Carnegie Mellon University
Heinz College

95-865
Unstructured Data Analytics
Lecture 1: Course Overview, Basic Text Analysis, Co-occurrence Analysis

George Chen

Big Data

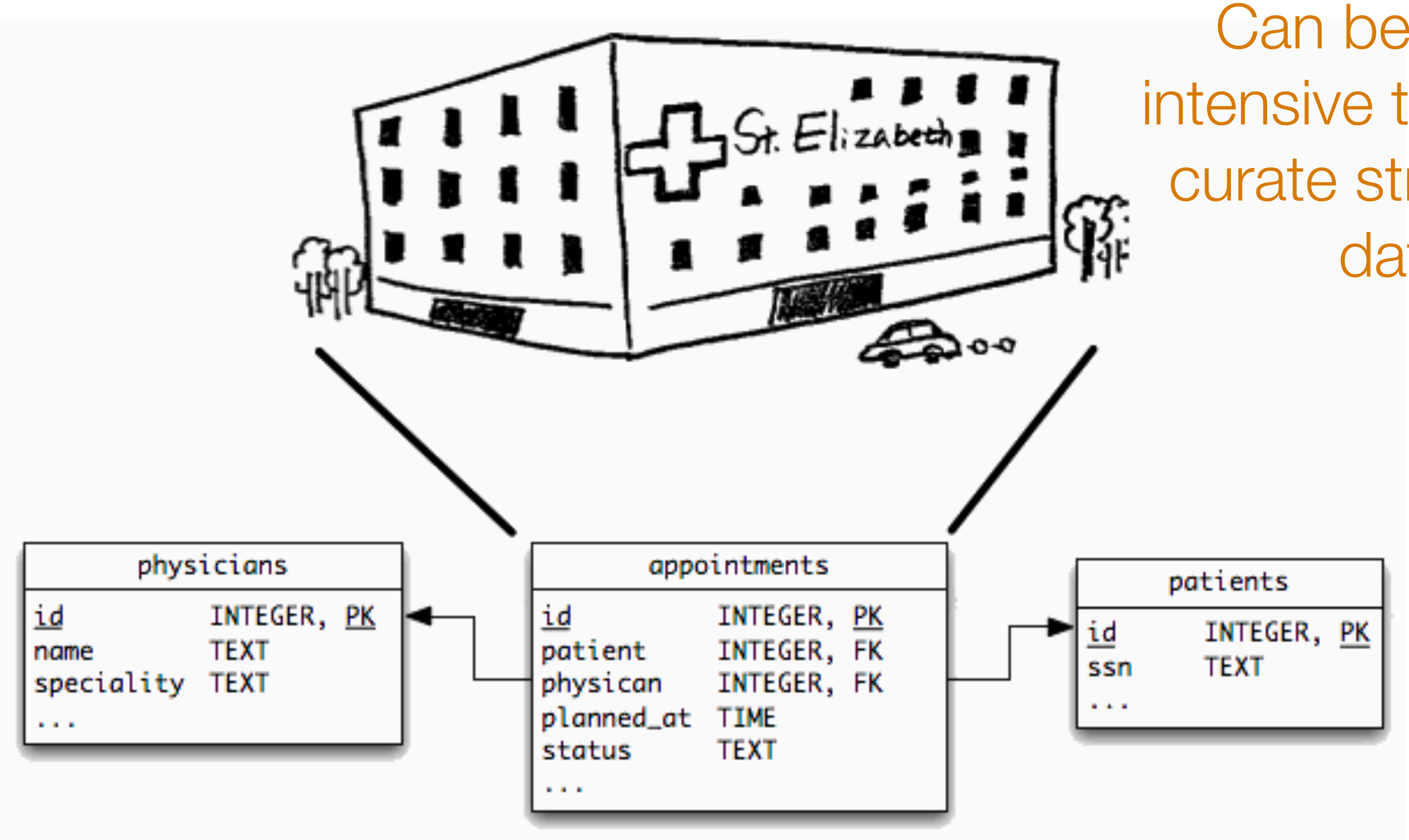
We’re now collecting data on virtually every human endeavor

How do we turn these data into actionable insights?

Two Types of Data

Structured Data

Well-defined elements, relationships between elements



physicians	
id	INTEGER, PK
name	TEXT
speciality	TEXT
...	

appointments	
id	INTEGER, PK
patient	INTEGER, FK
physican	INTEGER, FK
planned_at	TIME
status	TEXT
...	

patients	
id	INTEGER, PK
ssn	TEXT
...	

Can be labor-intensive to collect/curate structured data

Image source: http://revision-zero.org/images/logical_data_independence/hospital_appointments.gif

Unstructured Data

No pre-defined model—elements and relationships ambiguous

Examples:

- Text
- Images
- Videos
- Audio

Often: Want to use heterogeneous data to make decisions

Of course, there *is* structure in this data but the structure is not neatly spelled out for us

We have to extract what elements matter and figure out how they are related!

Example 1: Health Care

Forecast whether a patient is at risk for getting a disease?

Data

- Chart measurements (e.g., weight, blood pressure)
- Lab measurements (e.g., draw blood and send to lab)
- Doctor’s notes
- Patient’s medical history
- Family history
- Medical images

Example 2: Electrification

Where should we install cost-effective solar panels in developing countries?

Data

- Power distribution data for existing grid infrastructure
- Survey of electricity needs for different populations
- Labor costs
- Raw materials costs (e.g., solar panels, batteries, inverters)
- Satellite images

Example 3: Online Education

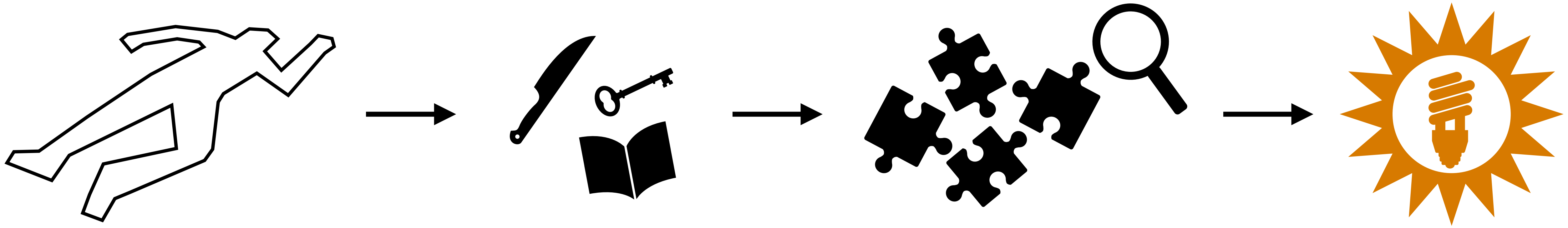
What parts of an online course are most confusing and need refinement?

Data

- Clickstream info through course website
- Video statistics
- Course forum posts
- Assignment submissions

Image source: African Reporter

Unstructured Data Analysis



Question

The dead body

This is provided by a practitioner

Data

The evidence

Some times you have to collect more evidence!

Finding Structure

Puzzle solving, careful analysis

Exploratory data analysis

Insights

When? Where? Why? How? Perpetrator catchable?

Answer original question

There isn't always a follow-up prediction problem to solve!

UDA involves *lots* of data → write computer programs to assist analysis

95-865

Prereq: Python programming

Students who ignore this prereq do poorly in the course

Part I: Exploratory data analysis

Part II: Predictive data analysis

95-865

Part I: Exploratory data analysis

Identify structure present in “unstructured” data

- Frequency and co-occurrence analysis
- Visualizing high-dimensional data/dimensionality reduction
- Clustering
- Topic modeling (a special kind of clustering)

Part II: Predictive data analysis

Make predictions using structure found in Part I

- Classical classification methods
- Neural nets and deep learning for analyzing images and text

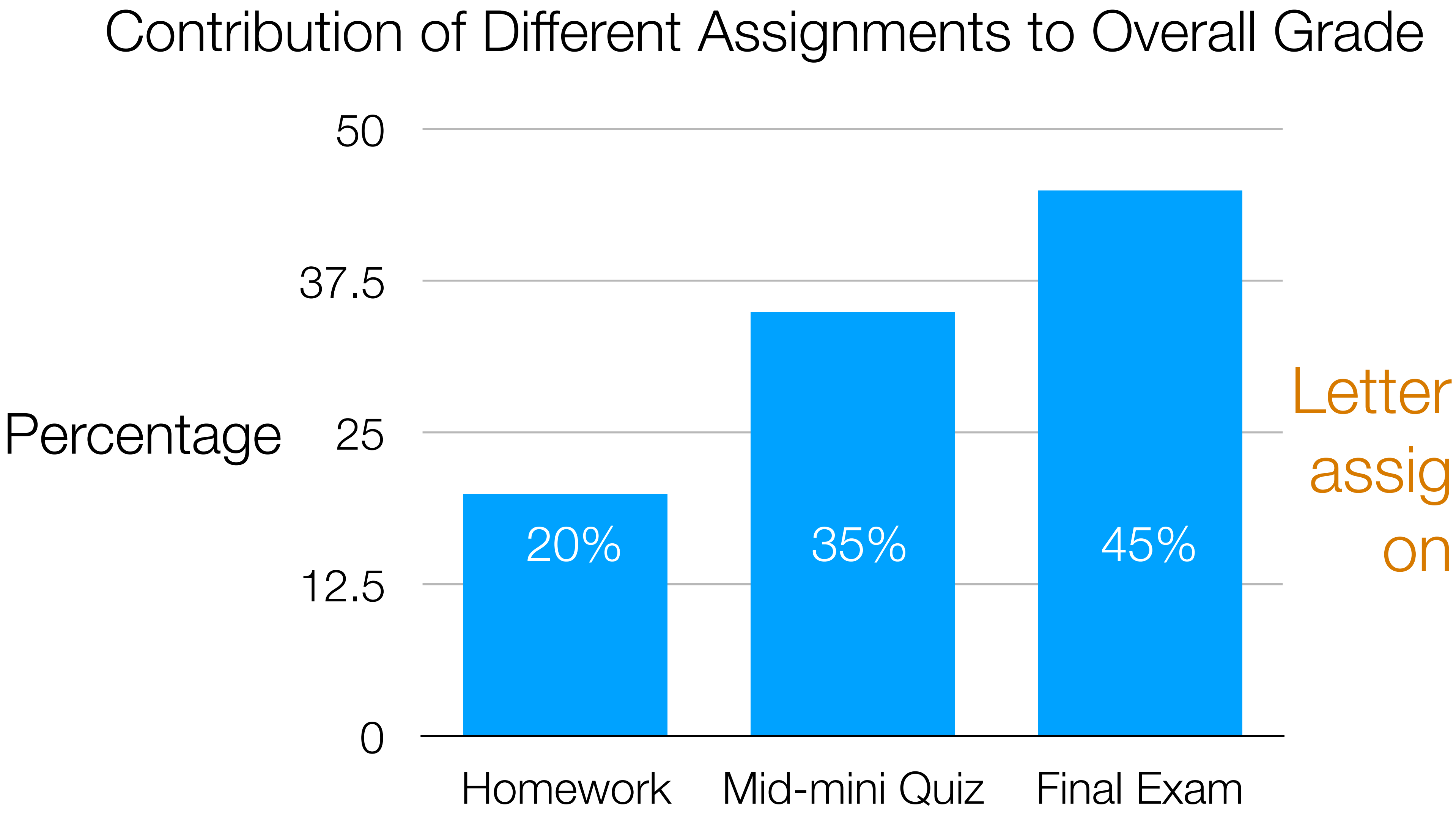
Course Goals

By the end of this course, you should have:

- Lots of hands-on programming experience with exploratory and predictive data analysis
- A high-level understanding of what methods are out there and which methods are appropriate for different problems
- A *very* high-level understanding of how these methods work *and what their limitations are*
- The ability to apply and interpret the methods taught to solve problems faced by organizations

I want you to leave the course with **practically useful** skills solving real-world problems with unstructured data analytics!

