Carnegie Mellon University
Heinz College

# 95-865 Unstructured Data Analytics

## Week 1: Course overview, analyzing text using frequencies

George Chen

# Big Data

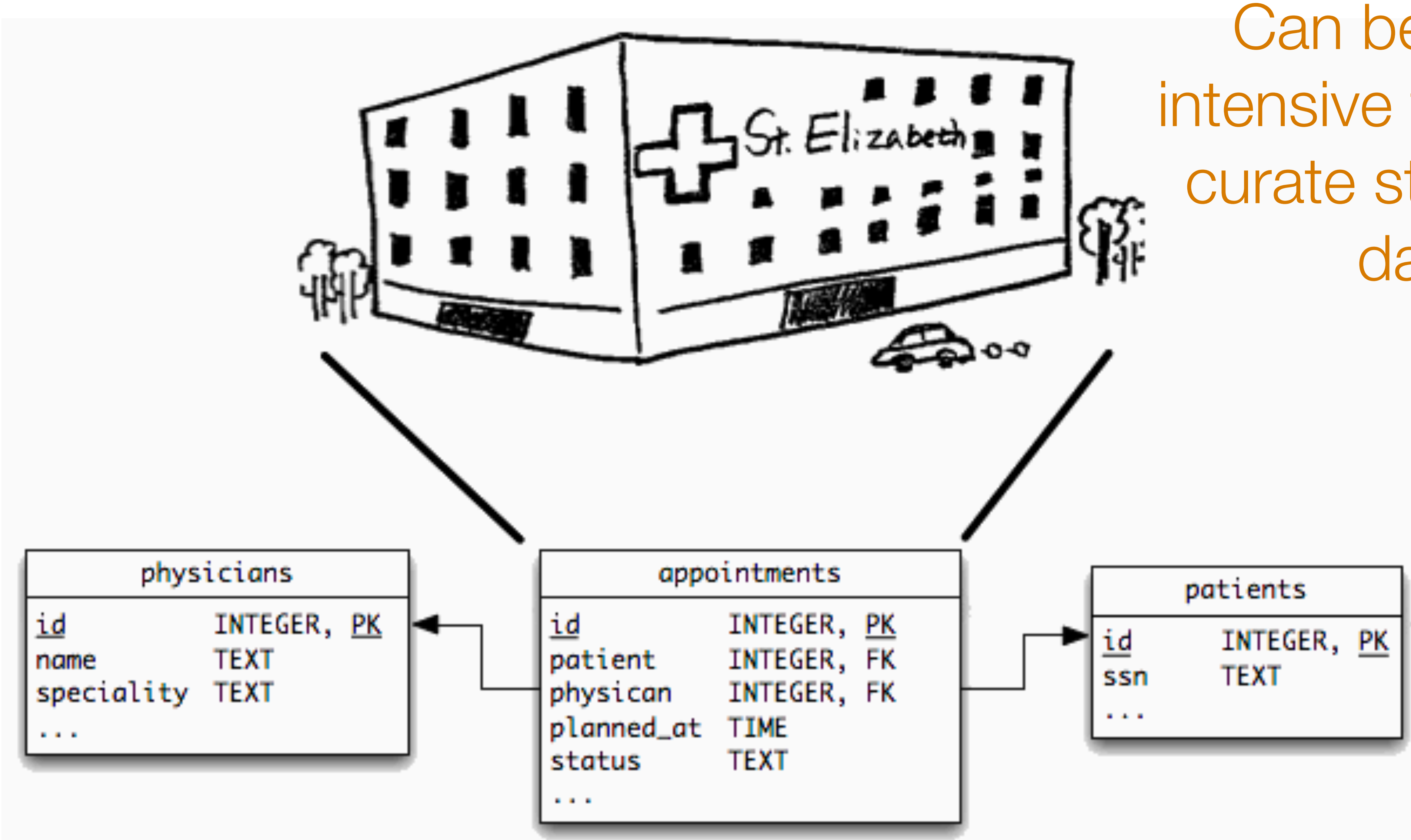
We're now collecting data on virtually every human endeavor

How do we turn these data into actionable insights?

# Two Types of Data

# Structured Data

Well-defined elements, relationships between elements




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?*

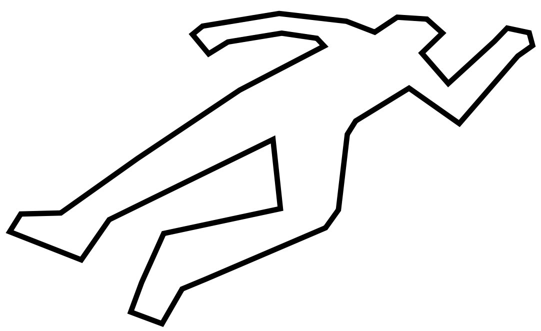
Data

- 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

→

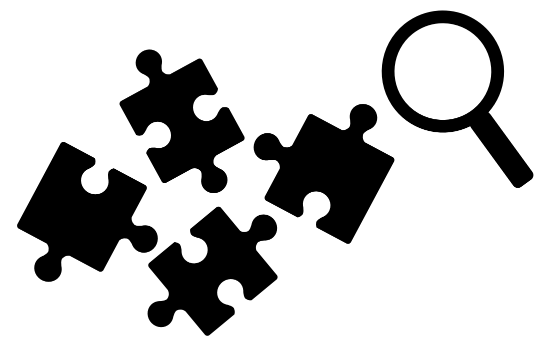
Data

*The evidence*

Some times you have to collect more evidence!

