Word Vectors

Princeton AI4ALL: NLP Group

How should we represent words?

https://www.fluentu.com/blog/educator-english/wp-content/uploads/sites/13/2014/11/beginner-esl-vocabulary-how-to-teach-100-words-in-one-lesson.jpg

© Can Stock Photo

https://foodscapesbristol.files.wordpress.com/2013/07/day-5.jpg
https://antarasdiary.com/30-brilliant-typography-inspirations-featuring-the-word-summer/
https://www.canstockphoto.com/music-word-cloud-34509746.html

To communicate this information to computers…

To communicate this information to computers…

quantitatively represent words,

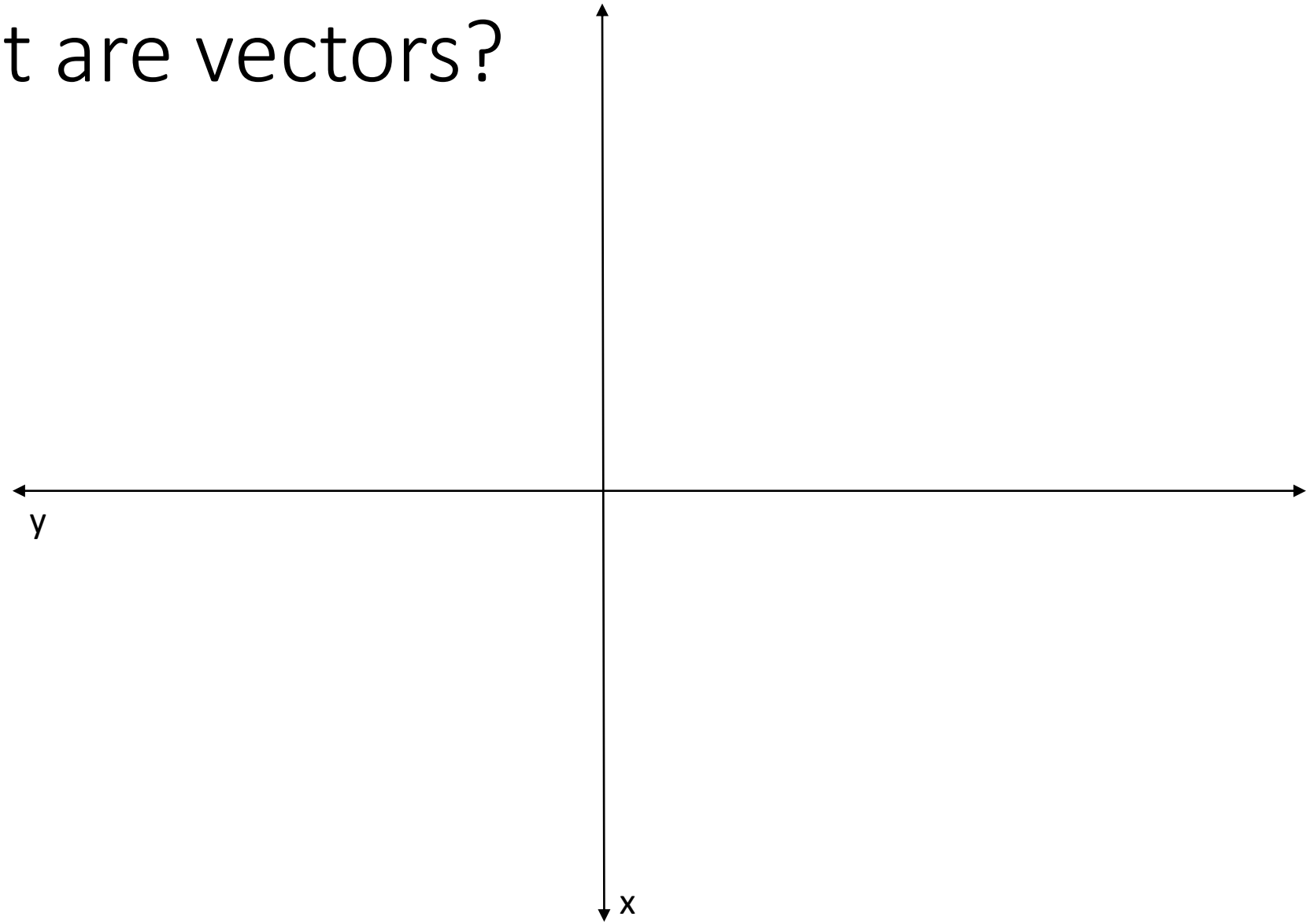
To communicate this information to computers…

