Carnegie Mellon University
Heinz College

95-865 Unstructured Data Analytics

Lecture 1: Course overview, analyzing text using frequencies

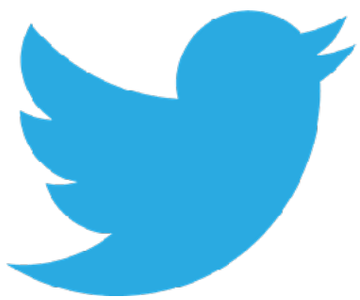
Slides by George H. Chen

What is this course about?

Big Data

We're now collecting data on virtually every human endeavor

amazon.com



NETFLIX

fitbit®

lyft

UPMC

LIFE CHANGING MEDICINE

How do we turn data into actionable insights?

Generative AI Technologies

- AI assistants like (Chat)GPT, Gemini, Claude, Llama, and DeepSeek will continue to get better over time
  - As of April 2023, GPT-4 can get a B on a quantum computing final exam! https://scottaaronson.blog/?p=7209
  - As of July 2024, AlphaProof (built on Gemini) can achieve silver for International Math Olympiad problems https://deepmind.google/discover/blog/ai-solves-imo-problems-at-silver-medal-level/
- For this class, I will view whether you get help from AI *to be the same as whether you got help from a human*
  - Regardless of whether you get info from an AI or a human, I want *you* to be able to tell whether this info is correct or not
    - Exams will be paper & pencil with no electronics allowed
- At the end of the semester, we will go over a basic version of GPT (generative pre-trained transformers)

Why is this class called “unstructured data analytics”?

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

Is a patient at risk of getting a nasty 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?

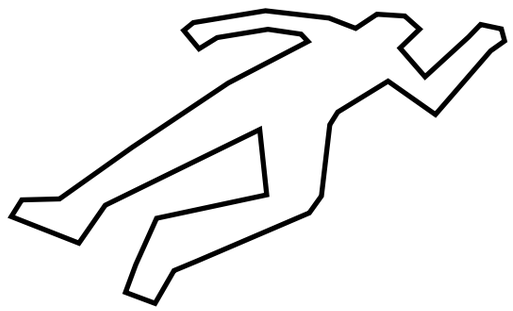
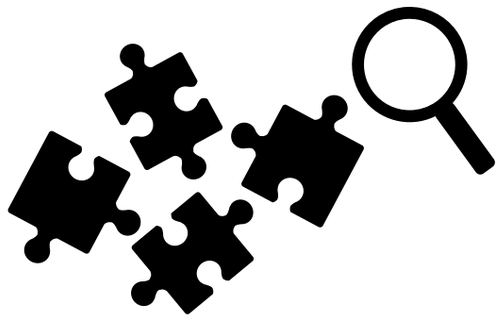
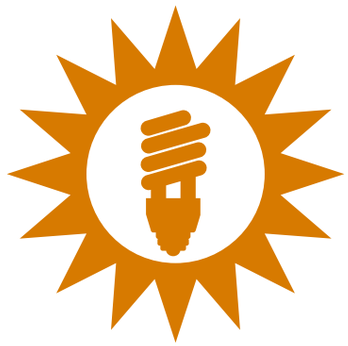
Data

- 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 Analytics (UDA)

Much like how many murder mysteries go unsolved,
many data analysis (unstructured or not) problems can be extremely difficult

Question	Data	Finding Structure	Insights
			
The dead body	*The evidence*	*Puzzle solving, careful analysis*	*When? Where? Why? How? Perpetrator catchable?*
This is provided by a domain expert	Some times you have to collect more evidence!	Exploratory data analysis	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!

As we go from covering classical methods to modern ones, it's good to understand *why* newer methods were developed

UDA technologies change very rapidly! GPTs might be hot today but be out of fashion tomorrow!

