

Word Vectors

Princeton AI4ALL: NLP Group

What's Next

Overview of the Week

- Monday (today)

1-3: Introduction to Word Vectors

- Tuesday

1-3: Word Vectors + FNC

- Wednesday

1-3: Lexical Overlaps

- Thursday

9:30-12: Putting together the FNC + Discussion

1-3: Minipresentations (10 minutes per person) + Free time to work on presentation.

- Friday

9:30-10:30: Free time to finish/practice presentation.

10:30-12, 1-3: Presentations with other groups.

Info about NLP & Ethics/Policy minipresentations

- Find a topic related to NLP and ethics/policy (can be a paper, news article, general topic of debate).
- Prepare either a presentation with slides, or a few questions for debate/discussion.
- Lead a mini-session about this topic!
- We can do sign-ups to avoid overlaps.

Info about presentations

Friday presentation sessions: 20 min to educate other campers about their project + 10 min for questions

Audience: other campers

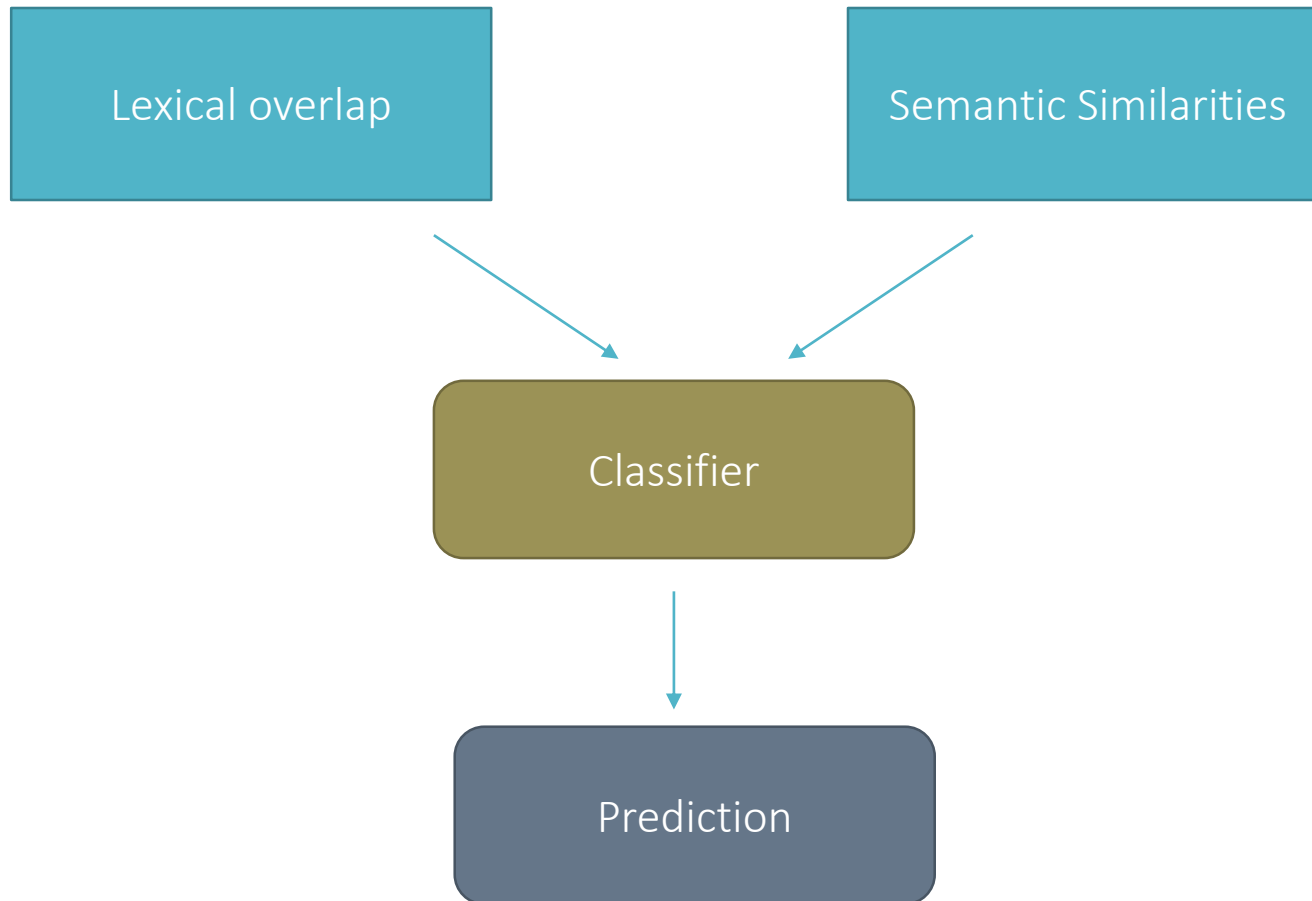
Visuals: Poster/slides/handouts/something creative

Friday banquet: 5 min to give a spotlight about project.

Audience: visitors

No visuals

Overview of the FNC Classifier



How should we
represent words?

To communicate this information to
computers...

