Carnegie Mellon University
Heinz College

95-865 Unstructured Data Analytics

Lecture 1: Course overview, analyzing text using frequencies

George Chen

What is this course about?

Big Data

We're now collecting data on virtually every human endeavor

How do we turn data into actionable insights?

Two Types of Data

Structured Data

Well-defined elements, relationships between elements

Patients Table

Patient ID	First Name	Last Name	Middle Initial
0	…	…	…
1	…	…	…
2	…	…	…

Doctors Table

Doctor ID	First Name	Last Name	Middle Initial
0	…	…	…
1	…	…	…
2	…	…	…

Appointments Table

Appointment ID	Patient ID	Doctor ID	Start time	End time
0	…	…	…	…
1	…	…	…	…
2	…	…	…	…

Can be labor-intensive to collect/curate structured data

Unstructured Data

No pre-defined model—elements and relationships ambiguous

Common examples:

- Text
- Images
- Videos
- Audio

Often: Want to make decisions using multiple types of unstructured data, or unstructured + structured data

Of course, there *is* structure in "unstructured" data but it is not neatly spelled out for us

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

Just because something *can* be stored as any of these doesn't mean that it must be unstructured!

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?

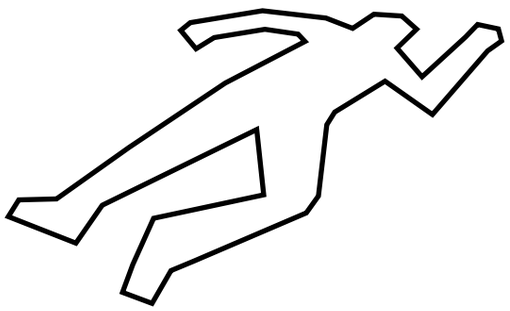
Data

- 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

Image source: African Reporter

Unstructured Data Analysis

Question



The dead body

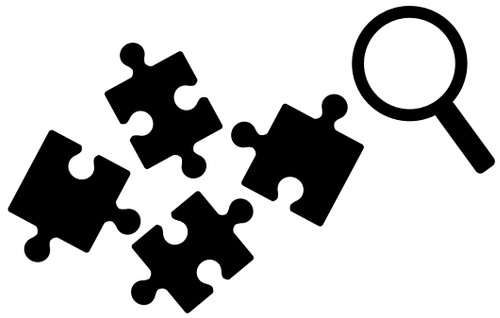
This is provided by a practitioner

Data

The evidence

Some times you have to collect more evidence!

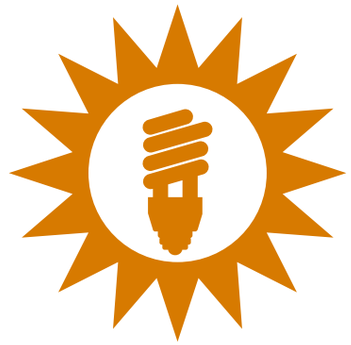
Finding Structure



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 known structure in data

- Basic concepts and how to assess quality of prediction models
- 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!

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 webpage**
http://www.andrew.cmu.edu/user/georgech/95-865/

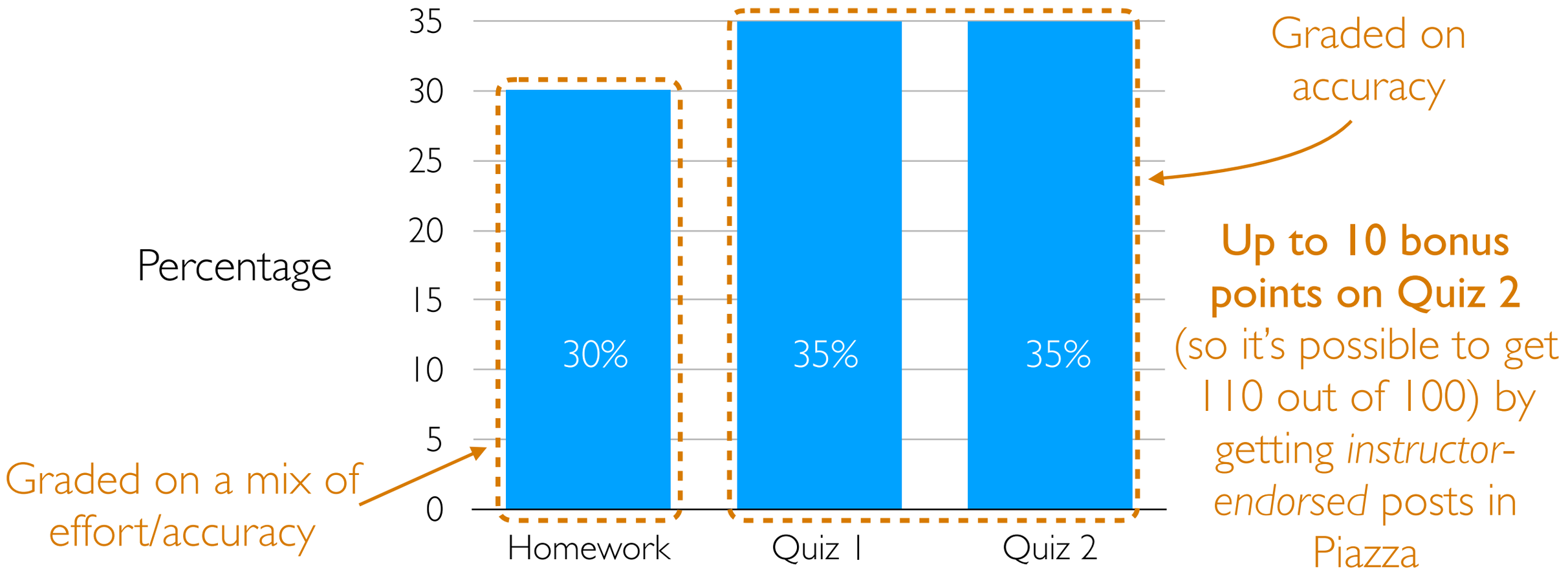
Assignments will be posted and submitted on **Canvas**