→

Finding Structure

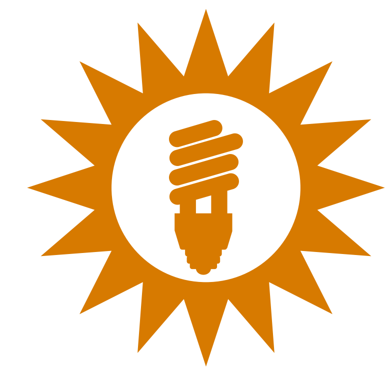


*Puzzle solving, careful analysis*

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

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

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

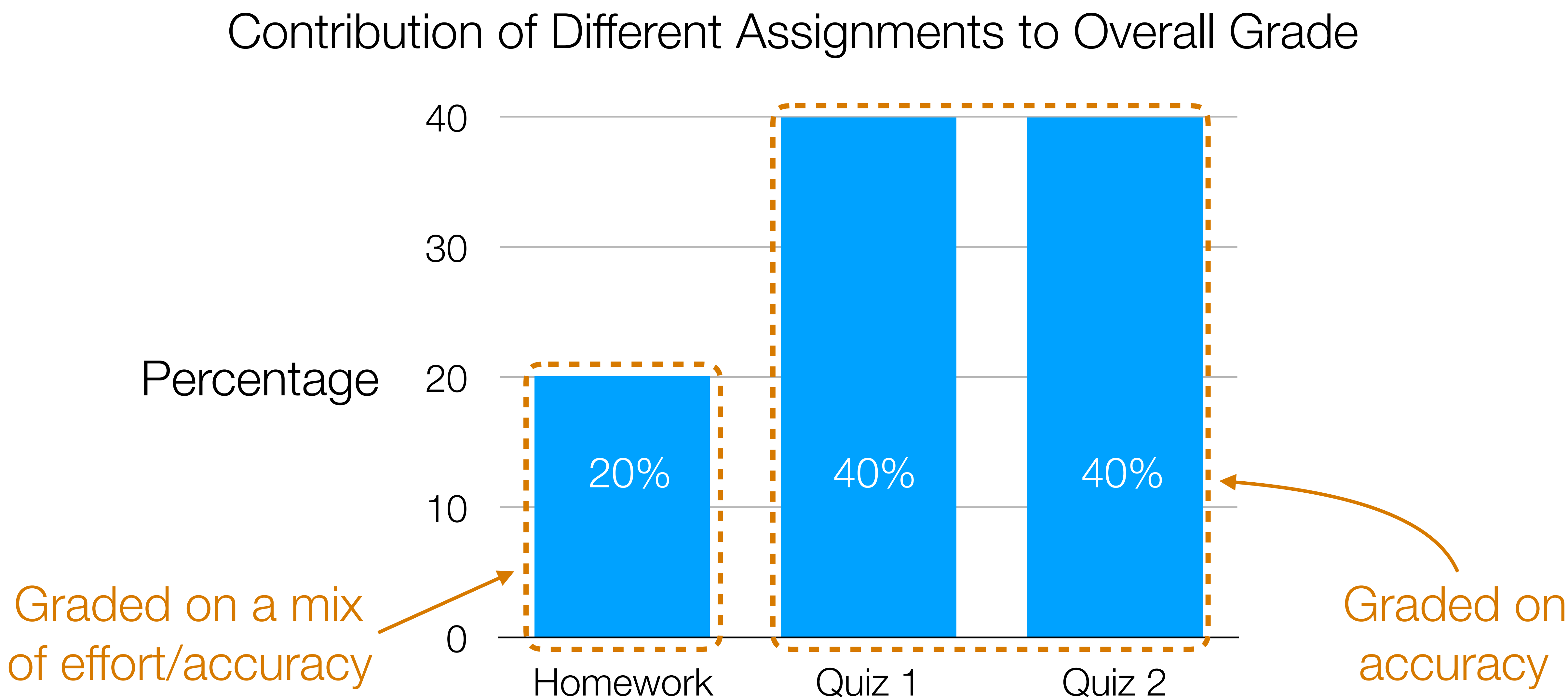
# Course Goals

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

**All assignments involve coding in Python**

(popular amongst machine learning/computer science community)

**HW3 uses Amazon Web Services for cloud computing**

(many real datasets too large to either fit or process on personal machine)

# 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 =(**

# 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)**

# The Two Quizzes

Format:

