

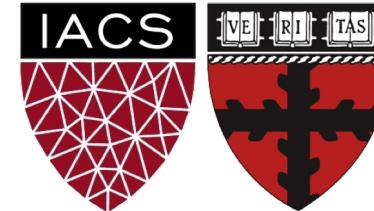
Lecture 2: Data

What it is, where to get it,
and factors to consider.

Harvard IACS

CS109B

Pavlos Protopapas, Kevin Rader, and Chris Tanner



Learning Objectives

- Understand different types and formats of data
- Be able to soundly select appropriate data
- Have awareness of biases that exist
- Be able to refine questions to suite your true inquiry
- Understand how to parse text with regular expressions

Agenda

- What is data?
- Aspects of data: formats, scope, biases, etc
- Asking precise questions
- Parsing data with Regular Expressions

Agenda

- What is data?
- Aspects of data: formats, scope, biases, etc
- Asking precise questions
- Parsing data with Regular Expressions

What is data?

What is data?

Def₁ Factual information (such as measurements or statistics) used as a basis for reasoning, discussion, or calculation

Def₂ Information in digital form that can be transmitted or processed

Def₃ Information output by a sensing device or organ that includes both useful and irrelevant or redundant information and must be processed to be meaningful

What is data?

Def₁ Factual information (such as measurements or statistics) used as a basis for reasoning, discussion, or calculation

Def₂ Information in digital form that can be transmitted or processed

Def₃ Information output by a sensing device or organ that includes both useful and irrelevant or redundant information and must be processed to be meaningful

What is data?

Def₁ Factual information (such as measurements or statistics) used as a basis for reasoning, discussion, or calculation

Def₂ Information in digital form that can be transmitted or processed

Def₃ Information output by a sensing device or organ that includes both useful and irrelevant or redundant information and must be processed to be meaningful

Factual information (such as measurements or statistics) used as a basis for reasoning, discussion, or calculation

Scenario₁ Measurements from a thermometer every hour for a year

Scenario₂ Counts from a person who tracks the days that a particular hummingbird visits his birdfeeder across an entire year

Scenario₃ Tweets from a politician

Scenario₄ Readouts from a mysterious sensor that was purchased from a local yard sale.

Factual information (such as measurements or statistics) used as a basis for reasoning, discussion, or calculation

- Scenario₁** Measurements from a thermometer every hour for a year
Probably inaccurate data
- Scenario₂** Counts from a person who tracks the days that a particular hummingbird visits his birdfeeder across an entire year
- Scenario₃** Tweets from a politician
Probably missing data
- Scenario₄** Readouts from a mysterious sensor that was purchased from a local yard sale.

Factual information (such as measurements or statistics) used as a basis for reasoning, discussion, or calculation

Scenario₁ Measurements from a thermometer every hour for a year

Scenario₂ Counts from a person who tracks the days that a particular hummingbird visits his birdfeeder across an entire year

Scenario₃ Tweets from a politician

Probably not 100% factually true

Scenario₄ Readouts from a mysterious sensor that was purchased from a local yard sale.

Factual information (such as measurements or statistics) used as a basis for reasoning, discussion, or calculation

Scenario₁ Measurements from a thermometer every hour for a year

Scenario₂ Counts from a person who tracks the days that a particular hummingbird visits his birdfeeder across an entire year

Scenario₃ Tweets from a political candidate

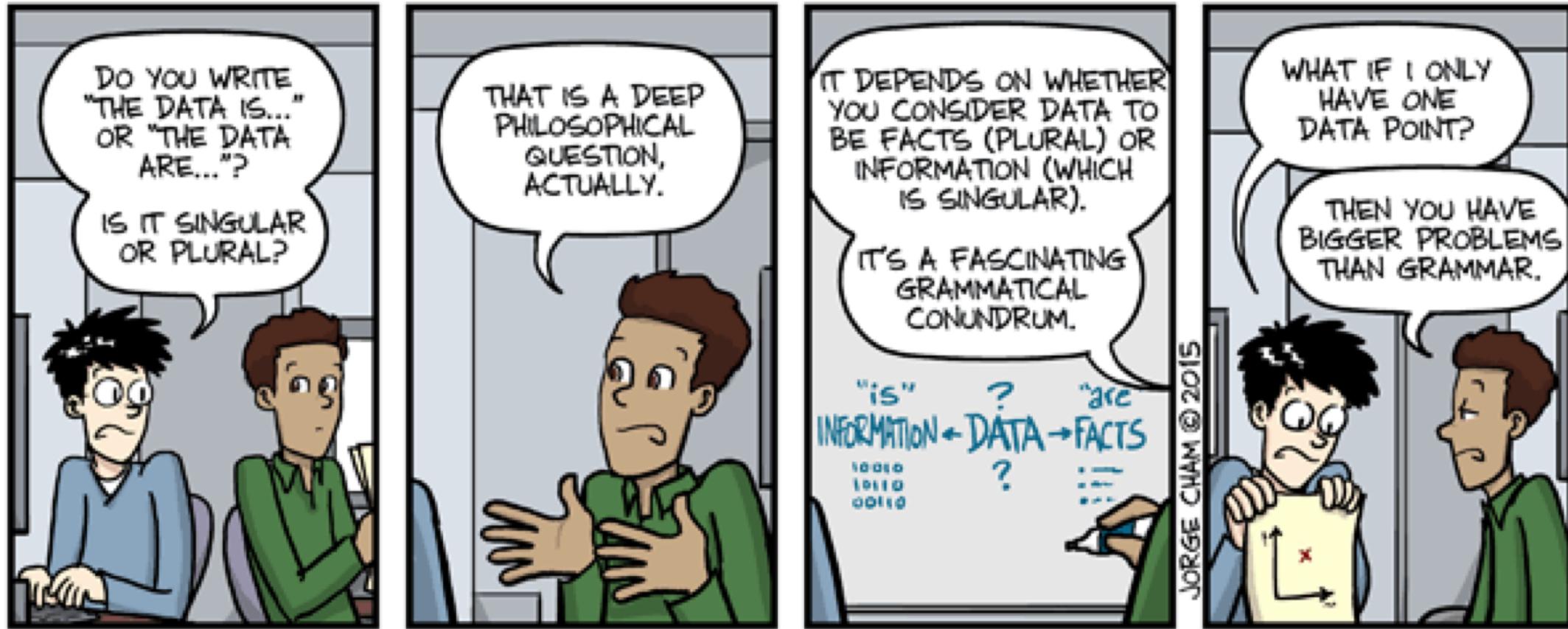
Don't know what it represents.
Just numbers. Still data.