Deliverables & Grading



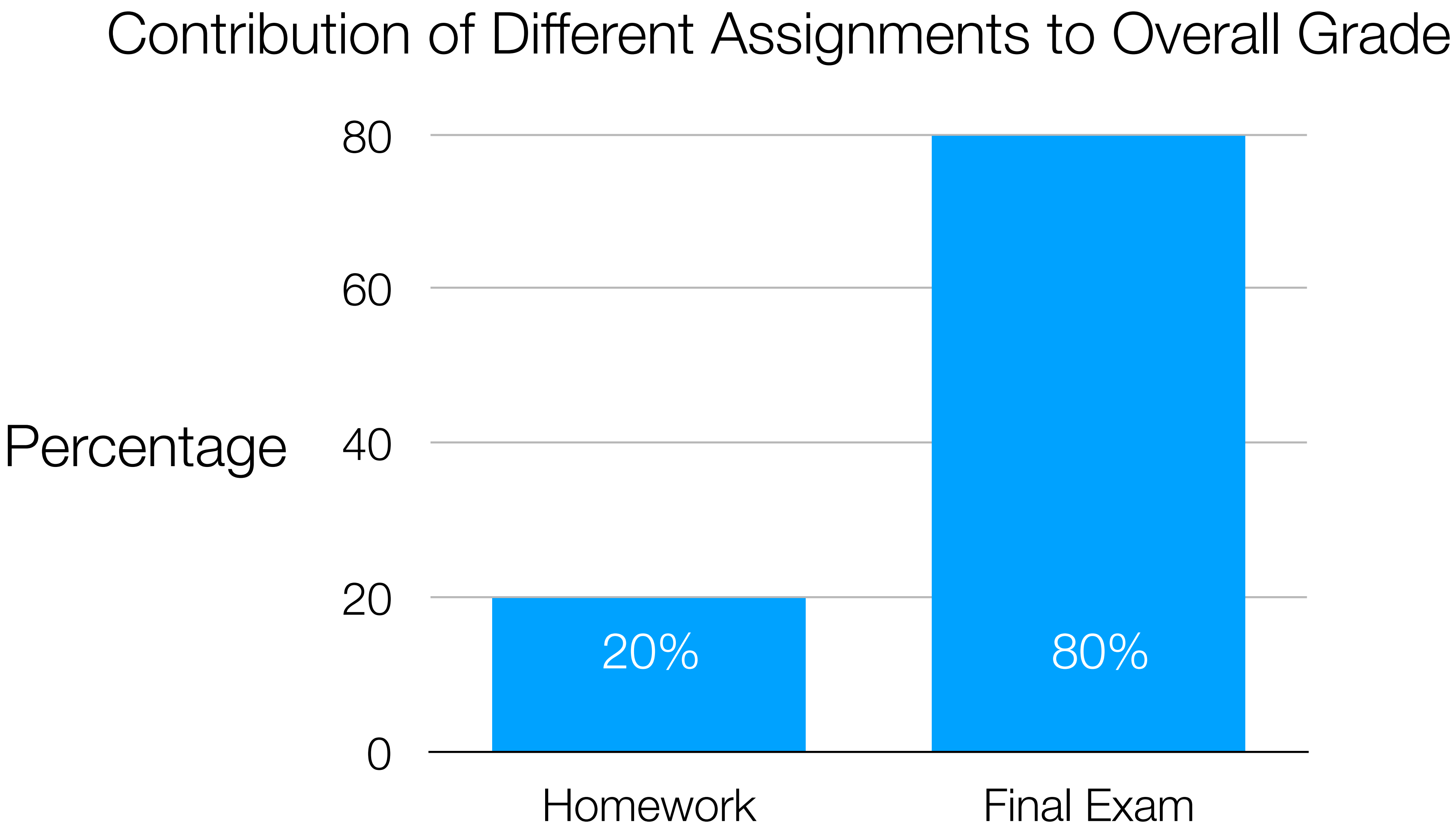
Letter grades are assigned based on a curve

Assignments involve coding in Python
(we use popular packages such as scikit-learn **and** keras**)**

Some problems require cloud computing
(we use Amazon Web Services)

1 Grading Exception

If you do better on the final exam than the mid-mini quiz:



Collaboration & Academic Integrity

- If you are having trouble, **ask for help!**
  - We will answer questions on Piazza and will also expect students to help answer questions!
  - **Do not post your candidate solutions on Piazza**
- In the real-world, you will unlikely be working alone
  - We encourage you to discuss concepts/how to approach problems
  - Please acknowledge classmates you talked to or resources you consulted (e.g., stackoverflow)
- **Do not share your code with classmates (instant message, email, Box, Dropbox, AWS, etc)**

Penalties for cheating are severe
e.g., 0 on assignment, F in course =(

Programming and Cloud Computing

python™

amazon webservices

- The data science/machine learning tools available have changed *drastically* over the last few years
  - Working with most of the latest innovations requires some programming (Python is common)
- Datasets encountered by many organizations are now often *massive*
  - Datasets often either won't fit or won't be processed fast enough on your personal machine but renting compute resources is now cheap (e.g., Amazon Web Services, Google Compute)

Course ~~Textbook~~ Materials

No existing textbook matches the course… =(

Main source of material: lectures slides

We'll post complimentary reading as we progress

Check **course website**
http://www.andrew.cmu.edu/user/georgech/95-865/

Assignments will be posted and submitted on **canvas**

Please post questions to **piazza (link is within canvas)**

Pittsburgh/Adelaide Weirdness

- Piazza is shared across Pittsburgh and Adelaide sections
- Homework due dates are identical across sections (so…some due dates are at weird hours)
- My Adelaide office hours are by appointment (email me to schedule) — we can do Skype or Google Hangouts
- I will physically be in Adelaide the week of Nov 12-16 (my schedule that week is still TBD)

Computing Environment

- We will be using **Anaconda (Python 3 version)** https://www.anaconda.com/

- We will give instructions for any third party packages to install and how to set up **Amazon Web Services** for cloud compute

- You will be submitting assignments in the form of **Jupyter notebooks**

Mid-mini Quiz and Final Exam

Format:

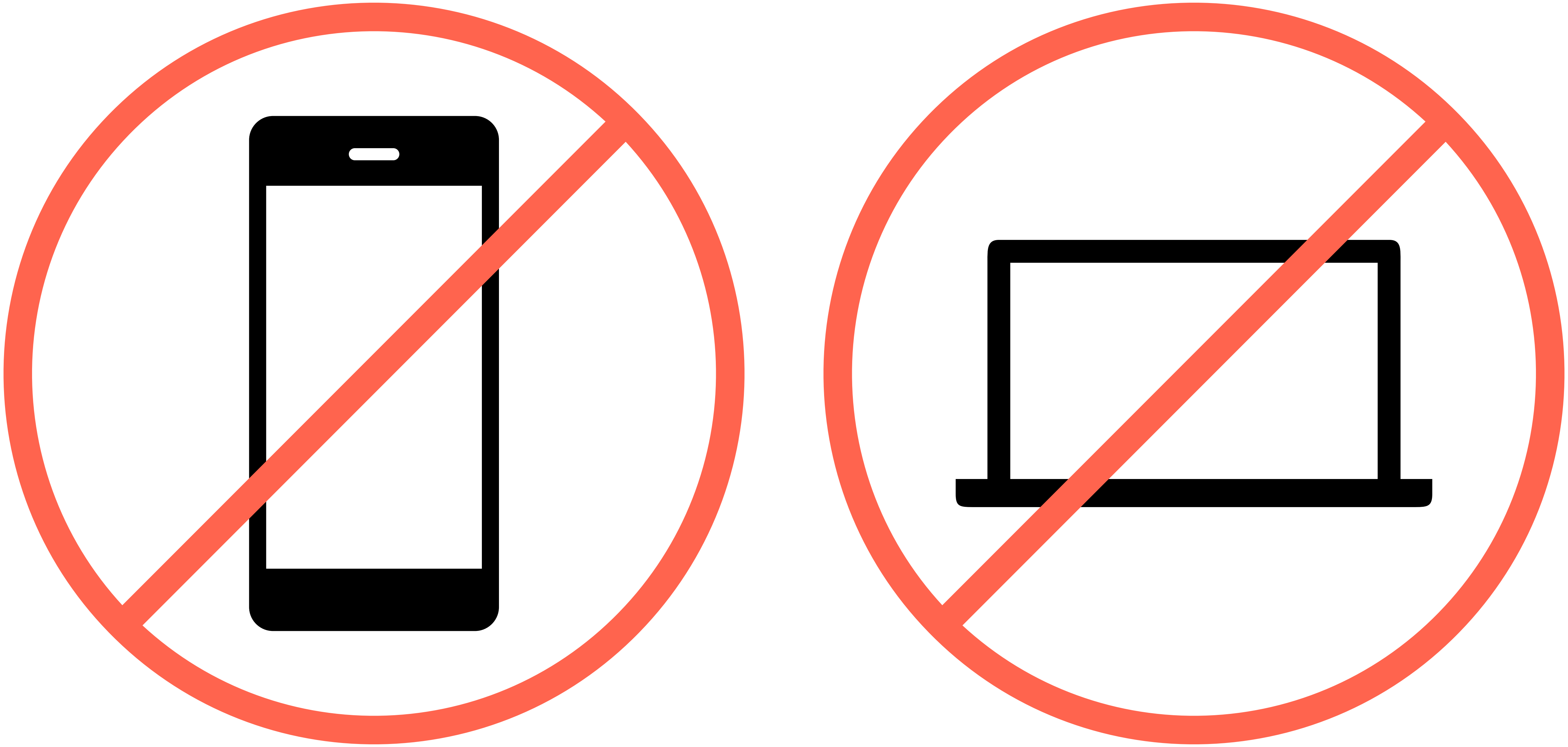
- **You have to bring a laptop computer and produce a Jupyter notebook** that answers a series of questions
- No collaboration (obviously)
- You are responsible for making sure your laptop has a compute environment set up appropriately and has enough battery life (or you sit close to a power outlet)
- Late exams will *not* be accepted

Late Homework Policy

- You are allotted 2 late days
  - If you use up a late day on an assignment, you can submit up to 24 hours late with no penalty
  - If you use up both late days on the same assignment, you can submit up to 48 hours late with no penalty
- Late days are *not* fractional
- This policy is in place precisely to account for various emergencies (health issues, etc) and you will not be given additional late days