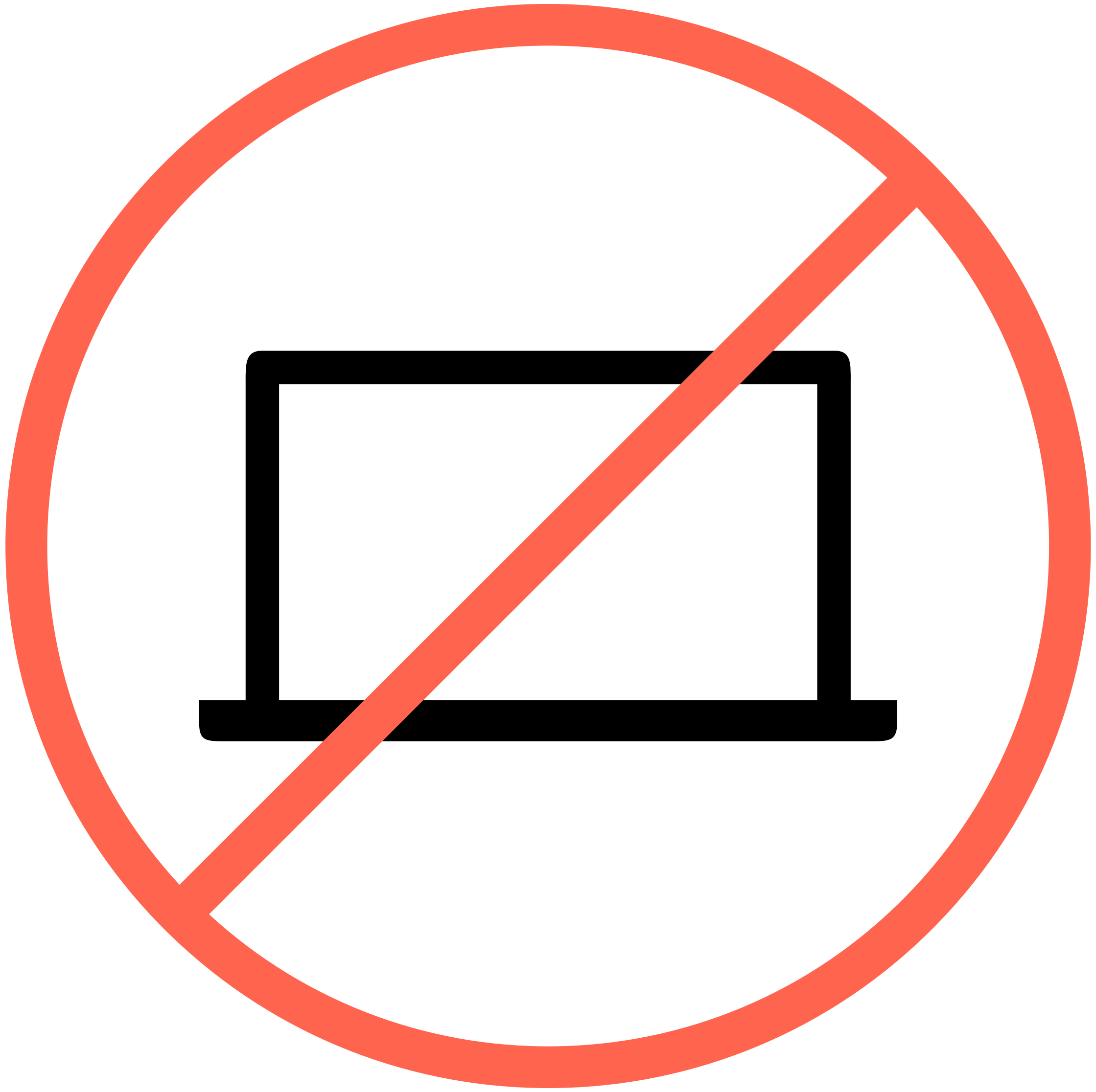
- **You bring a laptop computer and produce a Jupyter notebook** that answers a series of questions
- Each quiz is 80 minutes
- Open notes, open internet, **closed to collaboration**
- You are responsible for making sure your laptop has a compute environment set up appropriately and has enough battery life (or you sit close to a power outlet)
- Late exams will *not* be accepted
- **Quiz 1:** Feb 7 (recitation slot)
- **Quiz 2:** Feb 28 (recitation slot)

# 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

Tianyu
Huang

Xiaobin
Shen

Sachin K
Sunny

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?

WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia
Wikipedia store

Interaction

Help
About Wikipedia
Community portal
Recent changes
Contact page

Tools

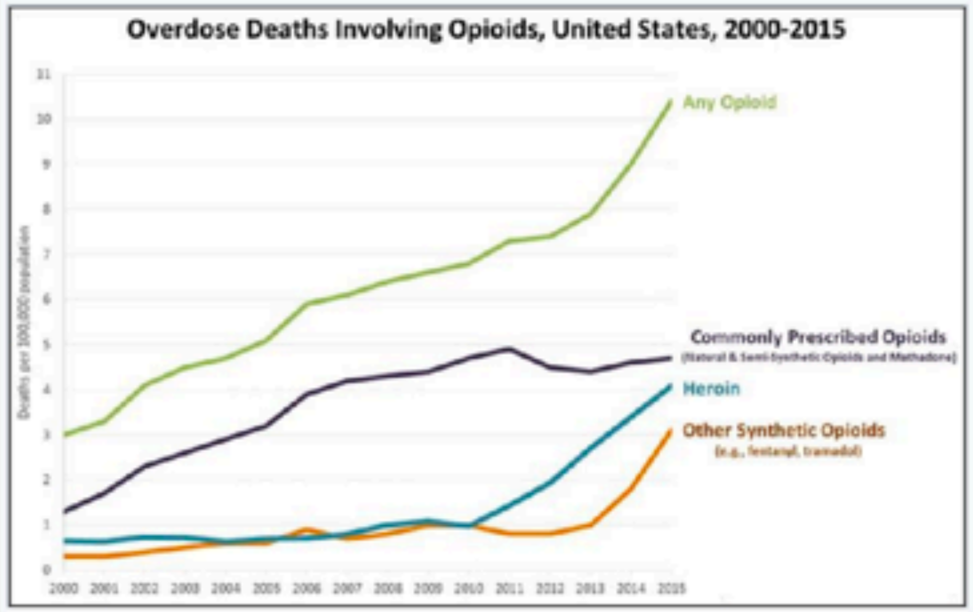
Not logged in Talk Contributions Create account Log in