To communicate this information to
computers...

quantitatively represent words,

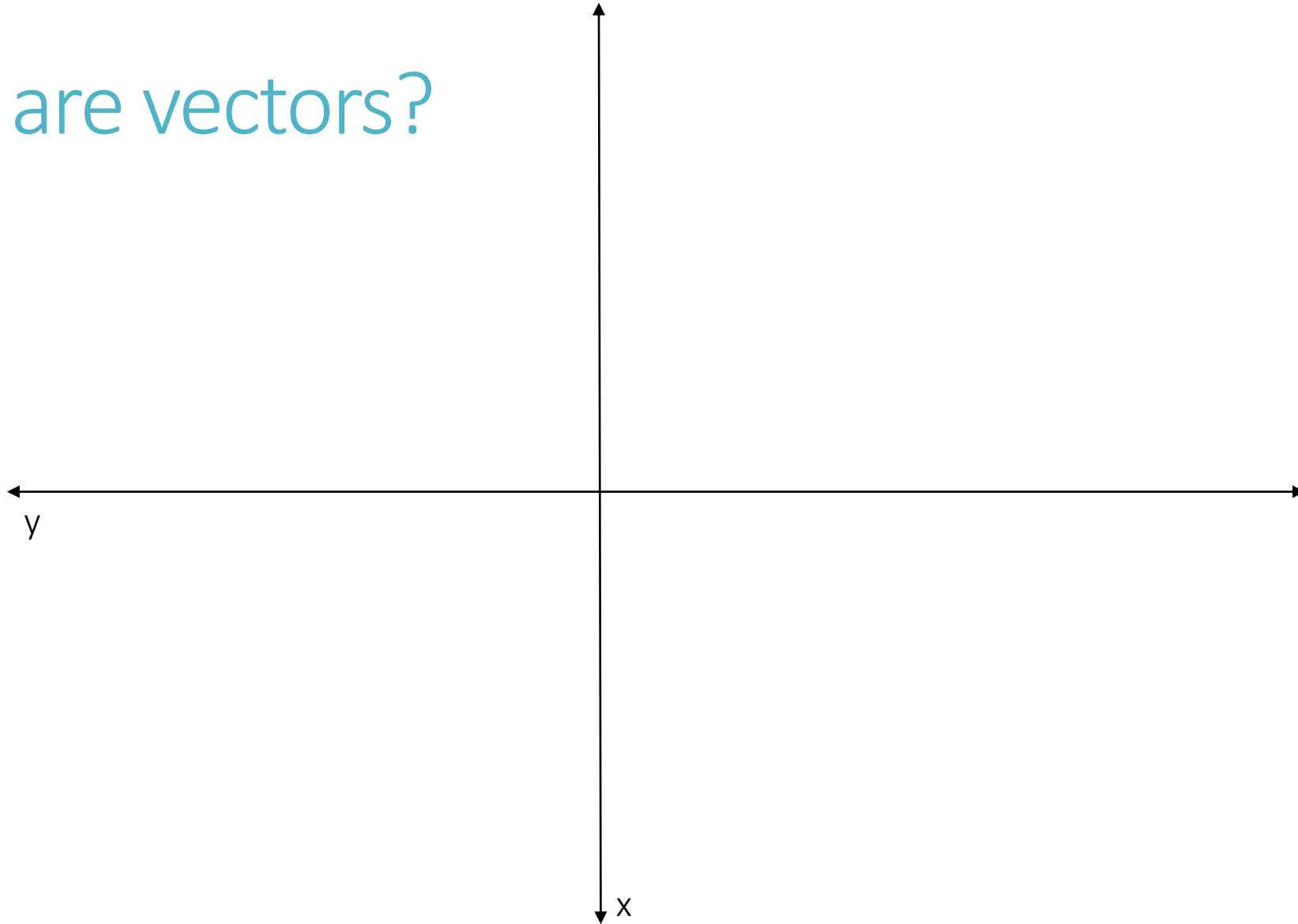
To communicate this information to
computers...

quantitatively represent words,

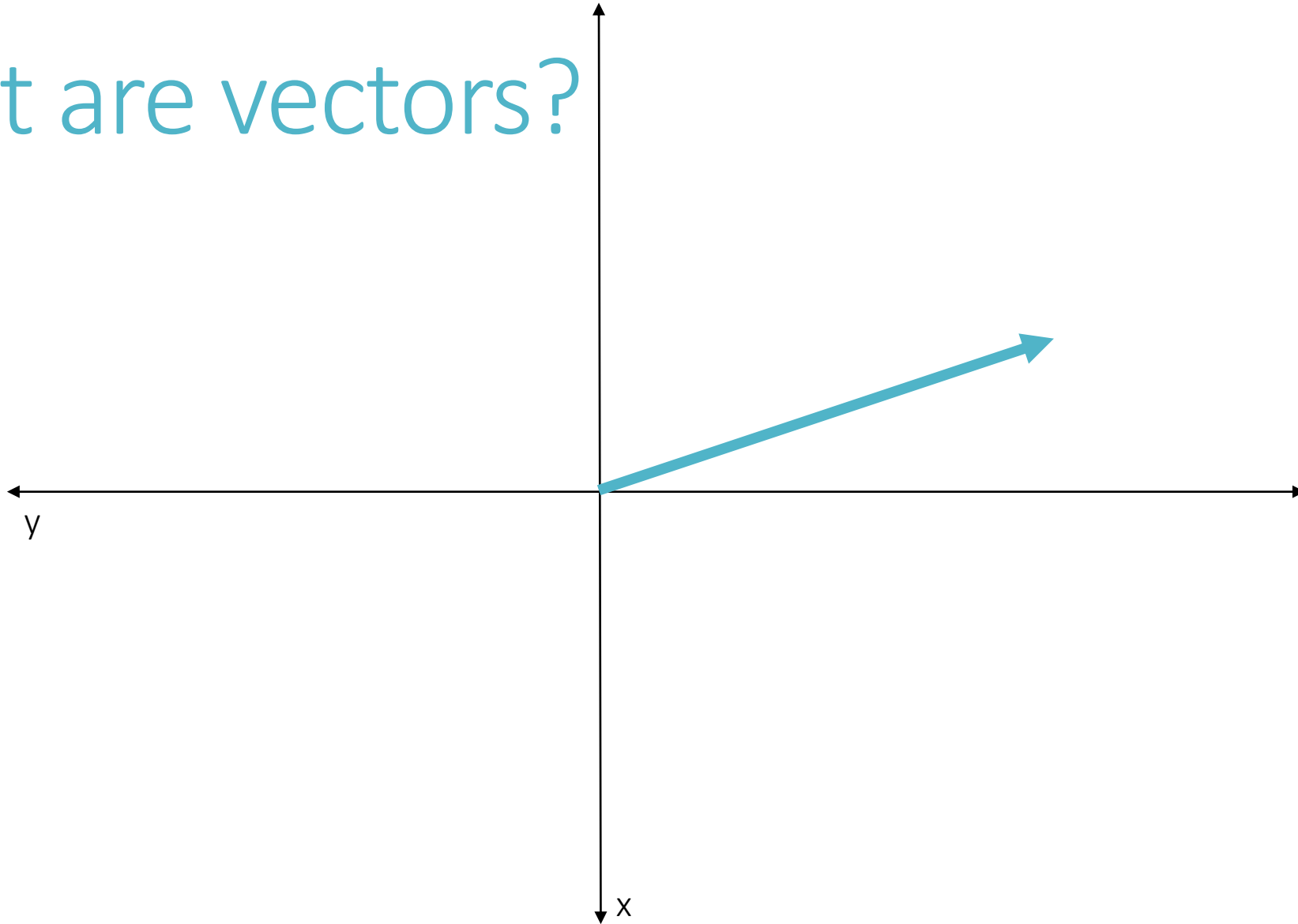
where more similar words have similar
representations

