Carnegie Mellon University
Heinz College

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

George Chen

Big Data

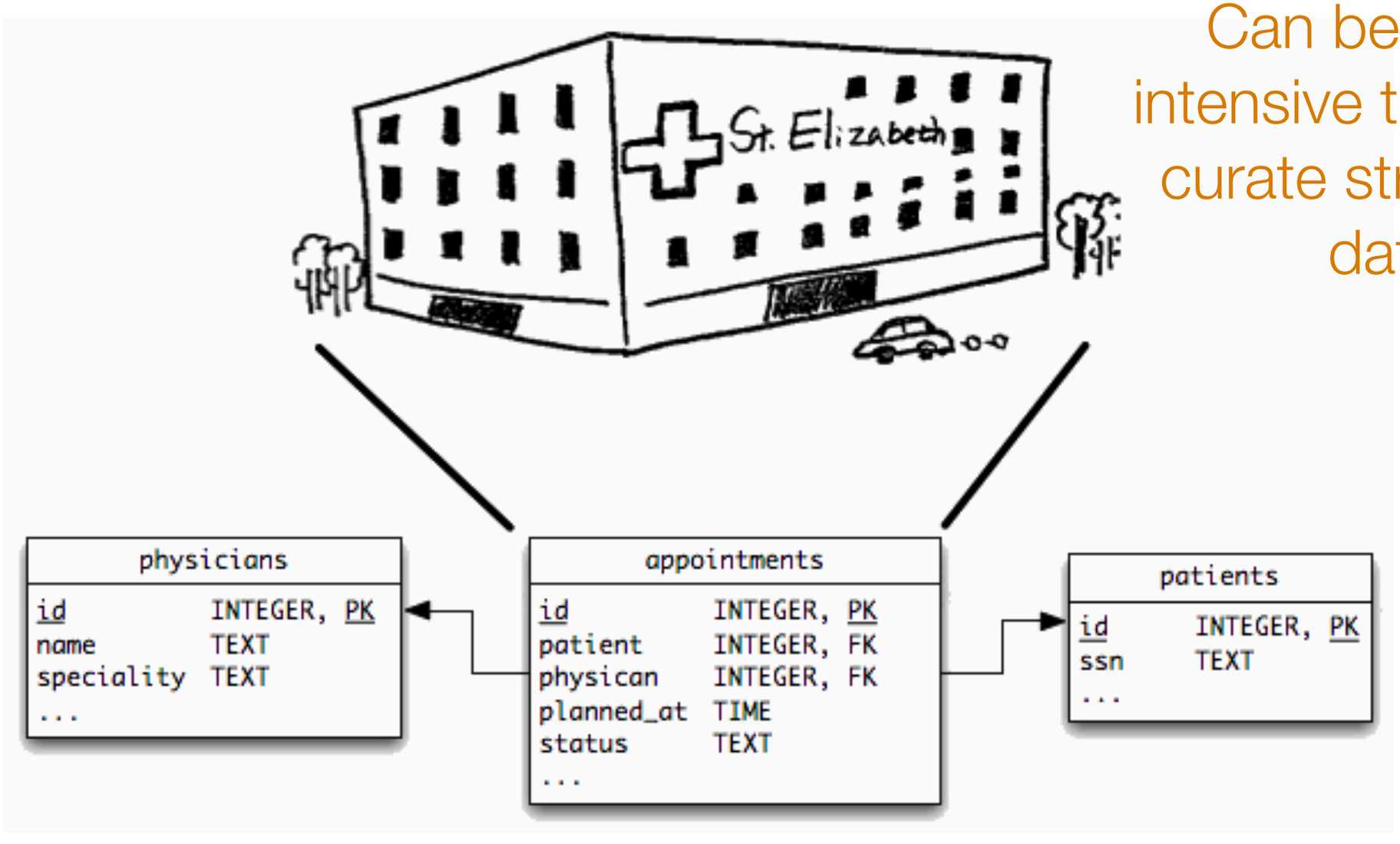
We're now collecting data on virtually every human endeavor

How do we turn these data into actionable insights?

Two Types of Data

Structured Data

Well-defined elements, relationships between elements



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?

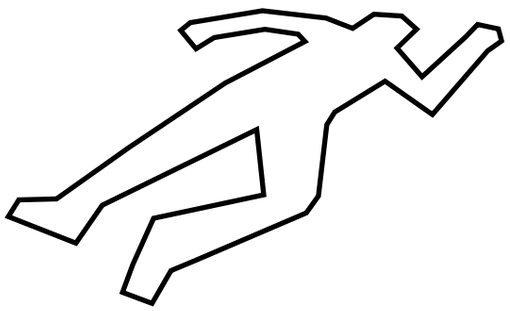
Data

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

Image source: African Reporter

Unstructured Data Analysis

Question



The dead body

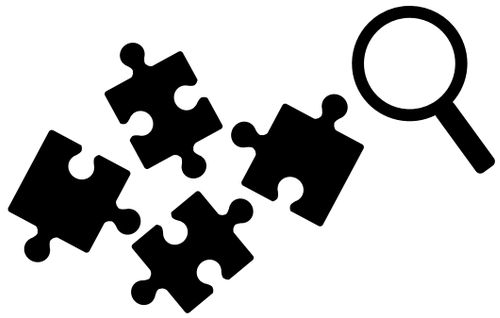
This is provided by a practitioner

Data

The evidence

Some times you have to collect more evidence!