Article Talk Read Edit View history Search Wikipedia

# 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.[1]

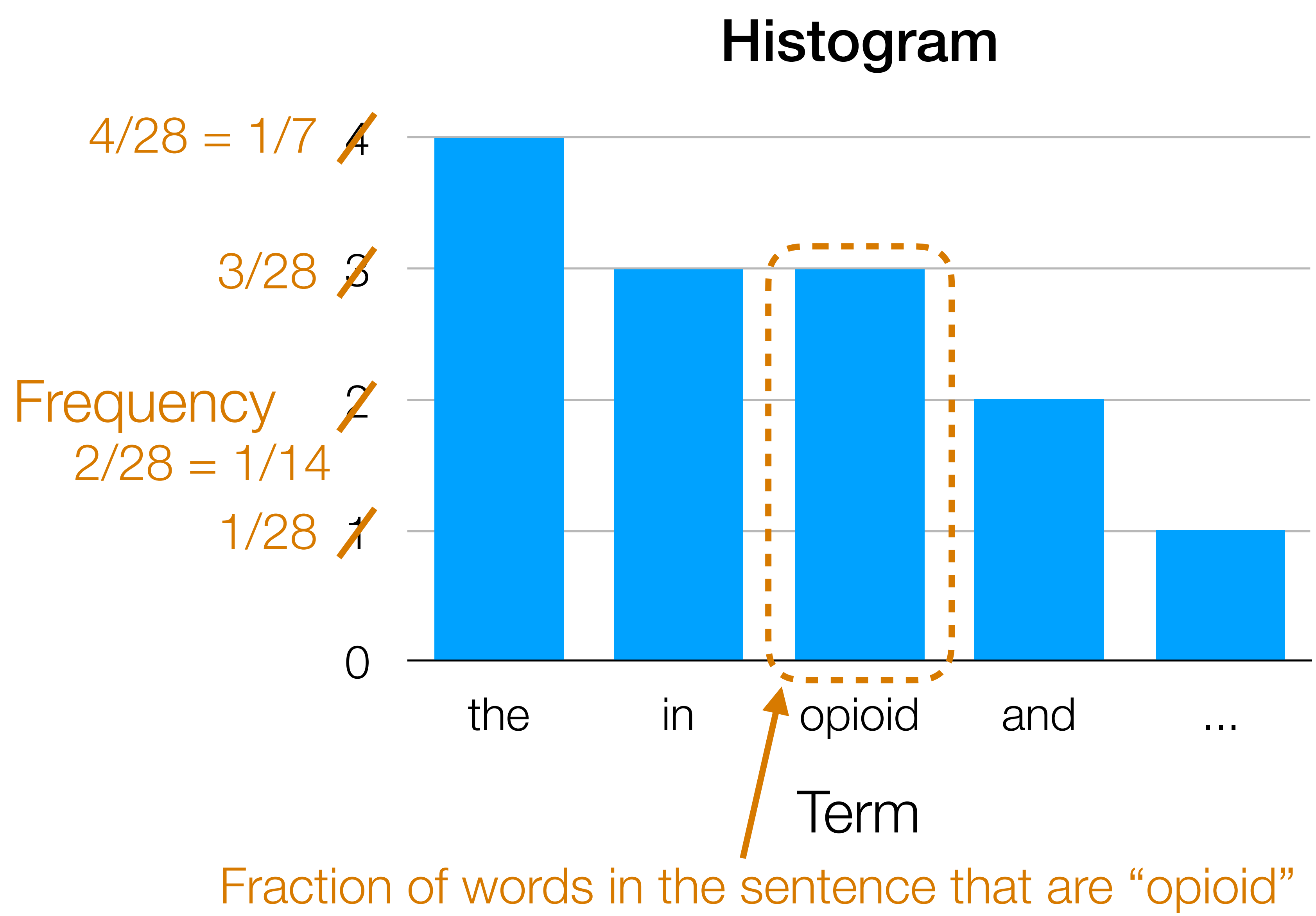
Source: Wikipedia, accessed 10/16/2017

**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*

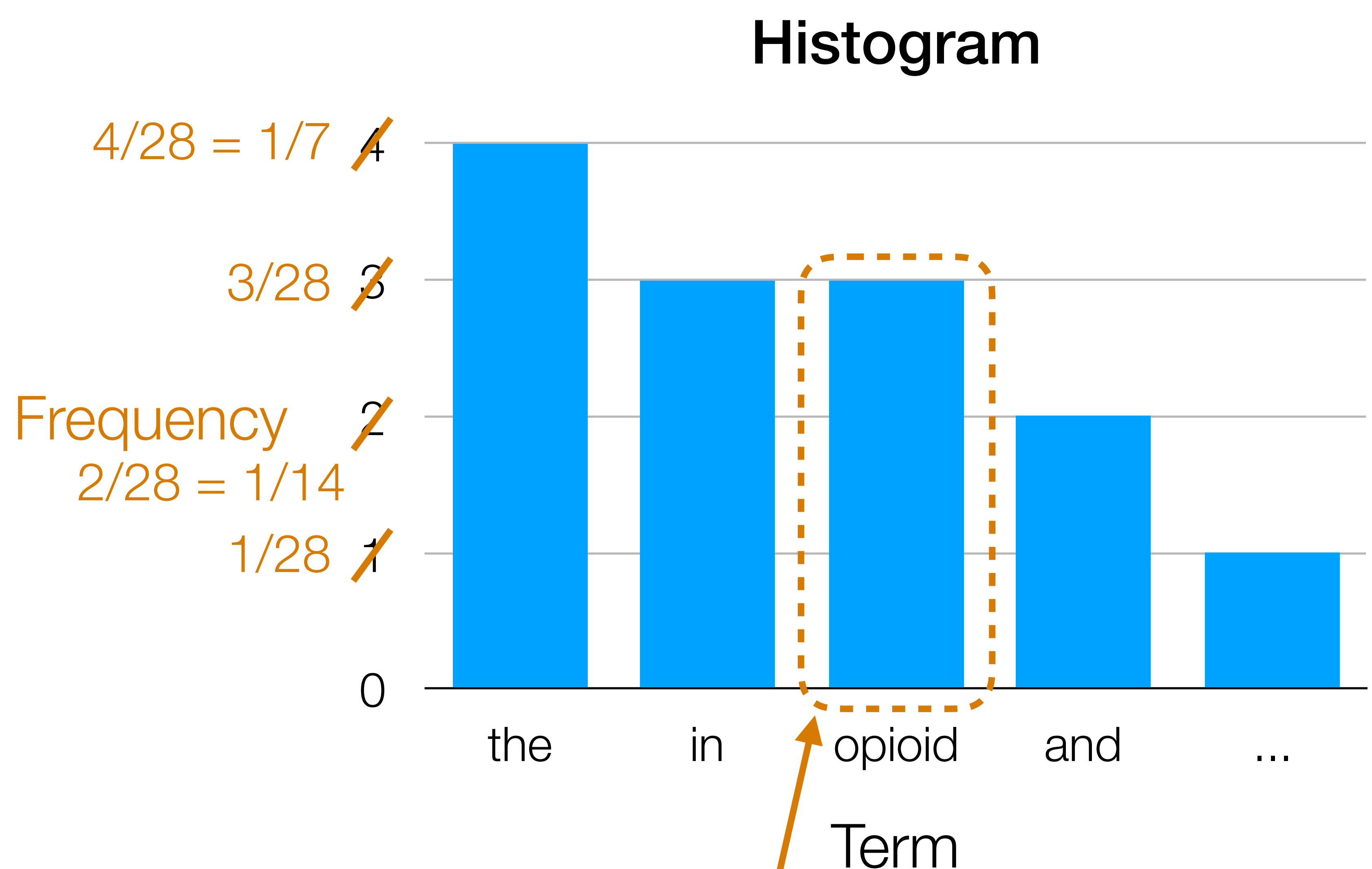