What are word vectors?

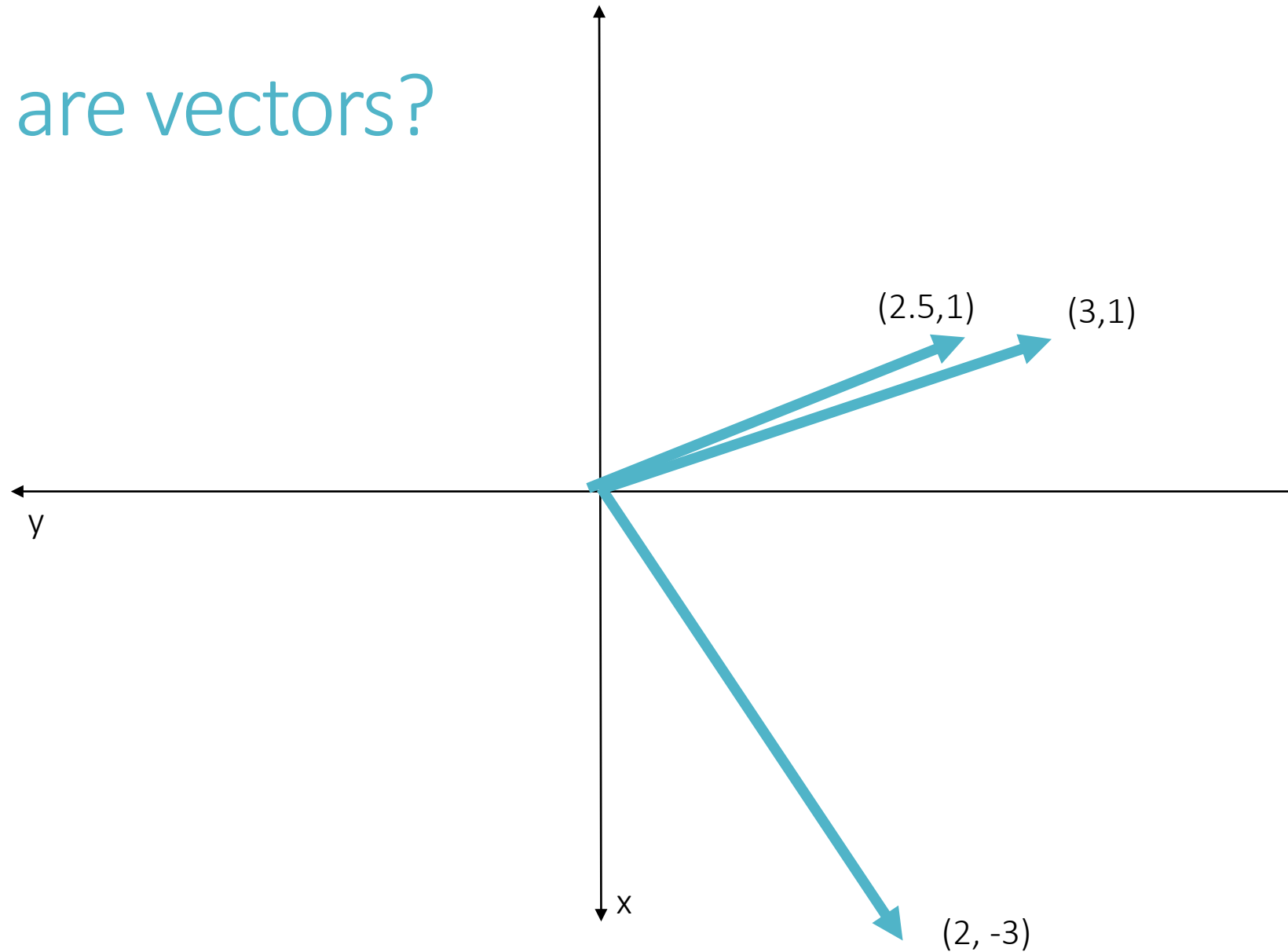
What are vectors?