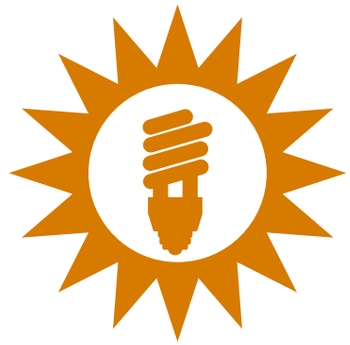
Finding Structure



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

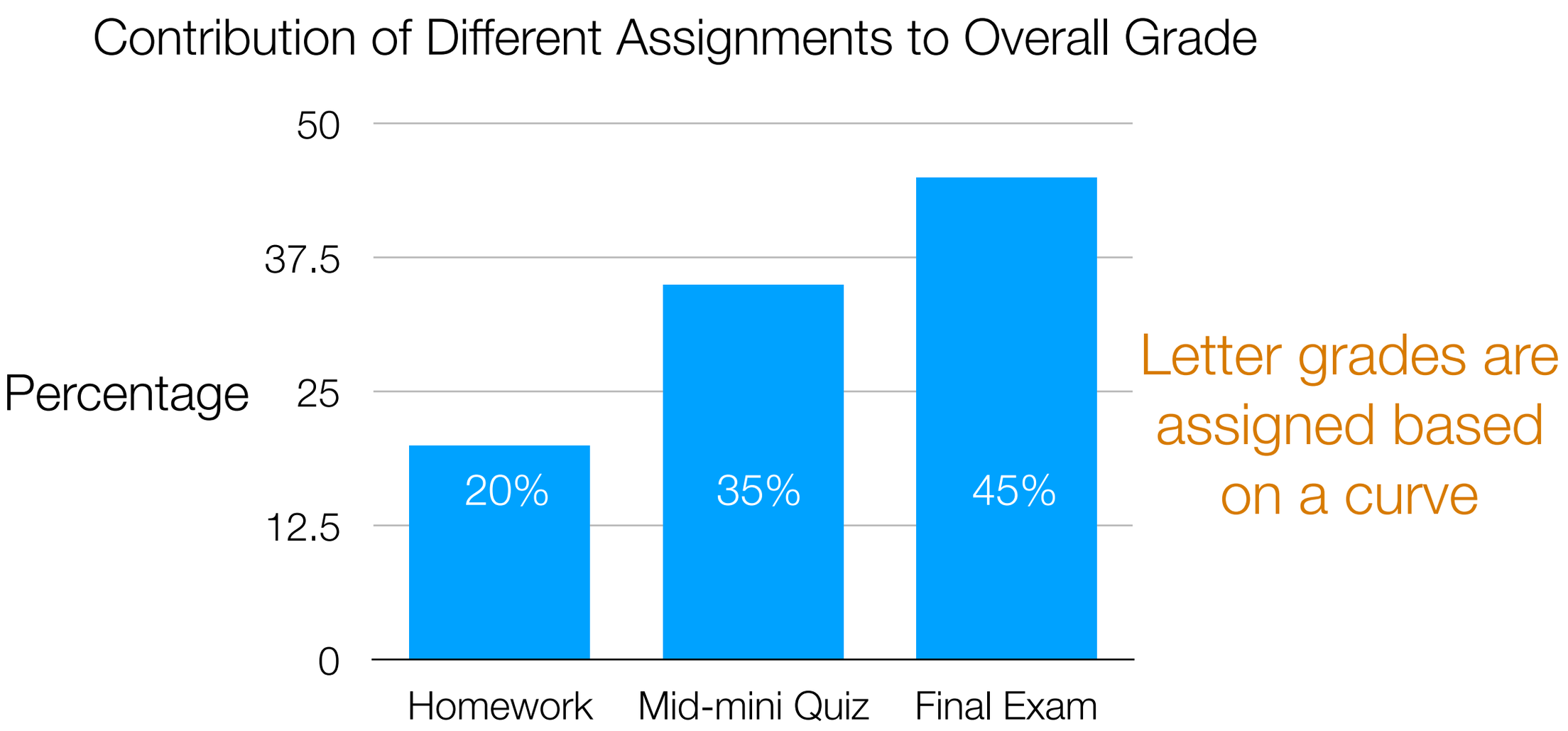
Course Goals

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



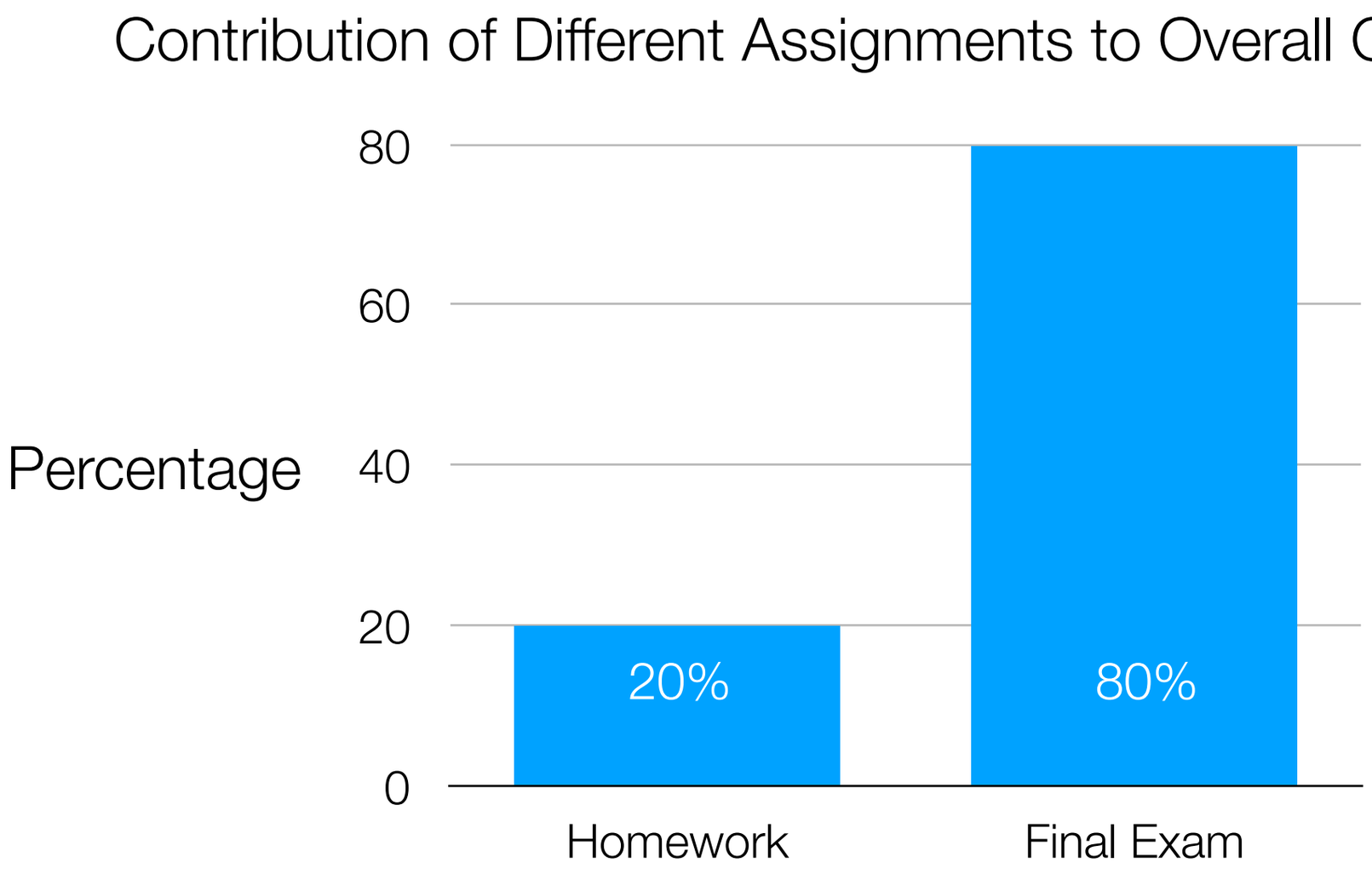
Letter grades are assigned based on a curve

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:



Homework vs Exam Grading

- We will only grade part of your homework for accuracy, and the rest on effort
    - Warning: you must have run your code already in Jupyter notebooks (more about this later)
- Exams are graded entirely on accuracy

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

- 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)**