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

canvas

Deliverables & Grading

Contribution of Different Assignments to Overall Grade



Letter grades are assigned based on a curve

All assignments involve coding in Python
(popular amongst machine learning/computer science community)

HW3 uses Google Colab for cloud computing
(many real datasets too large to either fit or process on a 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**
  - For code: post smallest snippet, how you know it's buggy (error message/etc), & what you've already tried to resolve the issue
- In the real world, you will unlikely be working alone
  - We encourage discussing concepts
  - 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: 0 on assignment, F in course =(

Pittsburgh/Adelaide Weirdness

This mini, I teach 95-865 to CMU's Pittsburgh 🇺🇸 and Adelaide 🇦🇺 campuses

- A4/B4/K4 sections have identical homework due dates and will share the same Piazza forum
- Pittsburgh has April 7 (Thursday) & April 8 (Friday) off…
- Adelaide has April 15 (Friday) off and also has 1 fewer week of instruction…
- I will physically be in Adelaide the week of April 18-22 and will teach remotely back to Pittsburgh (in middle of the night)
- See calendar on course webpage for how the lectures get roughly synced across A4/B4/K4 sections

The Two Quizzes (in-person for A4/B4)

Format:

- **You 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 that your laptop has a compute environment set up appropriately (check that course demos & HW run without issues) and has working internet
- Late exams will *not* be accepted
- **Quiz 1:** Friday Apr 1 during recitation slot
- **Quiz 2:** Friday Apr 29 during 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
- There is no need to tell us if you're using a late day or not (we'll figure it out from submission timestamps)

Course Staff

TAs:

- Aditya Malani
- Ashlyn Im
- Erick Rodriguez (the only Adelaide TA)
- Luping Fang
- Mingyuan Fu
- Yu Liu
- Ziyuan Liu

Instructor: George Chen

Office hours start next week (we're still sorting out the schedule): details will be posted on course webpage & Canvas

Part I.

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

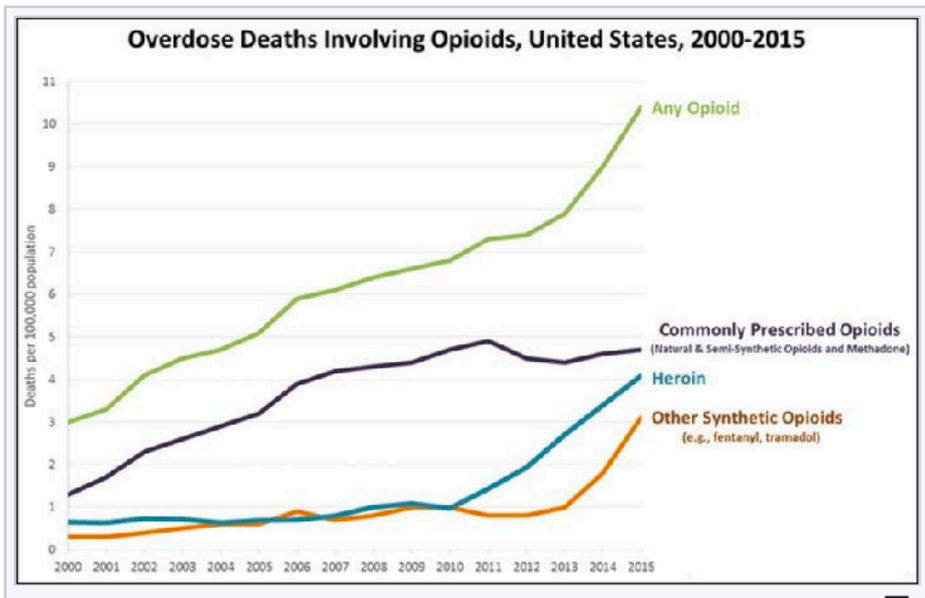
Help
About Wikipedia
Community portal
Recent changes
Contact page

