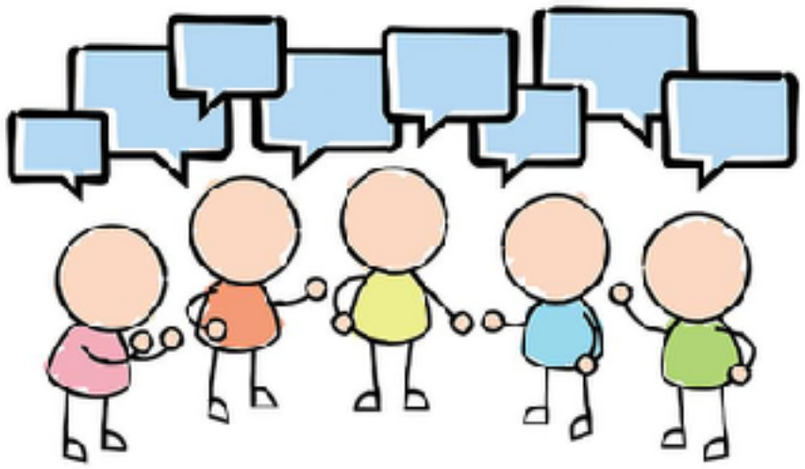
Introduction to Natural Language Processing

Laura Wendlandt
wenlaura@umich.edu

Ideas borrowed liberally from Reed Coke's slides.

What is NLP?

- Computers understanding language



- Computers generating language

Why do I care?

- What do I want to find out?

  Do my customers trust my business?

  How do people feel about this cultural trend?

- What kind of data do I have?

  Social media data

  Website comments

Why do I care?

- Is there information in my data that will help me answer these questions?
- How do I extract it?
- Are there tools that do this for me?

Outline

- **Why is NLP Hard?**
- Twitter data
- Preparing data
- Dataset Statistics
- NLP Tasks
  - Named Entity Recognition (NER)
  - Sentiment Analysis
  - Topic Modeling
  - Word Embeddings

Why is NLP Hard?

- Language is complex!
  - **Ambiguity**: Children make delicious snacks.

Why is NLP Hard?

- Language is complex!
  - **Ambiguity**: I ate a chocolate bar.
    I walked into a chocolate bar.

Why is NLP Hard?

- Language is complex!
    - **Humor and Sarcasm**: "Some cause happiness wherever they go; others whenever they go." - Oscar Wilde

Outline

- Why is NLP Hard?
- **Twitter data**
- Preparing data
- Dataset Statistics
- NLP Tasks
  - Named Entity Recognition (NER)
  - Sentiment Analysis
  - Topic Modeling
  - Word Embeddings

Let’s look at some real data!

Here's a cute panda to make your day! :) http://t.co/jeVWqXIK1r http://t.co/DIL4YjCadQ"

I just watched a video about a girl being "allergic" to the sun :(that's depressing