canvas

piazza

Computing Environment

- We will be using **Anaconda (Python 3.6 version)**
  https://www.anaconda.com/
  As of now, do not use the Python 3.7 version!

- 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:

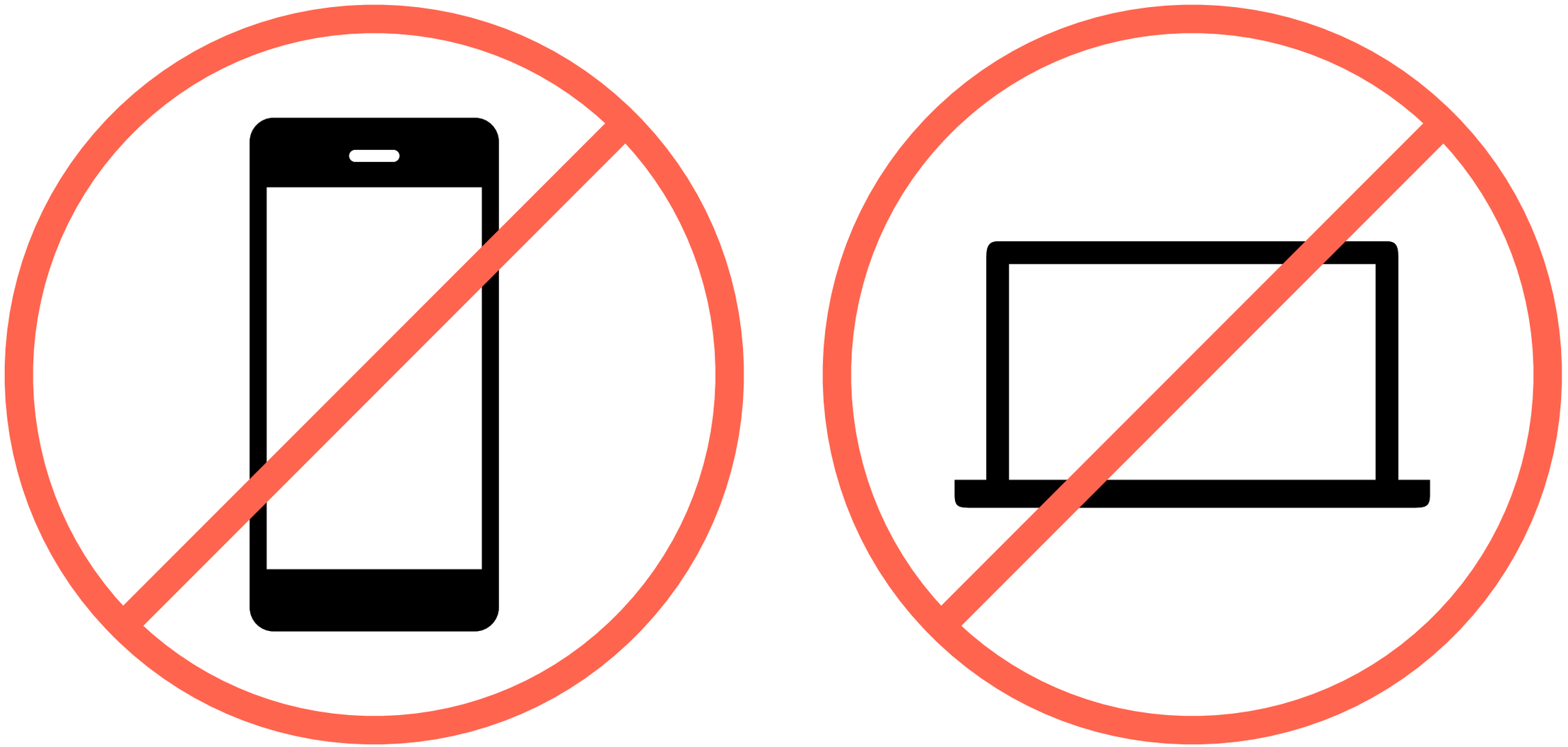
- **You have to bring a laptop computer and produce a Jupyter notebook** that answers a series of questions
- No collaboration (obviously)
- You are responsible for making sure your laptop has a compute environment set up appropriately and has enough battery life (or you sit close to a power outlet)
- Late exams will *not* be accepted
- **Quiz:** Friday Feb 15 at usual recitation time/location
- **Final:** Friday Mar 1 at usual recitation time/location