quantitatively represent words,

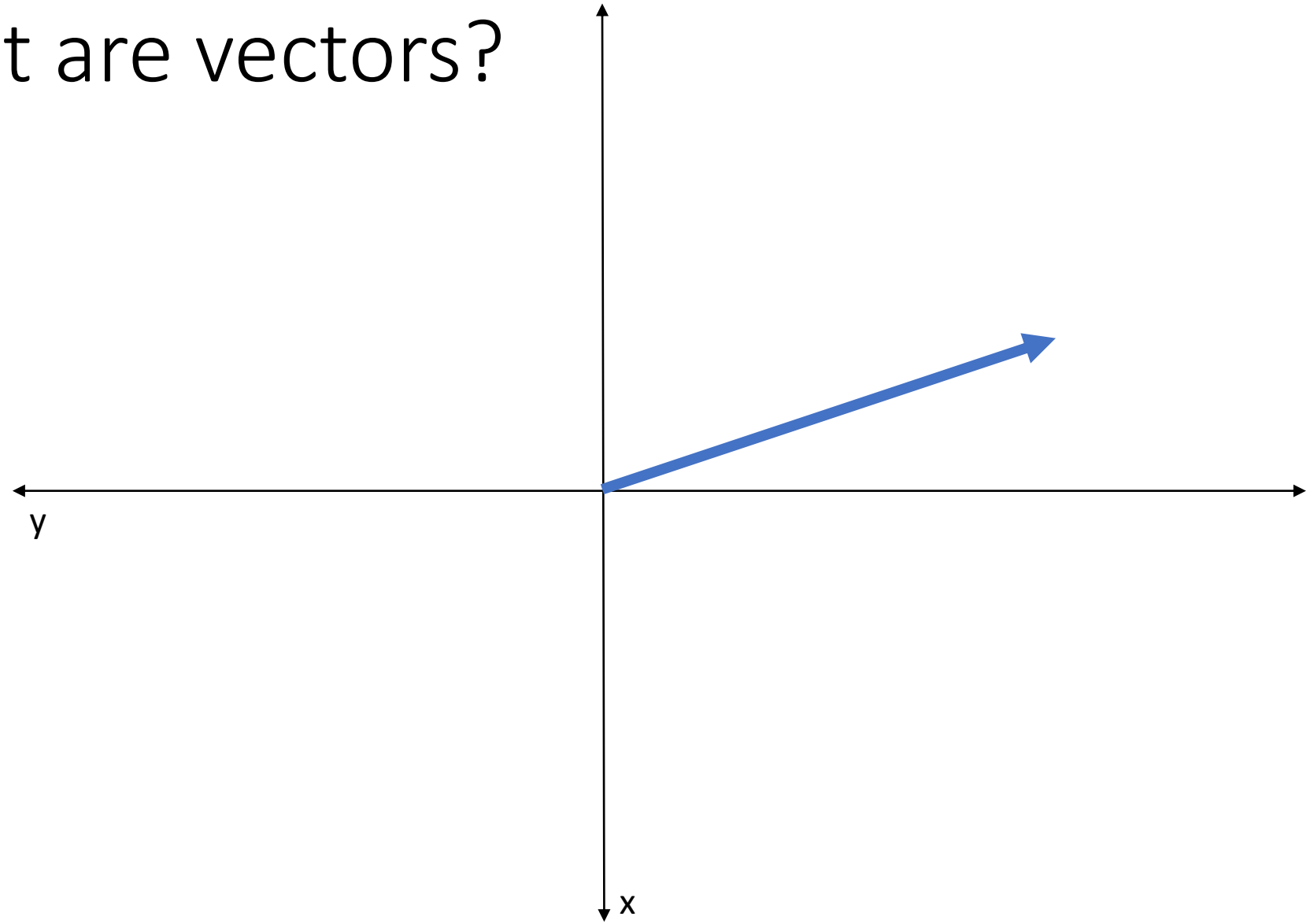
where more similar words have similar representations

What are word vectors?

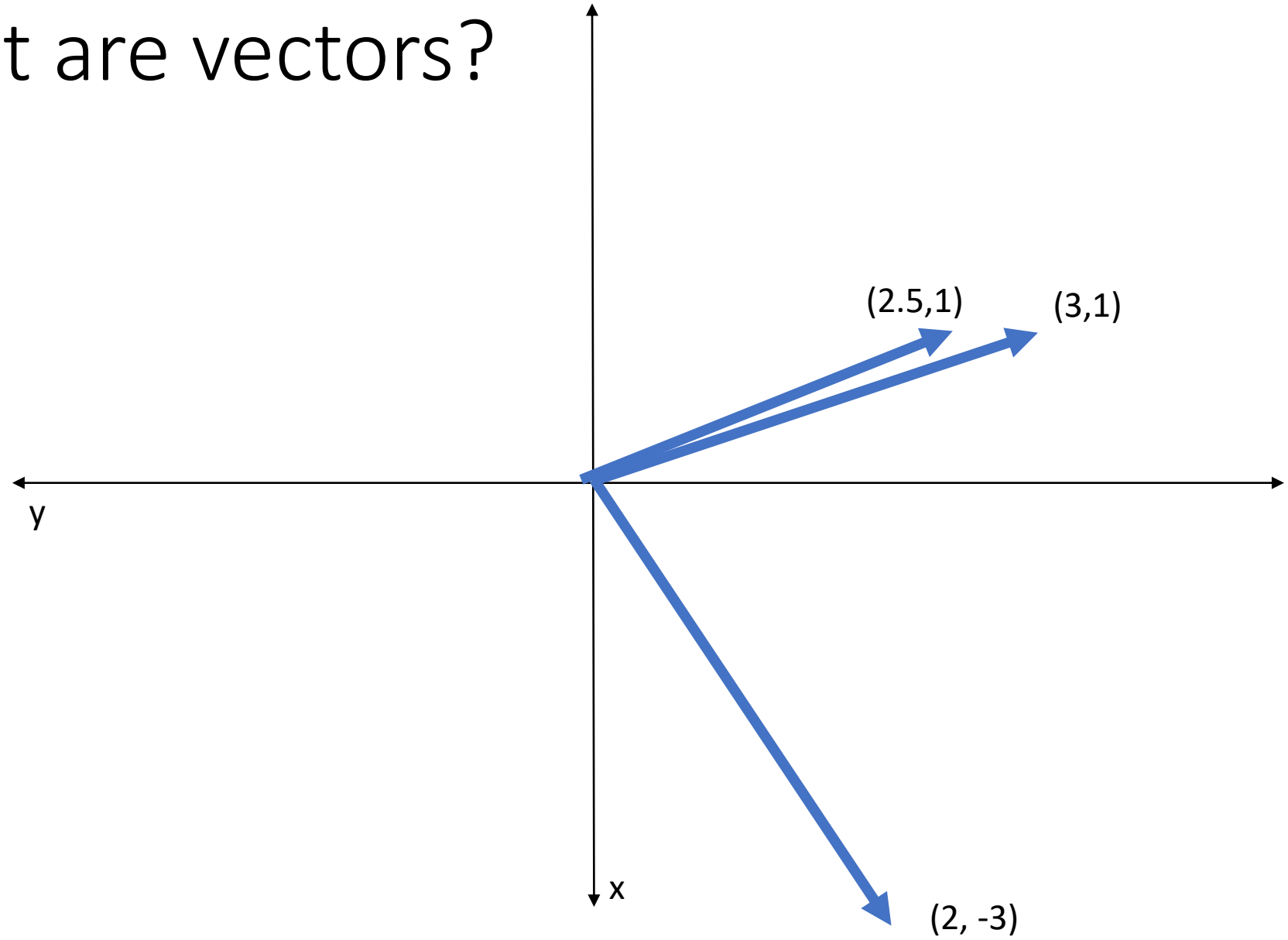
What are vectors?