Course ~~Textbook~~ Materials

No existing textbook matches the course… =(

Main source of material: lectures slides

We'll post supplemental reading as we progress

Check course webpage

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

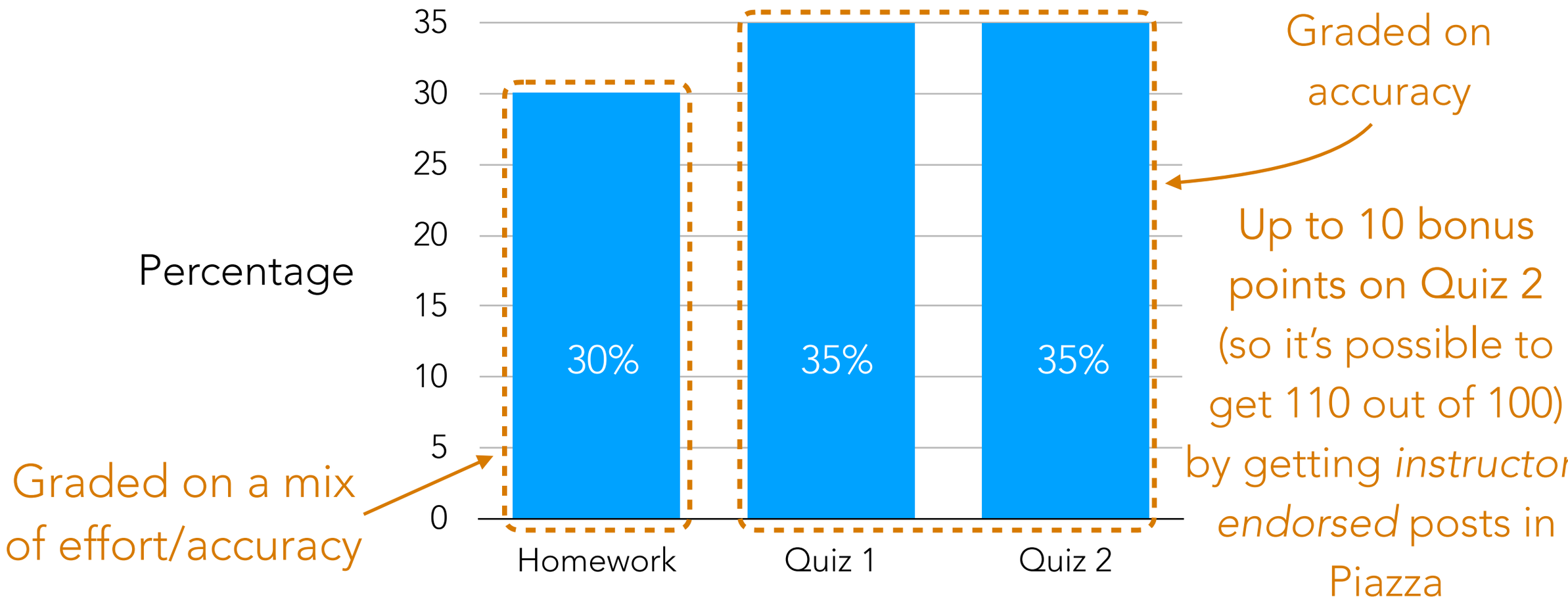
In general, announcements & links to all course-related things will be in Canvas

Homework assignments are submitted via Gradescope (link is within Canvas)

Please post questions to Piazza (link is within 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)

The Two Quizzes

Format:

- **In-person, on paper**
- Each quiz is **80 minutes**
- No electronics may be used during the exam
  (e.g., do <u>not</u> use a laptop, tablet, phone, calculator)
- Open notes (must be on paper and <u>not electronic</u>)

Quiz 1: Friday Mar 28 during recitation (5pm-6:20pm, HBH A301)

Quiz 2: Friday May 2 (1pm-2:20pm, location TBD)

Gradescope

- We're using Gradescope for grading everything (homework & quizzes)
- Your homework will involve coding *but we will ask that you save your code notebook as a PDF and submit only the PDF*
    - **Since we will not be re-running your code, make sure that your PDF includes all the code output!**
- We will scan your quizzes and grade them on Gradescope

Collaboration & Academic Integrity

- If you are having trouble, ***please 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., ChatGPT, Gemini, stackoverflow)
- **Do not share your code with classmates (instant message, email, Box, Dropbox, AWS, etc)**
- **Do not use solutions from past semesters**

Penalties for cheating are severe: 0 on assignment, possibly fail the course ☹️

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:

- Qianchi (Lily) Huang
- Shun Li
- Xiao Li
- Yubo Li

Instructor: George Chen

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

Part I.

Exploratory Data Analysis

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

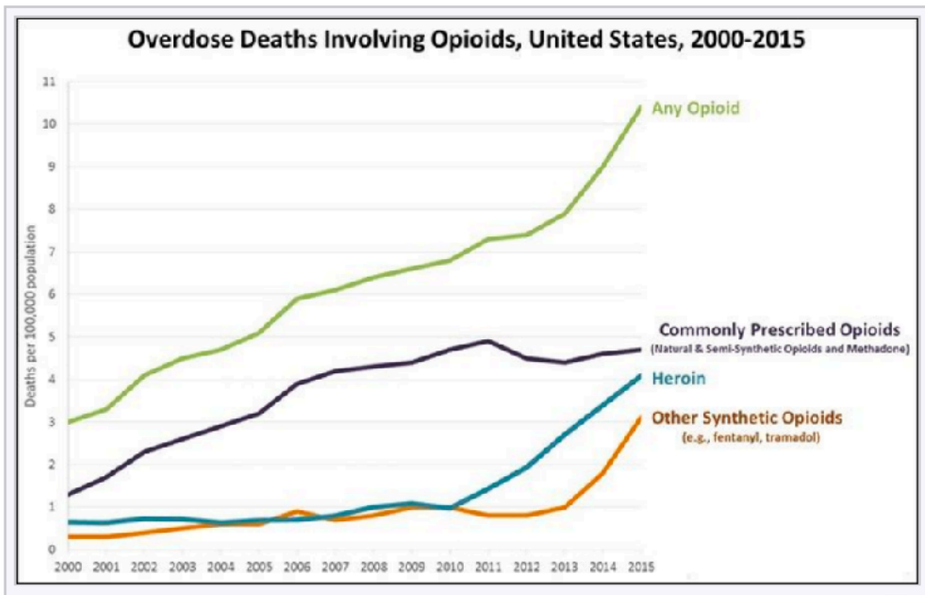
Help
About Wikipedia
Community portal
Recent changes
Contact page