**Term frequencies**

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

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*



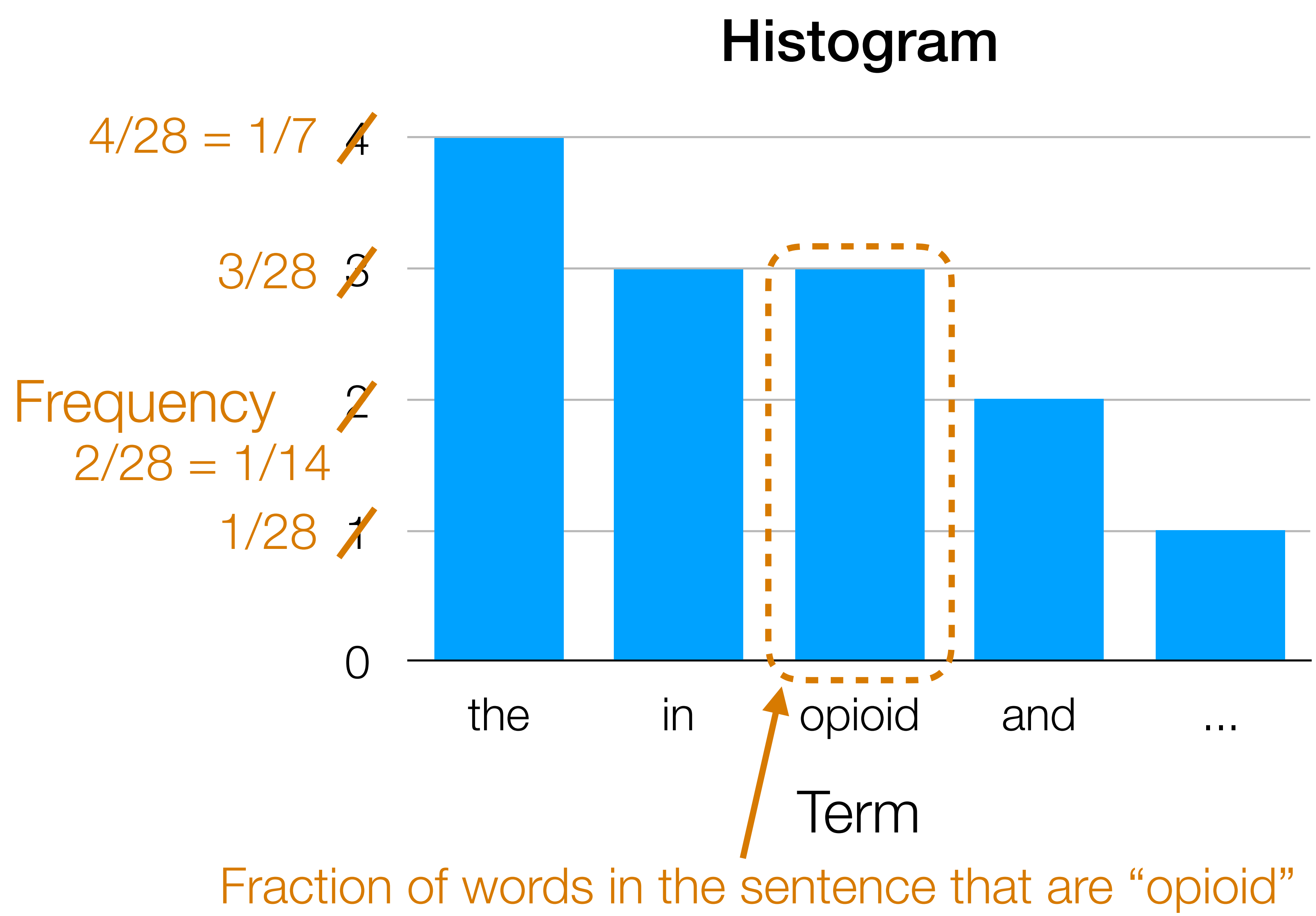

Fraction of words in the sentence that are “opioid”