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 October 16, 2017

Term frequencies

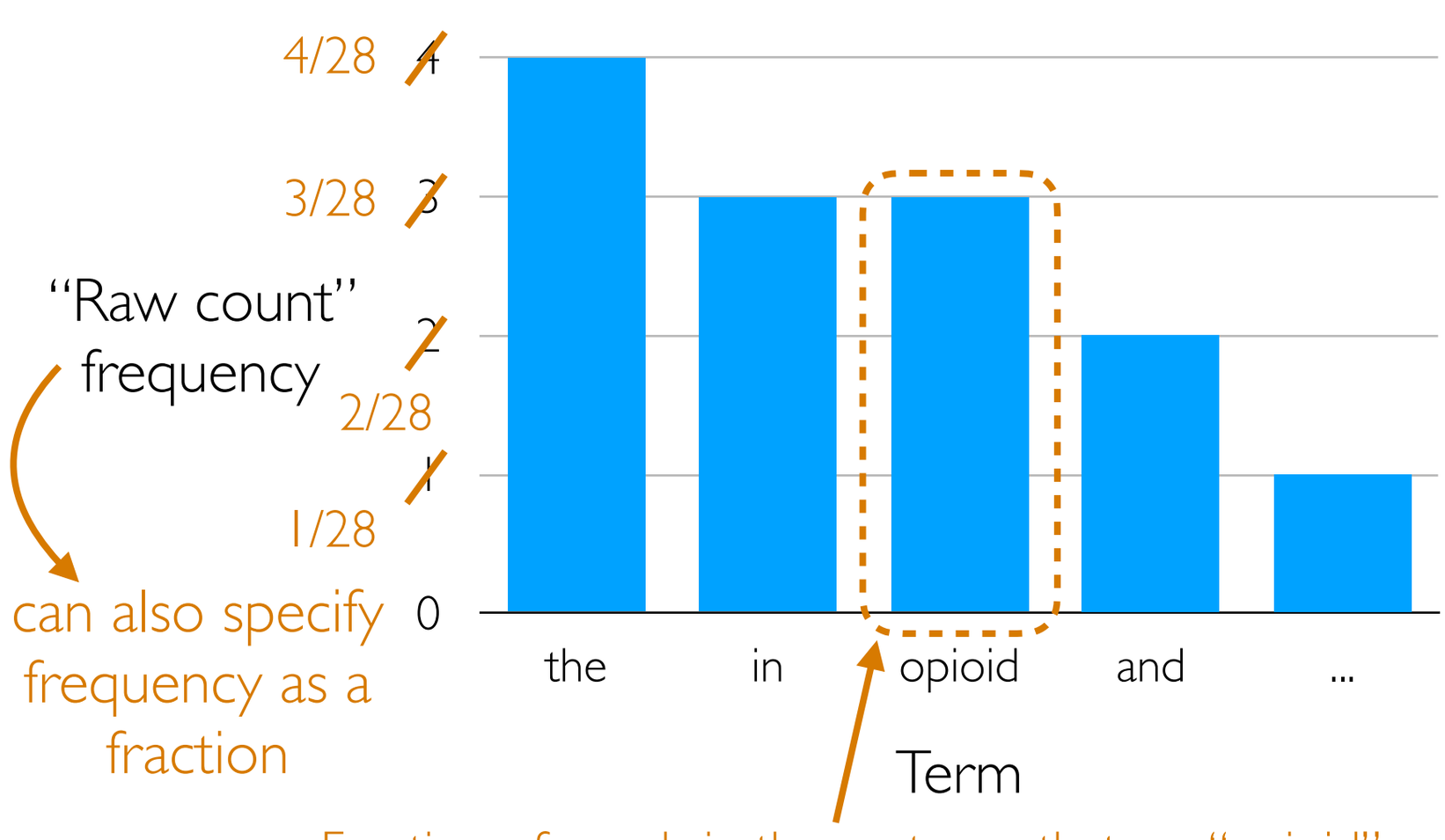
The: 1
opioid: 3
epidemic: 1
or: 1
crisis: 1
is: 1
the: 4
rapid: 1
increase: 1
in: 3
use: 1
of: 1
prescription: 1
and: 2
non-prescription: 1
drugs: 1
United: 1
States: 1
Canada: 1
2010s.: 1

/28
/28
/28
/28
/28
/28
/28
/28
/28
/28
/28
/28
/28
/28
/28
/28
/28
/28
/28
/28
/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

Histogram



"Raw count" frequency

can also specify frequency as a fraction

Fraction of words in the sentence that are "opioid"