Cell Phones and Laptops

Just like what you’d expect in a movie theater



We don’t want your device screens/sounds distracting classmates

Course Staff

Abhinav
Maurya

Mi
Zhou

Erick German
Rodriguez
Alvarez

Teaching Assistants

George
Chen

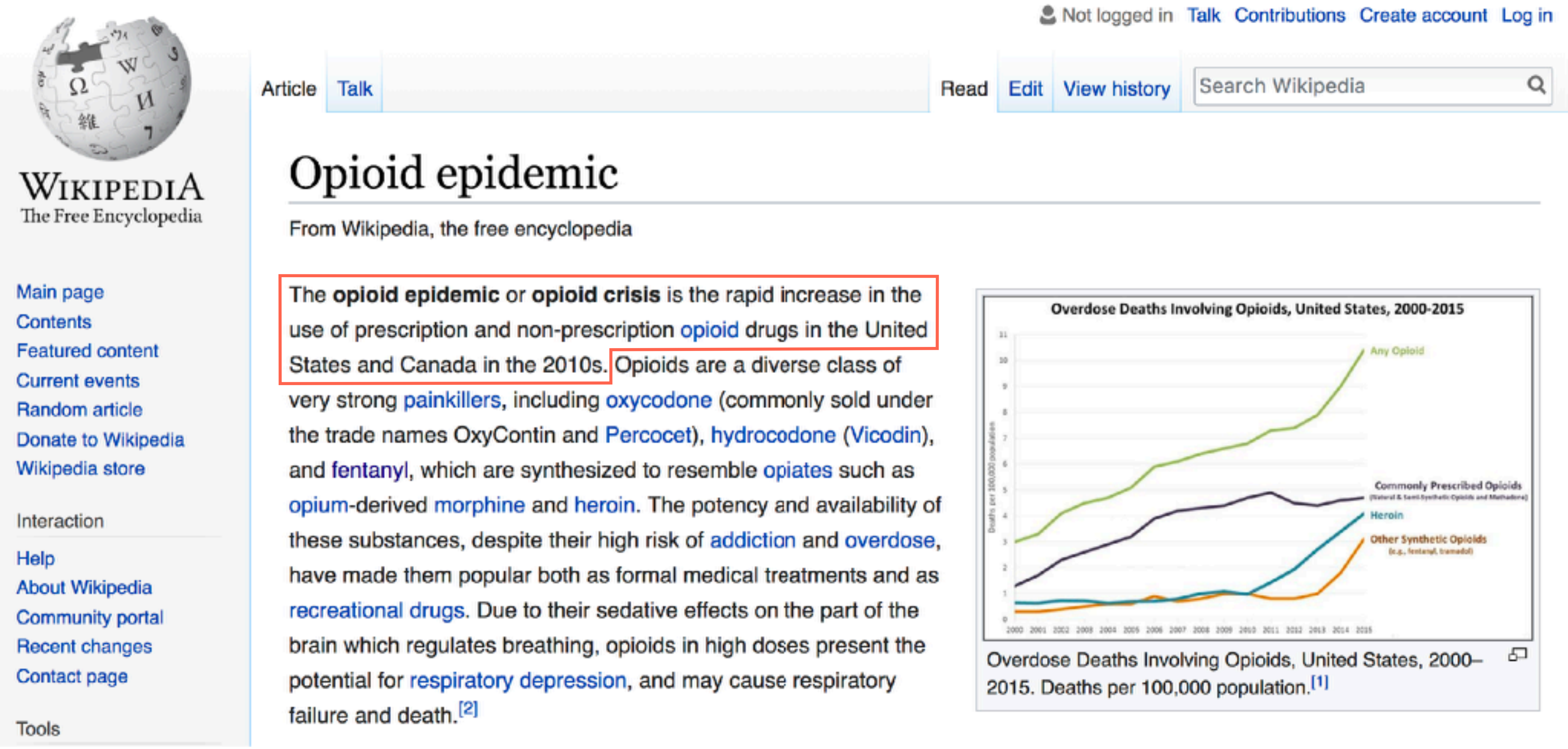
Instructor

Office hours:
Check course website
http://www.andrew.cmu.edu/user/georgech/95-865/

Part 1.
Exploratory Data Analysis

Play with data and make lots of visualizations to probe what structure is present in the data!

Basic text analysis: how do we represent text documents?



Opioid epidemic

From Wikipedia, the free encyclopedia

The **opioid epidemic** or **opioid crisis** is the rapid increase in the use of prescription and non-prescription opioid drugs in the United States and Canada in the 2010s. Opioids are a diverse class of very strong painkillers, including oxycodone (commonly sold under the trade names OxyContin and Percocet), hydrocodone (Vicodin), and fentanyl, which are synthesized to resemble opiates such as opium-derived morphine and heroin. The potency and availability of these substances, despite their high risk of addiction and overdose, have made them popular both as formal medical treatments and as recreational drugs. Due to their sedative effects on the part of the brain which regulates breathing, opioids in high doses present the potential for respiratory depression, and may cause respiratory failure and death.[2]

Overdose Deaths Involving Opioids, United States, 2000-2015
Deaths per 100,000 population
0
1
2
3
4
5
6
7
8
9
10
11
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
Any Opioid
Commonly Prescribed Opioids
(Natural & Semi-Synthetic Opioids and Methadone)
Heroin
Other Synthetic Opioids
(e.g., fentanyl, tramadol)

Overdose Deaths Involving Opioids, United States, 2000–2015. Deaths per 100,000 population.[1]

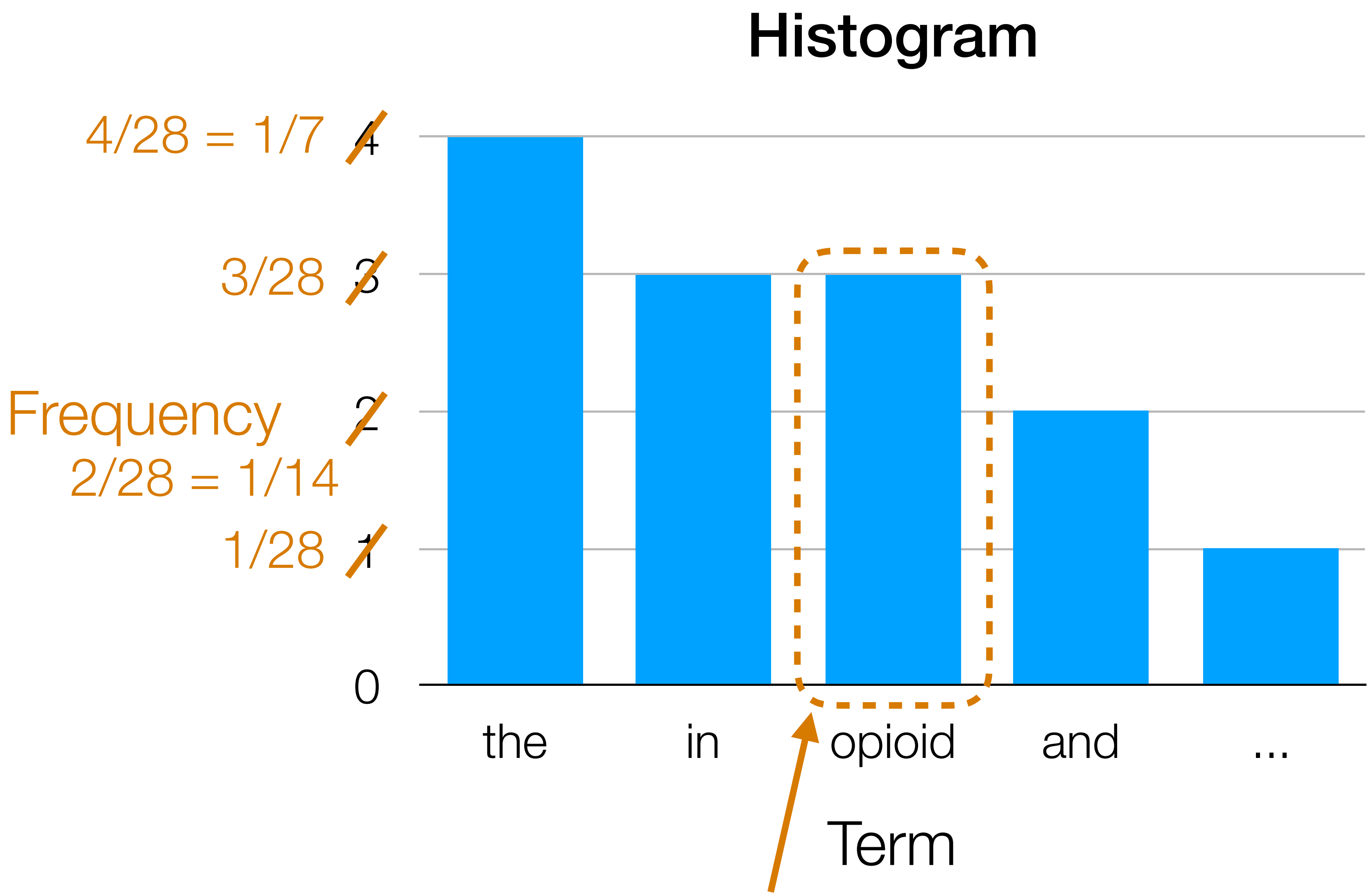
Source: Wikipedia, accessed 10/16/2017

Term frequencies

The: 1	/28
opioid: 3	/28
epidemic: 1	/28
or: 1	/28
crisis: 1	/28
is: 1	/28
the: 4	/28
rapid: 1	/28
increase: 1	/28
in: 3	/28
use: 1	/28
of: 1	/28
prescription: 1	/28
and: 2	/28
non-prescription: 1	/28
drugs: 1	/28
United: 1	/28
States: 1	/28
Canada: 1	/28
2010s.: 1	/28

The opioid epidemic or opioid crisis is the rapid increase in the use of prescription and non-prescription opioid drugs in the United States and Canada in the 2010s.

Total number of words in sentence: 28



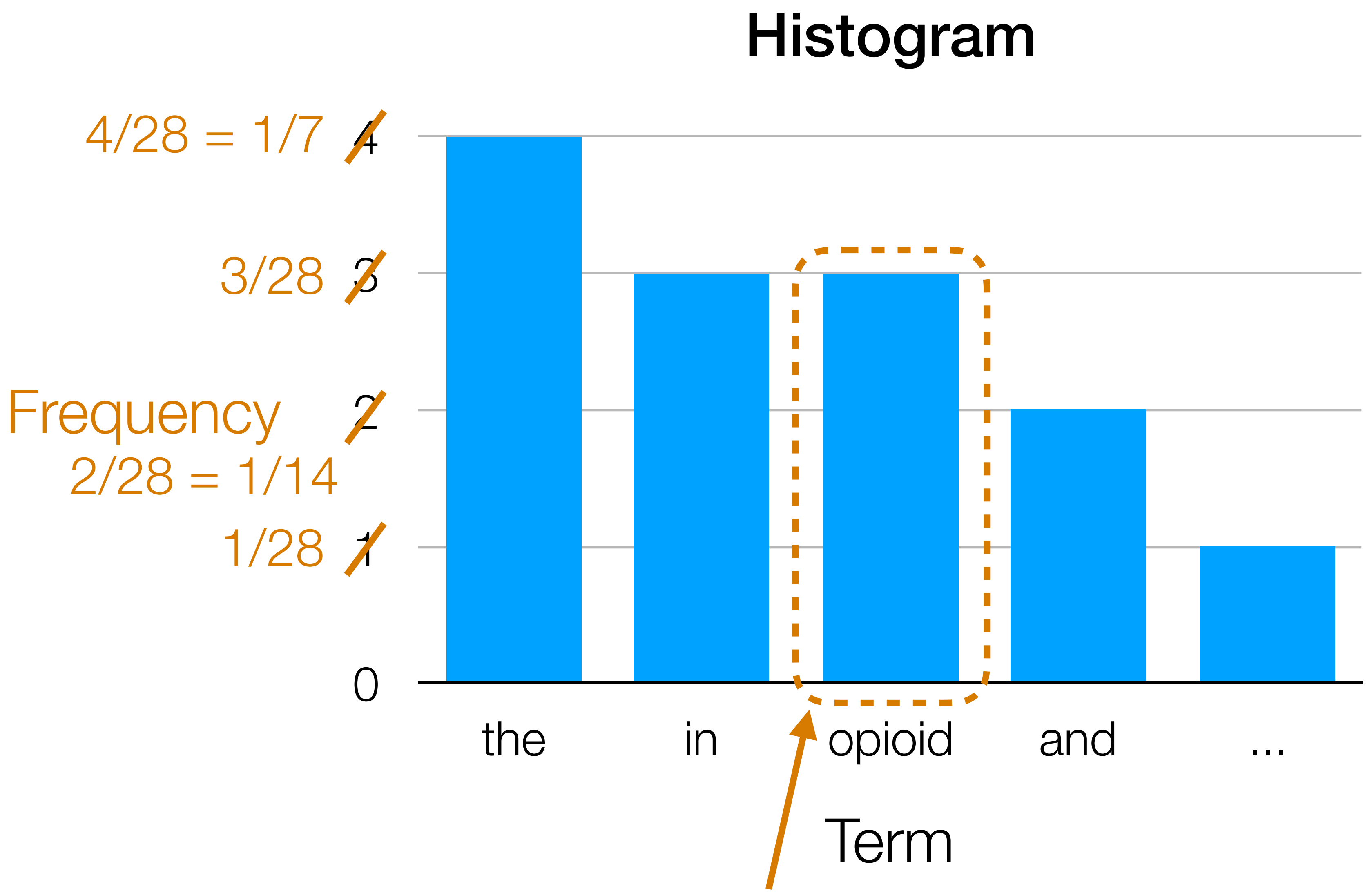
Fraction of words in the sentence that are “opioid”

opioid The epidemic or opioid crisis is the rapid increase in the use of prescription and non-prescription opioid drugs in the United States and Canada in the 2010s.