Tools

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

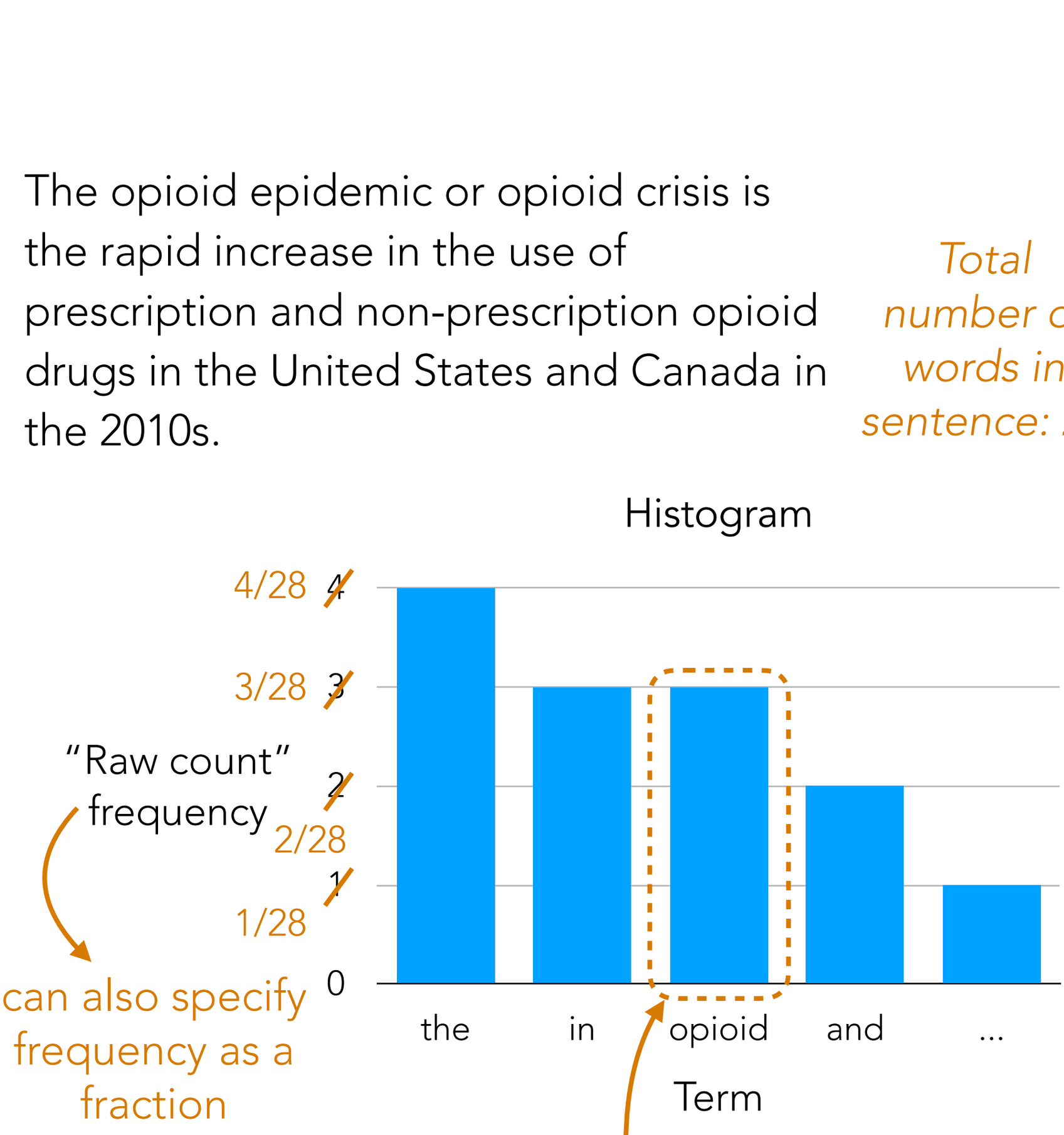
The opioid epidemic or opioid crisis is the rapid increase in the use of prescription and non-prescription opioid drugs in the United States and Canada in the 2010s.

