Carnegie Mellon University
Heinz College

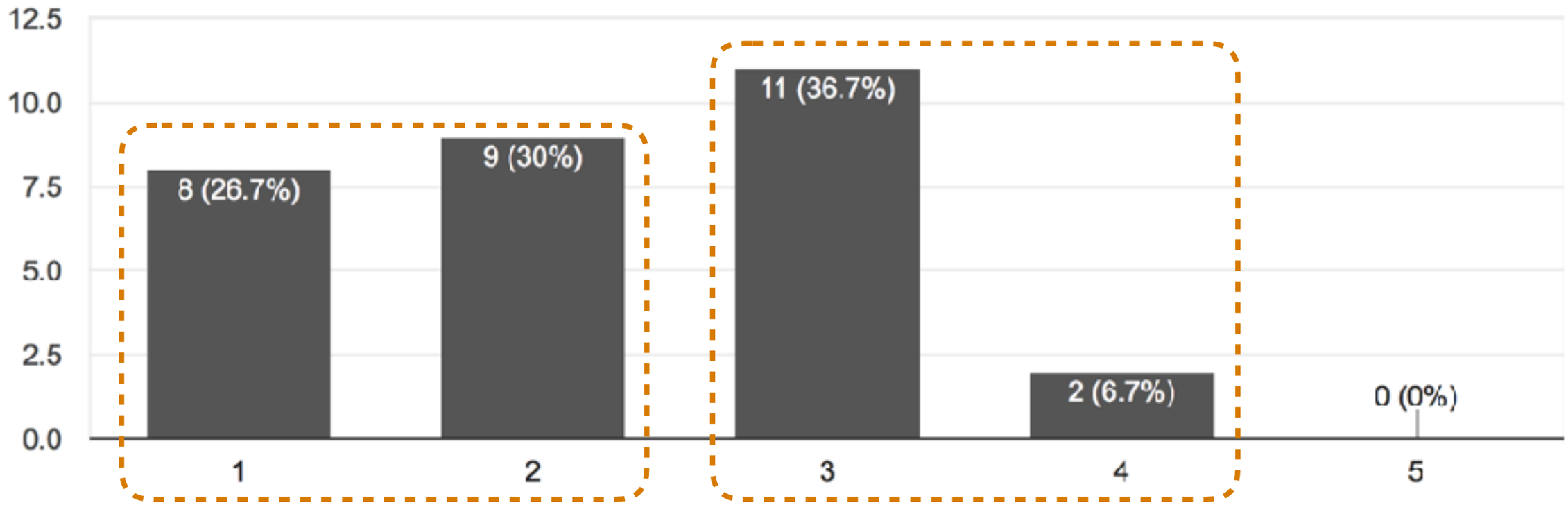
94-775 Lecture 4:
Basic Text Analysis

George Chen

Questionnaire Results

How much Python do you already know?

30 responses



~2/3 of the class has little to no Python experience

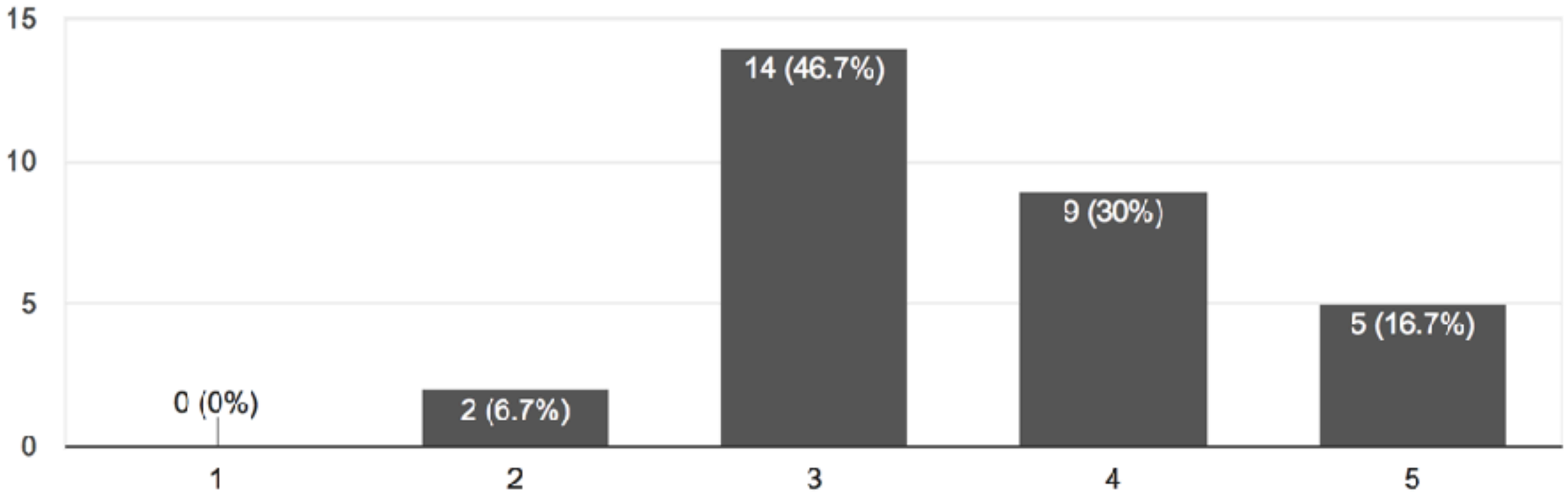
Various students here are concerned about course difficulty