Total number of words in sentence: 28

Term frequencies

The: 1	/28
opioid: 3	/28
epidemic: 1	/28
or: 1	/28
crisis: 1	/28
is: 1	/28
the: 4	/28
rapid: 1	/28
increase: 1	/28
in: 3	/28
use: 1	/28
of: 1	/28
prescription: 1	/28
and: 2	/28
non-prescription: 1	/28
drugs: 1	/28
United: 1	/28
States: 1	/28
Canada: 1	/28
2010s.: 1	/28



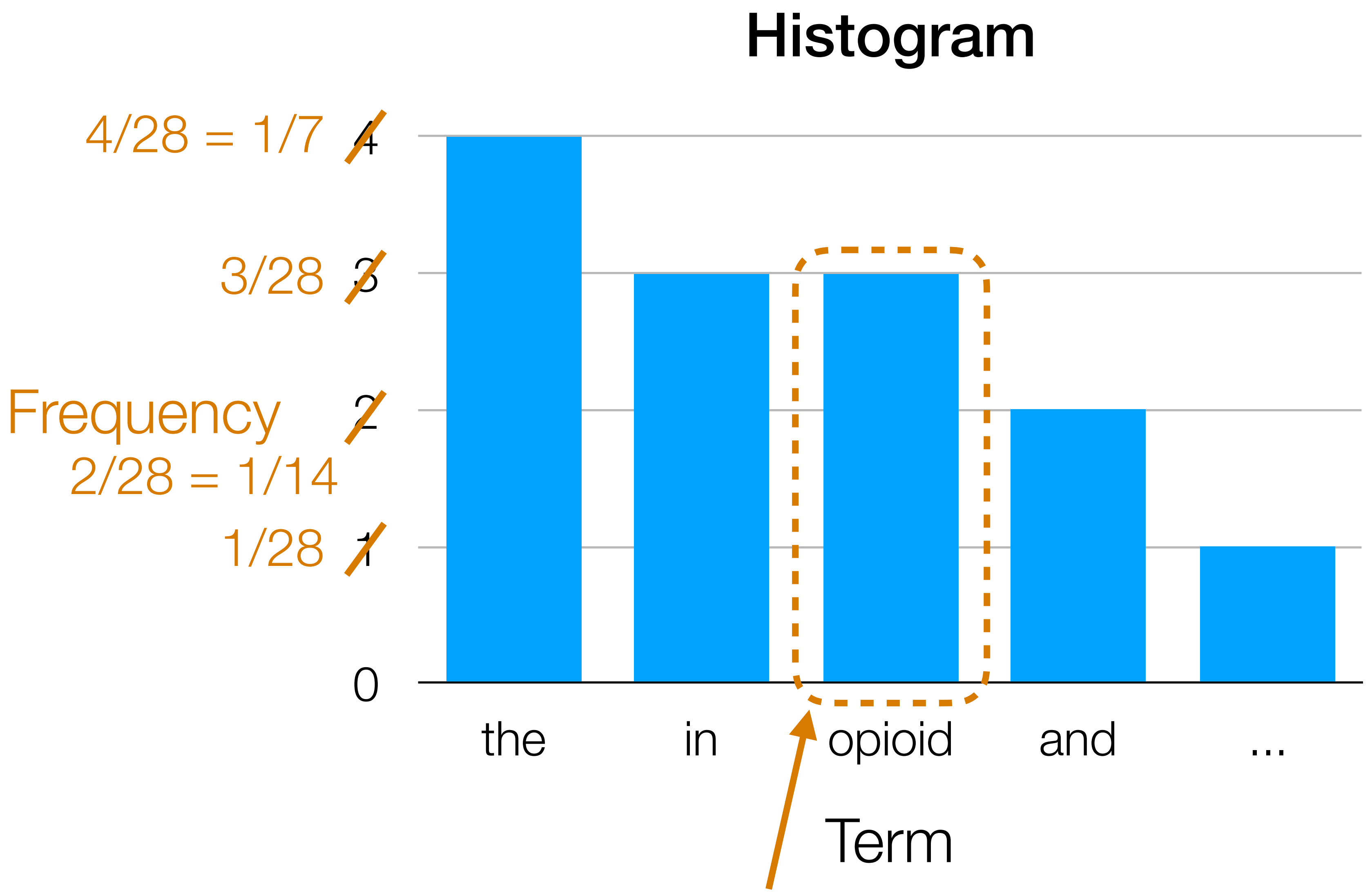
Fraction of words in the sentence that are "opioid"

Term frequencies

The: 1	/28
opioid: 3	/28
epidemic: 1	/28
or: 1	/28
crisis: 1	/28
is: 1	/28
the: 4	/28
rapid: 1	/28
increase: 1	/28
in: 3	/28
use: 1	/28
of: 1	/28
prescription: 1	/28
and: 2	/28
non-prescription: 1	/28
drugs: 1	/28
United: 1	/28
States: 1	/28
Canada: 1	/28
2010s.: 1	/28

increase the drugs opioid in The States or prescription opioid and of is rapid in opioid crisis the use non-prescription Canada 2010s. in United and the epidemic the

Total number of words in sentence: 28



4/28 = 1/7

3/28

Frequency

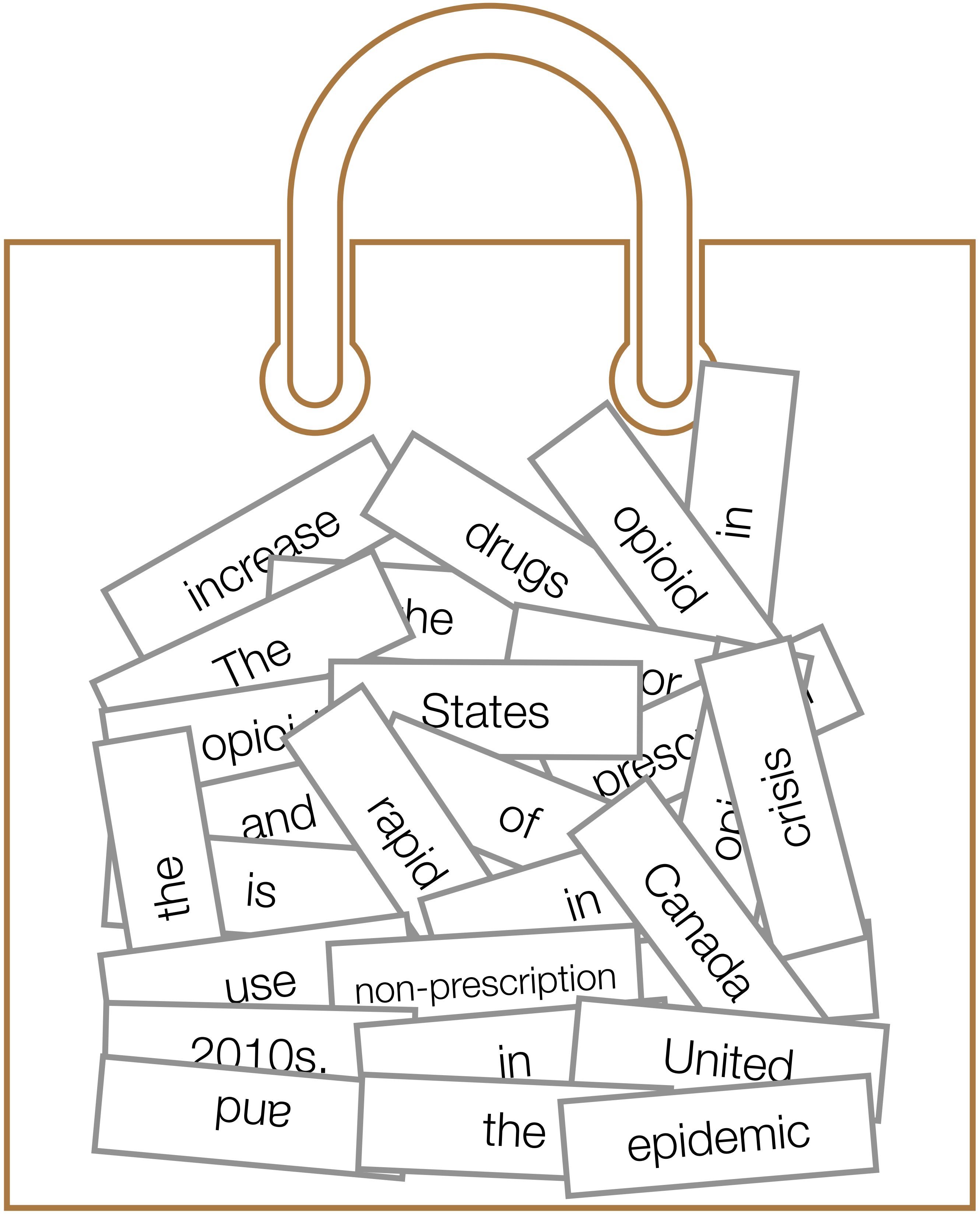
2/28 = 1/14

1/28

Fraction of words in the sentence that are “opioid”

increase the drugs opioid
in The States or
prescription opioid and of
is rapid in opioid crisis the
use non-prescription
Canada 2010s. in United
and the epidemic the

Bag of Words Model



Ordering of words doesn't matter

What is the probability of drawing the word "opioid" from the bag?