**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

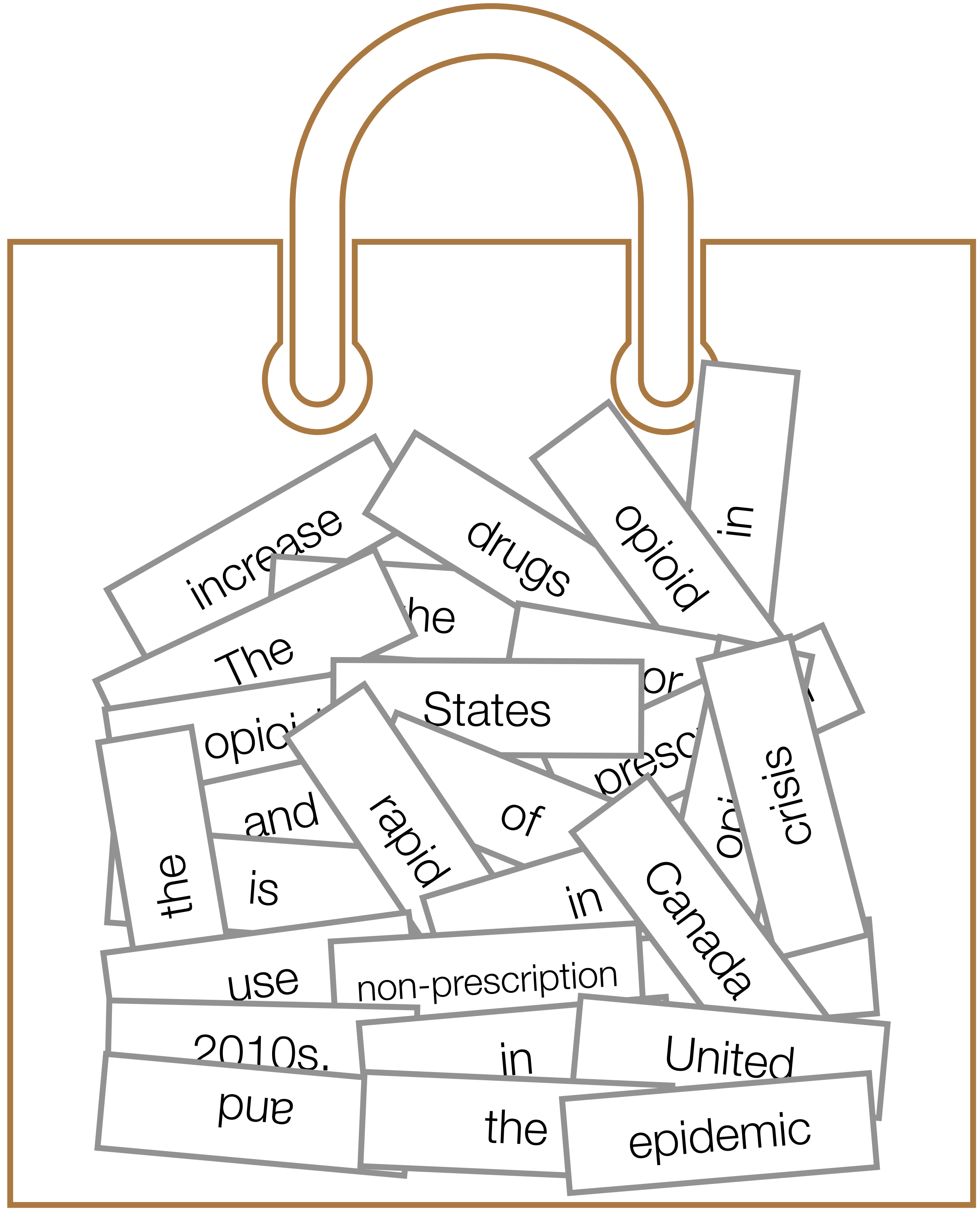
increase the drugs opioid in The States or prescription opioid and of is rapid in opioid crisis the use non-prescription Canada 2010s. in United and the epidemic the