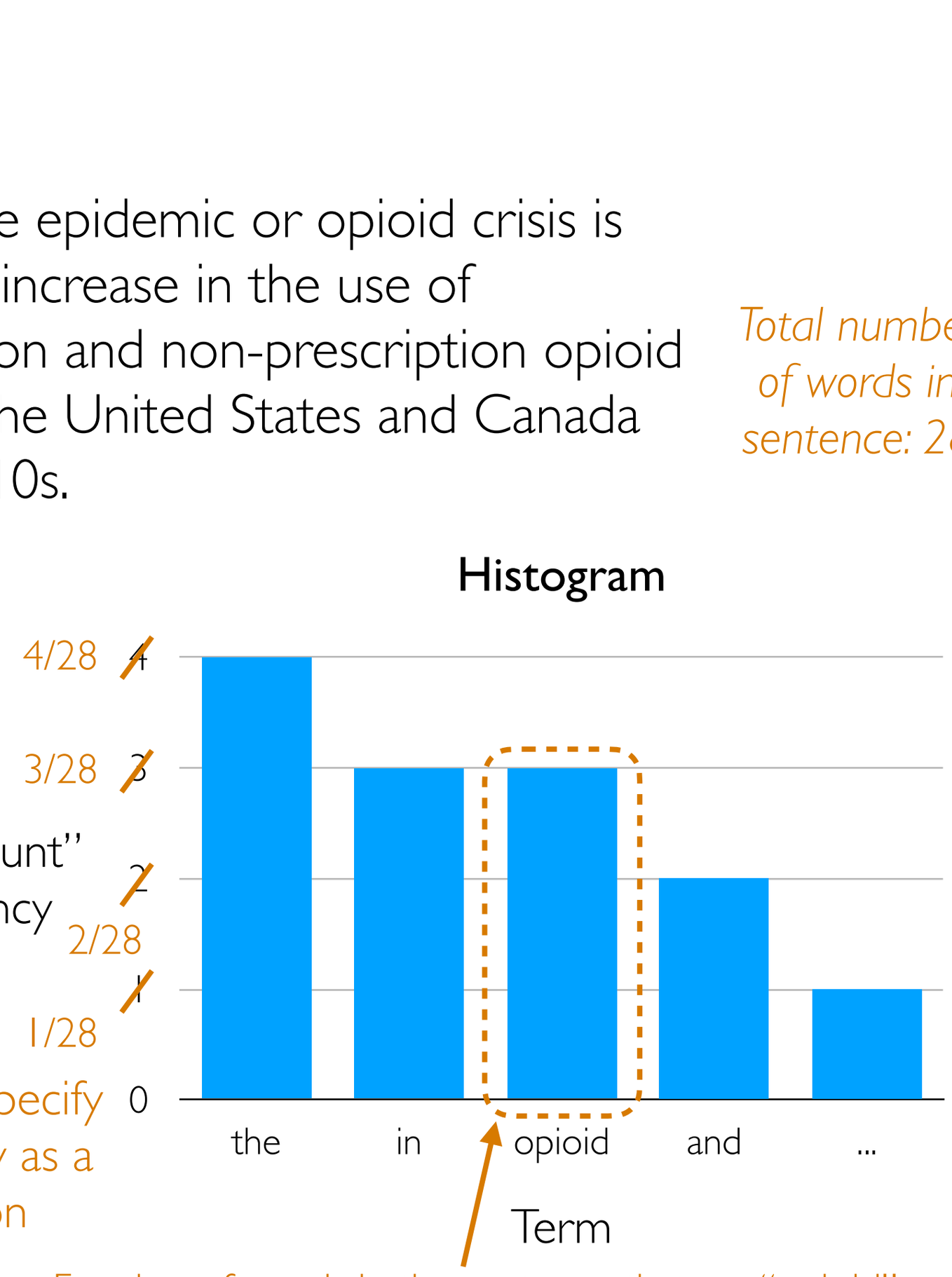
Term frequencies

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

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

Total number of words in sentence: 28

Histogram



"Raw count" frequency

can also specify frequency as a fraction

Fraction of words in the sentence that are "opioid"