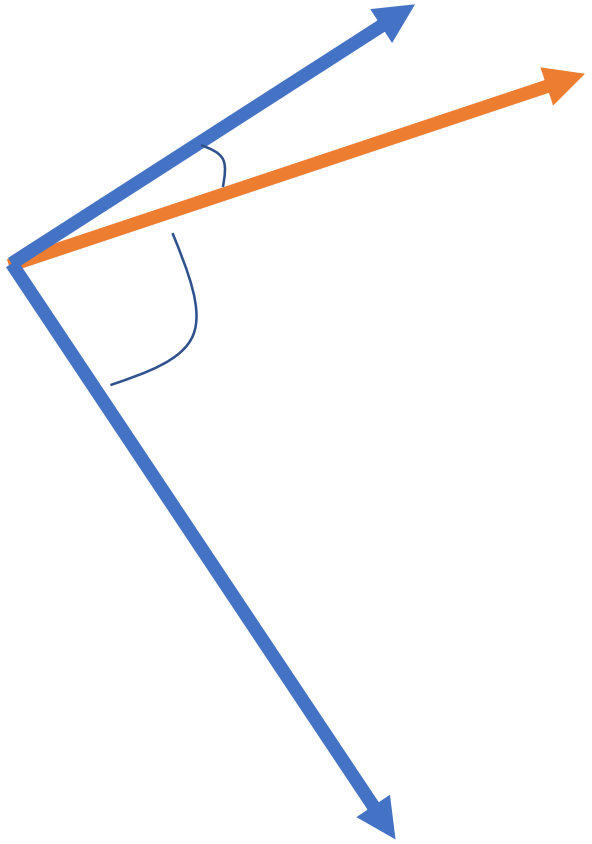
What are vectors?



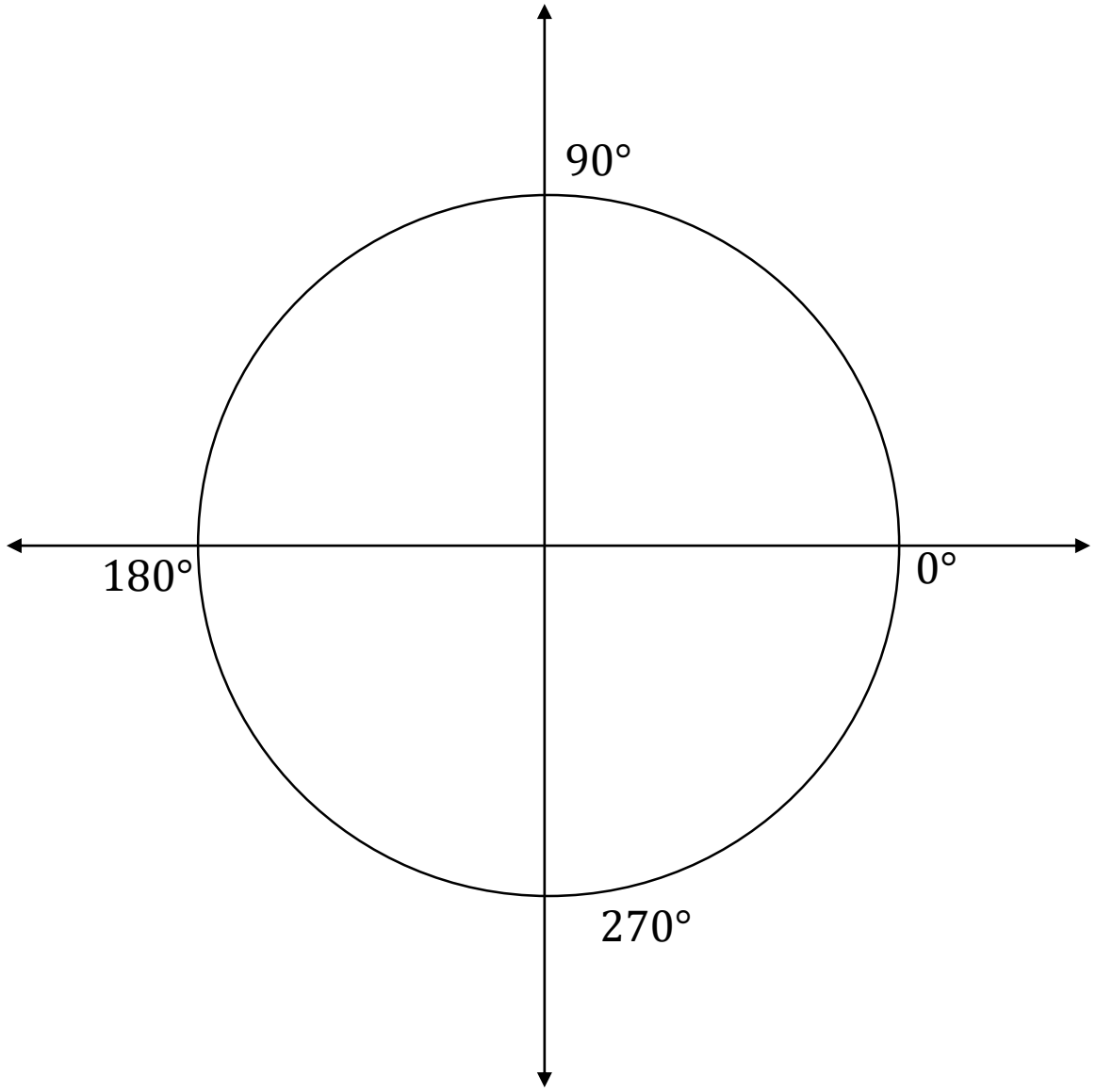
What are vectors?