*Total number of words in sentence: 28*



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)

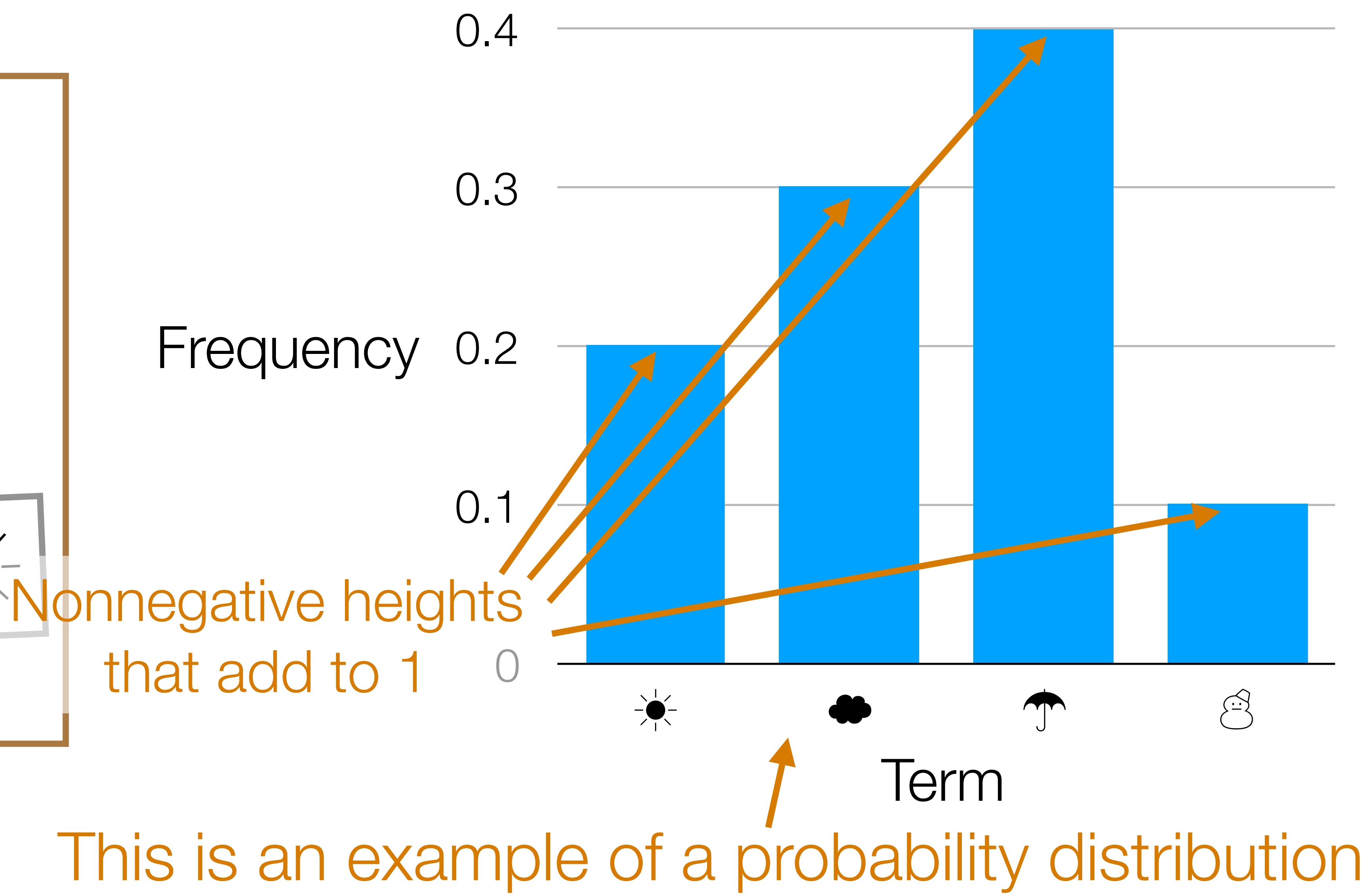
  What does the *ctf* of "opioid" for all of Wikipedia refer to?