Multiple students already comfortable with Python have complained that I'm spending too much time on basic Python

Questionnaire Results (cont'd)

How have you found the first two lectures?

30 responses



Overall I'd like this distribution to shift rightward!

Some Comments

- This course explicitly does not require Python as a prereq
  - Students who already know it: please be patient with students who aren't as fluent in Python!
- Certain students want to learn as much UDA as possible and they don't care for the final project
  - Please consider taking 95-865 instead in a future mini
  - This class is intentionally targeted toward policy students (as the course title suggests)
- Python software package coverage:
  - We will cover **numpy**, **spaCy**, **sklearn**, and **keras**
  - In the past, I have taught using **numpy** and **pandas** concurrently but many students found switching between them confusing…so I will not cover **pandas**

Some Comments (cont'd)

- What you should find out:
  - Core concepts taught in this class extend across software packages and even programming languages
  - If you already know R but are now just learning Python, you should realize that programming languages have a lot of similarities, and picking up more languages is not hard

- Mid-mini quiz:
  - Format is still TBD but likely will involve a mix of short conceptual questions (that could involve code) and a slightly longer coding question

Basic text analysis: how do we represent text documents?

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

Interaction

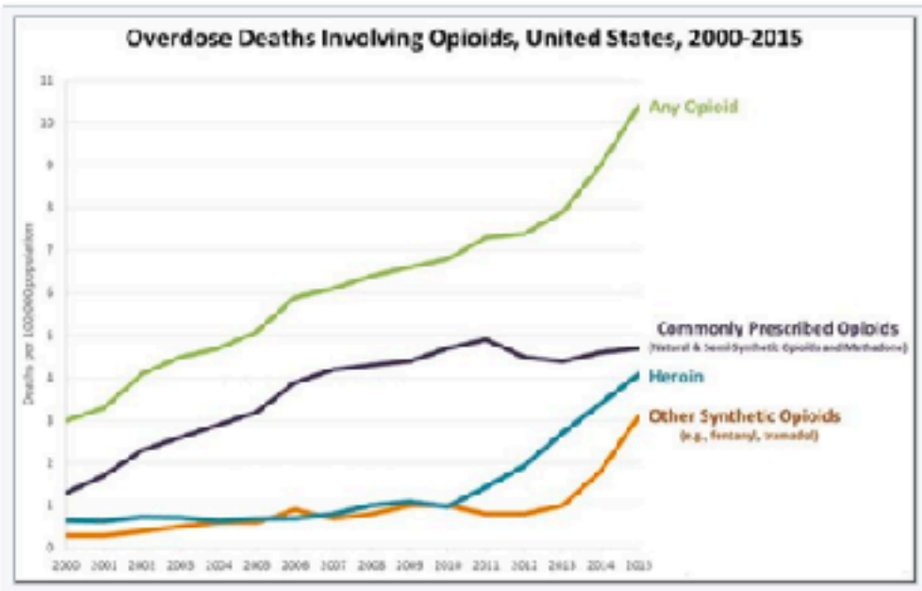
Help
About Wikipedia
Community portal
Recent changes
Contact page