Handling Many Documents

- We can of course apply this technique of word frequencies to an entire document and not just a single sentence

➔ For a collection of documents (e.g., all of Wall Street Journal between late 1980's and early 1990's, all of Wikipedia up until early 2015, etc), we call the resulting term frequency the **collection term frequency** (ctf)

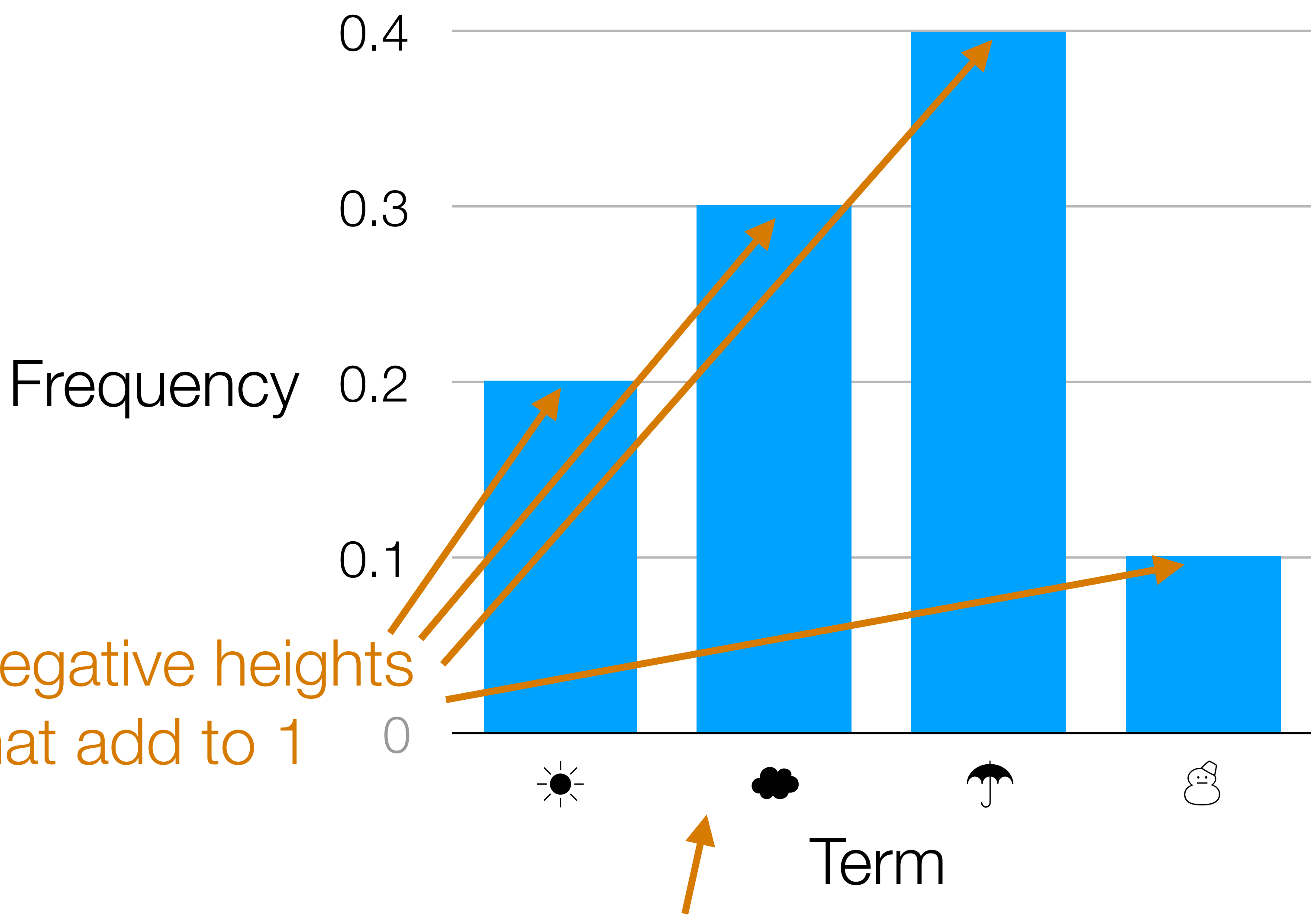
What does the *ctf* of "opioid" for all of Wikipedia refer to?

Many natural language processing (NLP) systems are trained on very large collections of text (also called **corpora**) such as the Wikipedia corpus and the Common Crawl corpus

So far did we use anything special about text?

Basic Probability in Disguise

"Sentence": ☀☂☁☁☁☂☃☂☂☀



Nonnegative heights that add to 1

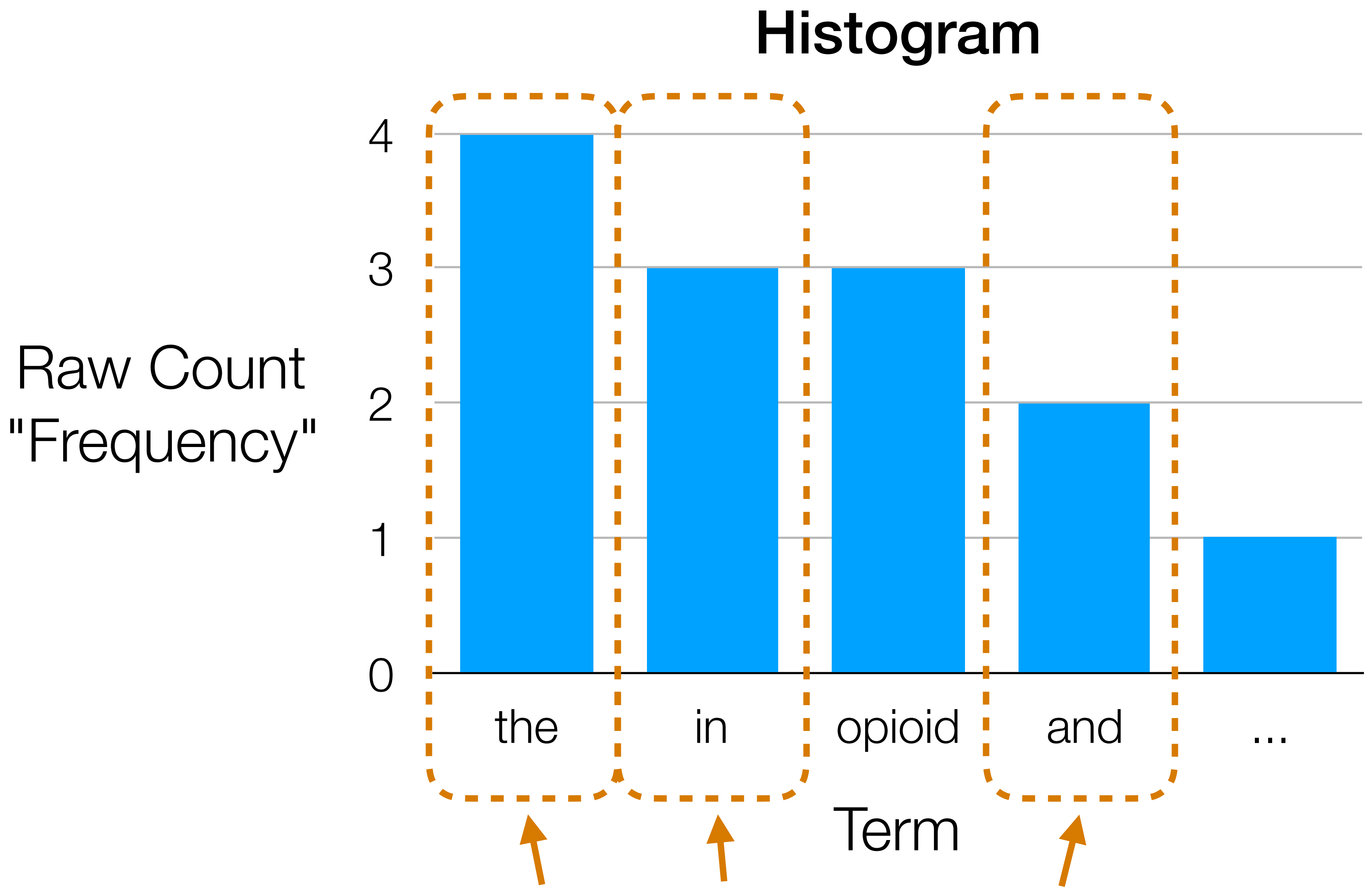
This is an example of a probability distribution

Probability distributions will appear throughout the course and are a **key component** to the success of many modern AI methods

Now let's take advantage of properties of text

In other words: natural language humans use has a lot of *structure* that we can exploit

Some Words Don't Help?



How helpful are these words to understanding semantics?

Bag-of-words models: many frequently occurring words unhelpful

We can remove these words first (remove them from the "bag")
➔ words that are removed are called **stopwords**

(determined by removing most frequent words or using curated stopword lists)

Example Stopword List (from spaCy)

'a', 'about', 'above', 'across', 'after', 'afterwards', 'again', 'against', 'all', 'almost', 'alone', 'along', 'already', 'also', 'although', 'always', 'am', 'among', 'amongst', 'amount', 'an', 'and', 'another', 'any', 'anyhow', 'anyone', 'anything', 'anyway', 'anywhere', 'are', 'around', 'as', 'at', 'back', 'be', 'became', 'because', 'become', 'becomes', 'becoming', 'been', 'before', 'beforehand', 'behind', 'being', 'below', 'beside', 'besides', 'between', 'beyond', 'both', 'bottom', 'but', 'by', 'ca', 'call', 'can', 'cannot', 'could', 'did', 'do', 'does', 'doing', 'done', 'down', 'due', 'during', 'each', 'eight', 'either', 'eleven', 'else', 'elsewhere', 'empty', 'enough', 'etc', 'even', 'ever', 'every', 'everyone', 'everything', 'everywhere', 'except', 'few', 'fifteen', 'fifty', 'first', 'five', 'for', 'former', 'formerly', 'forty', 'four', 'from', 'front', 'full', 'further', 'get', 'give', 'go', 'had', 'has', 'have', 'he', 'hence', 'her', 'here', 'hereafter', 'hereby', 'herein', 'hereupon', 'hers', 'herself', 'him', 'himself', 'his', 'how', 'however', 'hundred', 'i', 'if', 'in', 'inc', 'indeed', 'into', 'is', 'it', 'its', 'itself', 'just', 'keep', 'last', 'latter', 'latterly', 'least', 'less', 'made', 'make', 'many', 'may', 'me', 'meanwhile', 'might', 'mine', 'more', 'moreover', 'most', 'mostly', 'move', 'much', 'must', 'my', 'myself', 'name', 'namely', 'neither', 'never', 'nevertheless', 'next', 'nine', 'no', 'nobody', 'none', 'noone', 'nor', 'not', 'nothing', 'now', 'nowhere', 'of', 'off', 'often', 'on', 'once', 'one', 'only', 'onto', 'or', 'other', 'others', 'otherwise', 'our', 'ours', 'ourselves', 'out', 'over', 'own', 'part', 'per', 'perhaps', 'please', 'put', 'quite', 'rather', 're', 'really', 'regarding', 'same', 'say', 'see', 'seem', 'seemed', 'seeming', 'seems', 'serious', 'several', 'she', 'should', 'show', 'side', 'since', 'six', 'sixty', 'so', 'some', 'somehow', 'someone', 'something', 'sometime', 'sometimes', 'somewhere', 'still', 'such', 'take', 'ten', 'than', 'that', 'the', 'their', 'them', 'themselves', 'then', 'thence', 'there', 'thereafter', 'thereby', 'therefore', 'therein', 'thereupon', 'these', 'they', 'third', 'this', 'those', 'though', 'three', 'through', 'throughout', 'thru', 'thus', 'to', 'together', 'too', 'top', 'toward', 'towards', 'twelve', 'twenty', 'two', 'under', 'unless', 'until', 'up', 'upon', 'us', 'used', 'using', 'various', 'very', 'via', 'was', 'we', 'well', 'were', 'what', 'whatever', 'when', 'whence', 'whenever', 'where', 'whereafter', 'whereas', 'whereby', 'wherein', 'whereupon', 'wherever', 'whether', 'which', 'while', 'whither', 'who', 'whoever', 'whole', 'whom', 'whose', 'why', 'will', 'with', 'within', 'without', 'would', 'yet', 'you', 'your', 'yours', 'yourself', 'yourselves'

Is removing stop words always a good thing?