Scenario₄ Readouts from a mysterious sensor that was purchased from a local yard sale.

What is data?

Datum	A single piece of information, which can be treated as an observation
Data	The plural of datum; multiple observations
Dataset	A homogenous collection of data (each datum must have the same focus)

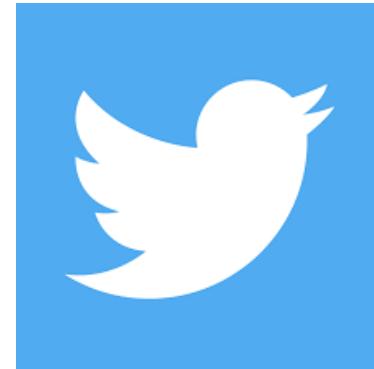
What is data?



Source: http://phdcomics.com/comics/archive_print.php?comicid=1816

What is data?

Everything can be data! Just requires making observations.



Before we dive too deep into the different aspects of data,
recall the *Data Science* process

Ask an interesting question

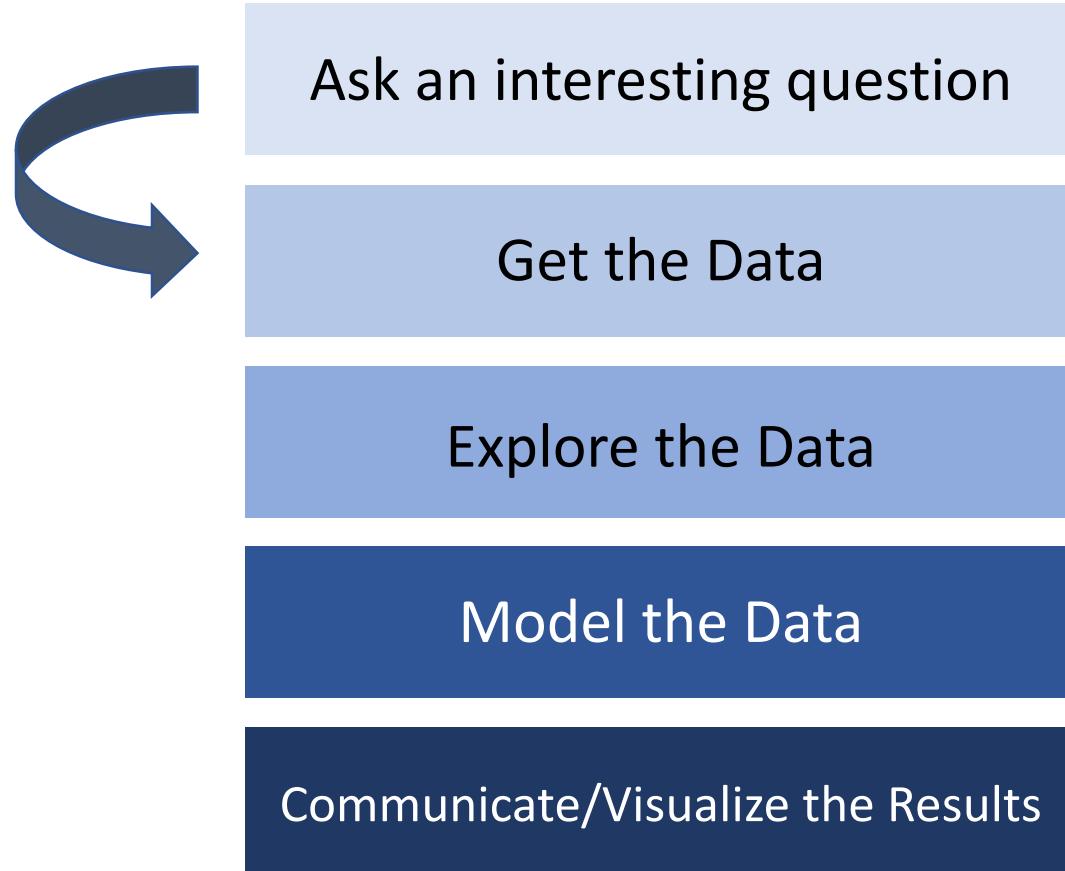
Get the Data

Explore the Data