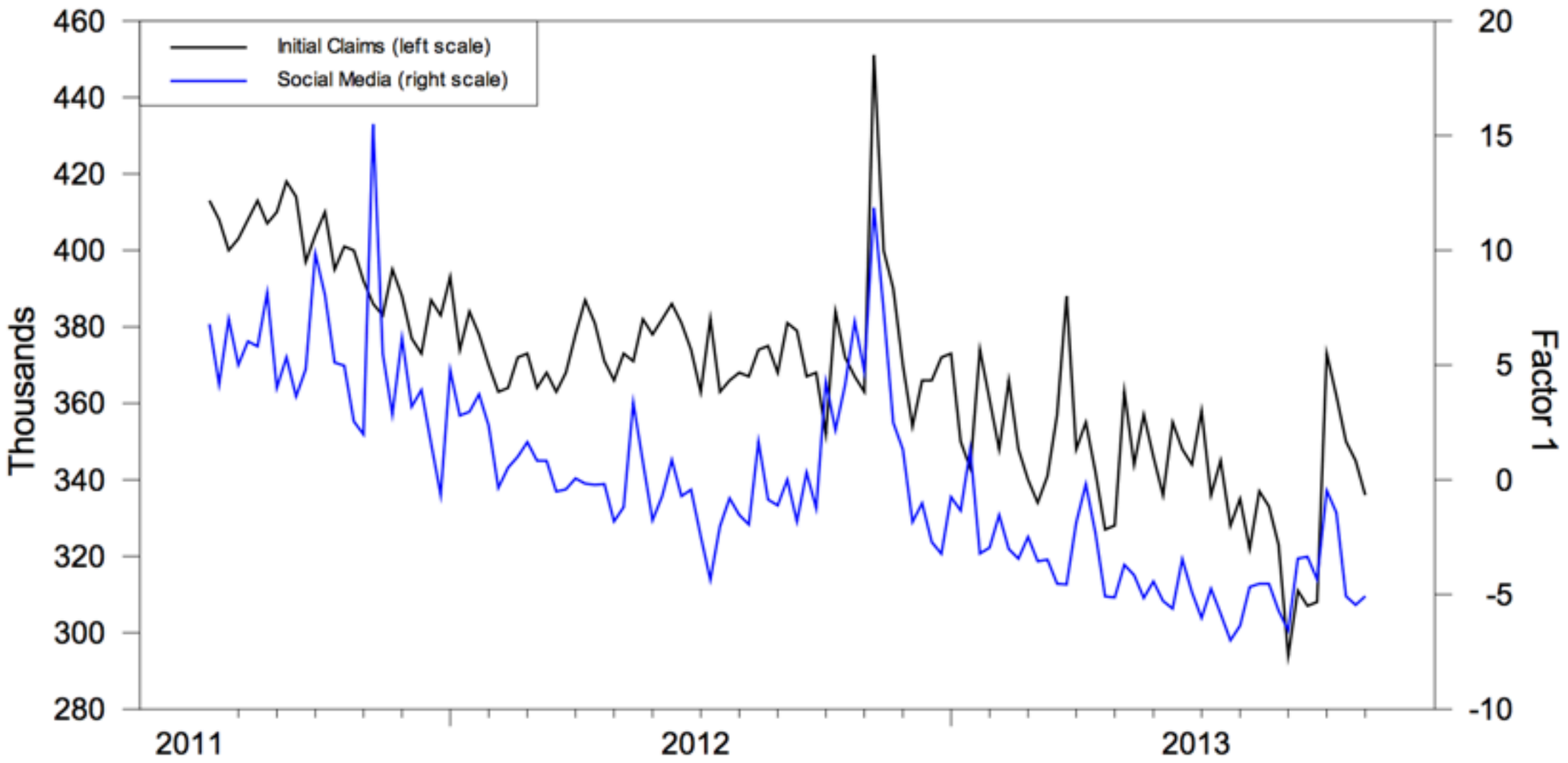
I WANT A WHITE FRENCH BULLDOG :(((

👑👑👑 》》》》 I LOVE YOU SO MUCH. I BELİEVE THAT HE WİLL FOLLOW. PLEASE FOLLOW ME PLEASE JUSTİN @justinbieber :(x15.337 》》》》 S E E M E 👑👑👑

RT @natalieben: #bbcqt - so would Miliband really let David Cameron back in rather than "do a deal" with the SNP?

What can you do with Twitter data?

- Predict unemployment



Antenucci, Dolan, et al. *Using social media to measure labor market flows*. No. w20010. National Bureau of Economic Research, 2014.

What can you do with Twitter data?

- Predict which tweet will be retweeted more

Tan, Chenhao, Lillian Lee, and Bo Pang. "The effect of wording on message propagation: Topic-and author-controlled natural experiments on Twitter." *ACL* (2014).

Outline

- Why is NLP Hard?
- Twitter data
- **Preparing data**
- Dataset Statistics
- NLP Tasks
  - Named Entity Recognition (NER)
  - Sentiment Analysis
  - Topic Modeling
  - Word Embeddings

Python Libraries

- **NLTK** = Natural Language Toolkit - good for preparing data
  - https://www.nltk.org/
  - Also, good tutorial book: http://www.nltk.org/book/
- **Gensim** - great for topic modeling and word embeddings
  - https://radimrehurek.com/gensim/
- **Stanford Core NLP** (not actually Python, but has Python wrappers available) - good for NER and sentiment analysis, among other things
  - https://nlp.stanford.edu/software/
  - Good Python wrapper: pycorenlp

Python Libraries

tldr;

```
pip install nltk

pip install gensim

pip install pycorenlp

#Go to https://nlp.stanford.edu/software/ and download NER and sentiment analysis packages
```

Preparing Data

- Data is messy!
- How can we clean it up?