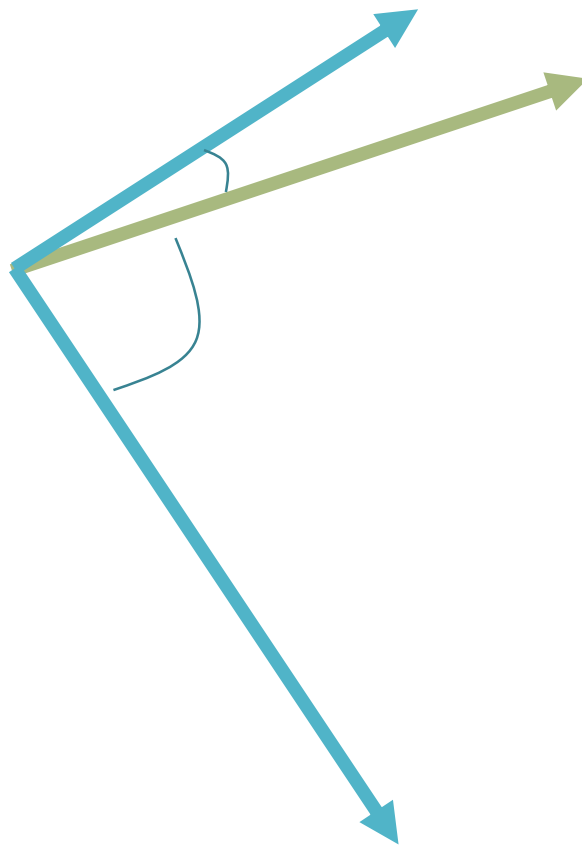
What are vectors?



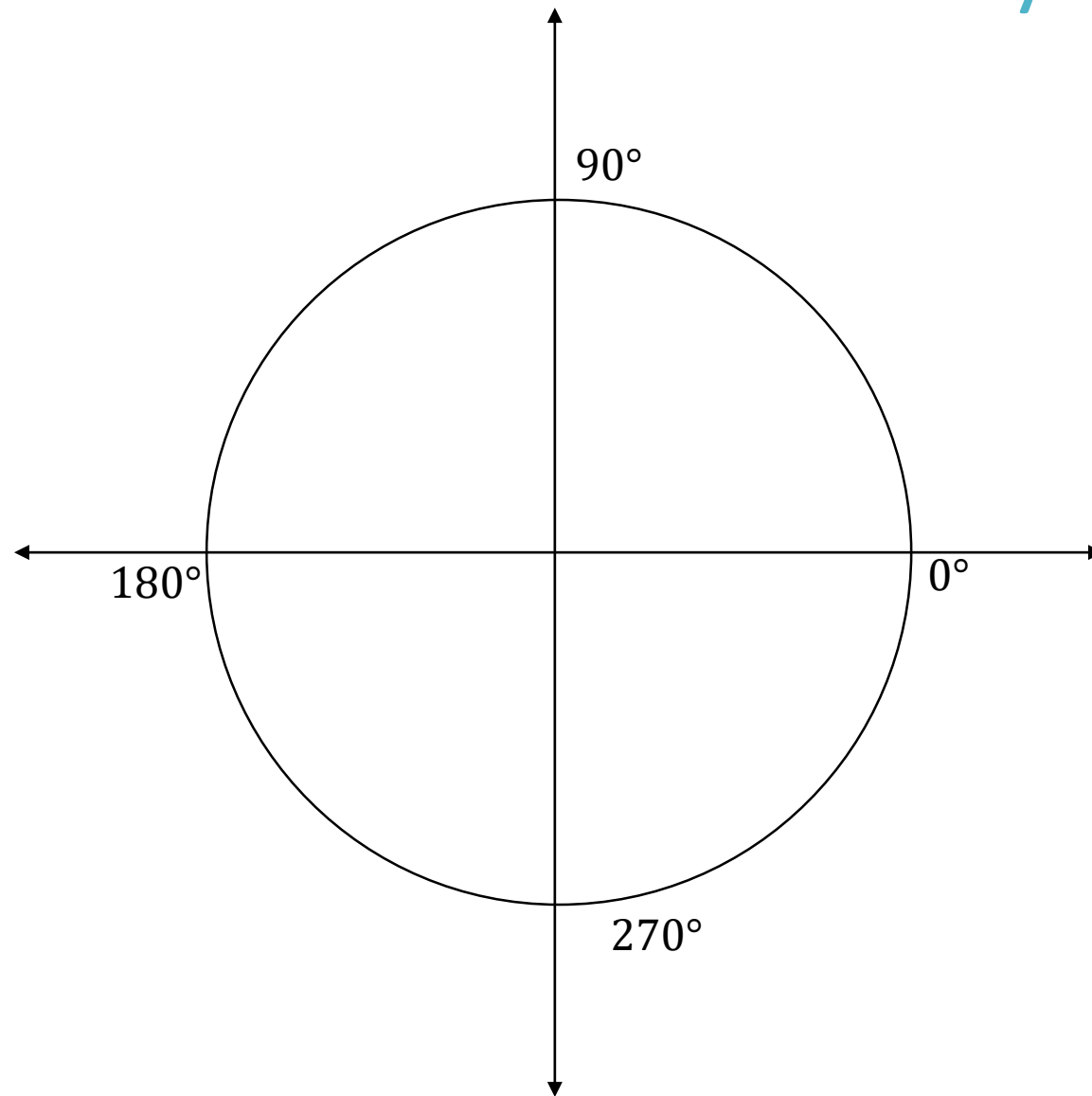
What are vectors?