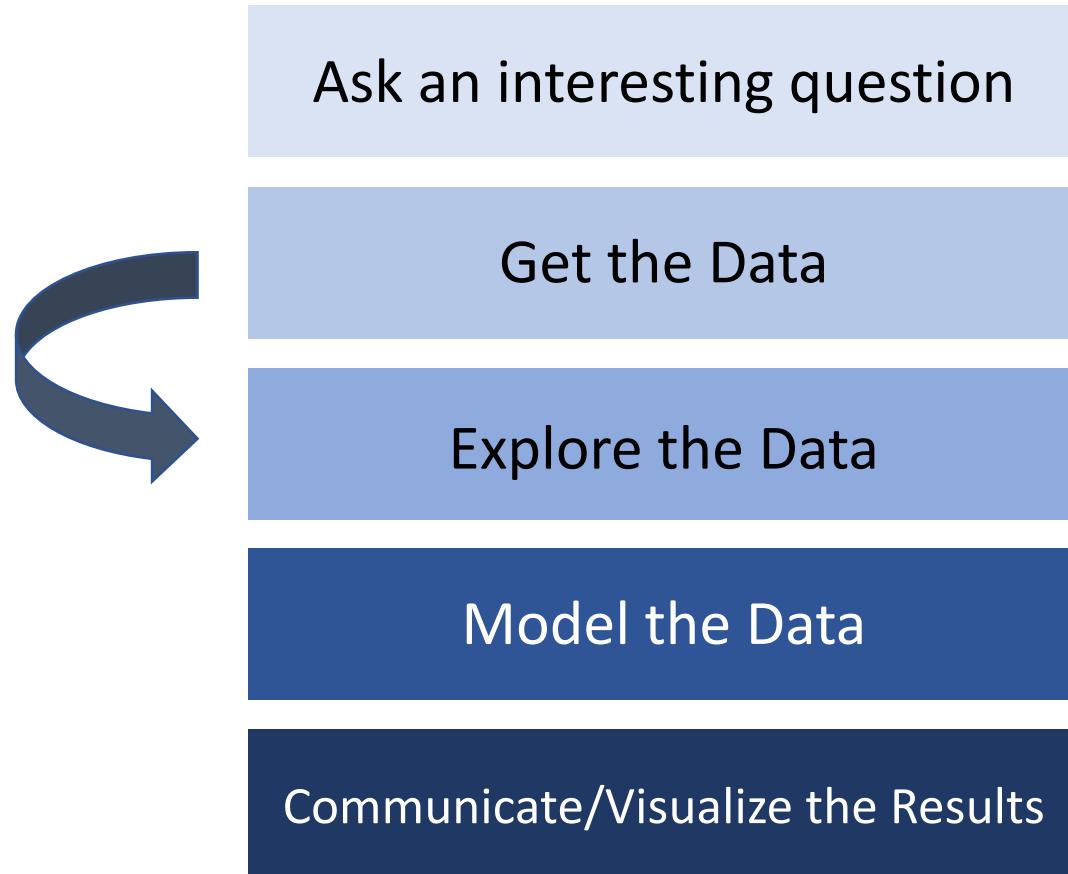
Model the Data

Communicate/Visualize the Results

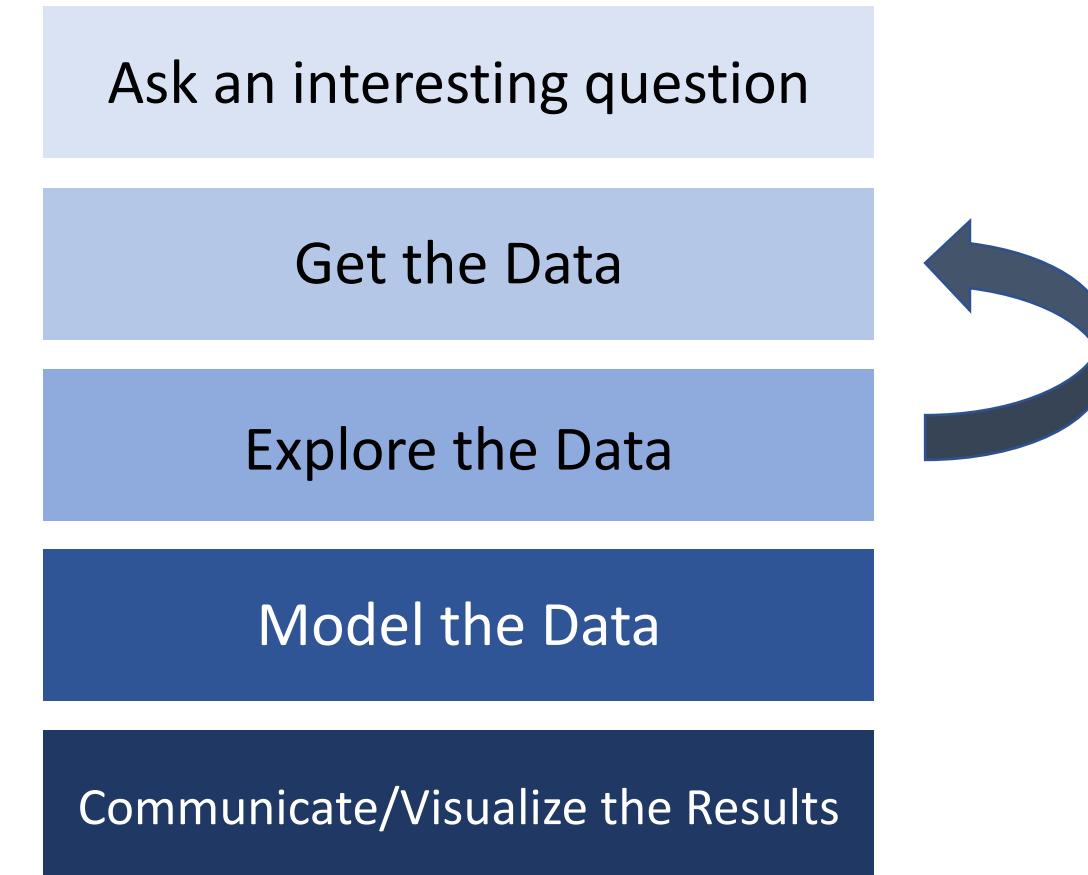
Before we dive too deep into the different aspects of data,
recall the *Data Science* process



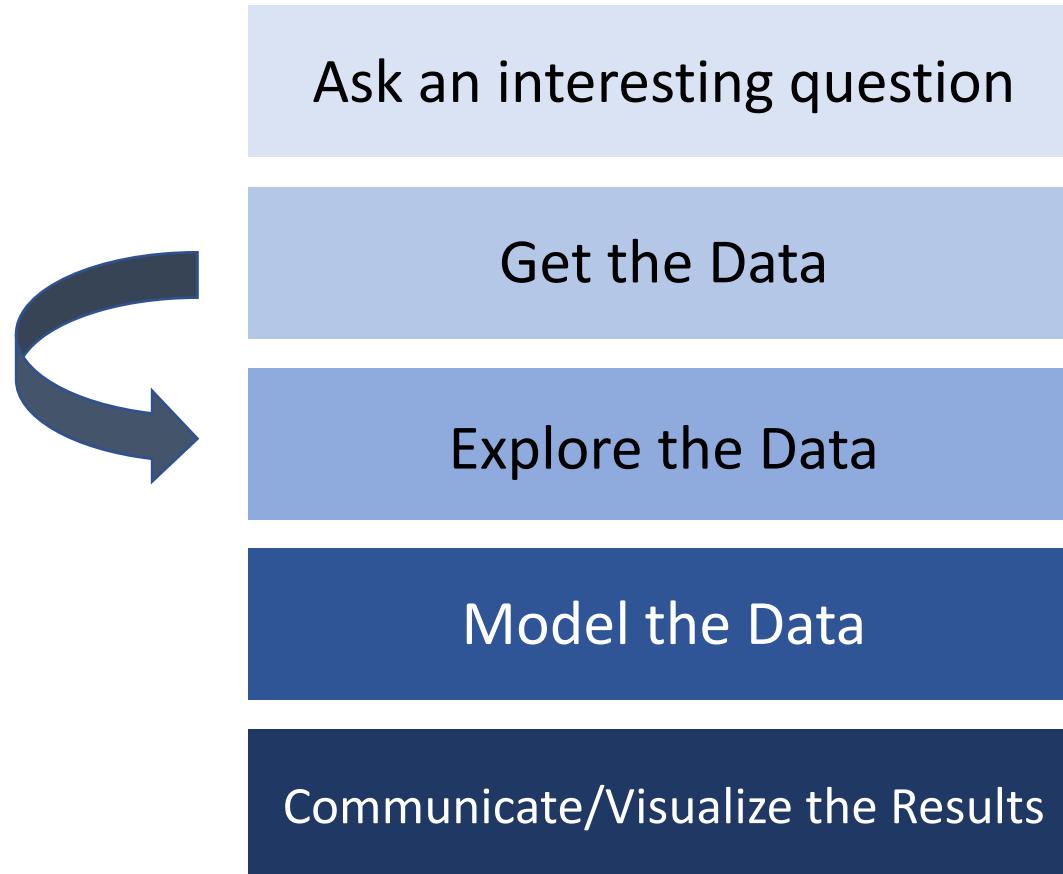
Before we dive too deep into the different aspects of data,
recall the *Data Science* process