	`Here's a cute panda to make your day! :) http://t.co/jeVWqXIK1r http://t.co/DIL4YjCadQ"`
Lowercase	`here's a cute panda to make your day! :) http://t.co/jevwqxik1r http://t.co/dil4yjcadq"`
Tokenize into words	`["here's",'a','cute','panda','to','make','your','day','!',':)','http://t.co/jevwqxik1r','http://t.co/dil4yjcadq']`
Remove links / rare words	`[“here’s",'a','cute','UNK','to','make','your','day','!',':)','LINK',LINK']`

Preparing Data

- Other data cleaning strategies (these depend on the scenario):
  - Tokenize into sentences (as well as tokenize into words)
  - Remove all punctuation
  - Remove digits (or replace digits with #)
  - Remove stop words (e.g., the, and, to, for)
  - Stem words

`run, running, runner` ⟶ `run`

Outline

- Why is NLP Hard?
- Twitter data
- Preparing data
- **Dataset Statistics**
- NLP Tasks
  - Named Entity Recognition (NER)
  - Sentiment Analysis
  - Topic Modeling
  - Word Embeddings

Dataset Statistics

- What are some ways that we can summarize such a big corpus of text?

Outline

- Why is NLP Hard?
- Twitter data
- Preparing data
- Dataset Statistics
- NLP Tasks
  - **Named Entity Recognition (NER)**
  - Sentiment Analysis
  - Topic Modeling
  - Word Embeddings

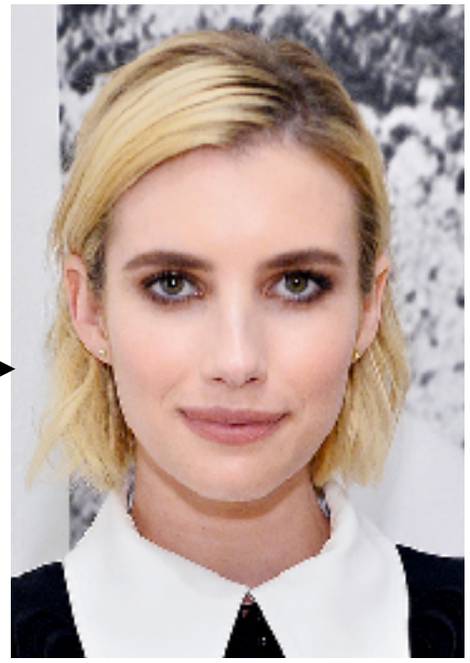
Named Entity Recognition (NER)

- Extract entities in the text
- PERSON, ORGANIZATION, LOCATION (time, date, monetary value, percentage)

```
[('rt', 'O'),
('@laboureoin', 'O'),
(':', 'O'),
('the', 'O'),
('economy', 'O'),
('was', 'O'),
('growing', 'O'),
('3', 'O'),
('times', 'O'),
('faster', 'O'),
('on', 'O'),
('the', 'O'),
('day', 'O'),
(‘david', 'PERSON'),
('cameron', 'PERSON'),
('became', 'O'),
('prime', 'O'),
('minister', 'O'),
('than', 'O'),
('it', 'O'),
('is', 'O'),
('today', 'O'),
('..', 'O'),
('#bbcqt', 'O'),
('LINK', 'O')]
```

Named Entity Recognition (NER)

jane
miss kang
michael woodford
bush
emma roberts
chris gayle
kath
jonah
jumma mubarak
lewis
miss dubai



Named Entity Recognition (NER)

- After identifying all of the entities, you may need to combine some
  - David, Cameron, David Cameron, Mr. Cameron
- There will always be some errors!

Outline

- Why is NLP Hard?
- Twitter data
- Preparing data
- Dataset Statistics
- NLP Tasks
  - Named Entity Recognition (NER)
  - **Sentiment Analysis**
  - Topic Modeling
  - Word Embeddings