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

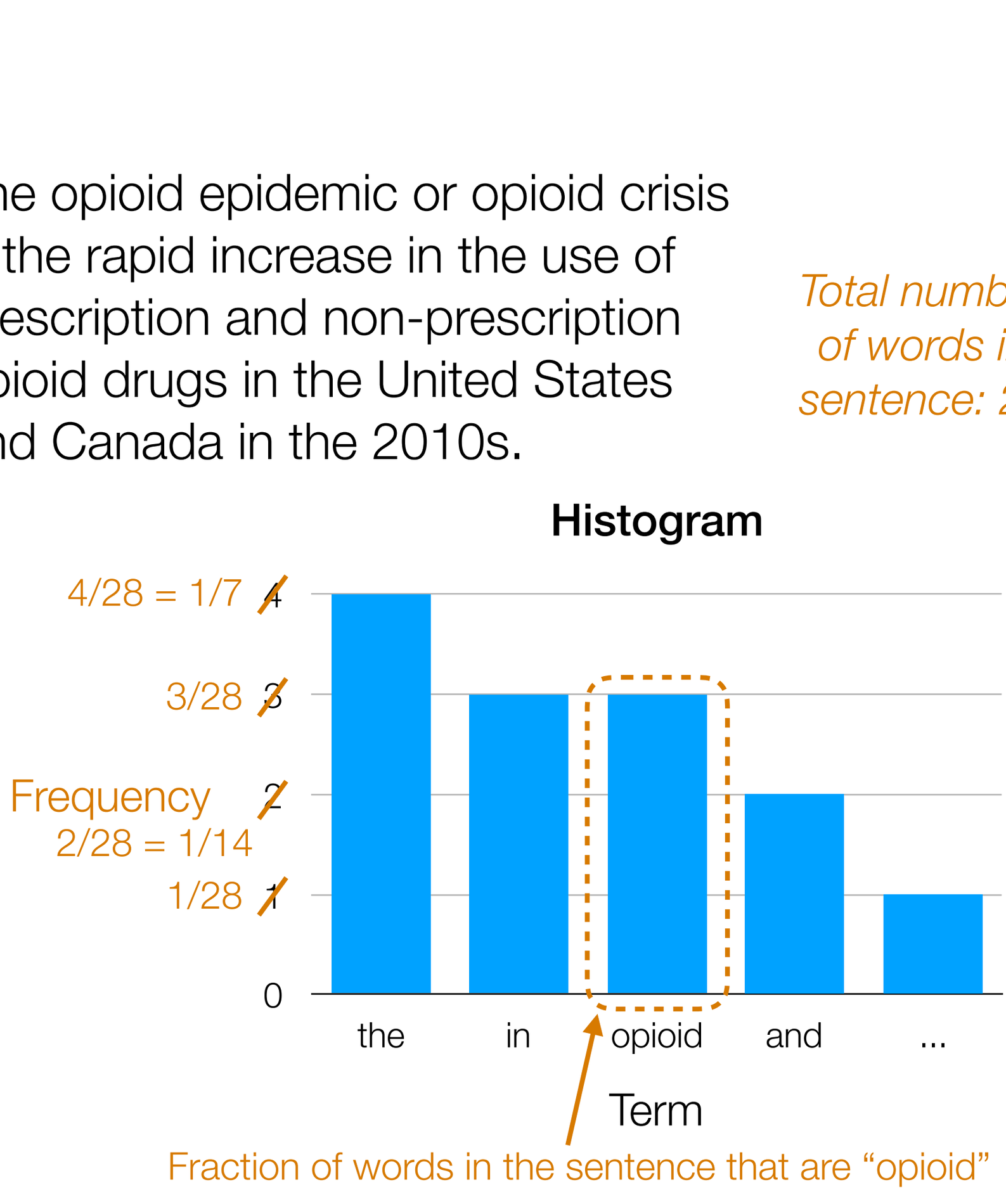
The opioid epidemic or opioid crisis is the rapid increase in the use of prescription and non-prescription opioid drugs in the United States and Canada in the 2010s.

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”