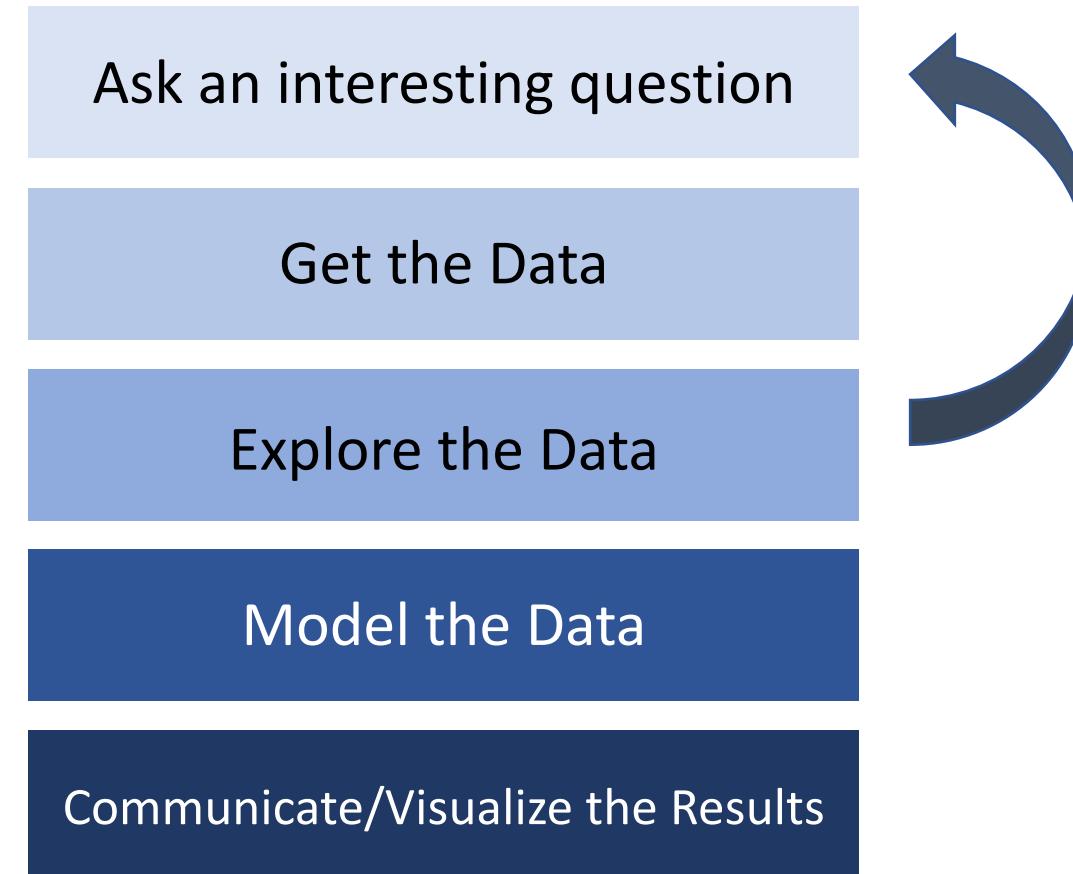
Before we dive too deep into the different aspects of data,
recall the *Data Science* process



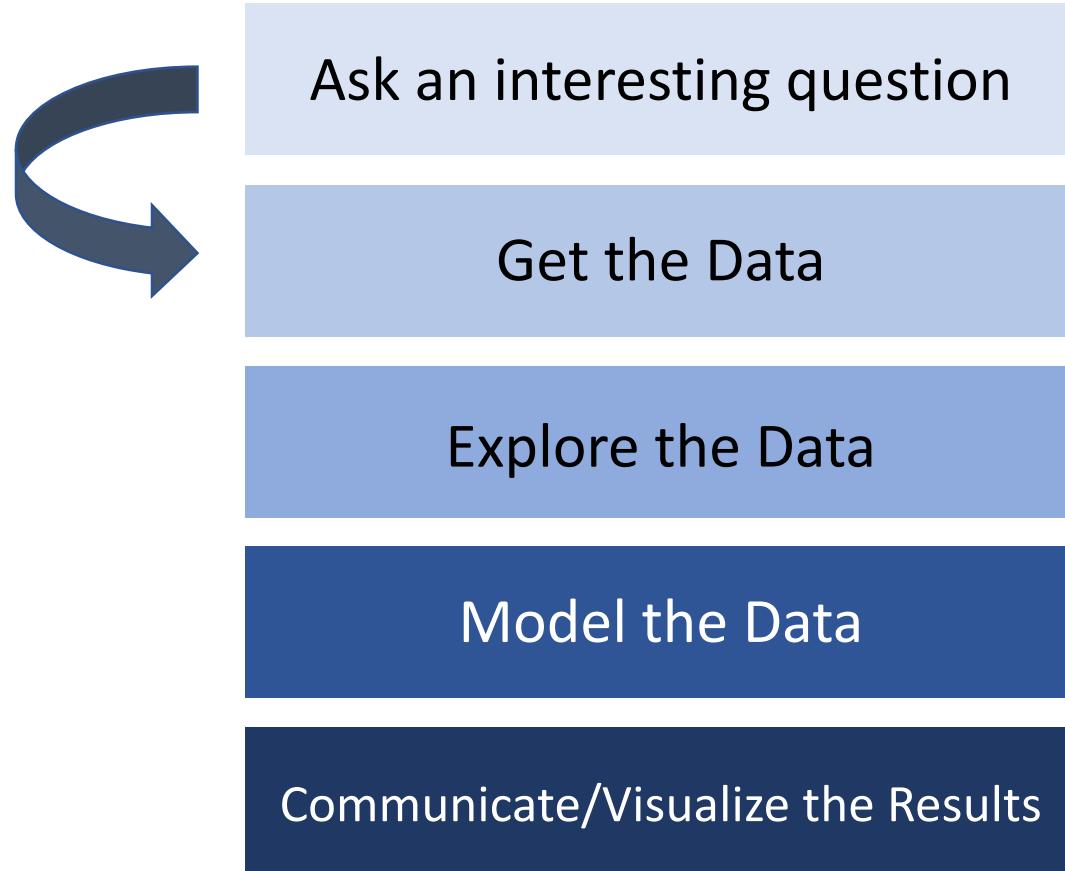
Before we dive too deep into the different aspects of data,
recall the *Data Science* process