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



"Raw count" frequency

can also specify frequency as a fraction

Fraction of words in the sentence that are "opioid"

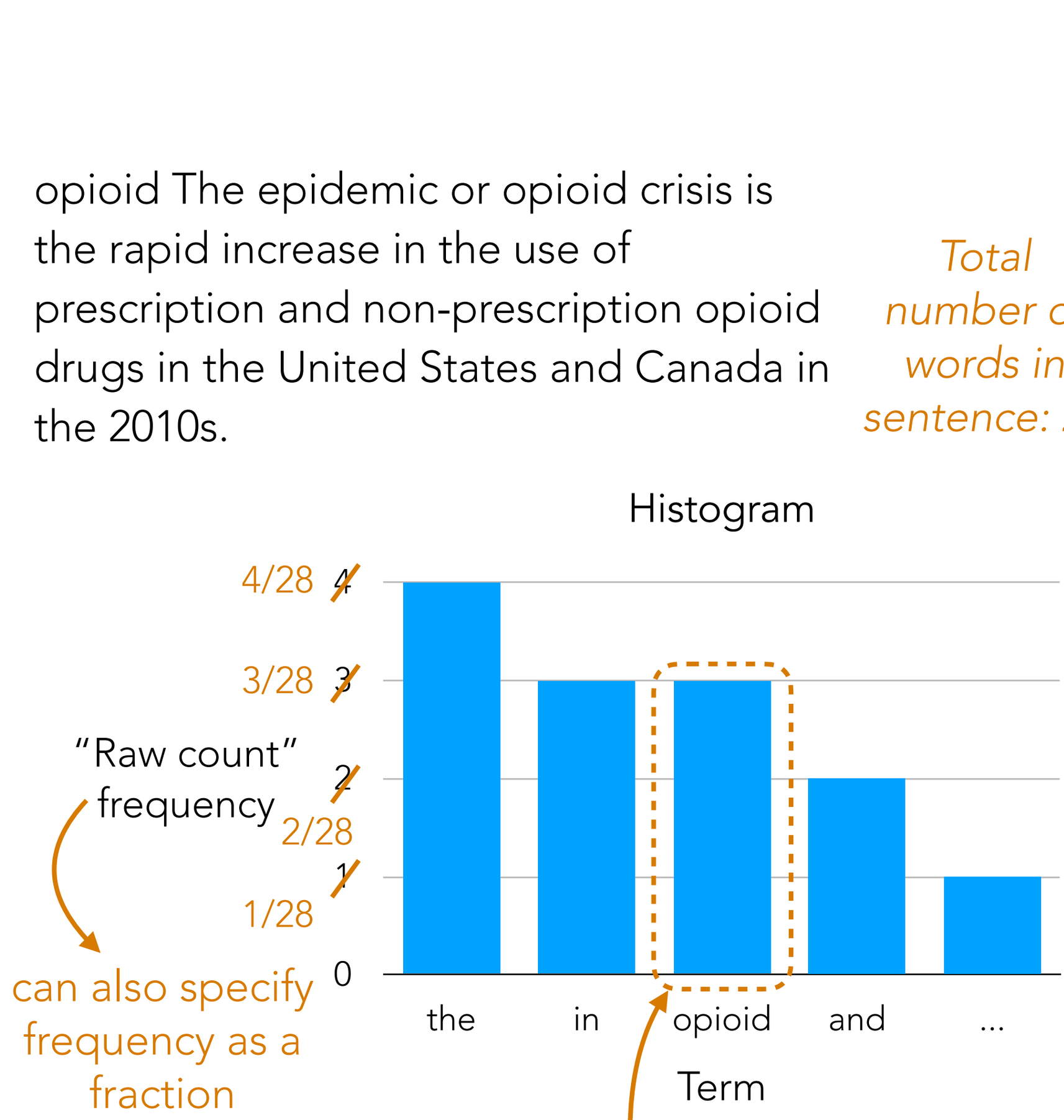
opioid The epidemic or opioid crisis is the rapid increase in the use of prescription and non-prescription opioid drugs in the United States and Canada in the 2010s.

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



"Raw count" frequency

can also specify frequency as a fraction

Fraction of words in the sentence that are "opioid"