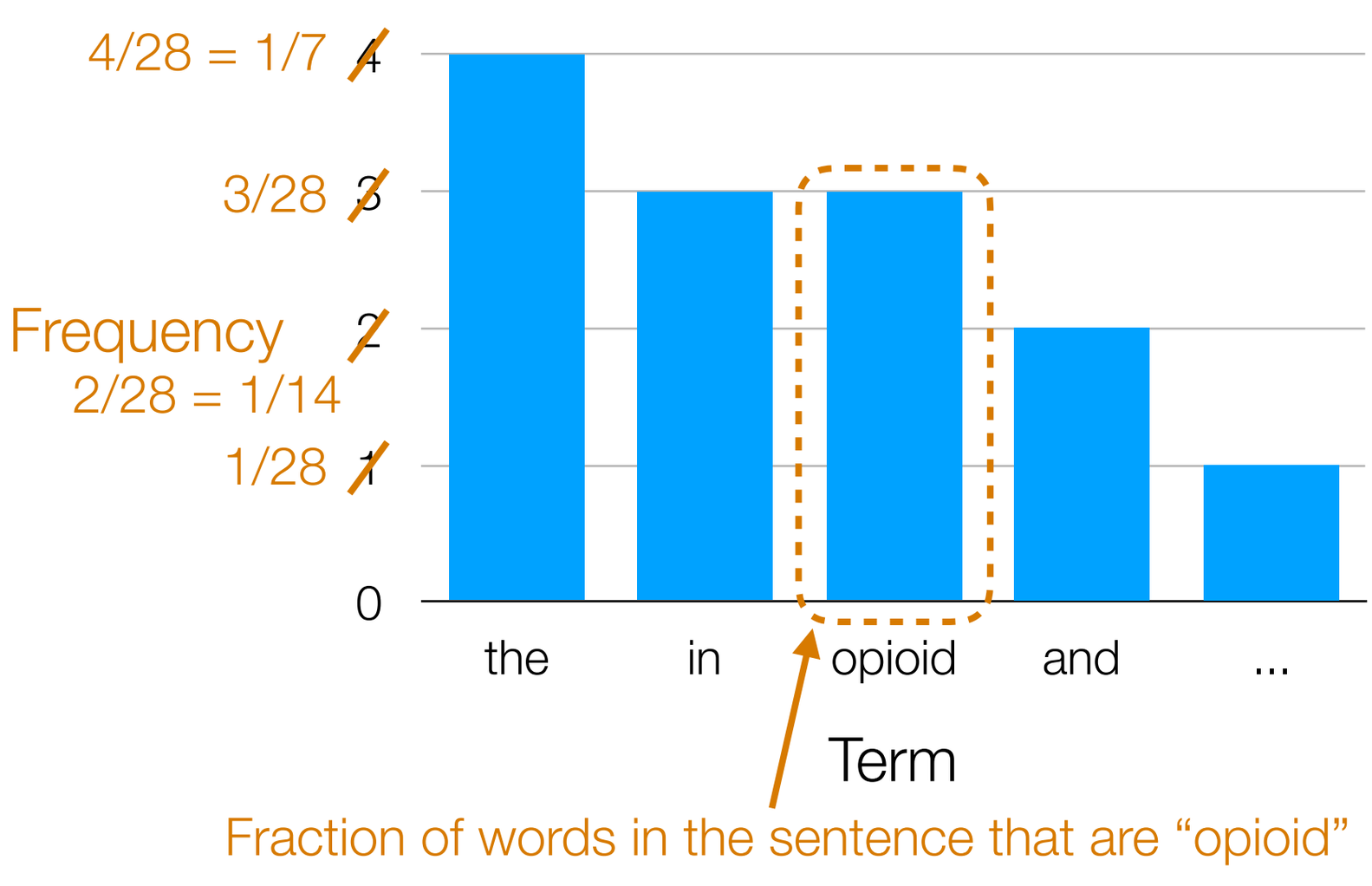
opioid The epidemic or opioid crisis is the rapid increase in the use of prescription and non-prescription opioid drugs in the United States and Canada in the 2010s.

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”