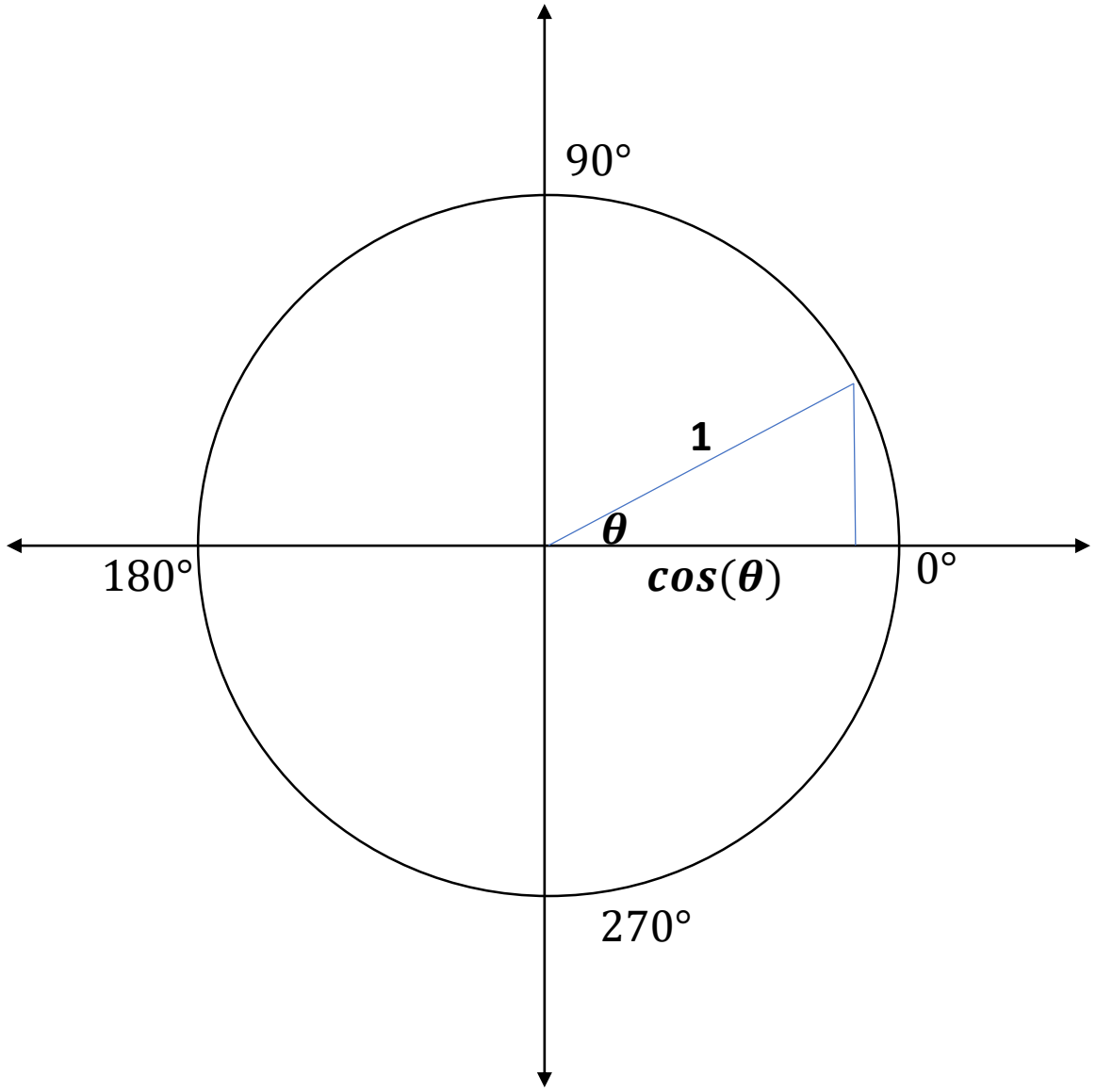
How do we measure similarity between vectors?