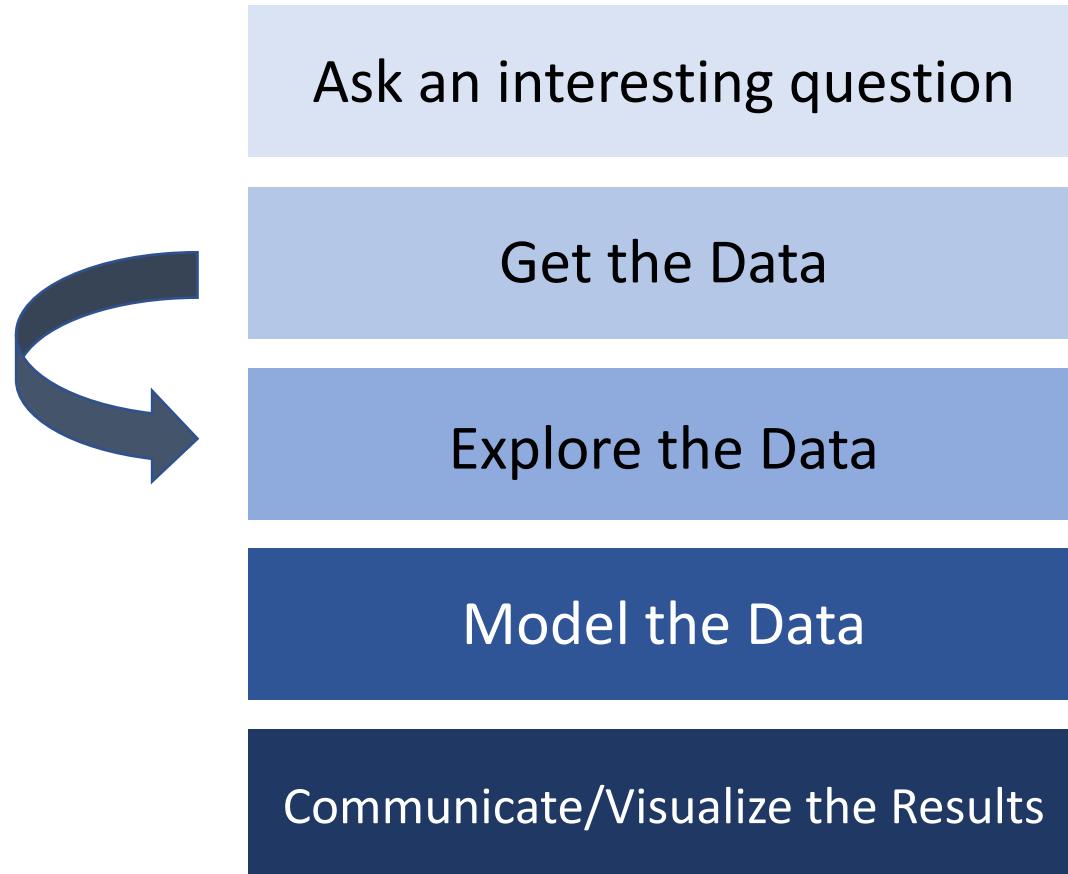
Before we dive too deep into the different aspects of data,
recall the *Data Science* process