Term frequencies

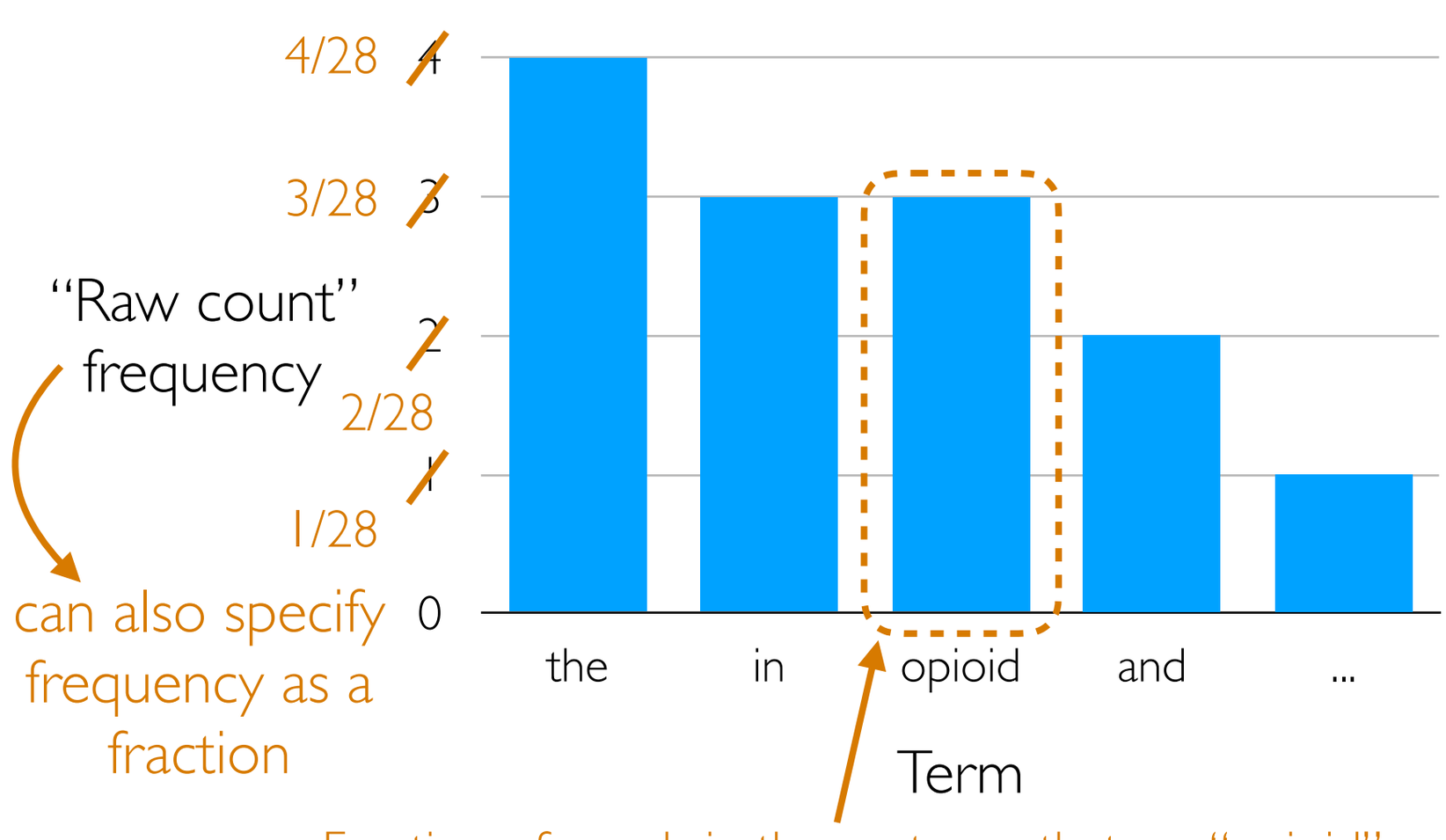
The: 1
opioid: 3
epidemic: 1
or: 1
crisis: 1
is: 1
the: 4
rapid: 1
increase: 1
in: 3
use: 1
of: 1
prescription: 1
and: 2
non-prescription: 1
drugs: 1
United: 1
States: 1
Canada: 1
2010s.: 1

/28
/28
/28
/28
/28
/28
/28
/28
/28
/28
/28
/28
/28
/28
/28
/28
/28
/28
/28
/28
/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