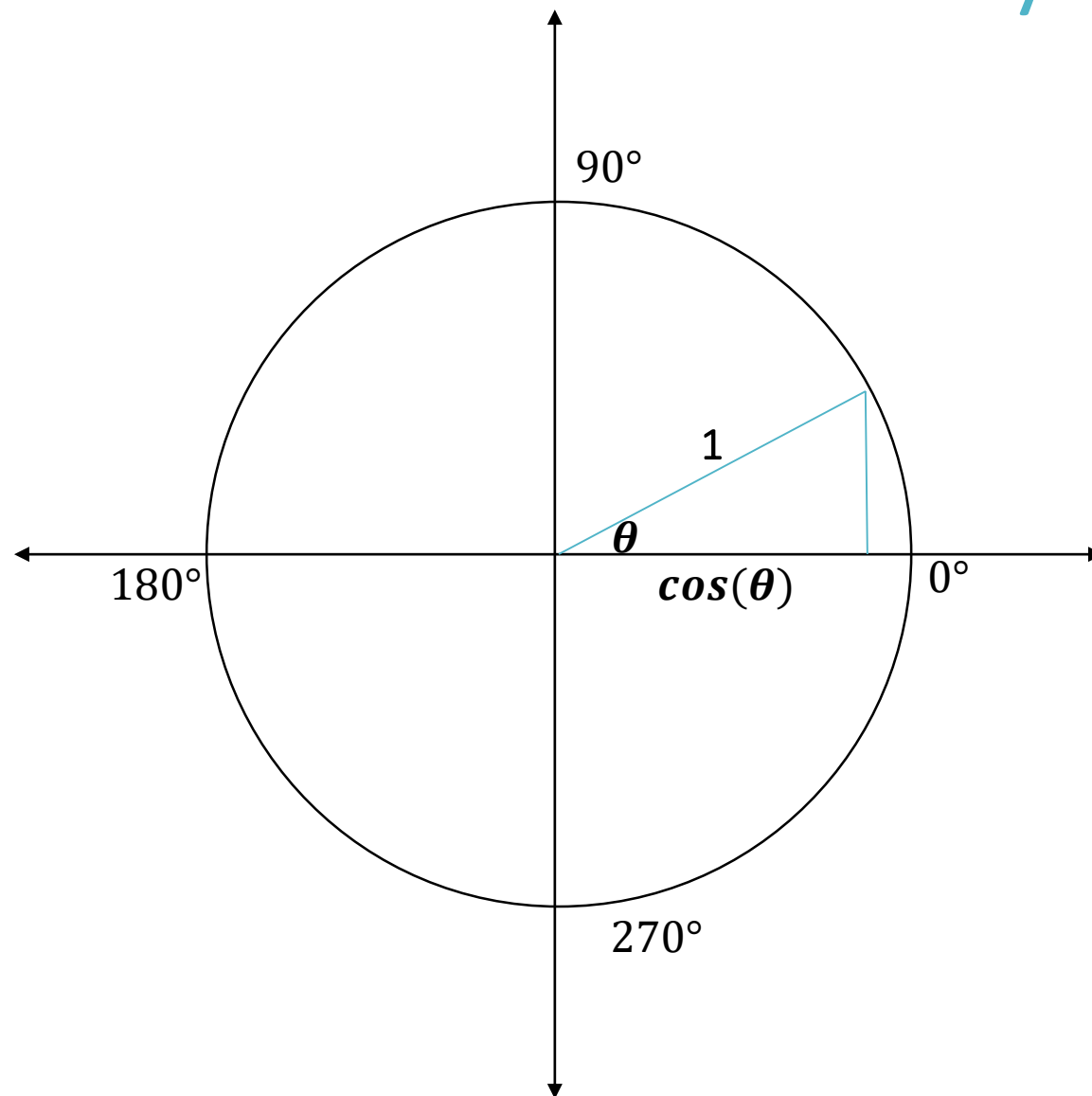
How do we measure similarity between vectors?