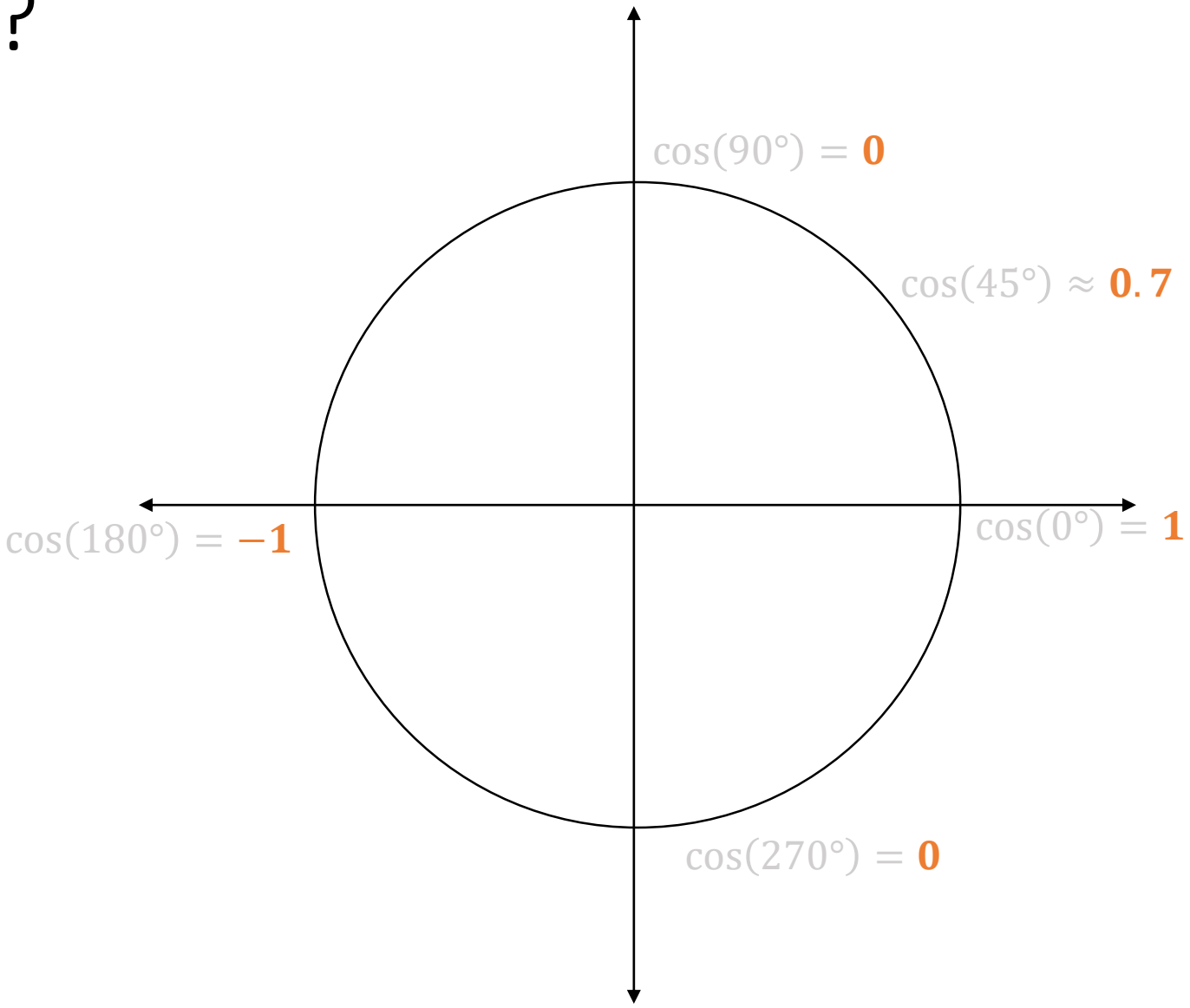
How do we measure similarity between vectors?



How do we measure similarity between vectors?



How do we measure similarity between vectors?



How do we measure similarity between vectors?

Use the cosine similarity between vectors

What if we want more information?

What if we want more information?

Vectors in 2 dimensions

3	2.5	2
1	1	-3

Vectors in 6 dimensions

3	2.5
1	1
-2	0.5
4	-2
1	9
0.7	1

What if we want more information?

Vectors in 6 dimensions

3	2.5
1	1
-2	0.5
4	-2
1	9
0.7	1

$$\text{cosine similarity (a, b)} = \frac{a \cdot b}{||a|| \cdot ||b||}$$

How do we create word vectors?

How do we know if words are similar?

How do we know if words are similar?