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

Emaad Manzoor

Yucheng Huang

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?

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

Interaction

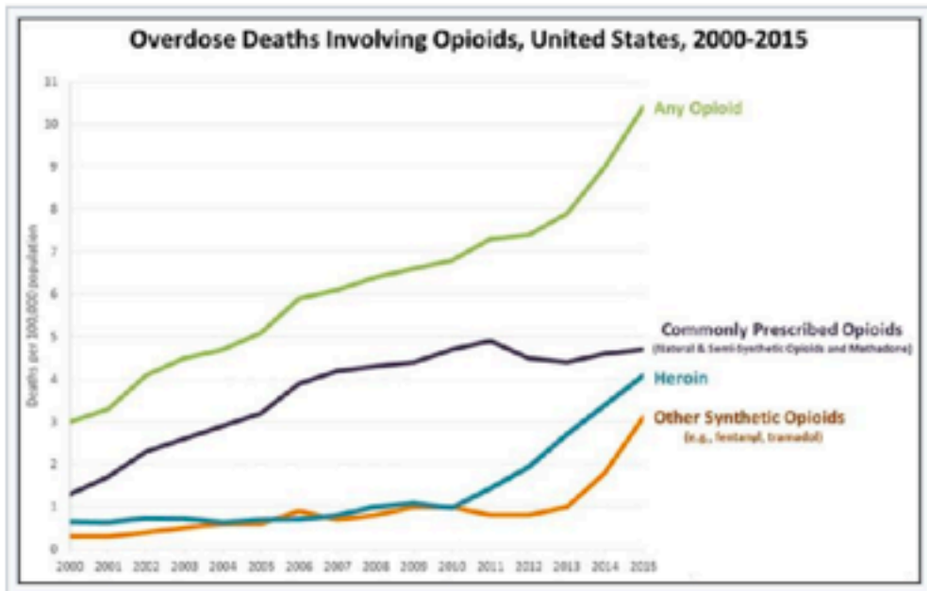
Help
About Wikipedia
Community portal
Recent changes
Contact page