“To be or not to be”

Some Words Mean the Same Thing?

Term frequencies

The:	1
opioid:	3
epidemic:	1
or:	1
crisis:	1
is:	1
the:	4
rapid:	1
increase:	1
in:	3
use:	1
of:	1
prescription:	1
and:	2
non-prescription:	1
drugs:	1
United:	1
States:	1
Canada:	1
2010s.:	1

Should capitalization matter?

What about:

- walk, walking
- democracy, democratic, democratization
- good, better

Merging modified versions of "same" word to be analyzed as a single word is called **lemmatization**

(we'll see software for doing this shortly)

What about a word that has multiple meanings?

Challenging: try to split up word into multiple words depending on meaning (requires inferring meaning from context)

This problem is called **word sense disambiguation** (WSD)

Treat Some Phrases as a Single Word?

Term frequencies

The: 1
opioid: 3
epidemic: 1
or: 1
crisis: 1
is: 1
the: 4
rapid: 1
increase: 1
in: 3
use: 1
of: 1
prescription: 1
and: 2
non-prescription: 1
drugs: 1
United: 1
States: 1
Canada: 1
2010s.: 1

First need to detect what are "named entities": called **named entity recognition**

(we'll see software for doing this shortly)

Treat as single 2-word phrase “United States”?

Some Other Basic NLP Tasks

- **Tokenization:** figuring out what are the atomic "words" (including how to treat punctuation)
- **Part-of-speech tagging:** figuring out what are nouns, verbs, adjectives, etc
- **Sentence recognition:** figuring out when sentences actually end rather than there being some acronym with periods in it, etc

Bigram Model

The opioid epidemic or opioid crisis is the rapid increase in the use of prescription and non-prescription opioid drugs in the United States and Canada in the 2010s.

The opioid	opioid epidemic	epidemic or
or opioid	opioid crisis	crisis is

Ordering of words now matters (a little)

...

"Vocabulary size" (# unique cards) dramatically increases!

If using stopwords, remove any phrase with at least 1 stopword

1 word at a time: **unigram** model

2 words at a time: **bigram** model

n words at a time: ***n*-gram** model

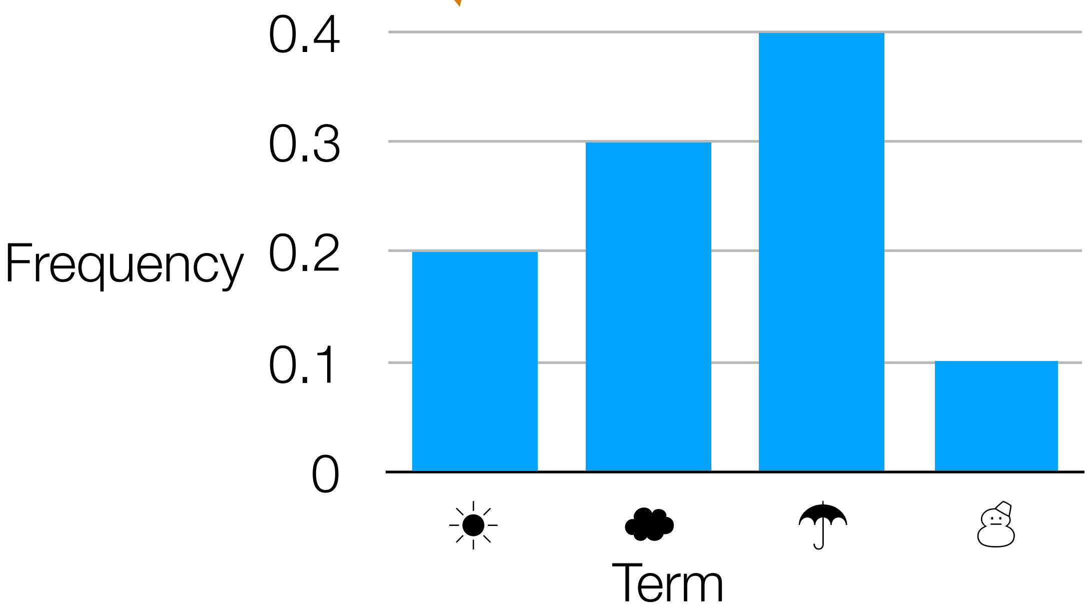
The spaCy Python Package

Demo

Recap: Basic Text Analysis

- Represent text in terms of “features”
(e.g., how often each word/phrase appears, whether it’s a named entity, etc)
  - Can repeat this for different documents:
  represent each document as a “feature vector”

"Sentence": ☀☂☁☁☁☂☃☂☂☀



$$\begin{bmatrix} 0.2 \\ 0.3 \\ 0.4 \\ 0.1 \end{bmatrix}$$

This is a point in 4-dimensional space, $\mathbb{R}^4$

# dimensions = number of terms

In general (not just text): first represent data as feature vectors

Example: Representing an Image

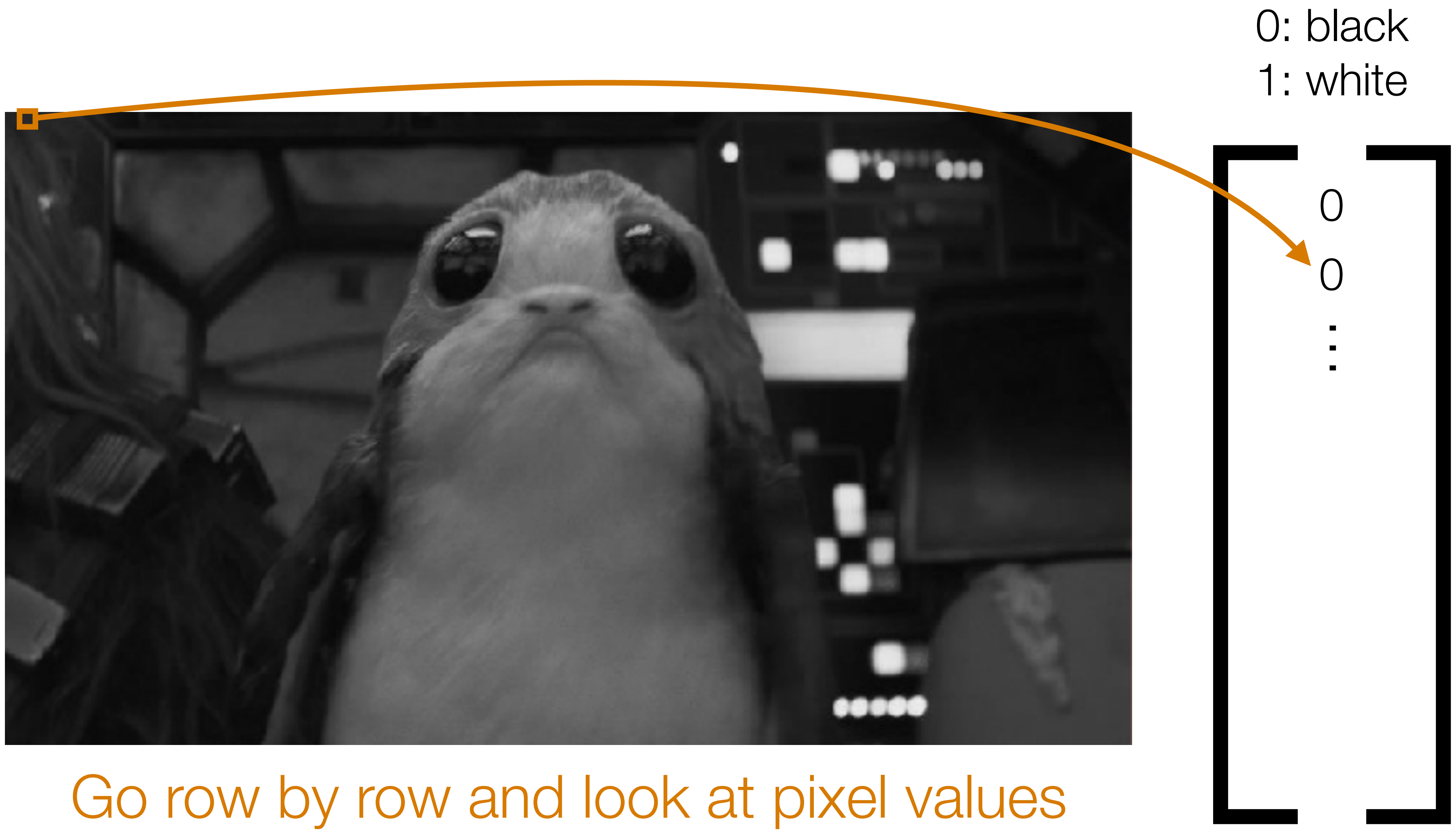
0: black
1: white

$$\begin{bmatrix} 0 \\ \\ \\ \\ \end{bmatrix}$$

Go row by row and look at pixel values

Image source: starwars.com

Example: Representing an Image



0: black
1: white

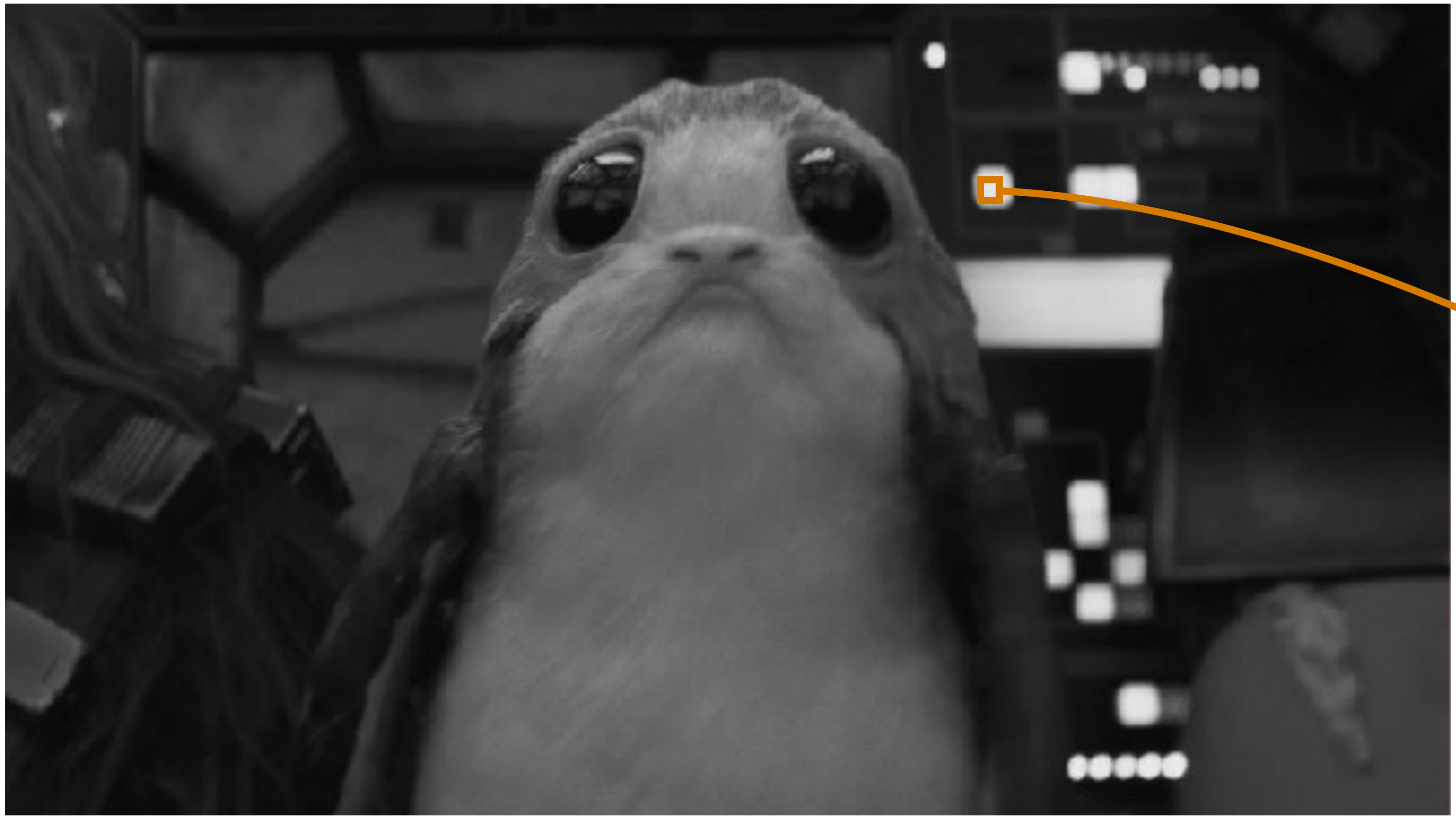
$$\begin{bmatrix} 0 \\ 0 \\ \vdots \\ \\ \\ \\ \end{bmatrix}$$

Go row by row and look at pixel values

Image source: starwars.com

Example: Representing an Image

0: black
1: white



$$\begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0.9 \\ \vdots \\ \\ \end{bmatrix}$$

Go row by row and look at pixel values

Image source: starwars.com

Example: Representing an Image

0: black
1: white

$$\begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0.9 \\ \vdots \\ \\ 0.3 \end{bmatrix}$$

Go row by row and look at pixel values

# dimensions = image width × image height

Very high dimensional!