PRINCETON UNIVERSITY

WordNet

A Lexical Database for English

- What is WordNet
- People
- News
- Use Wordnet Online
- Download
- Citing WordNet
- License and Commercial Use
- Related Projects
- Documentation
- Publications

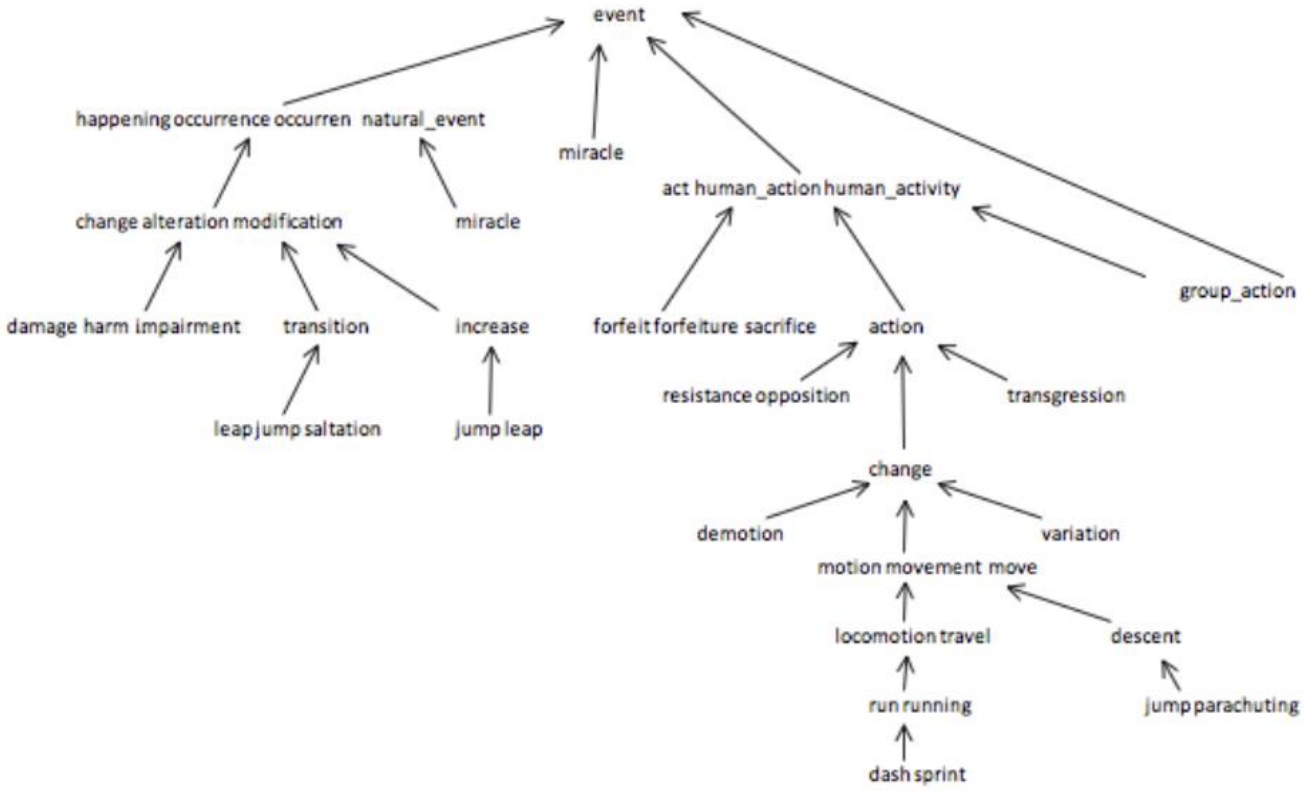
What is WordNet?

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the creators of WordNet and do not necessarily reflect the views of any funding agency or Princeton University.

When writing a paper or producing a software application, tool, or interface based on WordNet, it is necessary to properly cite the source. Citation figures are critical to WordNet funding.

About WordNet

WordNet® is a large lexical database of English. Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept. Synsets are interlinked by means of conceptual-semantic and lexical relations. The resulting network of meaningfully related words and concepts can be navigated with the browser. WordNet is also freely and publicly available for download. WordNet's structure makes it a useful tool for computational linguistics and natural language processing.



https://wordnet.princeton.edu/
https://www.cs.princeton.edu/courses/archive/fall12/cos226/assignments/wordnet.html

How do we know if words are similar?

Natural language processing

From Wikipedia, the free encyclopedia

This article is about language processing by computers. For the processing of language by the human brain, see Language processing in the brain.

Natural language processing (**NLP**) is an area of computer science and artificial intelligence concerned with the interactions between computers and human (natural) languages, in particular how to program computers to process and analyze large amounts of natural language data.

Challenges in natural language processing frequently involve speech recognition, natural language understanding, and natural language generation.

How do we know if words are similar?

Natural language processing

From Wikipedia, the free encyclopedia

This article is about language processing by computers. For the processing of language by the human brain, see Language processing in the brain.

Natural language processing (**NLP**) is an area of computer science and artificial intelligence concerned with the interactions between computers and human (natural) languages, in particular how to program computers to process and analyze large amounts of natural language data.

Challenges in natural language processing frequently involve speech recognition, natural language understanding, and natural language generation.

Distributional Hypothesis

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

- John Rupert Firth (1957)

How do we find the distribution of nearby words?

1. Count cooccurrence within a window of size n
2. Create a vector cooccurrences

How do we find the distribution of nearby words?

1. Count cooccurrence within a window of size n
2. Create a vector cooccurrences

Any problems with this?