Tools

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

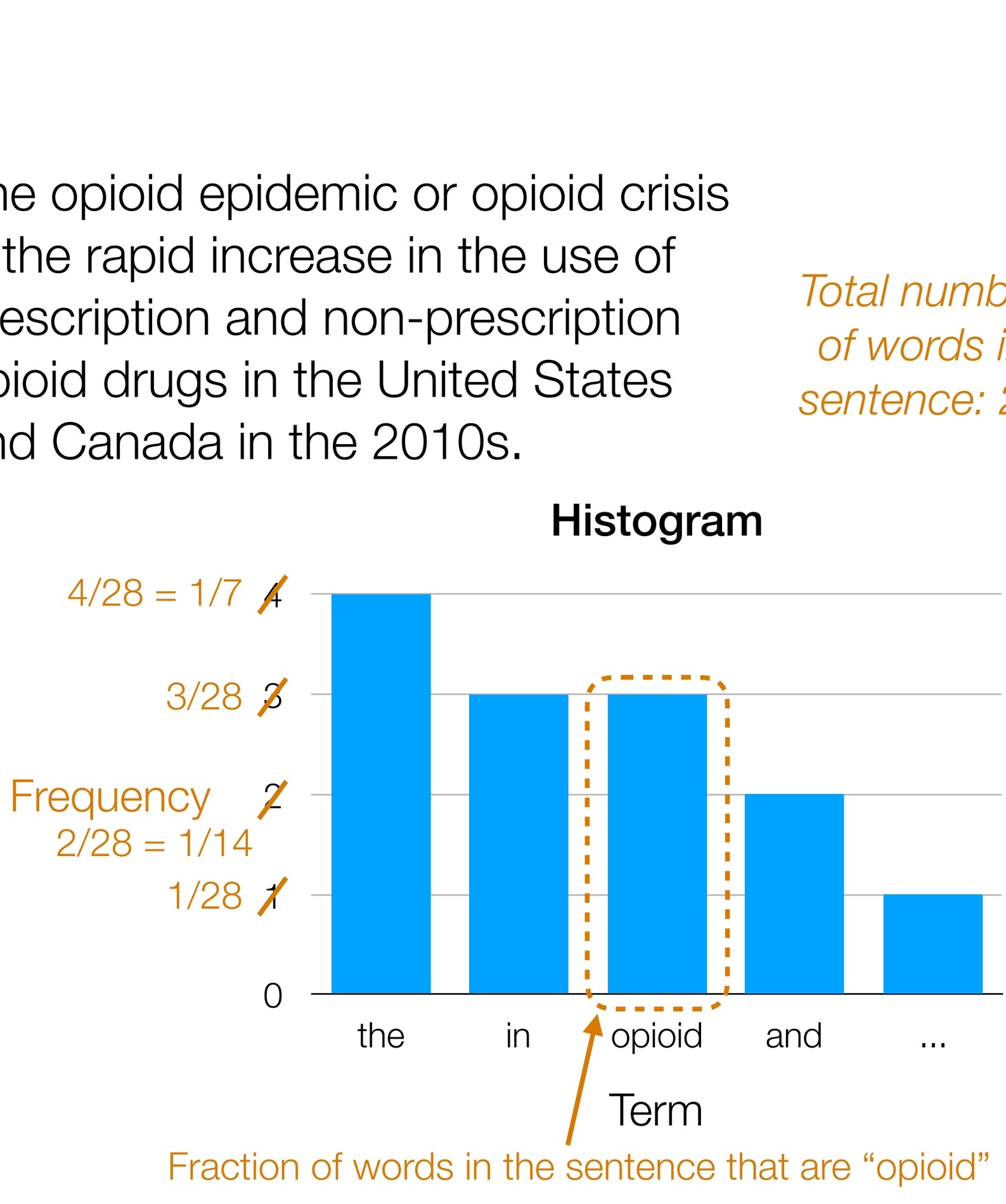
The opioid epidemic or opioid crisis is the rapid increase in the use of prescription and non-prescription opioid drugs in the United States and Canada in the 2010s.

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

Histogram



Fraction of words in the sentence that are “opioid”

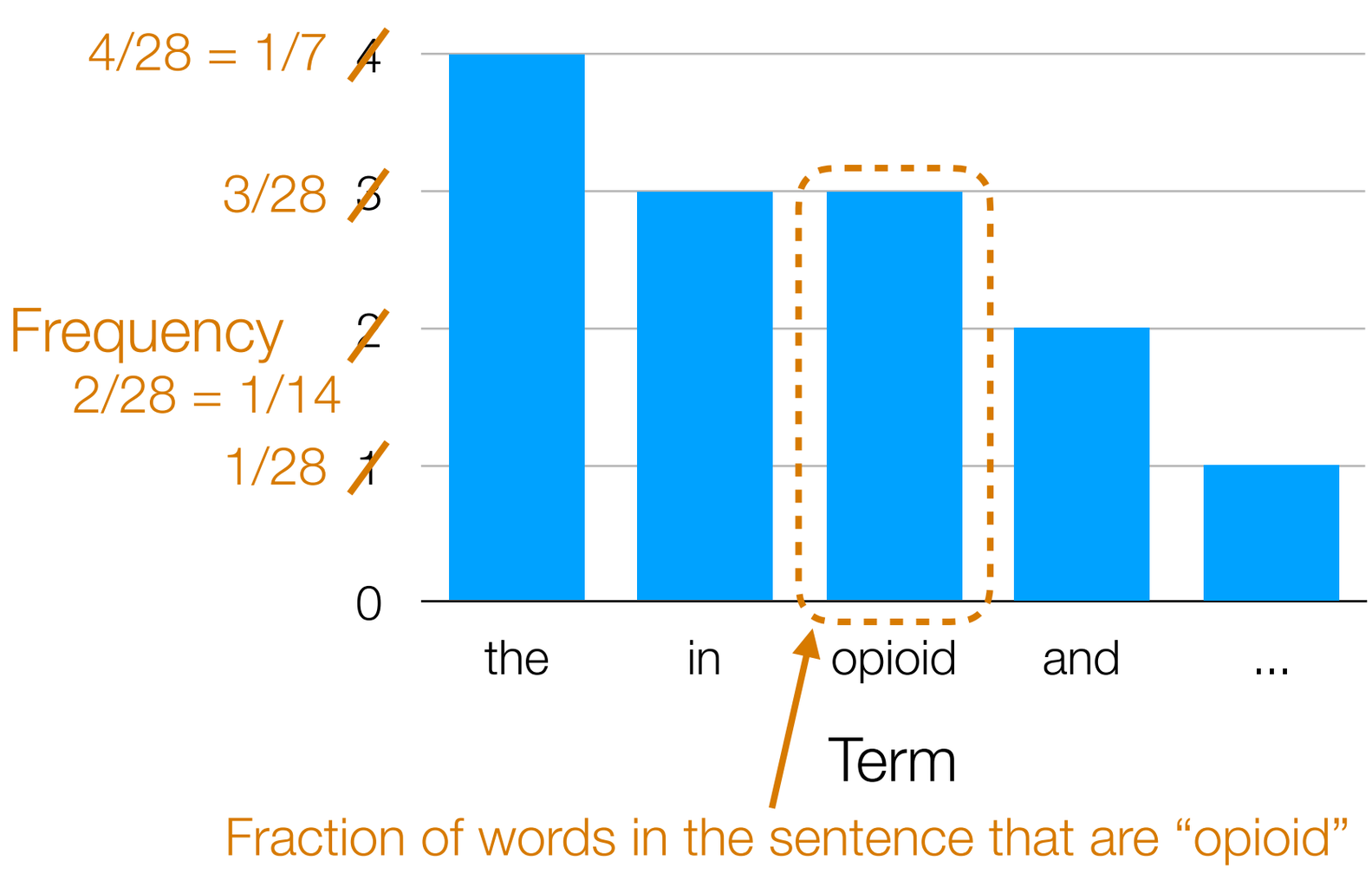
opioid The epidemic or opioid crisis is the rapid increase in the use of prescription and non-prescription opioid drugs in the United States and Canada in the 2010s.