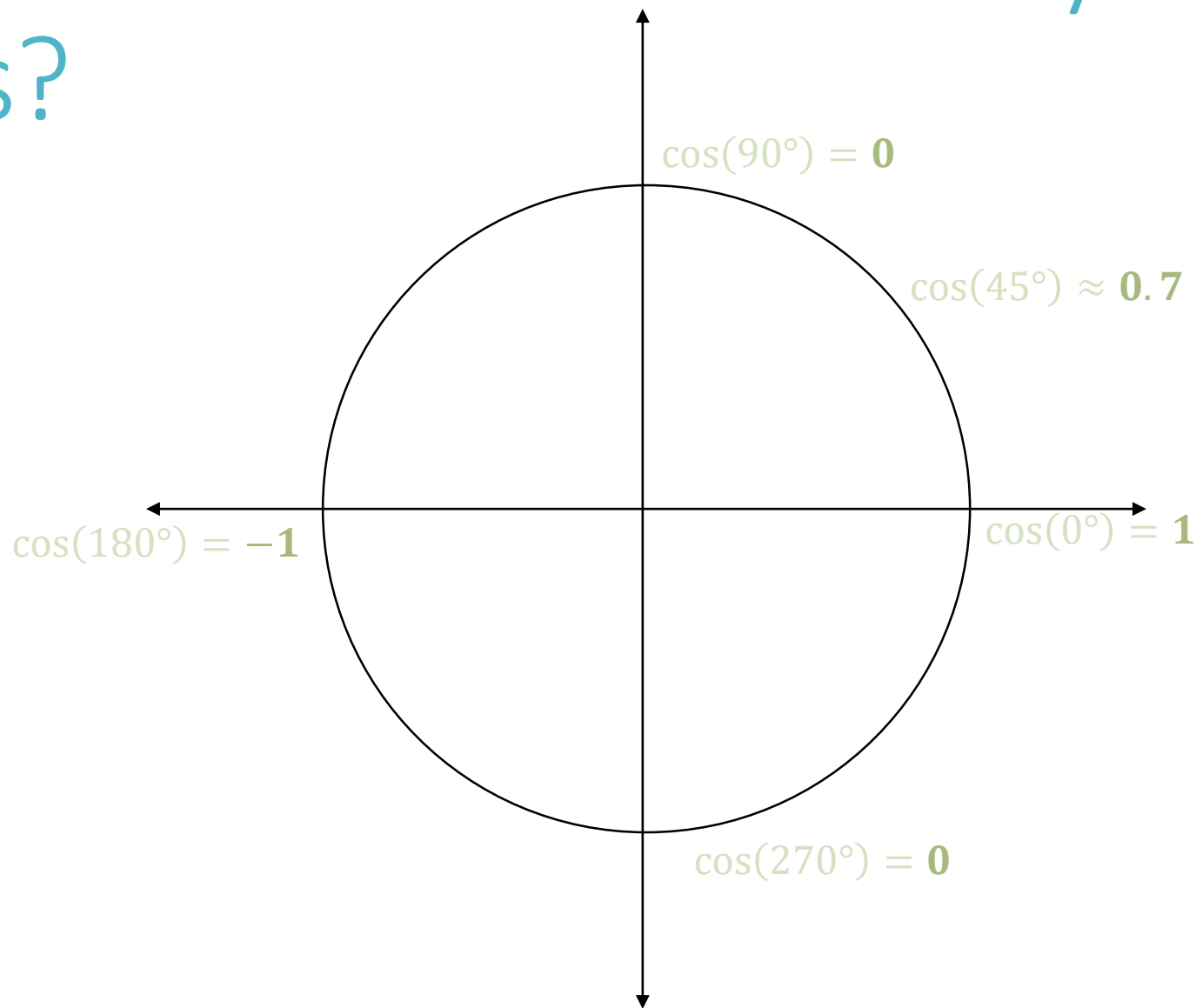
How do we measure similarity between vectors?



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?

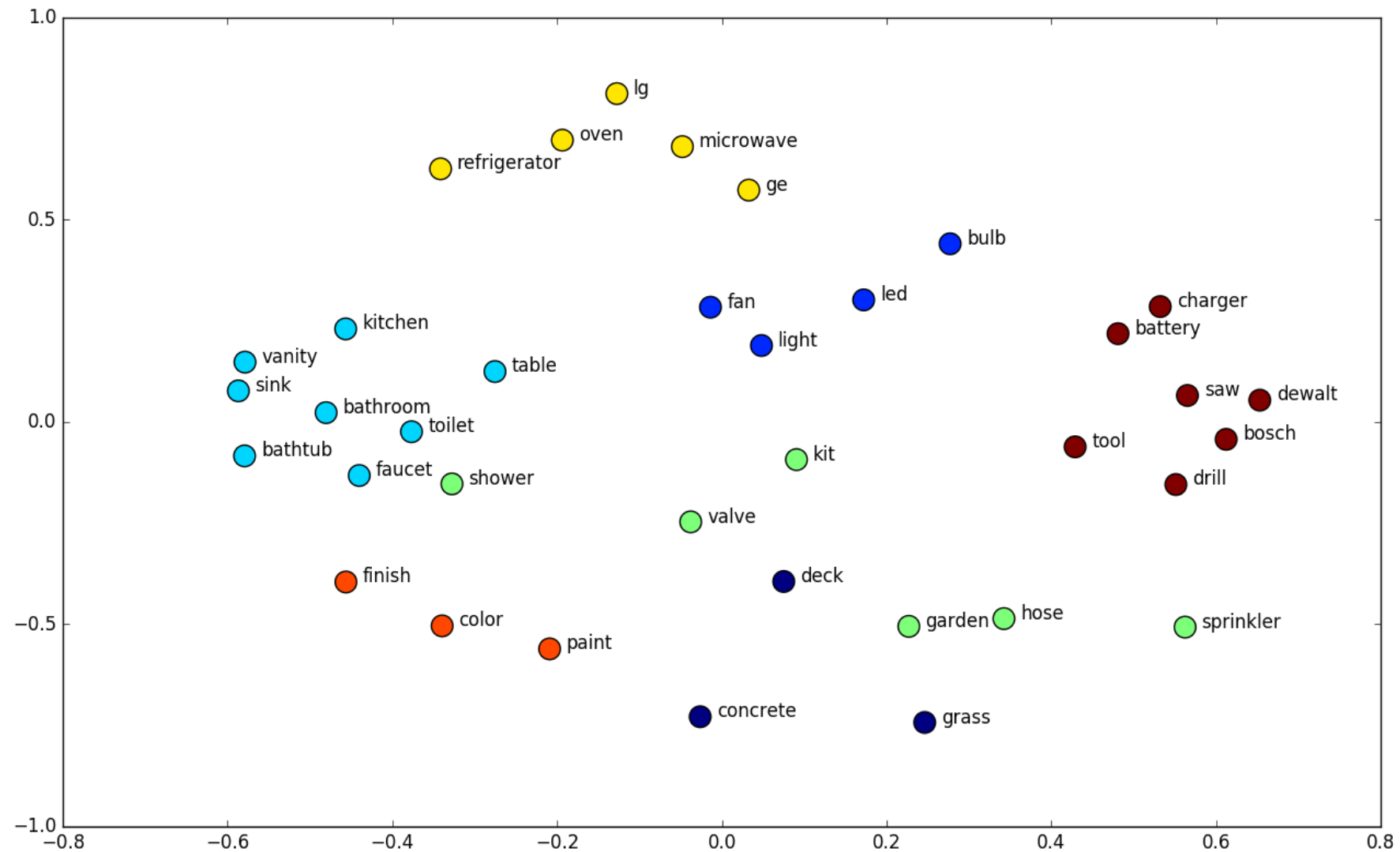
Some words are more informative than others...