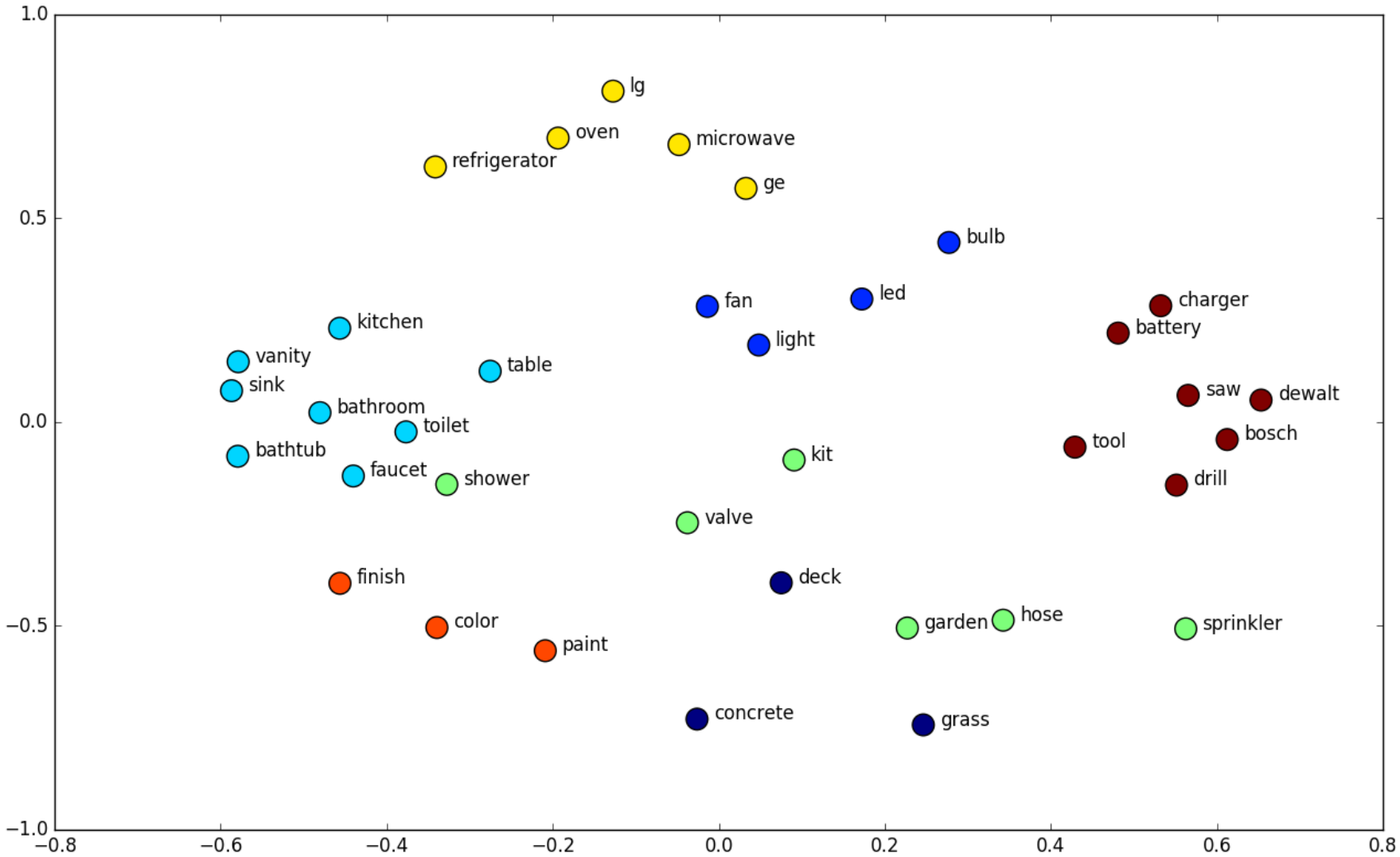
Some words are more informative than others…

Pointwise Mutual Information (PMI)