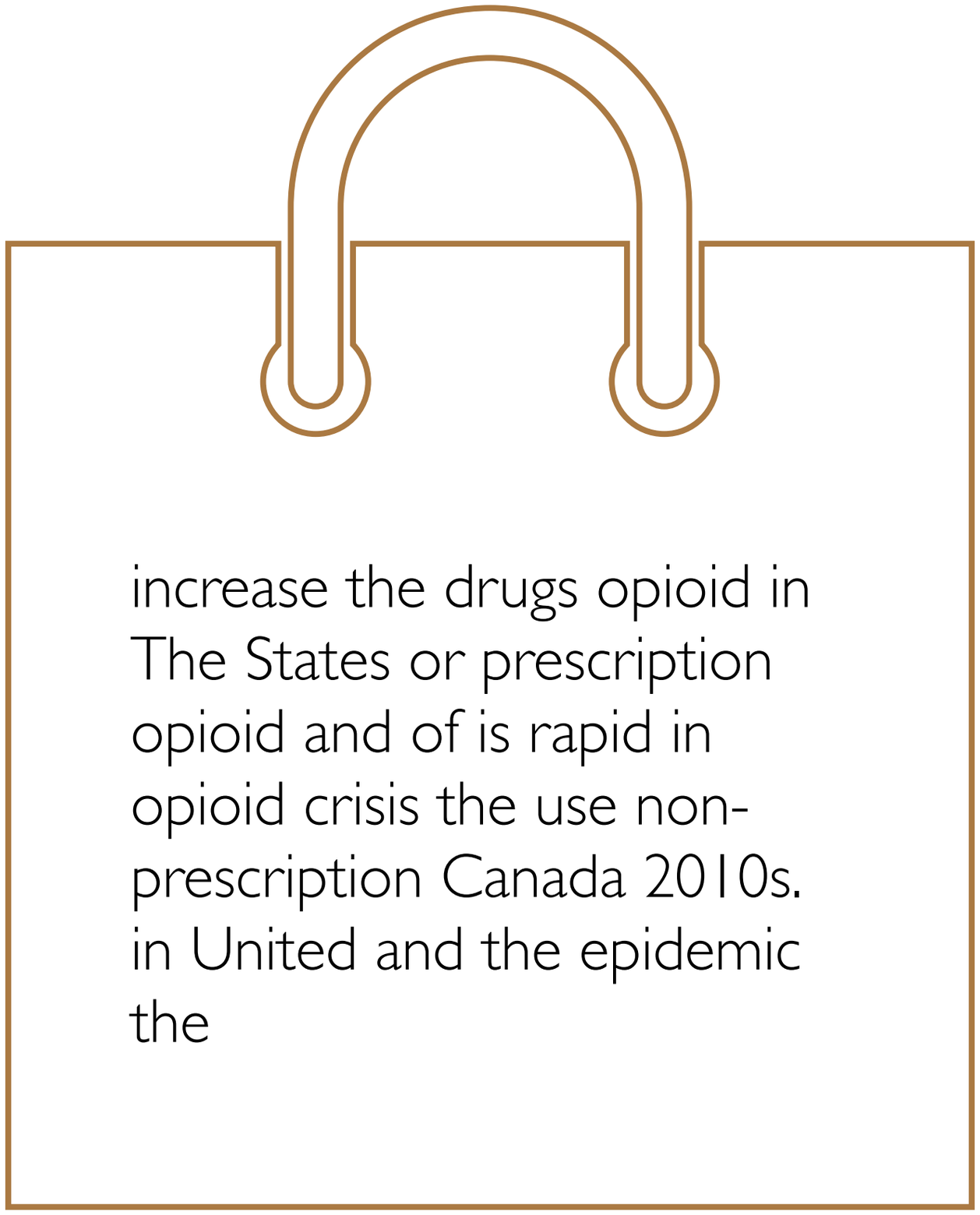
Histogram



"Raw count" frequency

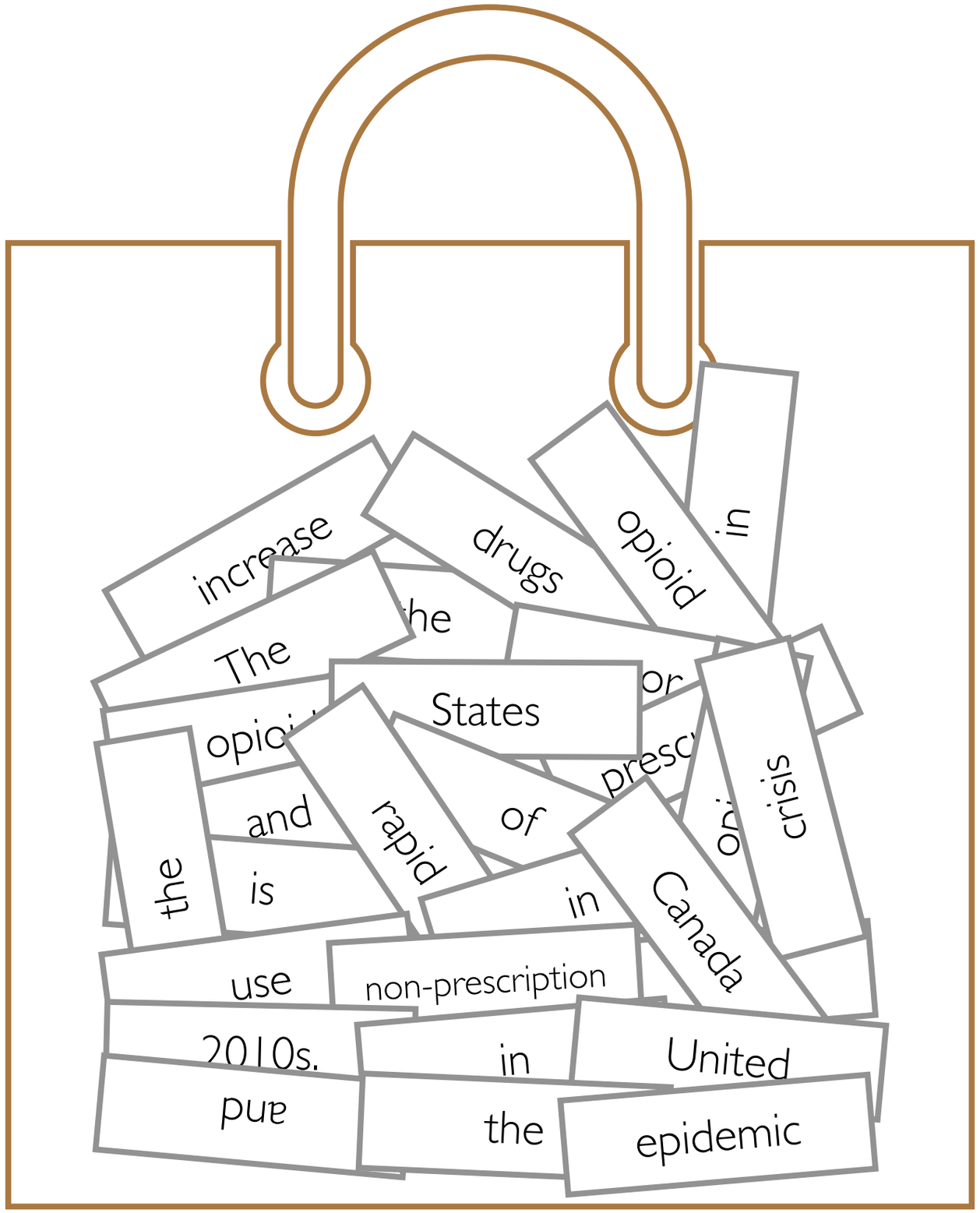
can also specify frequency as a fraction