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

Histogram



Fraction of words in the sentence that are “opioid”

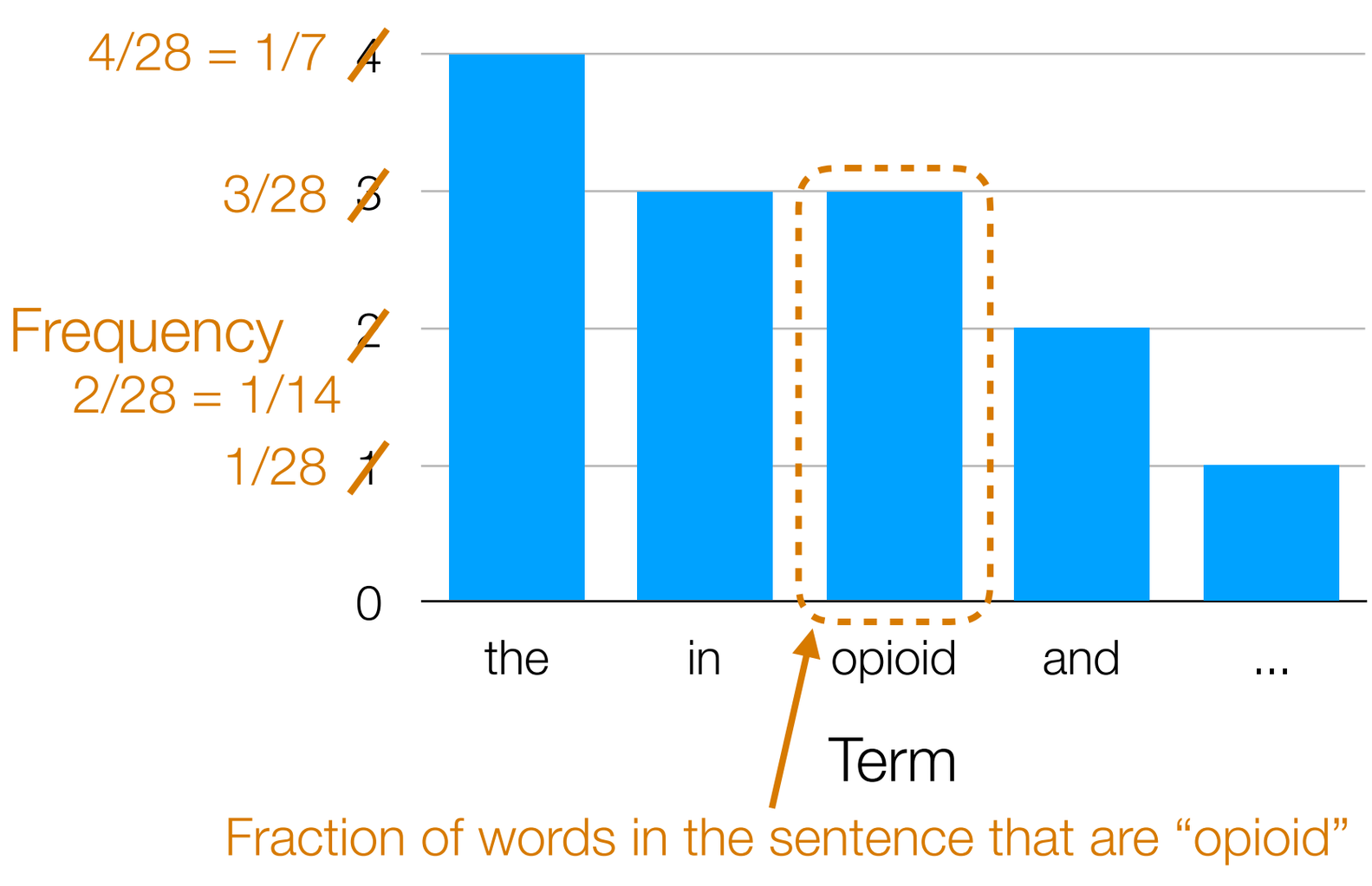
increase the drugs opioid in The States or prescription opioid and of is rapid in opioid crisis the use non-prescription Canada 2010s. in United and the epidemic the