Image source: starwars.com

Back to Text

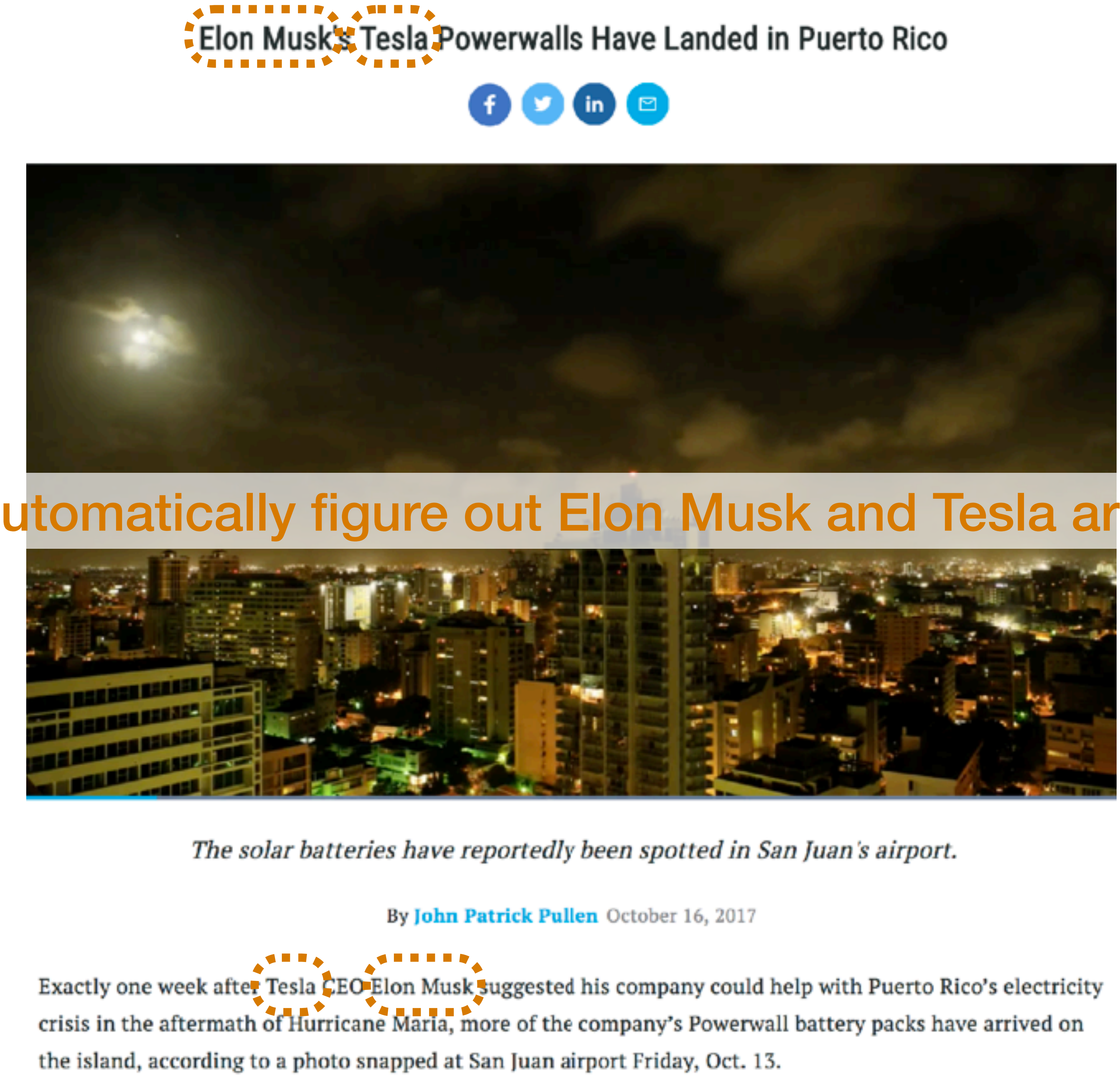
Unigram bag of words model is already quite powerful:

- Enough to learn topics
  (each text doc: raw word counts without stopwords)

- Enough to learn a simple detector for email spam

These are HW2 problems

Finding Possibly Related Entities



Elon Musk's Tesla Powerwalls Have Landed in Puerto Rico

The solar batteries have reportedly been spotted in San Juan's airport.

By John Patrick Pullen October 16, 2017

Exactly one week after Tesla CEO Elon Musk suggested his company could help with Puerto Rico's electricity crisis in the aftermath of Hurricane Maria, more of the company's Powerwall battery packs have arrived on the island, according to a photo snapped at San Juan airport Friday, Oct. 13.

How to automatically figure out Elon Musk and Tesla are related?

Source: http://fortune.com/2017/10/16/elon-musks-tesla-powerwalls-have-landed-in-puerto-rico/

Co-Occurrences

For example: count # news articles that have different named entities co-occur

	Apple	Facebook	Tesla
Elon Musk	10	15	300
Mark Zuckerberg	500	10000	500
Tim Cook	200	30	10

Big values ➔ *possibly* related named entities

Different Ways to Count

- Just saw: for all doc’s, count # of doc’s in which two named entities co-occur
  - This approach ignores # of co-occurrences *within a specific document* (e.g., if 1 doc has “Elon Musk” and “Tesla” appear 10 times, we count this as 1)
  - Could instead add # co-occurrences, not just whether it happened in a doc
- Instead of looking at # doc’s, look at co-occurrences within a *sentence*, or a *paragraph*, etc

Bottom Line

- There are many ways to count co-occurrences
- You should think about what makes the most sense/is reasonable for the problem you’re looking at

Co-Occurrences

For example: count # news articles that have different named entities co-occur

	Apple	Facebook	Tesla
Elon Musk	10	15	300
Mark Zuckerberg	500	10000	500
Tim Cook	200	30	10

Big values ➔ *possibly* related named entities

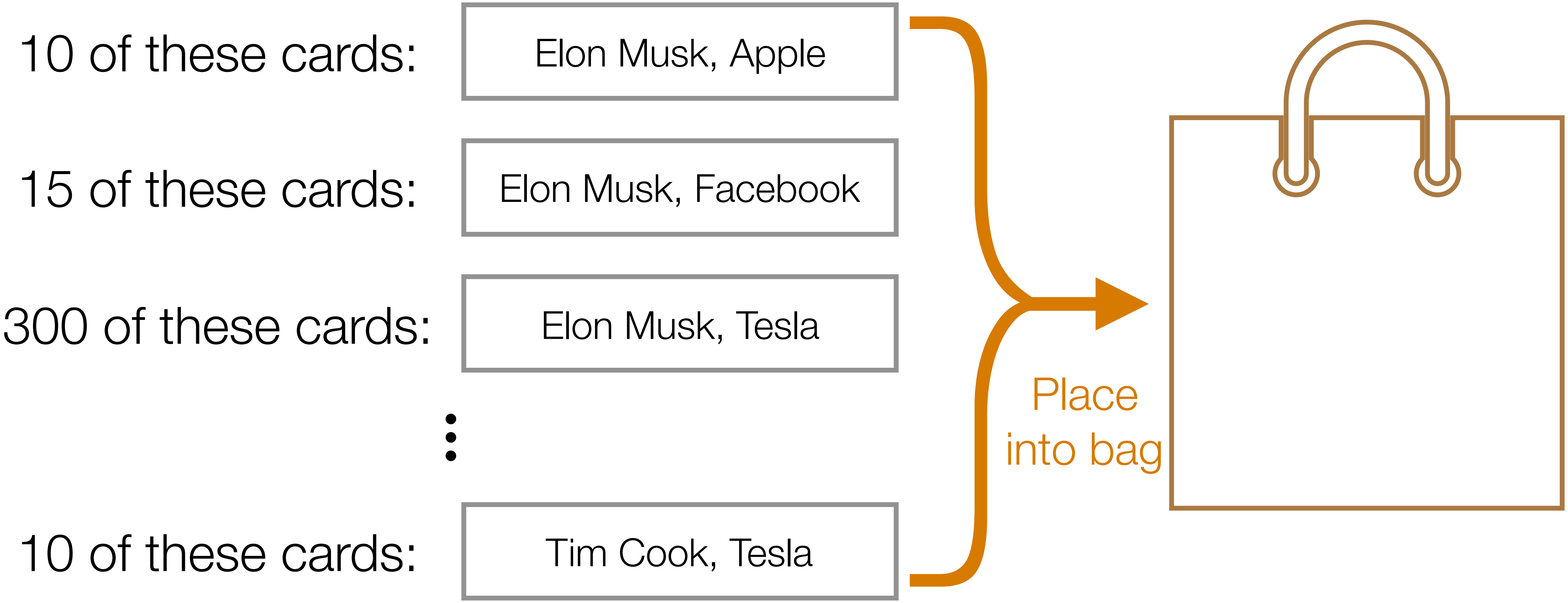
How to downweight “Mark Zuckerberg” if there are just way more articles that mention him?

Key idea: what would happen if people and companies had nothing to do with each other?

	Apple	Facebook	Tesla
Elon Musk	10	15	300
Mark Zuckerberg	500	10000	500
Tim Cook	200	30	10

Probability of drawing “Elon Musk, Apple”?

Probability of drawing a card that says “Apple” on it?