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

- Can of course compute word frequencies for an entire document and not just a single sentence
- Can also compute word frequencies for a collection of documents (e.g., all of Wikipedia, etc), resulting in what is called the **collection term frequency** (ctf)

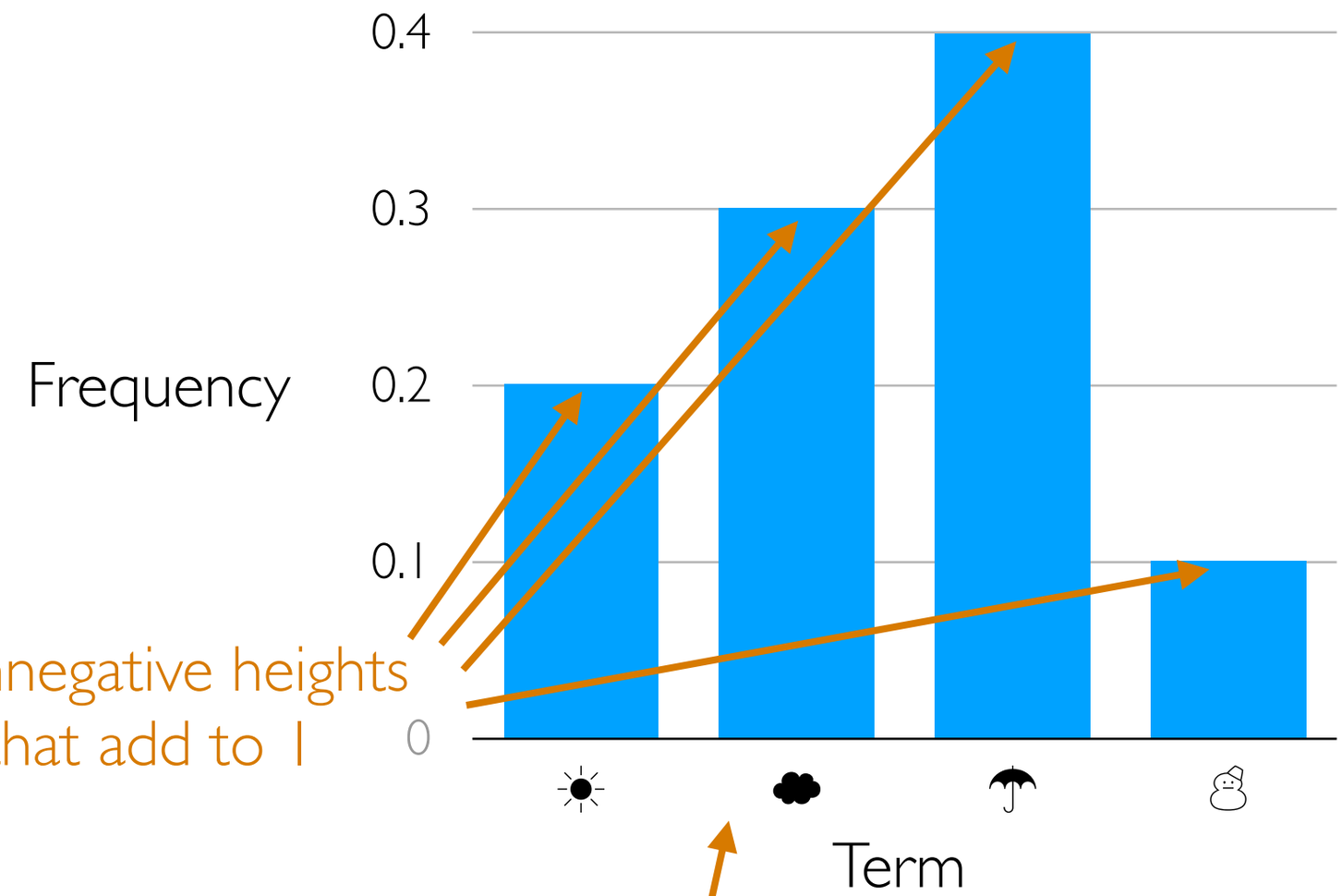
What does the *ctf* of "opioid" for all of Wikipedia refer to?