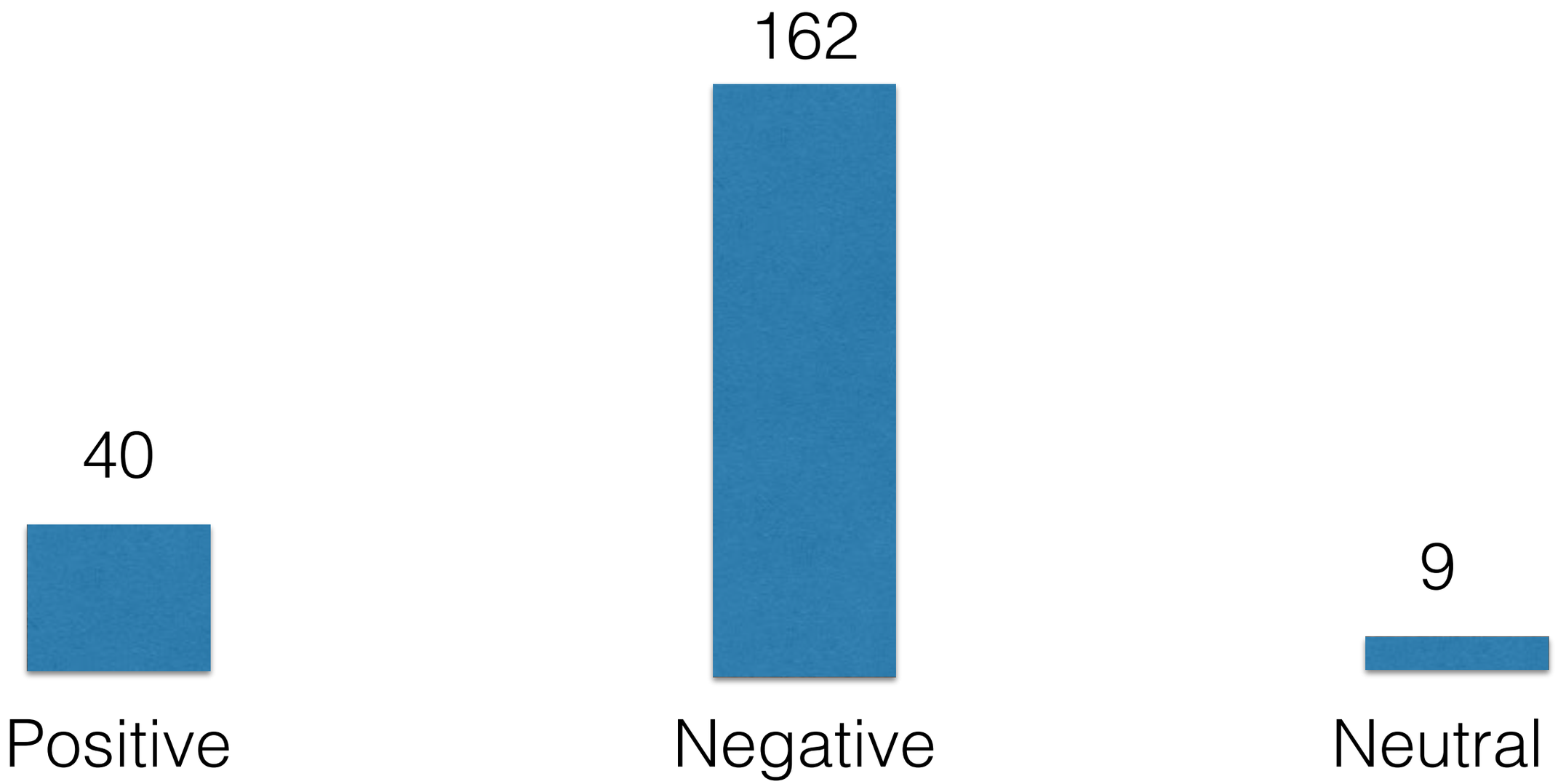
Sentiment Analysis

Sentiment Analysis

-	#followfriday @france_inte @pkuchly57 @milipol_paris for being top engaged members in my community this week :
-	@lamb2ja hey james !
:)	@despiteofficial we had a listen last night : as you bleed is an amazing track .'
:(	we do n't like to keep our lovely customers waiting for long !
:(	having boring time : do n't know what to do ...

Sentiment Analysis + NER!

Sentences including Justin Beiber



Sentiment Analysis

- Domain matters

She's a great athlete and she was not afraid to be aggressive.
This is a terrible restaurant. The wait staff were very aggressive.

Outline

- Why is NLP Hard?
- Twitter data
- Preparing data
- Dataset Statistics
- NLP Tasks
  - Named Entity Recognition (NER)
  - Sentiment Analysis
  - **Topic Modeling**
  - Word Embeddings