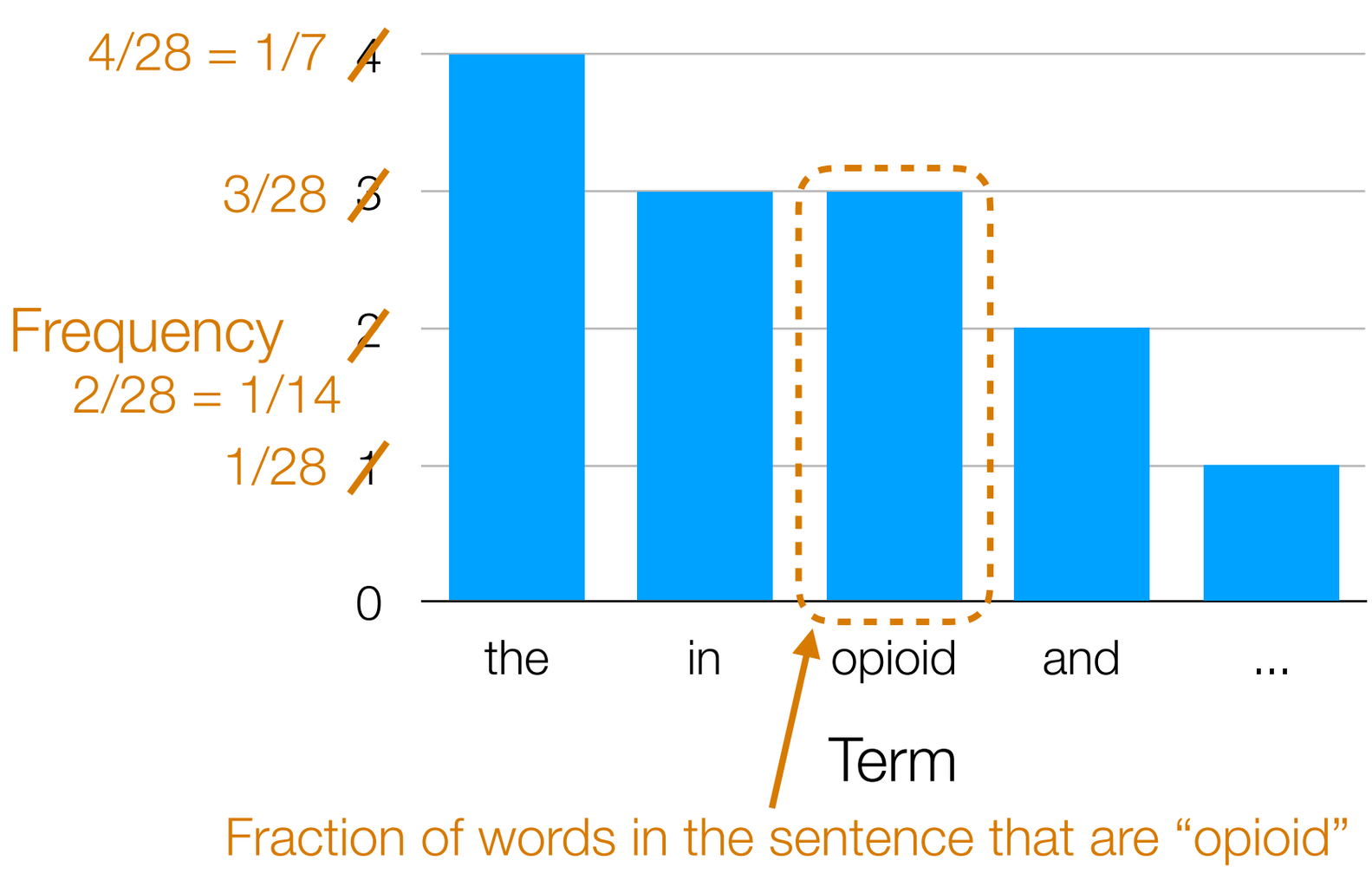
increase the drugs opioid in The States or prescription opioid and of is rapid in opioid crisis the use non-prescription Canada 2010s. in United and the epidemic the

