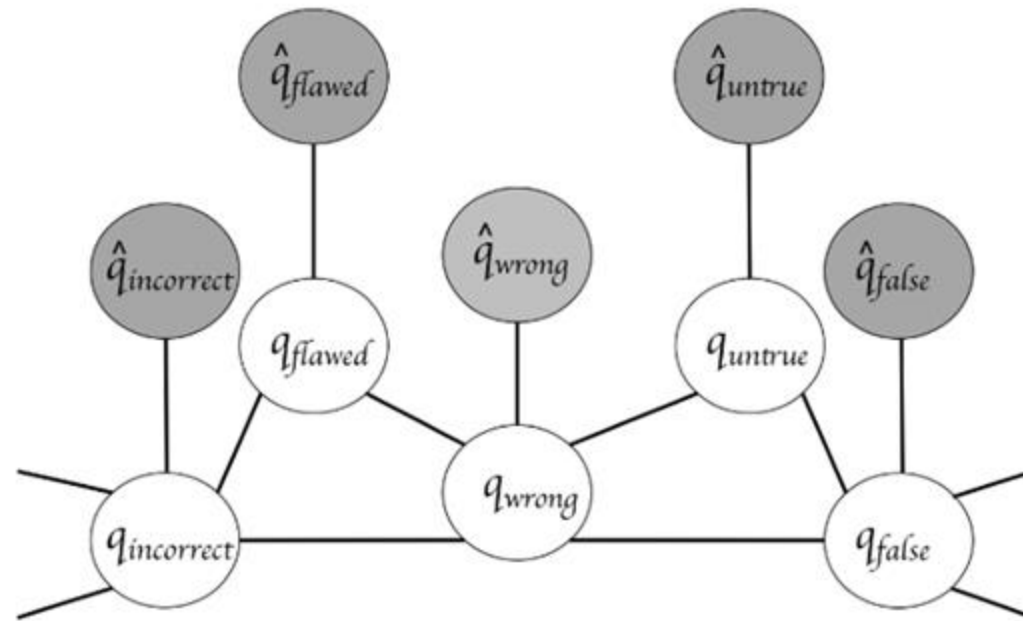
Unsupervised Coordination of Embeddings

- 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)



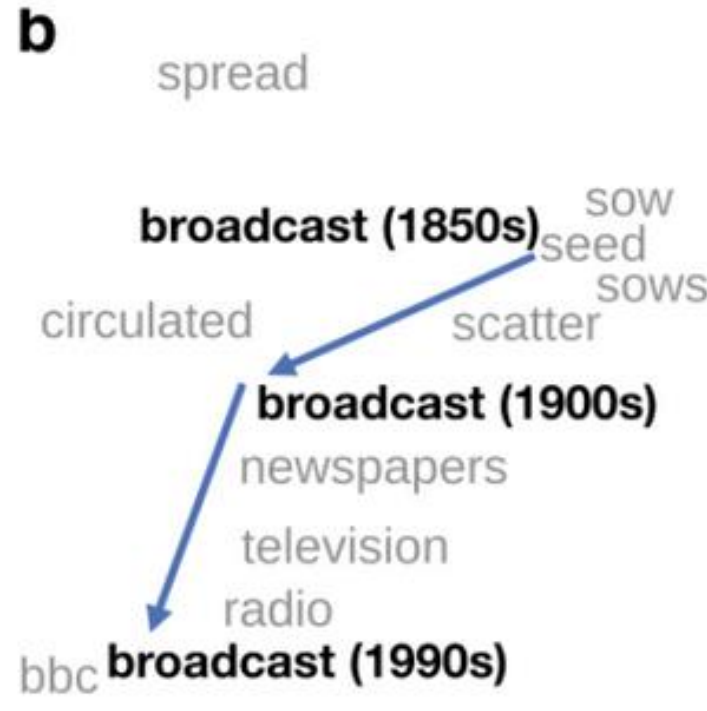
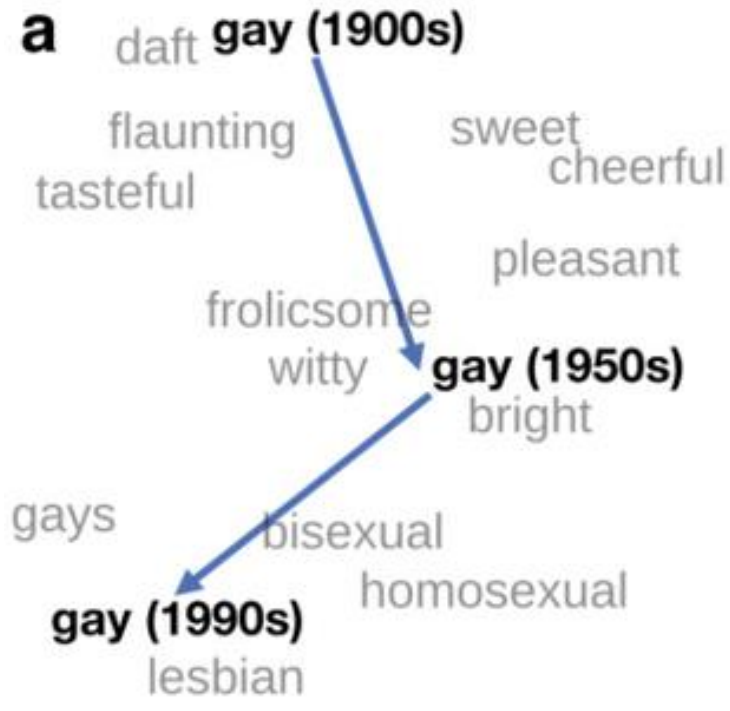
Retrofitting of Embeddings to Existing Lexicons

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



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

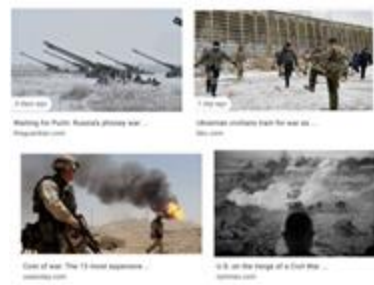




semantic to model temporal word analogy or relatedness (Szymanski, 2017; Rosin et al., 2017) or to capture the dynamics of semantic relations (Kutuzov et al., 2017)



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	<i>shape</i>
1.3	<i>color</i>
-4.5	<i>texture</i>