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

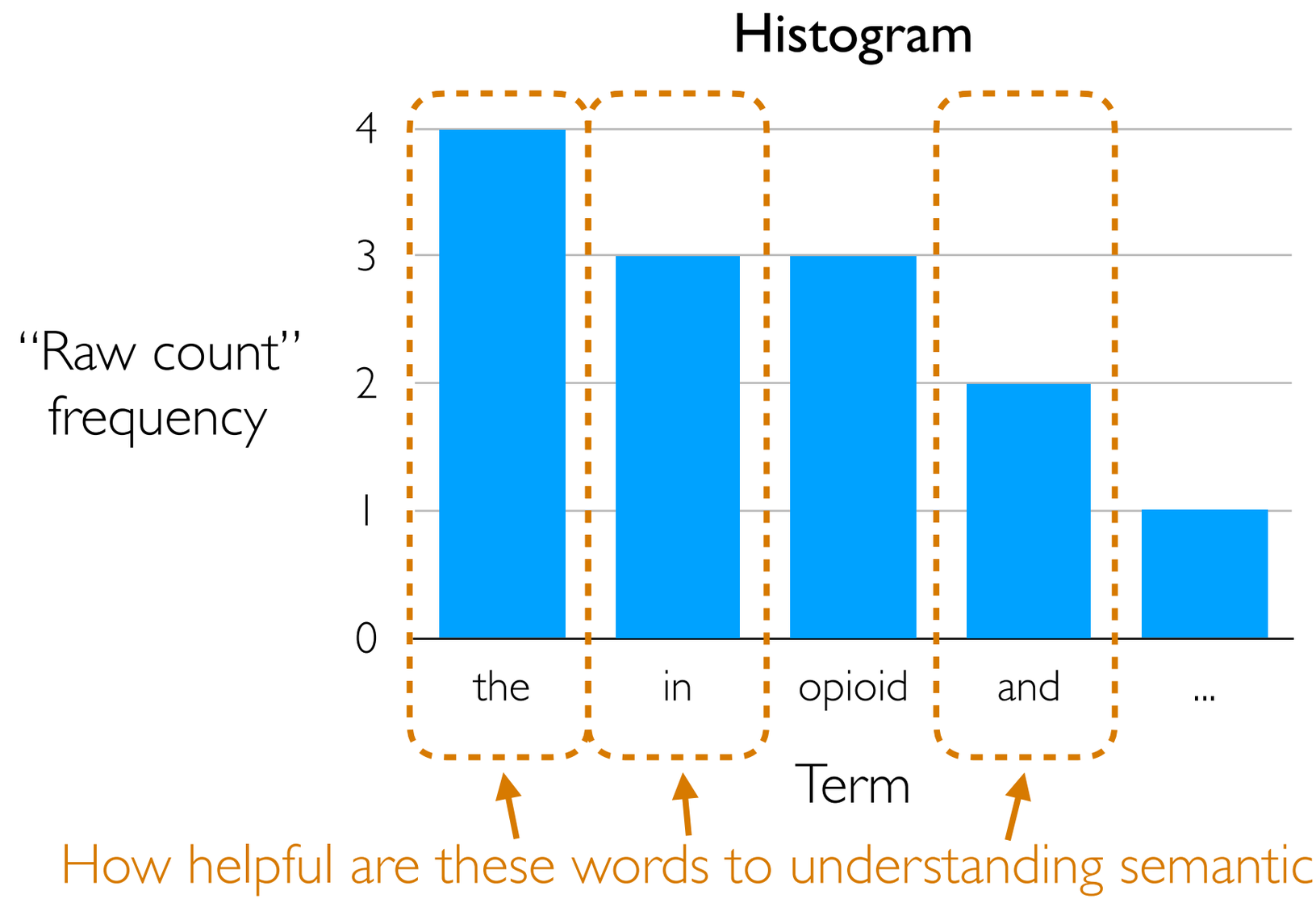
This is an example of a probability distribution

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

Let's take advantage of other properties of text

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

Some Words Don't Help?



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)

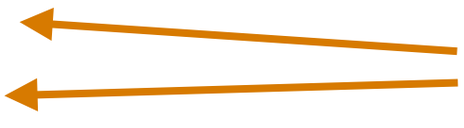
This problem is called **word sense disambiguation** (WSD)

Treat Some Phrases as a Single Word?

<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

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) ... # unique cards changes dramatically

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

3 words at a time: **trigram** model

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