Total number of words in sentence: 28

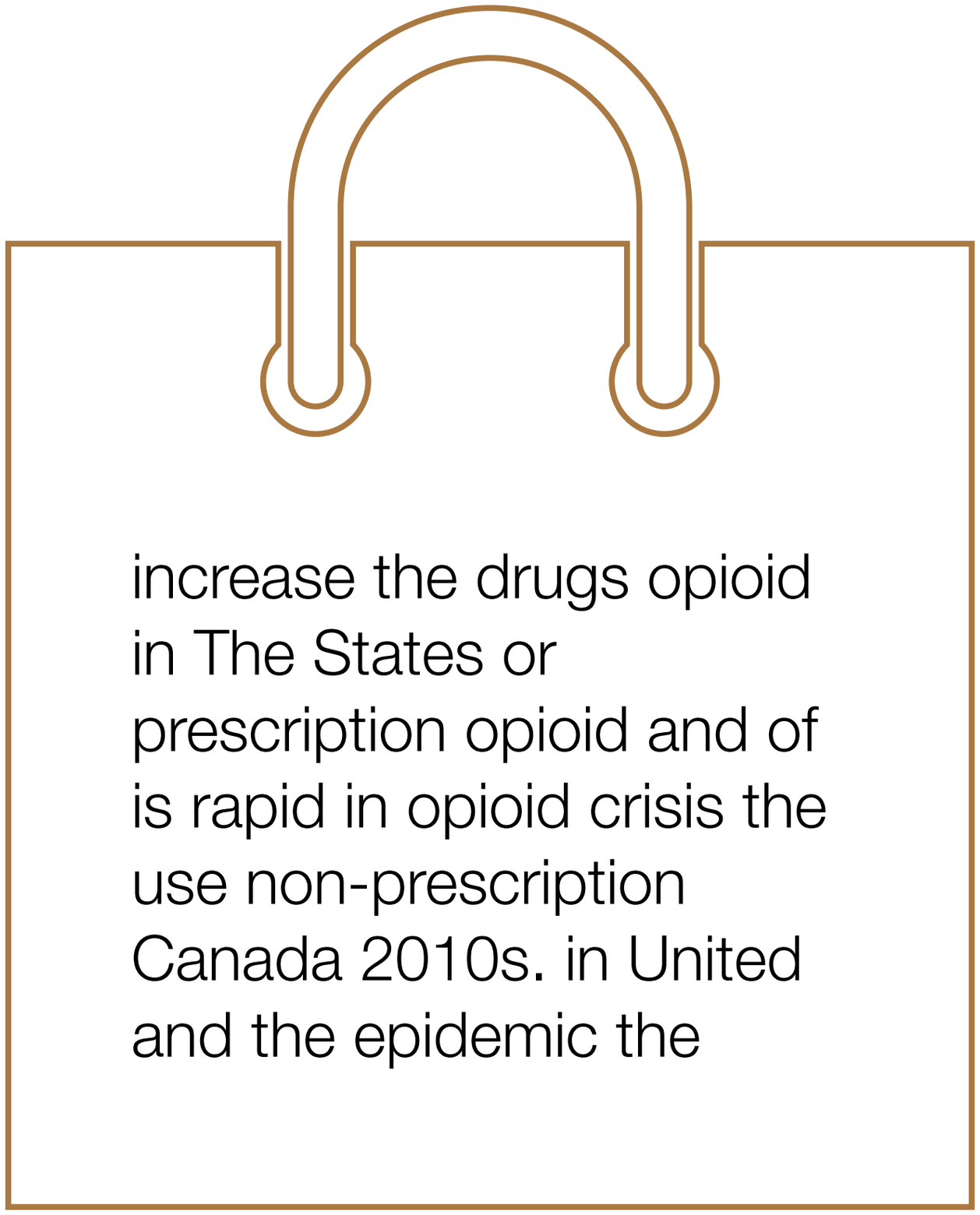
Term frequencies

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

Histogram

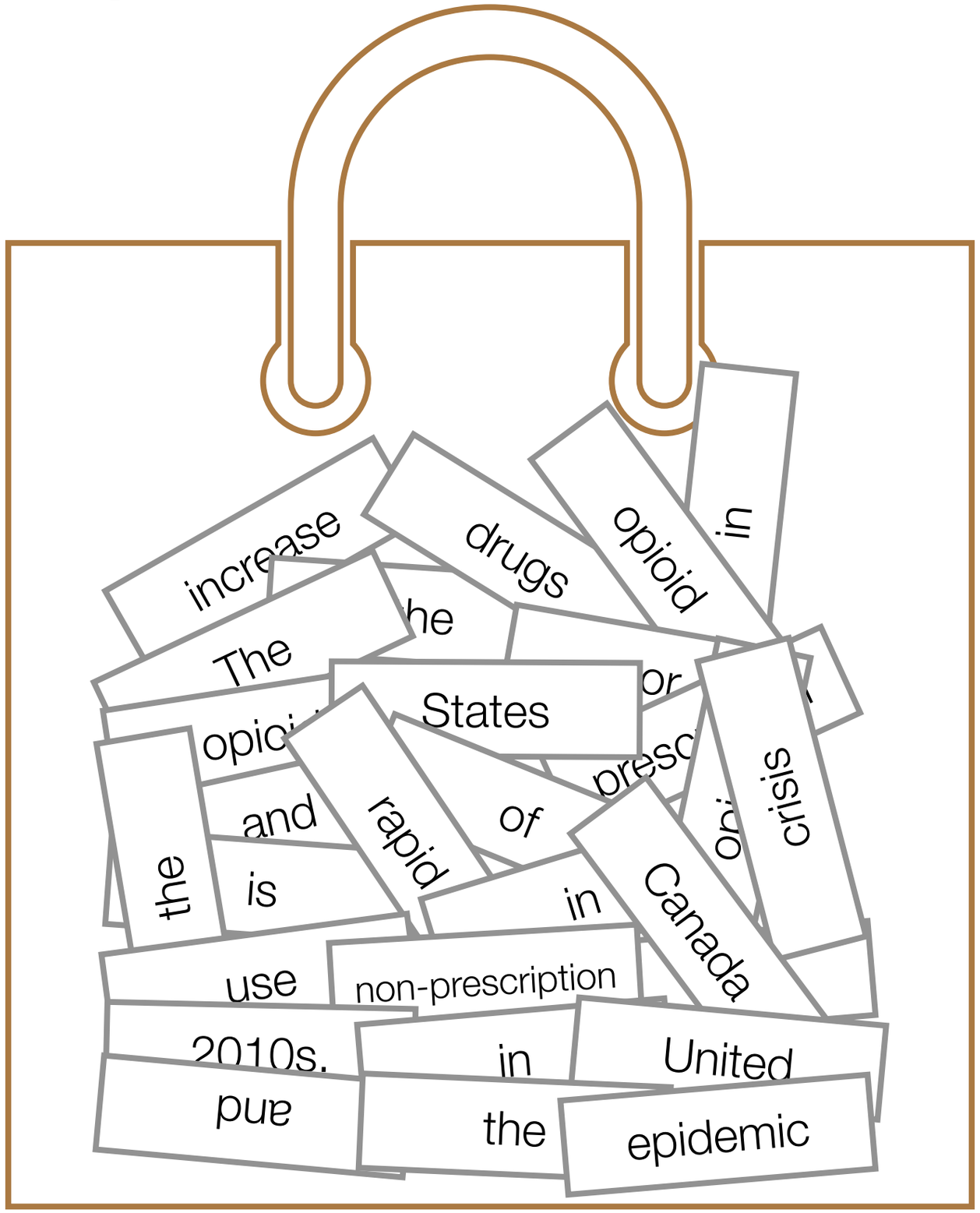


Fraction of words in the sentence that are “opioid”