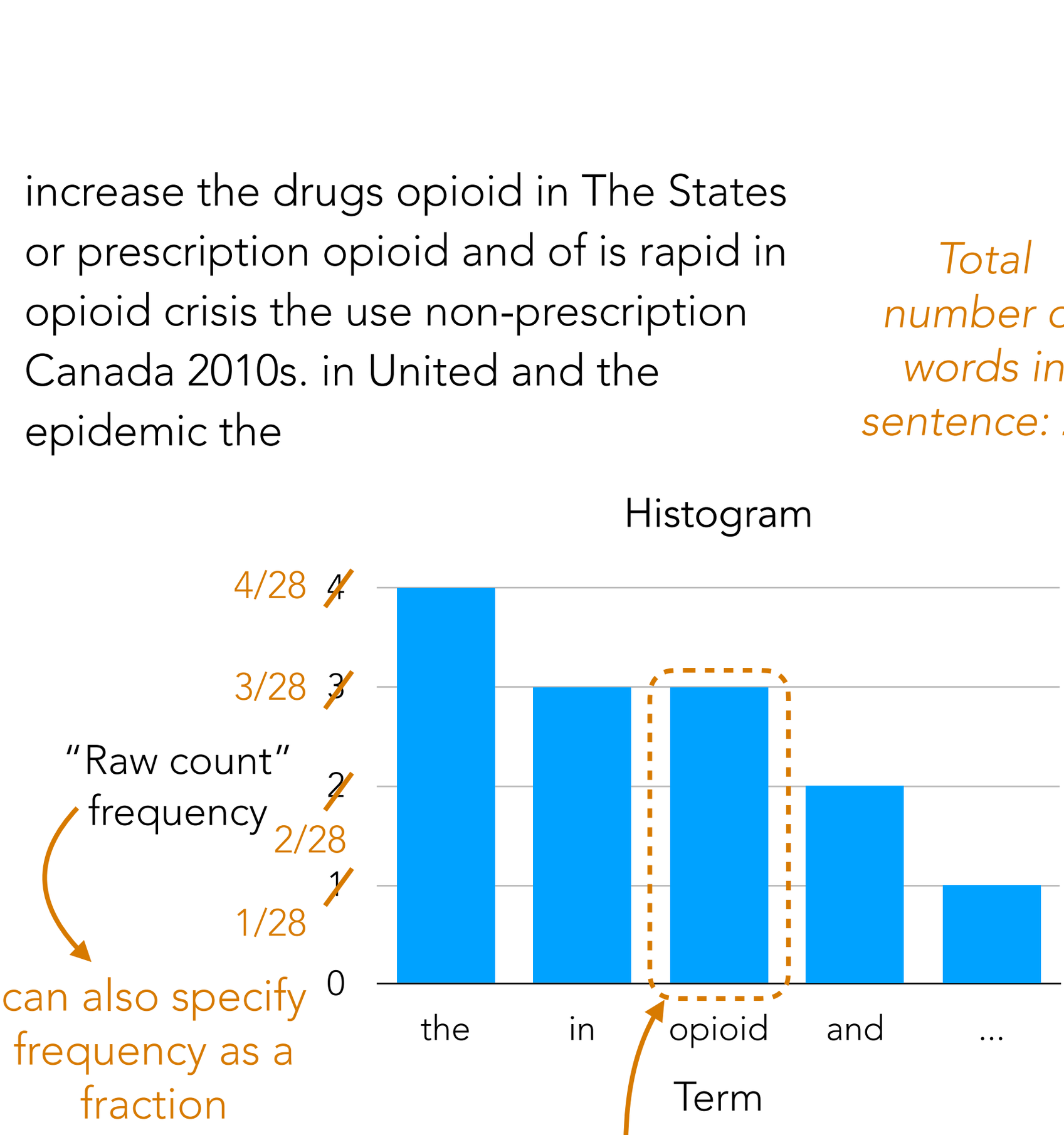
increase the drugs opioid in The States or prescription opioid and of is rapid in opioid crisis the use non-prescription Canada 2010s. in United and the epidemic the