Total number of words in sentence: 28

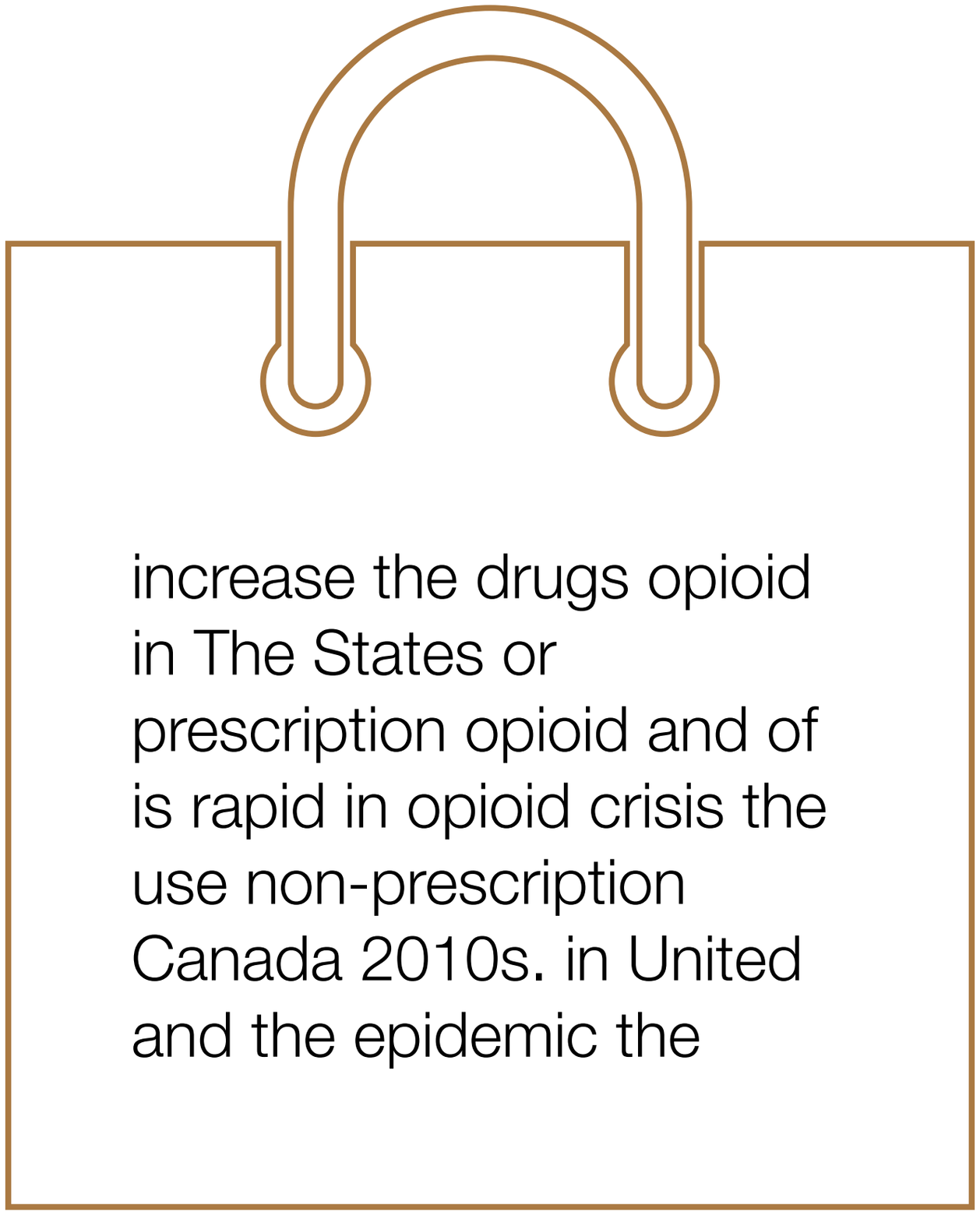
Term frequencies

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

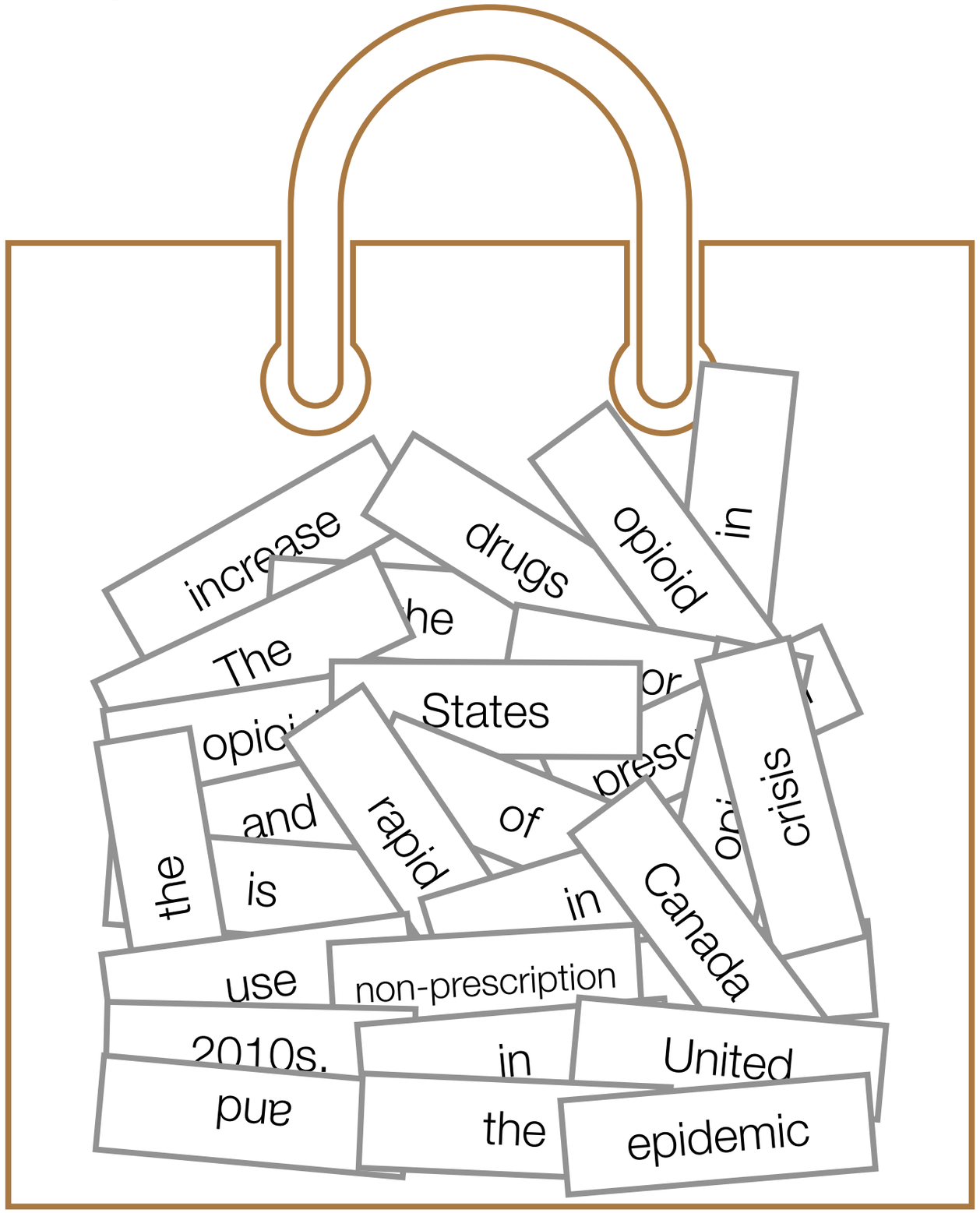