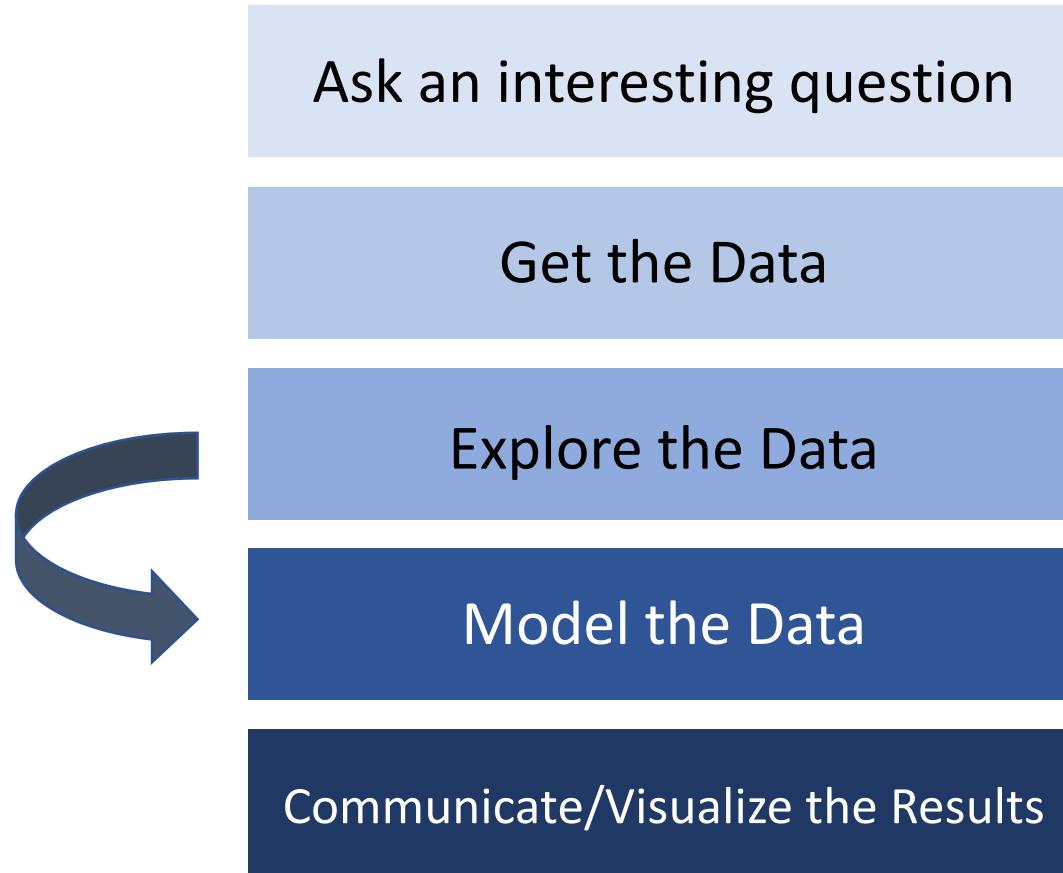
Before we dive too deep into the different aspects of data,
recall the *Data Science* process



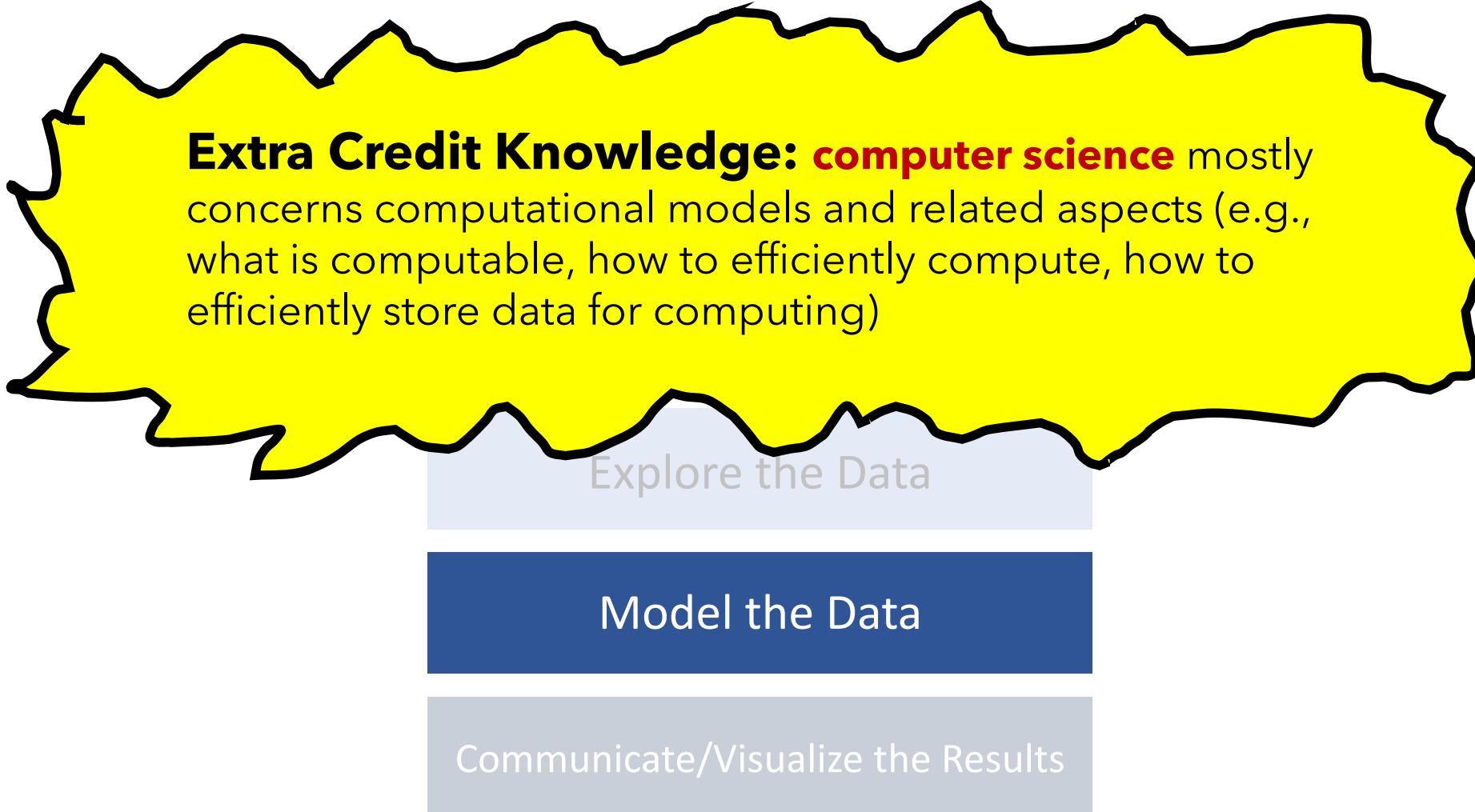
Before we dive too deep into the different aspects of data,
recall the *Data Science* process