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

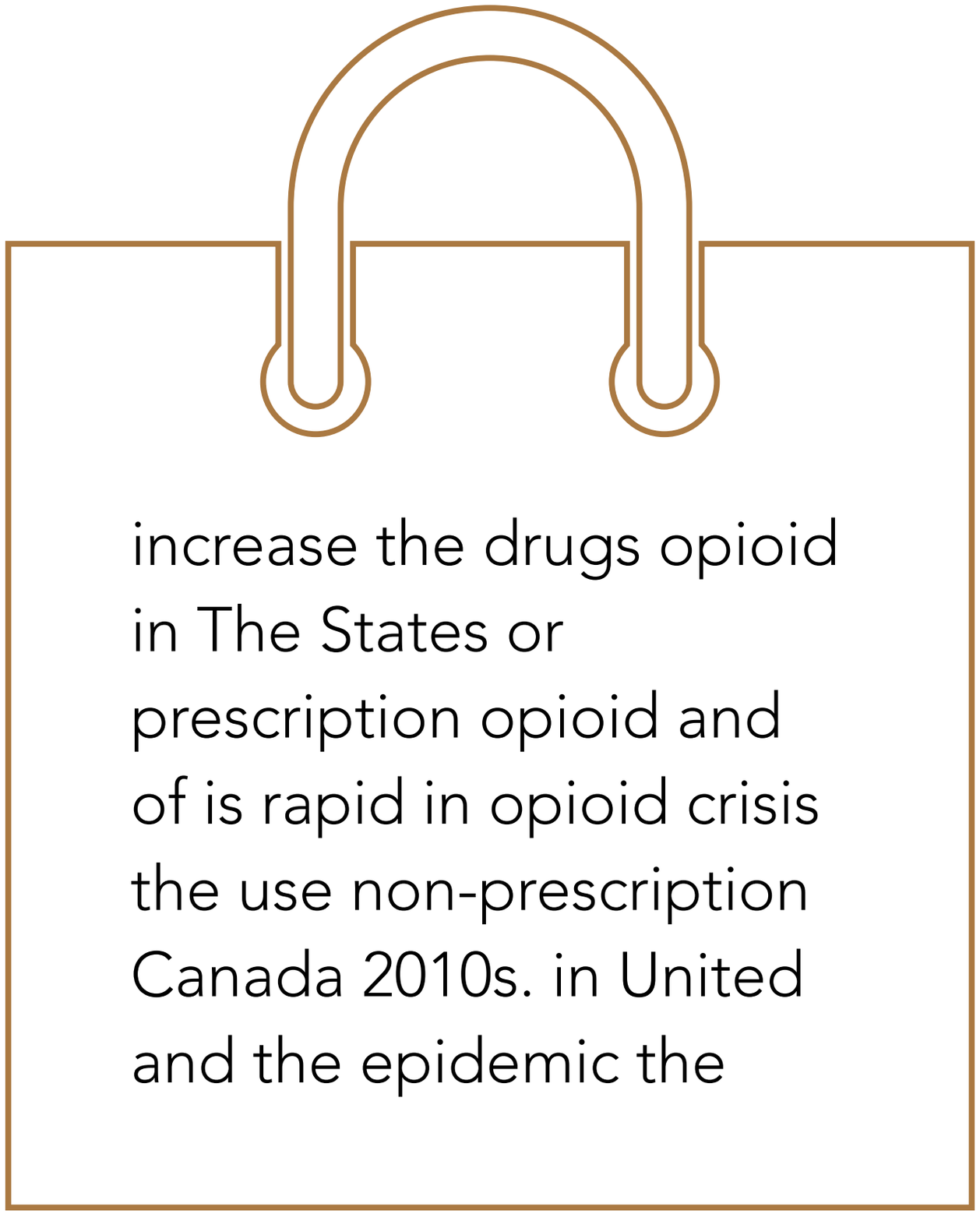


"Raw count" frequency

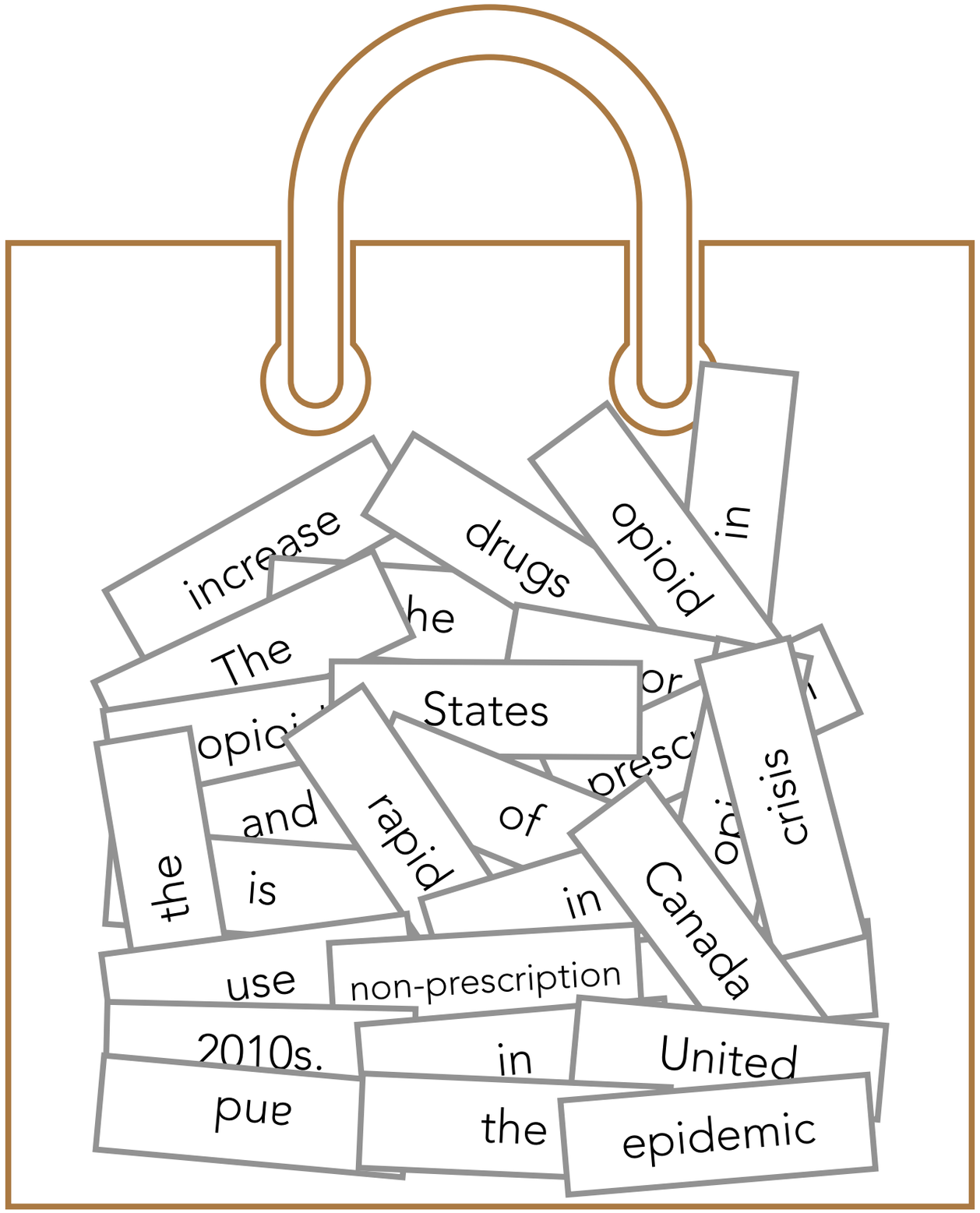
can also specify frequency as a fraction