Fraction of words in the sentence that are “opioid”



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)

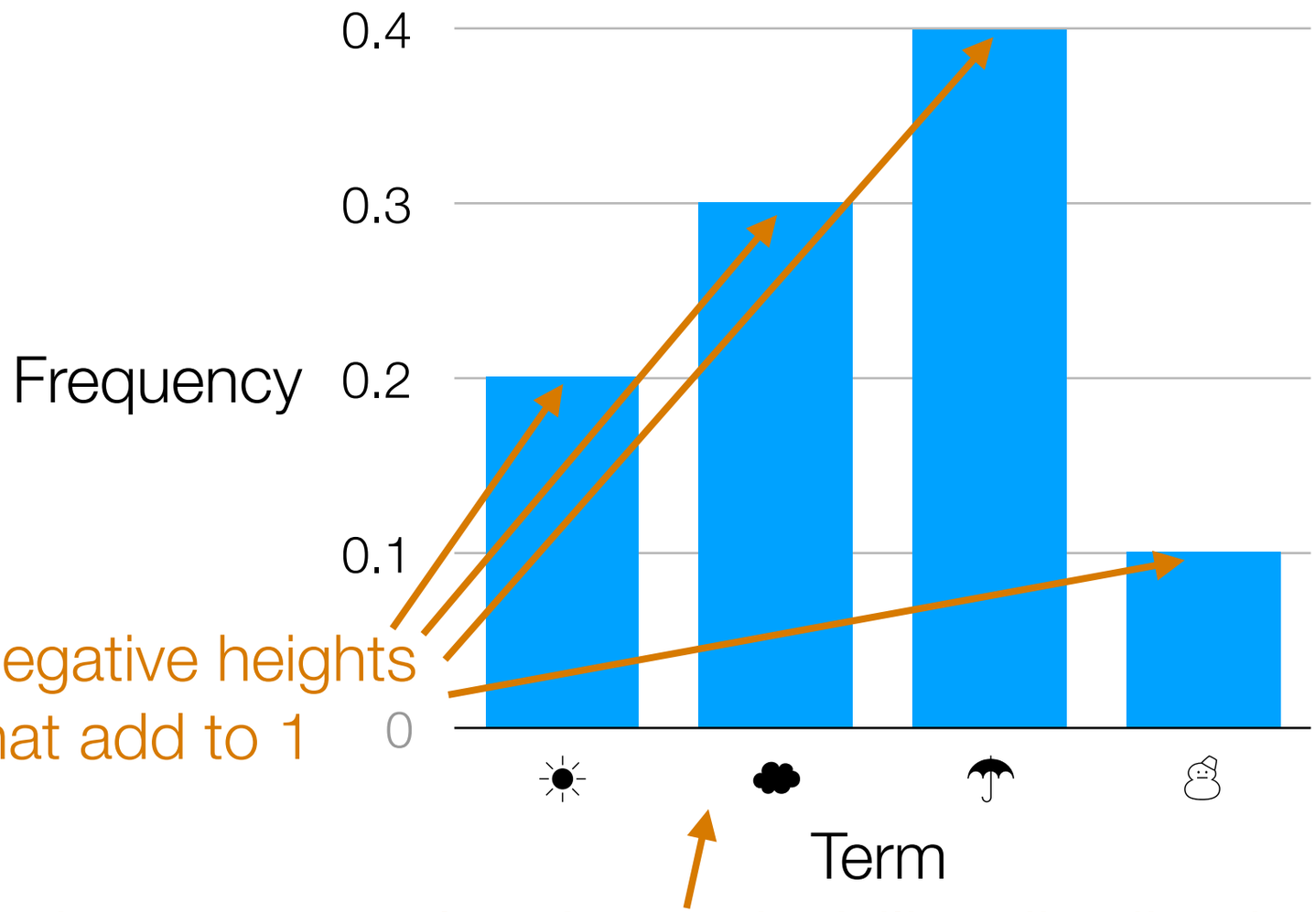
  What does the *ctf* of "opioid" for all of Wikipedia refer to?