10 of these cards: Elon Musk, Apple

15 of these cards: Elon Musk, Facebook

300 of these cards: Elon Musk, Tesla

⋮

10 of these cards: Tim Cook, Tesla

Place into bag

Co-occurrence table

	Apple	Facebook	Tesla
Elon Musk	10	15	300
Mark Zuckerberg	500	10000	500
Tim Cook	200	30	10

Total: 11565

Joint probability table

	Apple	Facebook	Tesla
Elon Musk	10 /11565	15 /11565	300 /11565
Mark Zuckerberg	500 /11565	10000 /11565	500 /11565
Tim Cook	200 /11565	30 /11565	10 /11565

sum to get P(Elon Musk)

Total: 11565

Joint probability table

	Apple	Facebook	Tesla	
Elon Musk	0.00086	0.00130	0.02594	**0.02810**
Mark Zuckerberg	0.04323	0.86468	0.04323	**0.95115**
Tim Cook	0.01729	0.00259	0.00086	**0.02075**
	0.06139	**0.86857**	**0.07004**	

Recall: if events A and B are independent, P(A, B) = P(A)P(B)

Joint probability table **if people and companies were independent**

	Apple	Facebook	Tesla	
Elon Musk	0.00173	0.02441	0.00197	**0.02810**
Mark Zuckerberg	0.05839	0.82614	0.06662	**0.95115**
Tim Cook	0.00127	0.01802	0.00145	**0.02075**
	0.06139	**0.86857**	**0.07004**	

Recall: if events A and B are independent, $P(A, B) = P(A)P(B)$

What we actually observe

	Apple	Facebook	Tesla
Elon Musk	0.00086	0.00130	0.02594
Mark Zuckerberg	0.04323	0.86468	0.04323
Tim Cook	0.01729	0.00259	0.00086

What should be the case if people are companies are independent

	Apple	Facebook	Tesla
Elon Musk	0.00173	0.02441	0.00197
Mark Zuckerberg	0.05839	0.82614	0.06662
Tim Cook	0.00127	0.01802	0.00145

Pointwise Mutual Information (PMI)

Probability of A and B co-occurring

$$\frac{\boxed{P(A, B)}}{\boxed{P(A)\ P(B)}}$$

if equal to 1
➔ A, B are indep.

Probability of A and B co-occurring *if they were independent*

PMI(A, B) is defined as the log of the above ratio

PMI measures (the log of) a ratio that says how far A and B are from being independent

Example PMI Calculation

Demo

Looking at All Pairs of Outcomes

- PMI measures how P(A, B) differs from P(A)P(B) using a **log ratio**
- **Log ratio** isn't the only way to compare!
- Another way to compare:

$$\frac{[\ P(A, B) - P(A)\ P(B)\ ]^2}{P(A)\ P(B)}$$

Phi-square is between 0 and 1

0 ➔ pairs are all indep.

$$\text{Phi-square} = \sum_{A,\ B} \frac{[\ P(A, B) - P(A)\ P(B)\ ]^2}{P(A)\ P(B)}$$

$$\text{Chi-square} = N \times \text{Phi-square}$$

Measures how close *all* pairs of outcomes are close to being indep.

N = sum of all co-occurrence counts

Phi-Square/Chi-Square Calculation

Demo

Summary: Co-Occurrences

- Joint probability P(A, B) can be poor indicator of whether A and B co-occurring is “interesting”
- Find interesting relationships between pairs of items by looking at PMI
  - Intuition: “Interesting” co-occurring events should occur more frequently than if they were to co-occur independently

Co-occurrence Analysis Applications

- If you're an online store/retailer:
  anticipate *when* certain products are likely to be purchased/rented/consumed more
  - Products & dates
- If you have a bunch of physical stores:
  anticipate *where* certain products are likely to be purchased/rented/consumed more
  - Products & locations
- If you're the police department:
  create "heat map" of where different criminal activity occurs
  - Crime reports & locations

Co-occurrence Analysis Applications

- If you're an online store/retailer: anticipate *when* certain products are likely to be purchased/ re
    -
- If y an sed/ re
    -
- If y cr curs
    - Crime reports & locations

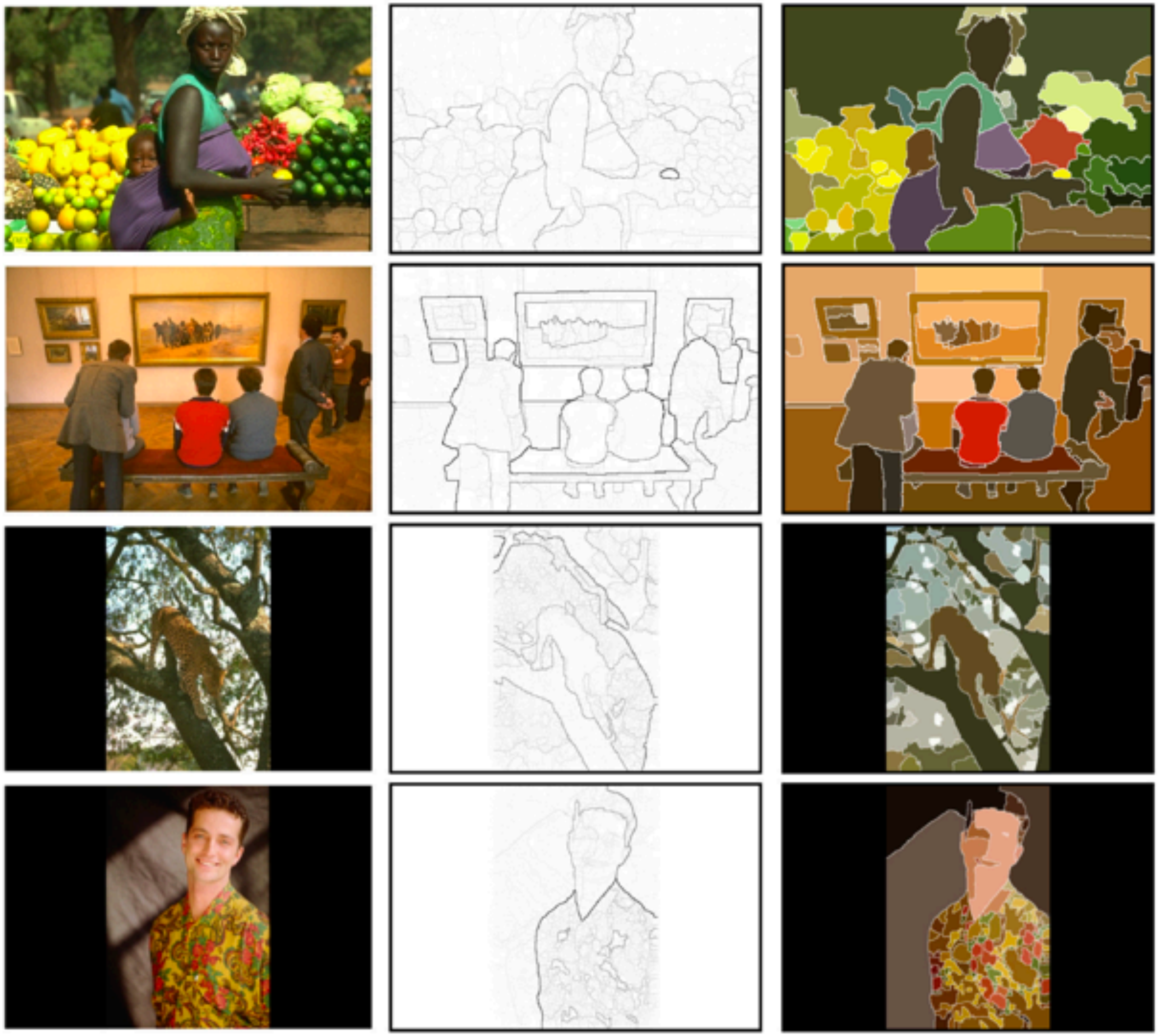
Examples of data to take advantage of:

- data collected by your organization
- social networks
- news websites
- blogs

Web scraping frameworks can be helpful:

- Scrapy
- Selenium (great with JavaScript-heavy pages)

Example Application of PMI: Image Segmentation



Phillip Isola, Daniel Zoran, Dilip Krishnan, and Edward H. Adelson. Crisp boundary detection using pointwise mutual information. ECCV 2014.

Example Application of PMI: Word Embeddings

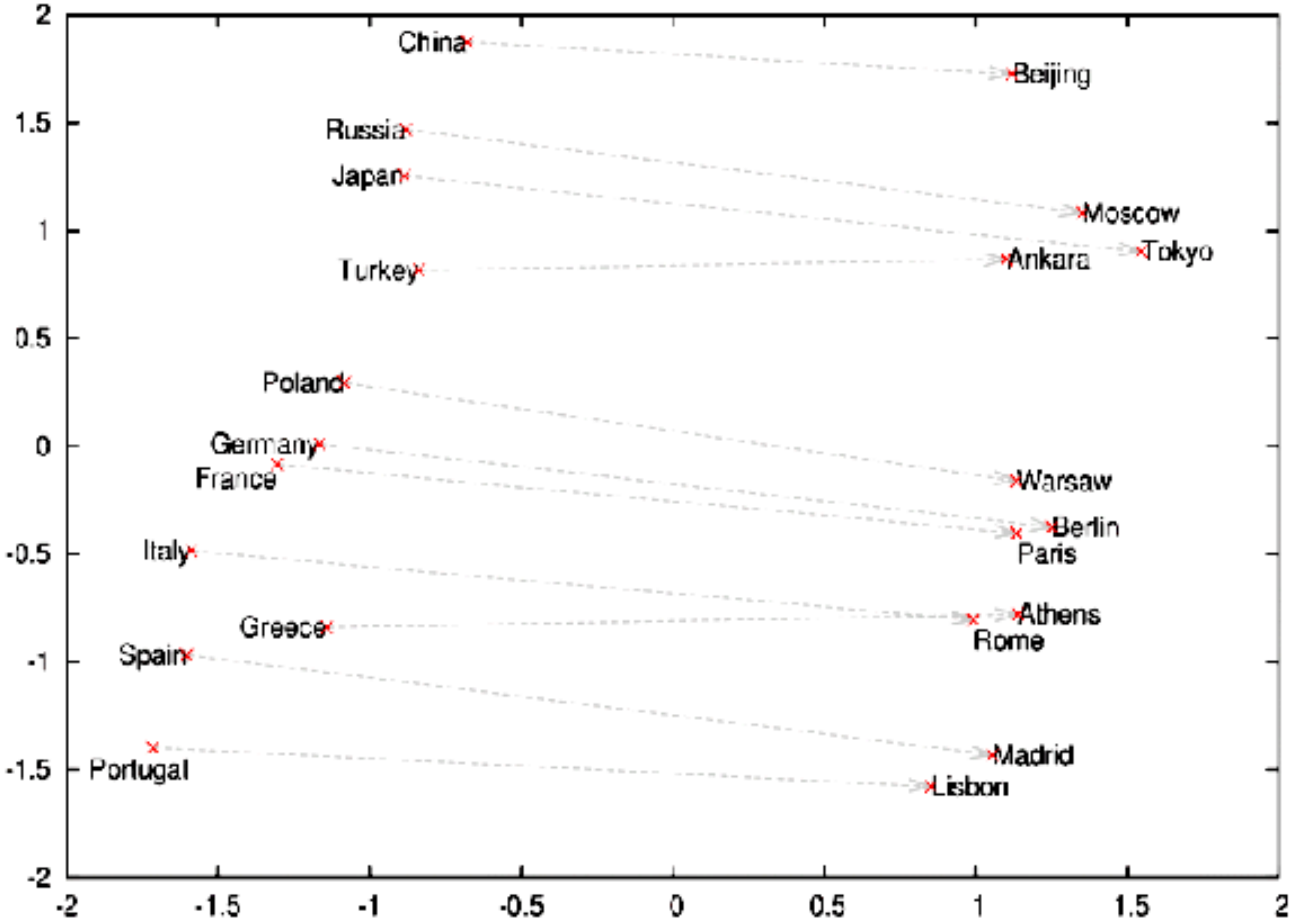


Image source: https://deeplearning4j.org/img/countries_capitals.png

Omer Levy and Yoav Goldberg. Neural word embeddings as implicit matrix factorization. NIPS 2014.