Fraction of words in the sentence that are "opioid"

Bag of Words Model



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), resulting in what is called the collection term frequency (ctf)

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

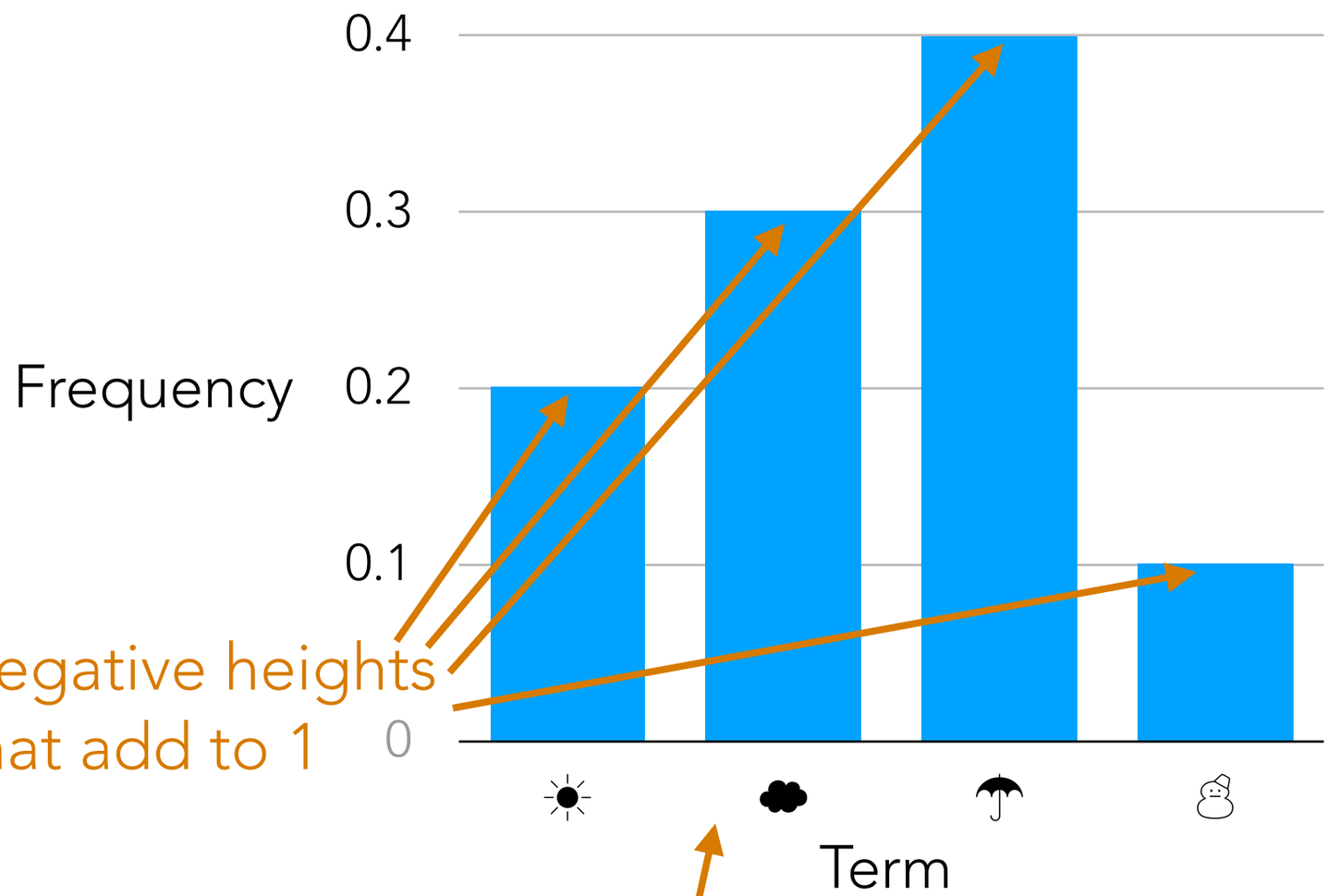
Terminology:

- **Corpus:** collection of text (e.g., Wikipedia corpus, Common Crawl corpus); plural form of corpus is **corpora**
- **Natural language processing (NLP):** field of linguistics, computer science, and AI focusing on automatic analysis of human languages
  - NLP systems are regularly trained on large corpora

So far did we use anything special about text?

Basic Probability in Disguise

"Sentence":

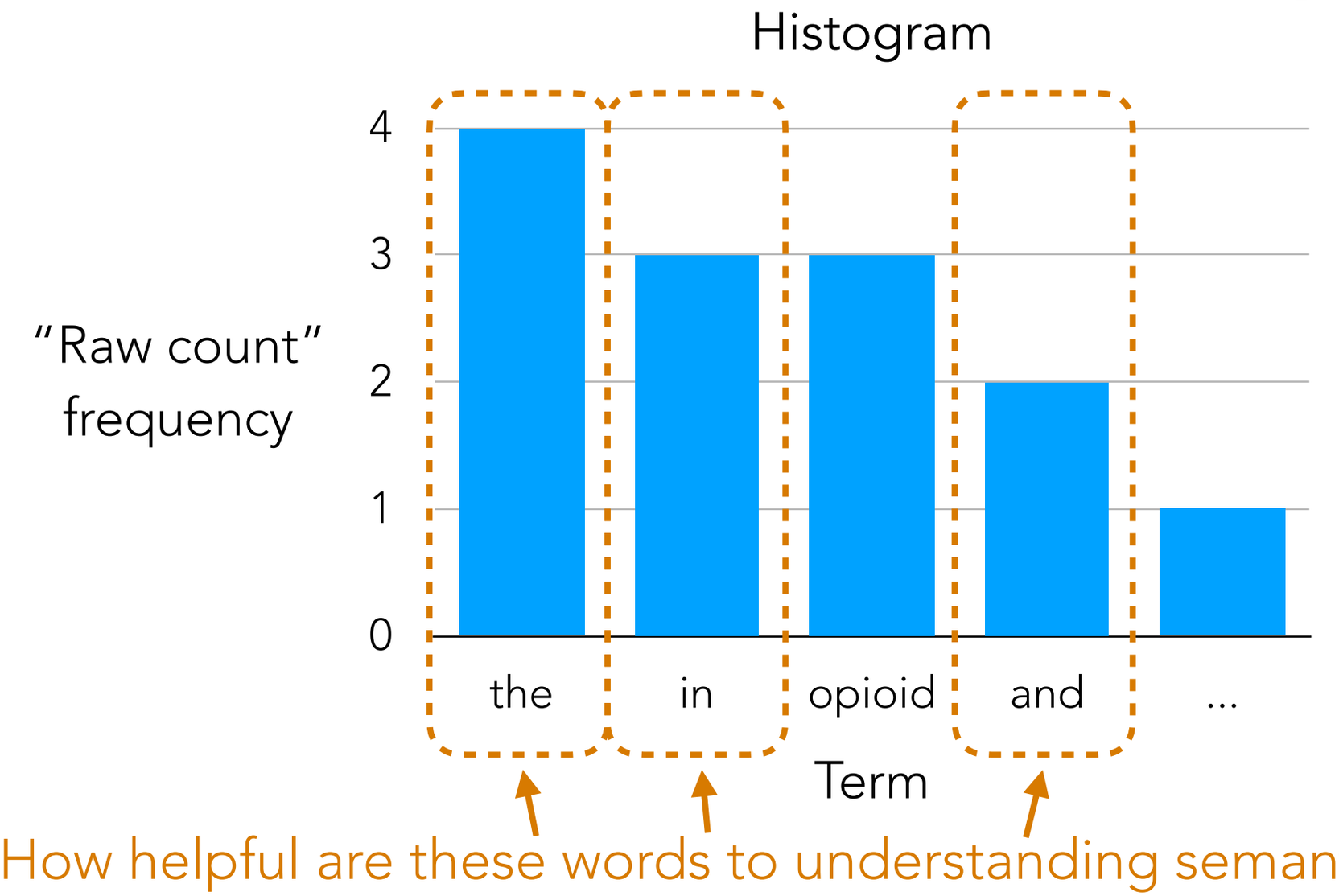


Probability distributions will appear throughout the course
and are essential to many modern AI methods

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?



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 English Stopword List (from NLP package 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"