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": ☀☂☁☁☁☂☃☂☂☀

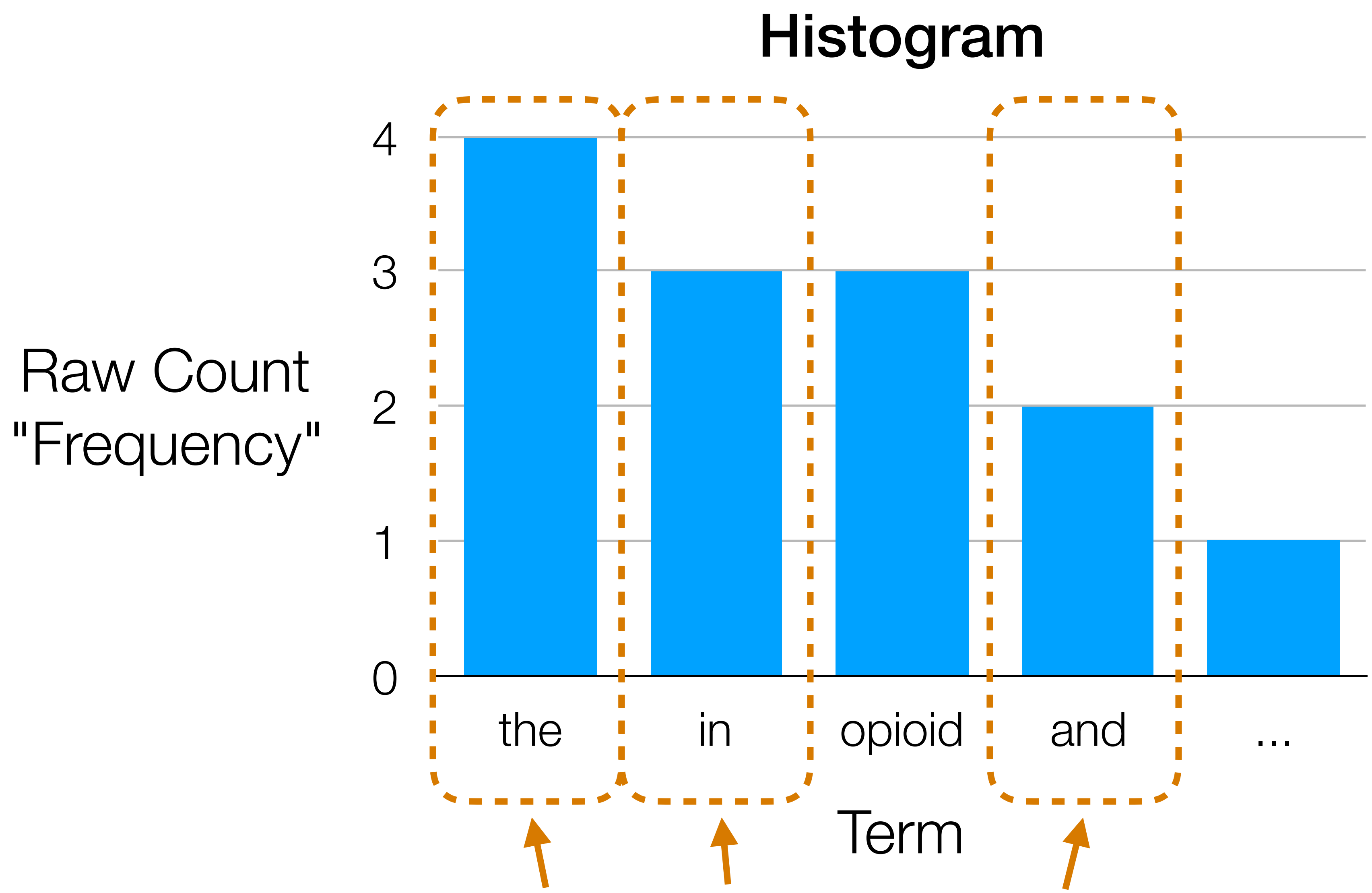



Probability distributions will appear throughout the course and are a **key component** to the success of many modern AI methods

# 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?

<u>**Term frequencies**</u>
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

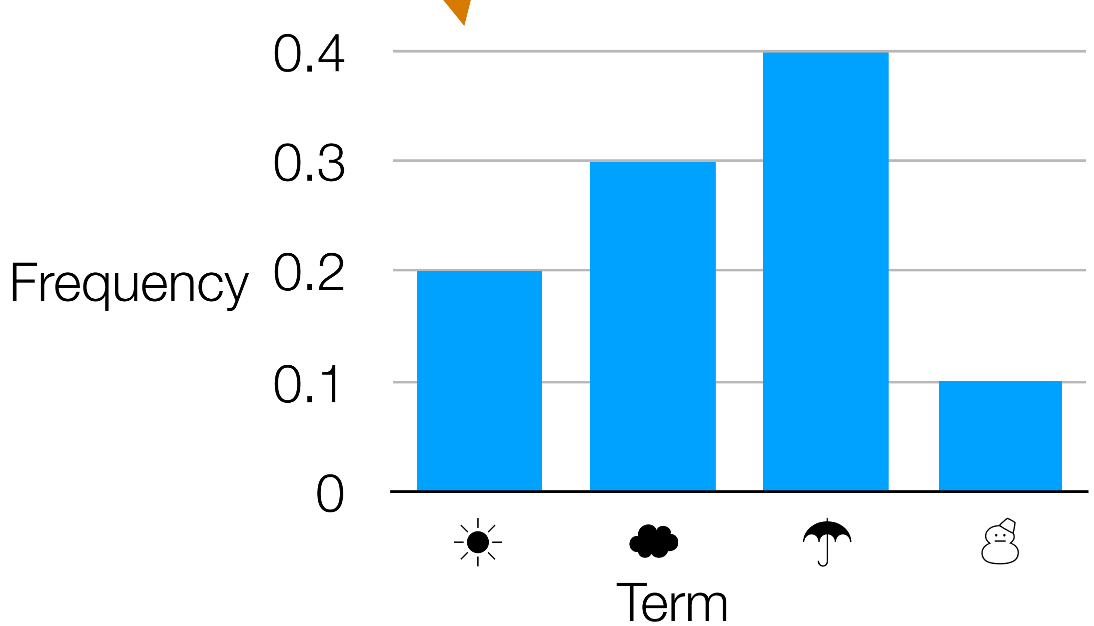
# **The** spaCy **Python Package**

Demo

# Recap: Basic Text Analysis

- Represent text in terms of “features”
  (such as how often each word/phrase appears)
  - 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

# Finding Possibly Related Entities

## Elon Musk's Tesla Powerwalls Have Landed in Puerto Rico

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

*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.

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?