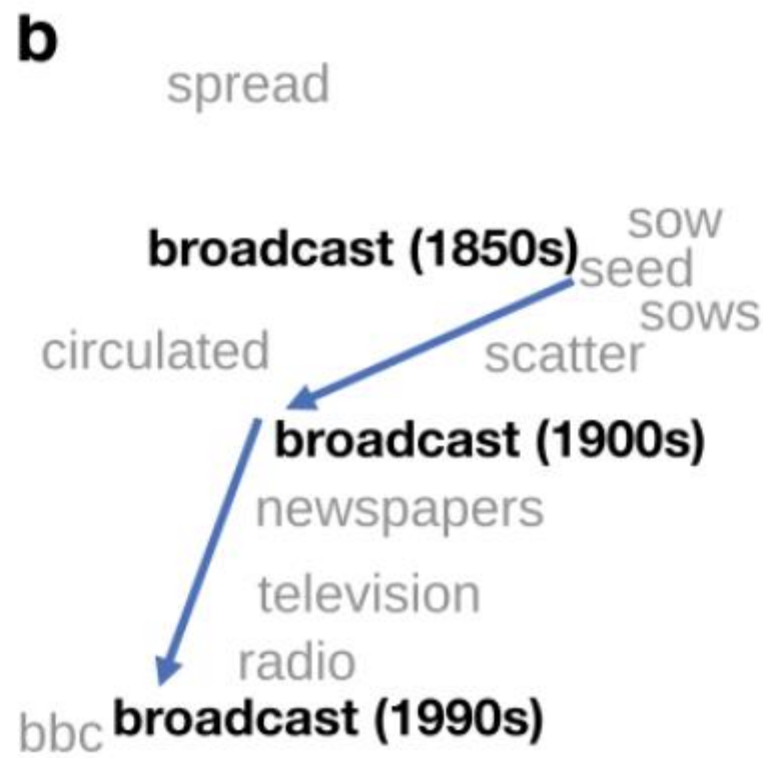
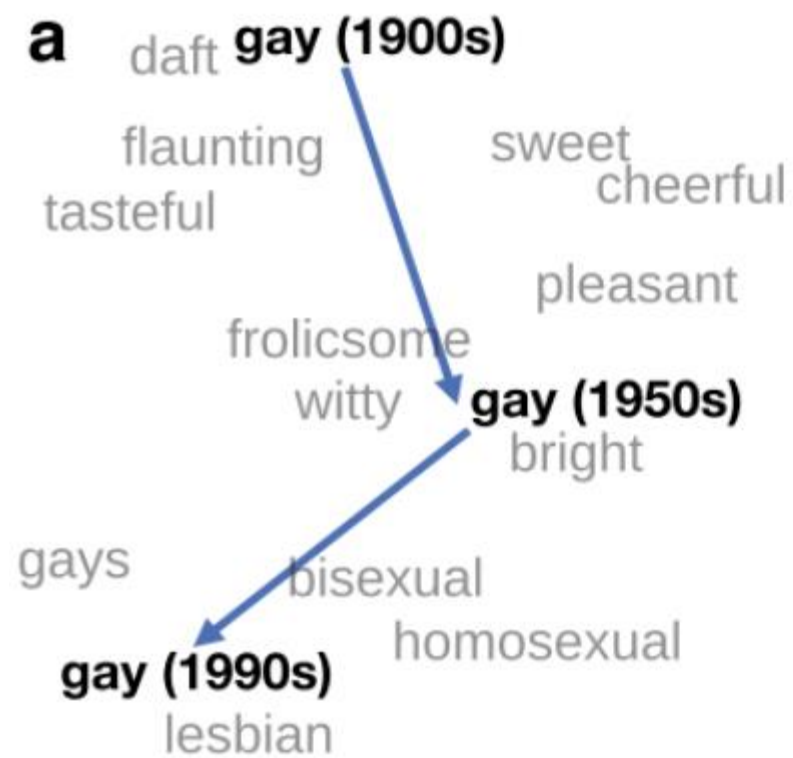
Pointwise Mutual Information (PMI)