Topic Modeling

- Automatically identify topics in a document

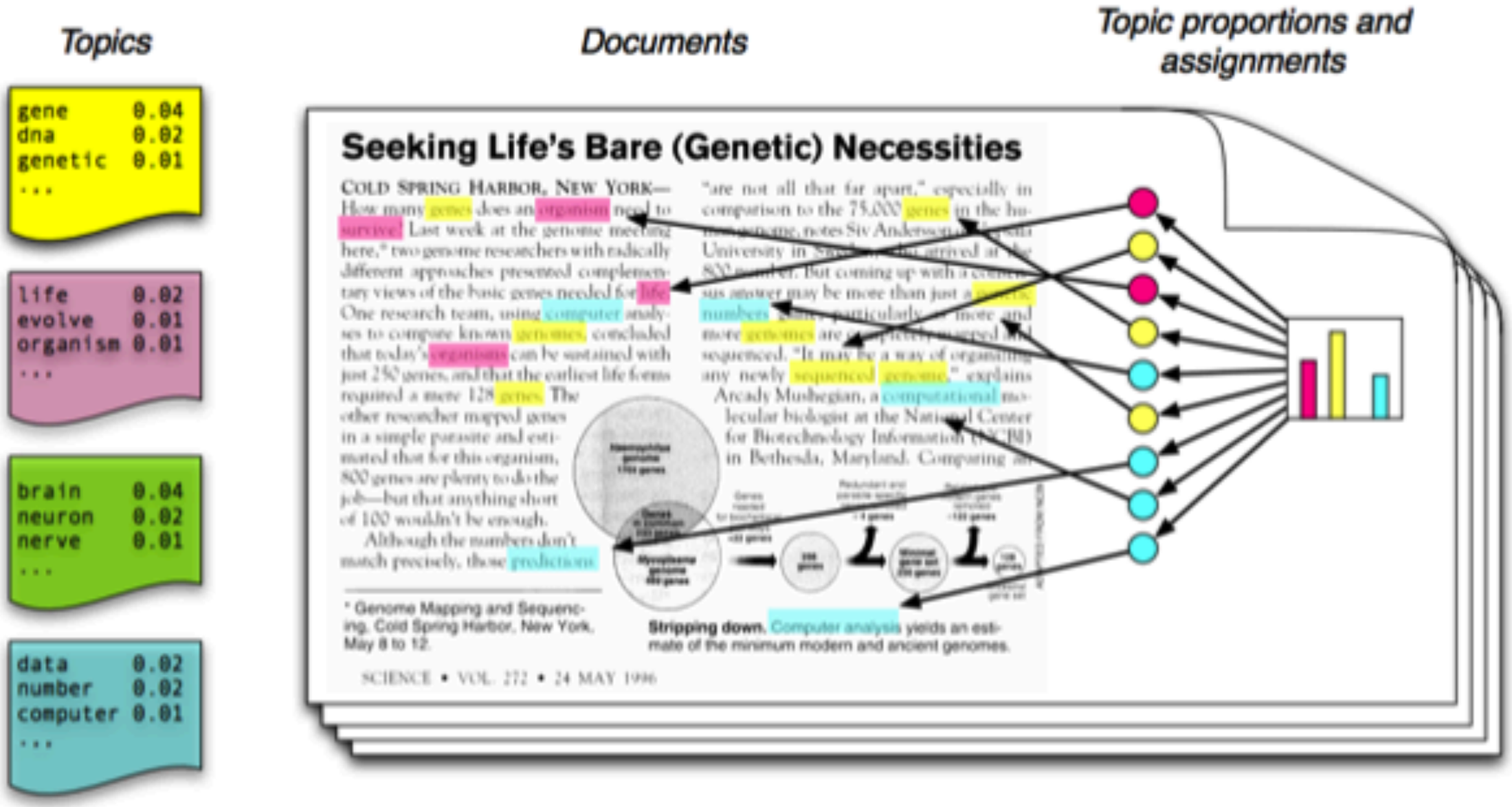


Figure source: Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77-84.

Topic Modeling

cant	im	miss	please	im
d	want	sad	follow	followed
like	like	much	u	get
today	kik	cant	》	go
feel	know	get	♛	thanks
want	snapchat	thank	back	still
see	thanks	good	love	3
okay	tired	u	justin bieber	see
im	pls	one	day	wanna
makes	sick	oh	sorry	please
2.9%	2.9%	88.4%	2.9%	3.0%

@joyster2012 @cathstaincliffe good for you, girl!! best wishes :-)

Topic Modeling

- LDA is a common topic modeling algorithm
- Good for exploration
- Sometimes topics are hard to interpret
- The topic model depends heavily on the number of topics you choose