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

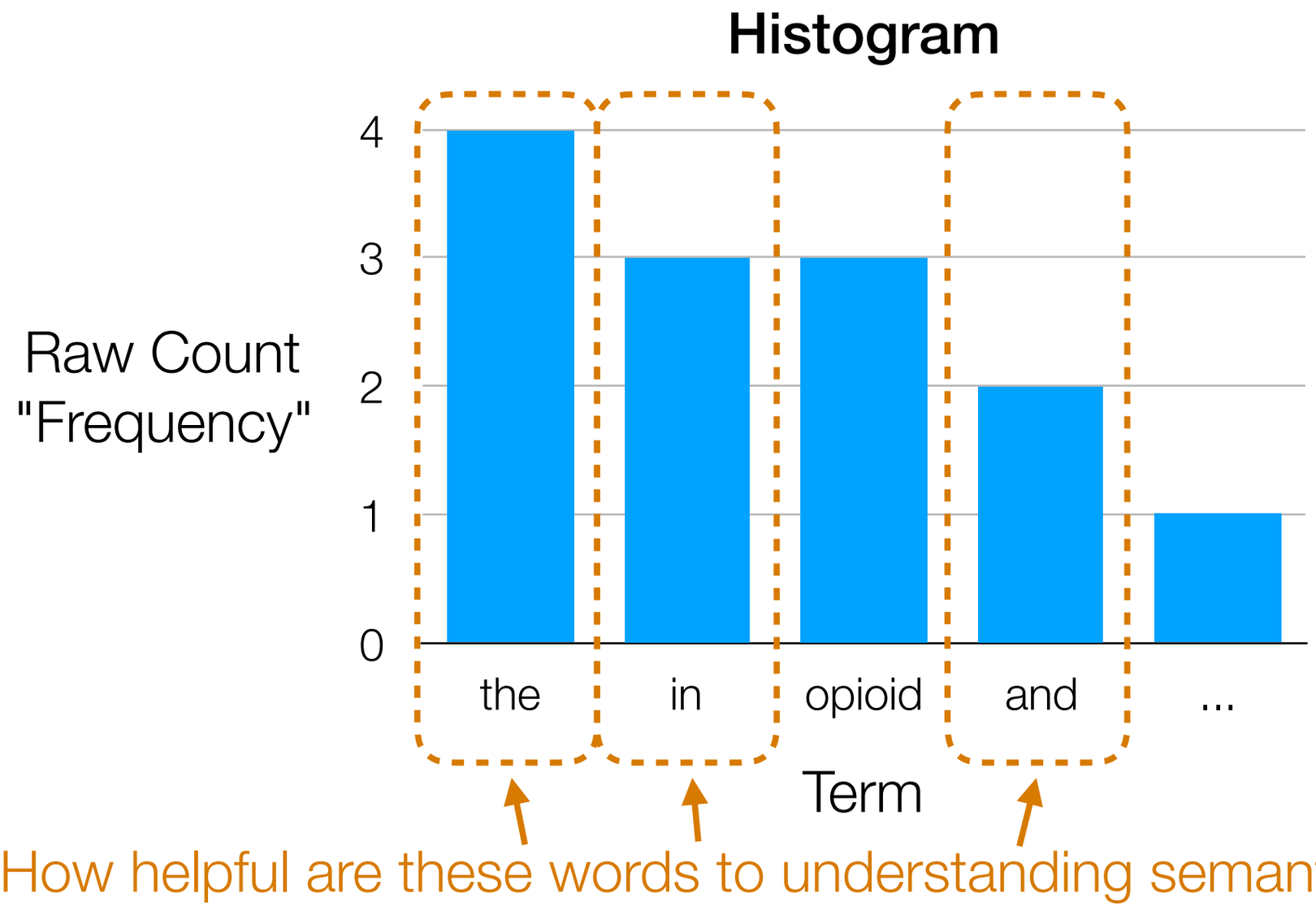
This is an example of a probability distribution

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

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?



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