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)

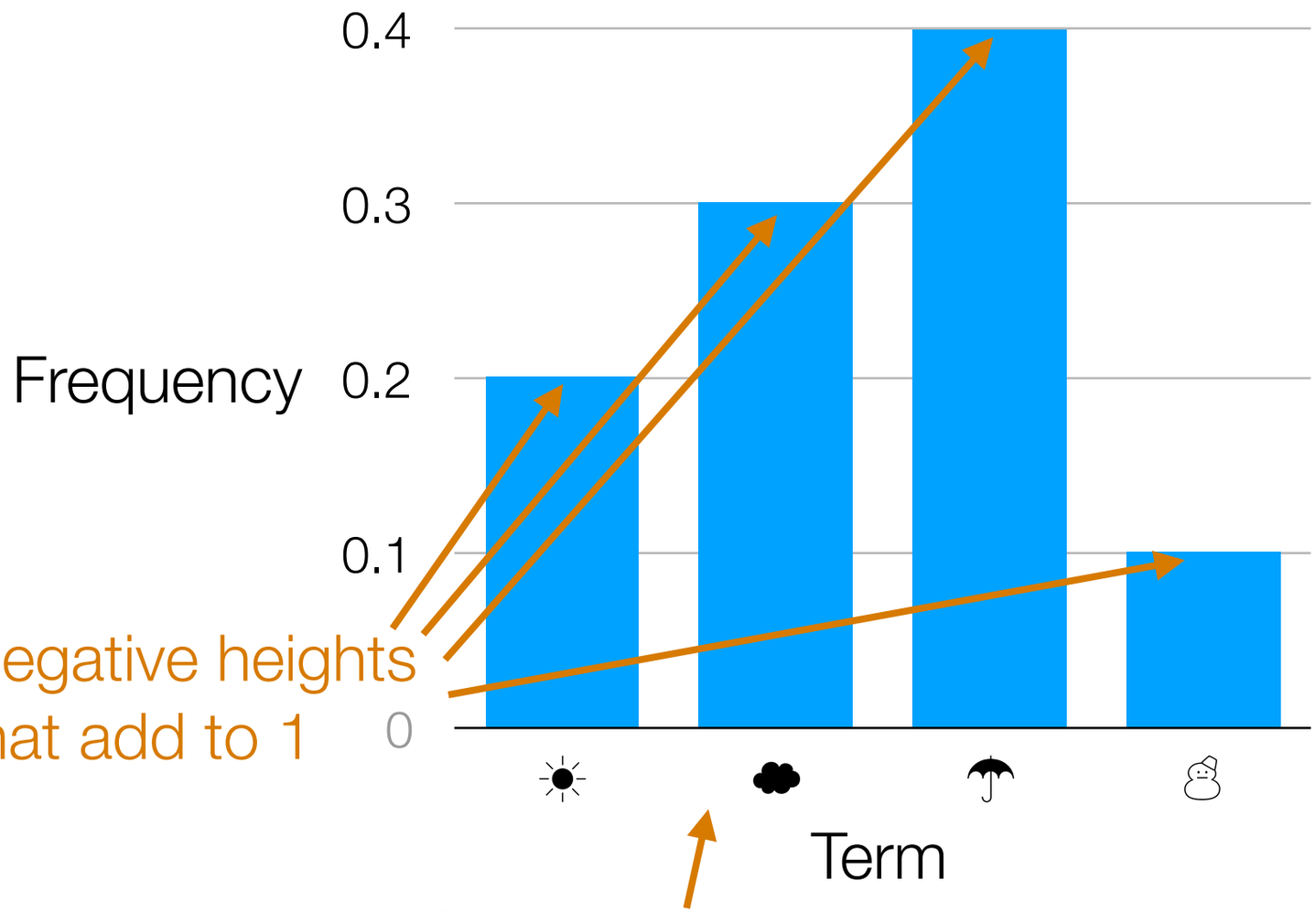
  What does the *ctf* of "opioid" for all of Wikipedia refer to?

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