$$PMI(a,b) = \log \frac{P(a,b)}{P(a)P(b)}$$

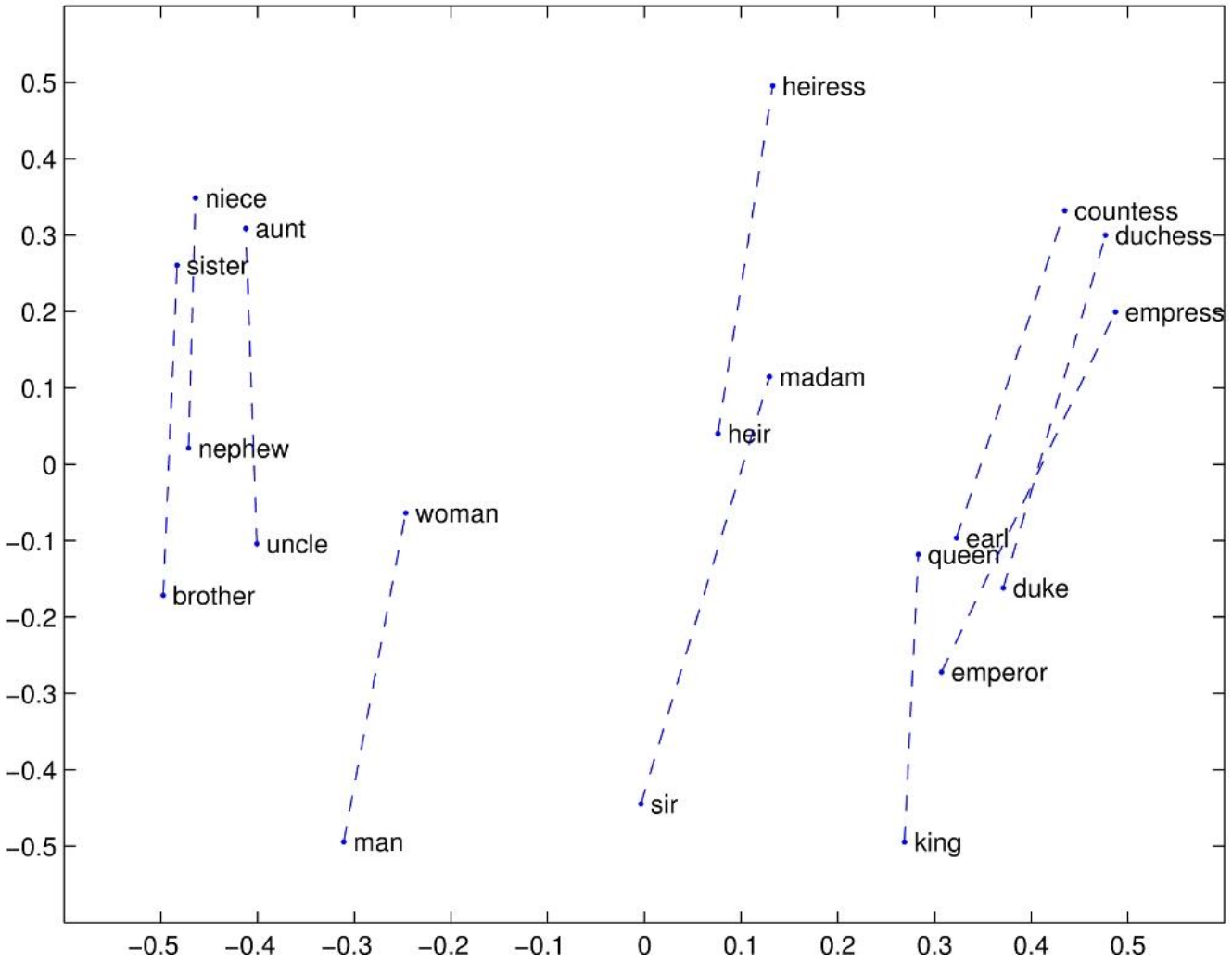
How can we represent sentences given word embeddings?

How can we use word vectors?



https://towardsdatascience.com/deep-learning-structured-data-8d6a278f3088

http://ruder.io/word-embeddings-2017/



https://nlp.stanford.edu/projects/glove/images/man_woman.jpg

Man is to Computer Programmer

Tolga Bolukbasi[1], **Kai-Wei Chang**[2], **James Zou**[2], **Venkatesh Saligrama**[1,2], **Adam Kalai**[2]
[1]Boston University, 8 Saint Mary's Street, Boston, MA
[2]Microsoft Research New England, 1 Memorial Drive, Cambridge, MA
tolgab@bu.edu, kw@kwchang.net, jamesyzou@gmail.com, srv@bu.edu, adam.kalai@microsoft.com

$$\overrightarrow{\text{man}} - \overrightarrow{\text{woman}} \approx \overrightarrow{\text{king}} - \overrightarrow{\text{queen}}$$

https://papers.nips.cc/paper/6228-man-is-to-computer-programmer-as-woman-is-to-homemaker-debiasing-word-embeddings.pdf

Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings

Tolga Bolukbasi[1], **Kai-Wei Chang**[2], **James Zou**[2], **Venkatesh Saligrama**[1,2], **Adam Kalai**[2]
[1]Boston University, 8 Saint Mary's Street, Boston, MA
[2]Microsoft Research New England, 1 Memorial Drive, Cambridge, MA
tolgab@bu.edu, kw@kwchang.net, jamesyzou@gmail.com, srv@bu.edu, adam.kalai@microsoft.com

$\overrightarrow{\text{man}} - \overrightarrow{\text{woman}} \approx \overrightarrow{\text{king}} - \overrightarrow{\text{queen}}$.

$\overrightarrow{\text{man}} - \overrightarrow{\text{woman}} \approx \overrightarrow{\text{computer programmer}} - \overrightarrow{\text{homemaker}}$.

https://papers.nips.cc/paper/6228-man-is-to-computer-programmer-as-woman-is-to-homemaker-debiasing-word-embeddings.pdf

Extreme *she*	Extreme *he*
1. homemaker	1. maestro
2. nurse	2. skipper
3. receptionist	3. protege
4. librarian	4. philosopher
5. socialite	5. captain
6. hairdresser	6. architect
7. nanny	7. financier
8. bookkeeper	8. warrior
9. stylist	9. broadcaster
10. housekeeper	10. magician

Gender stereotype *she-he* analogies

sewing-carpentry	registered nurse-physician	housewife-shopkeeper
nurse-surgeon	interior designer-architect	softball-baseball
blond-burly	feminism-conservatism	cosmetics-pharmaceuticals
giggle-chuckle	vocalist-guitarist	petite-lanky
sassy-snappy	diva-superstar	charming-affable
volleyball-football	cupcakes-pizzas	lovely-brilliant

Gender appropriate *she-he* analogies

queen-king	sister-brother	mother-father
waitress-waiter	ovarian cancer-prostate cancer	convent-monastery

https://papers.nips.cc/paper/6228-man-is-to-computer-programmer-as-woman-is-to-homemaker-debiasing-word-embeddings.pdf