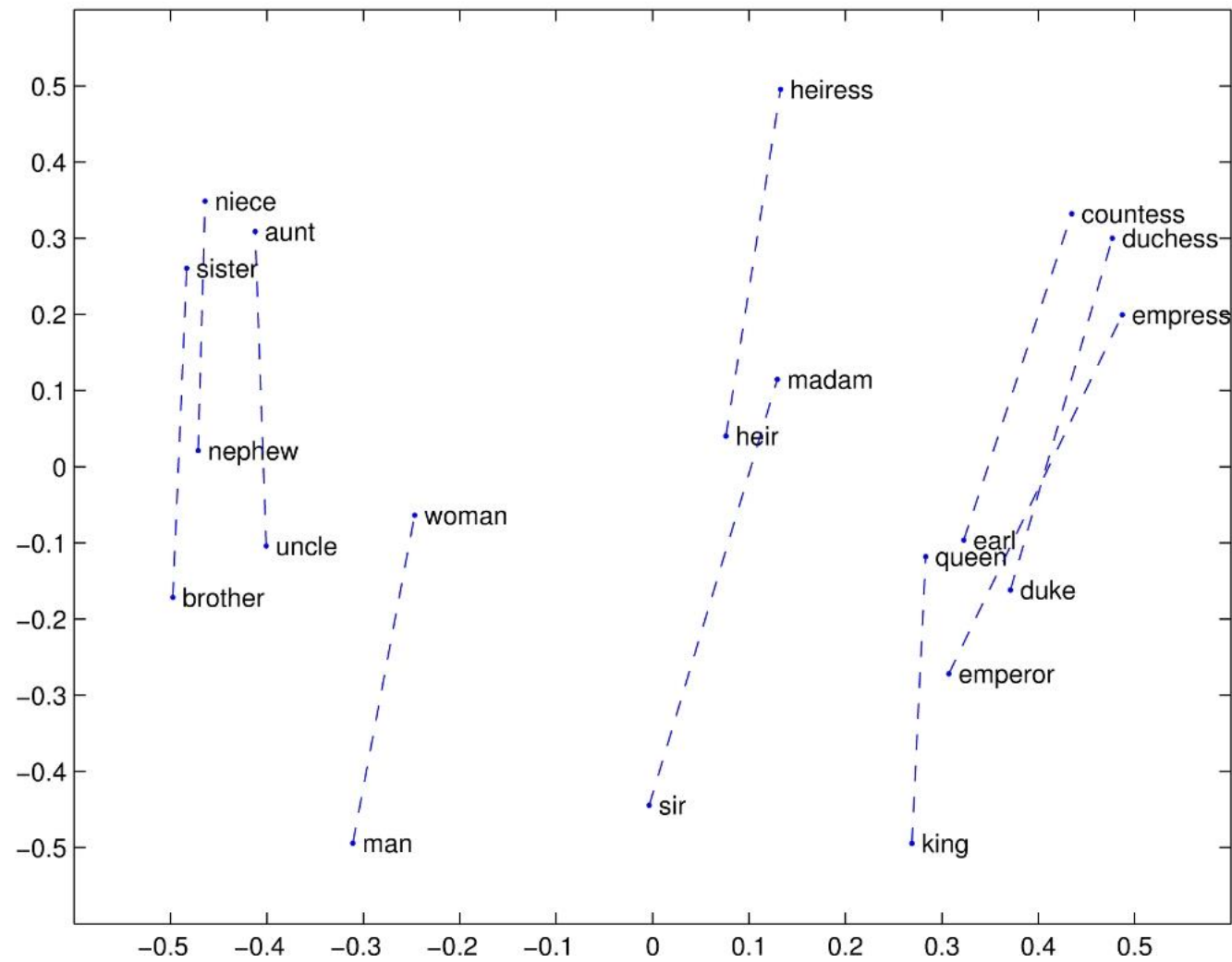
$$PMI(a, b) = \log \frac{P(a, b)}{P(a)P(b)}$$

How can we represent sentences given word embeddings?

How can we use word
vectors?







Man is to Computer Programmer

Tolga Bolukbasi¹, Kai-Wei Chang², James Zou², Venkatesh Saligrama^{1,2}, Adam Kalai²

¹Boston University, 8 Saint Mary's Street, Boston, MA

²Microsoft Research New England, 1 Memorial Drive, Cambridge, MA

tolgab@bu.edu, kw@kwchang.net, jamesyzou@gmail.com, srv@bu.edu, adam.kalai@microsoft.com

$$\vec{\text{man}} - \vec{\text{woman}} \approx \vec{\text{king}} - \vec{\text{queen}},$$

<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 Bolukbasi¹, Kai-Wei Chang², James Zou², Venkatesh Saligrama^{1,2}, Adam Kalai²

¹Boston University, 8 Saint Mary's Street, Boston, MA

²Microsoft Research New England, 1 Memorial Drive, Cambridge, MA

tolgab@bu.edu, kw@kwchang.net, jamesyzou@gmail.com, srv@bu.edu, adam.kalai@microsoft.com

$$\vec{\text{man}} - \vec{\text{woman}} \approx \vec{\text{king}} - \vec{\text{queen}},$$

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

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

| Extreme <i>she</i> | Extreme <i>he</i> | | | |
|---------------------------|--------------------------|--|---|---------------------------|
| 1. homemaker | 1. maestro | sewing-carpentry nurse-surgeon blond-burly giggle-chuckle sassy-snappy volleyball-football queen-king waitress-waiter | Gender stereotype <i>she-he</i> analogies | |
| 2. nurse | 2. skipper | | registered nurse-physician | housewife-shopkeeper |
| 3. receptionist | 3. protege | | interior designer-architect | softball-baseball |
| 4. librarian | 4. philosopher | | feminism-conservatism | cosmetics-pharmaceuticals |
| 5. socialite | 5. captain | | vocalist-guitarist | petite-lanky |
| 6. hairdresser | 6. architect | | diva-superstar | charming-affable |
| 7. nanny | 7. financier | | cupcakes-pizzas | lovely-brilliant |
| 8. bookkeeper | 8. warrior | | Gender appropriate <i>she-he</i> analogies | |
| 9. stylist | 9. broadcaster | | sister-brother | mother-father |
| 10. housekeeper | 10. magician | | 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>

[Github.com/cchen23/AI4ALL_NLP_Student](https://github.com/cchen23/AI4ALL_NLP_Student)

Day10_WordVectors.zip

source ~/miniconda3/bin/activate

jupyter notebook

Poll for tomorrow's review:
<http://bit.ly/2ANGceX>

Mini-presentation sign-ups

<http://bit.ly/2M6dJ8K>