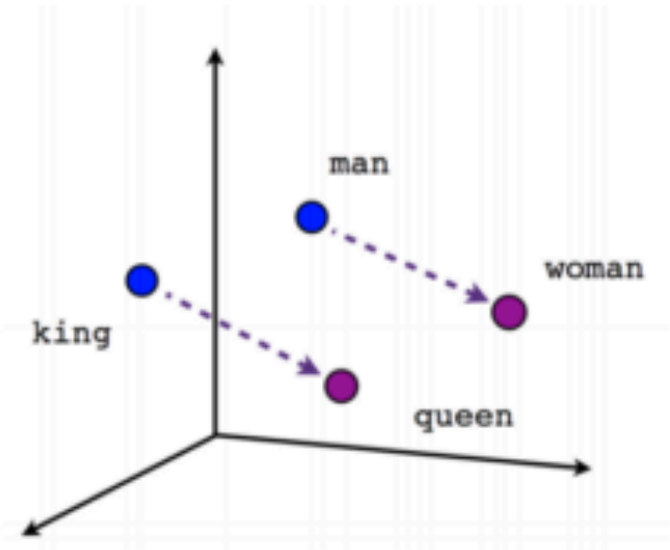
Outline

- Why is NLP Hard?
- Twitter data
- Preparing data
- Dataset Statistics
- NLP Tasks
  - Named Entity Recognition (NER)
  - Sentiment Analysis
  - Topic Modeling
  - **Word Embeddings**

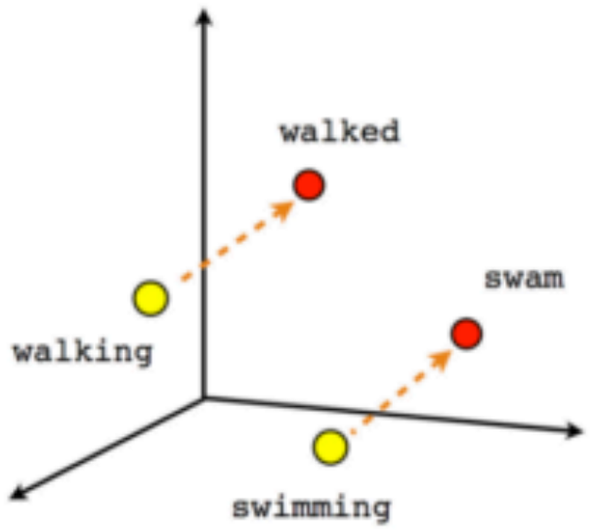
Word Embeddings

- **Word representations** that try to capture some of the meaning of the word

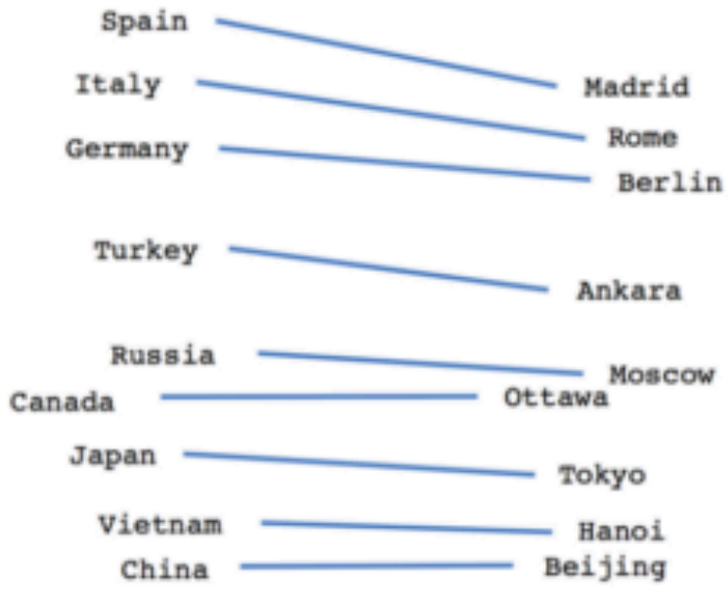
- You can think of them as high-dimensional points for each word (usually dimension = 300)



Male-Female



Verb tense



Country-Capital

Word Embeddings

- What can you do with word embeddings?

1. Use them as features in a machine learning algorithm (e.g., classification, regression)

2. Calculate the similarity between two words

3. Find similar words

Word Embeddings

- A flexible way to represent the meaning of words!
- If you have enough data, you can train your own (see Gensim's word2vec)
- If not, you can download pre-trained word embeddings
- https://www.quora.com/Where-can-I-find-some-pre-trained-word-vectors-for-natural-language-processing-understanding