Before we dive too deep into the different aspects of data,
recall the *Data Science* process



Before we dive too deep into the different aspects of data,
recall the *Data Science* process



Agenda

- What is data?
- Aspects of data: formats, scope, biases, etc
- Asking precise questions
- Parsing data with Regular Expressions

Agenda



What is data?



Aspects of data: formats, scope, biases, etc



Asking precise questions



Parsing data with Regular Expressions

Considerations when choosing a dataset

We want data that can answer our question(s) and is preferably easy to work with.

Data comes in all shapes and sizes though.

Considerations when choosing a dataset

- What data is necessary to answer our question?
- How difficult is it to analyze a dataset?
- Is the source authoritative? (.com, .net, .org, .gov, .name)
- Comprehensive data vs sampled data?
- Biases
- What is the allowed usage of data under its license?
- Who collected the data?
- When was the data collected?
- How was the data collected?
- How is the data formatted?
- Does your data collection procedures need to be approved by an IRB?
- Confidentiality Concerns

Considerations when choosing a dataset

- What data is necessary to answer our question?
- How difficult is it to analyze a dataset?
- Is the source authoritative? (.com, .net, .org, .gov, .name)
- Comprehensive data vs sampled data?
- Biases
- What is the allowed usage of data under its license?
- Who collected the data?
- When was the data collected?
- How was the data collected?
- How is the data formatted?
- Does your data collection procedures need to be approved by an IRB?
- Confidentiality Concerns

Considerations when choosing a dataset: format difficulty

hard for computers

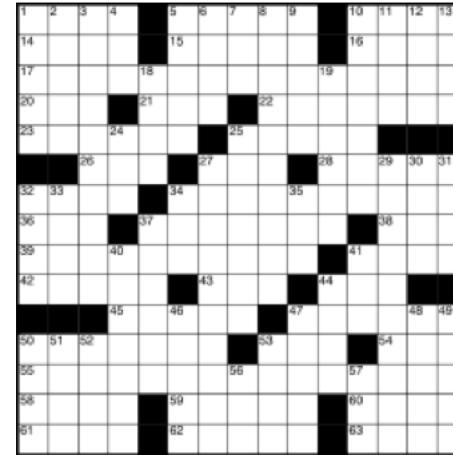


easy for computers



	A	B	C
1	name	age	height
2	Michael	46	5'9"
3	Jim	31	6'0"
4	Pam	29	5'7"
5	Meredith	53	5'6"
6	Dwight	35	5'10"

easy for people



Confusion at Palm Beach County polls

Some Al Gore supporters may have mistakenly voted for Pat Buchanan because of the ballot's design.

Although the Democrats are listed second in the column on the left, they are the third hole on the ballot.

Punching the second hole casts a vote for the Reform party.

The ballot card shows the following layout:

- Electors for President and Vice President:** A note states: "A vote for the candidates will actually be a vote for their electors." Below this, it says "(Votes for Group)".
- Column 1 (Republican):** GEORGE W. BUSH - PRESIDENT, DICK CHENEY - VICE PRESIDENT.
- Column 2 (Democratic):** AL GORE - PRESIDENT, JOE LIEBERMAN - VICE PRESIDENT.
- Column 3 (Libertarian):** HARRY BROWNE - PRESIDENT, ART OLIVIER - VICE PRESIDENT.
- Column 4 (Reform):** PAT BUCHANAN - PRESIDENT, EZOLA FOSTER - VICE PRESIDENT.
- Column 5 (Socialist):** DAVID McREYNOLDS - PRESIDENT, MARY CAL HOLLIS - VICE PRESIDENT.
- Column 6 (Constitution):** HOWARD PHILLIPS - PRESIDENT, J. CURTIS FRAZIER - VICE PRESIDENT.
- Column 7 (Green):** RALPH NADER - PRESIDENT, WINONA LaDUKE - VICE PRESIDENT.
- Column 8 (Socialist Workers):** JAMES HARRIS - PRESIDENT, MARGARET TROWE - VICE PRESIDENT.
- Column 9 (Workers World):** MONICA MOOREHEAD - PRESIDENT, GLORIA La RIVA - VICE PRESIDENT.
- Column 10 (Write-in Candidate):** JOHN HAGELIN - PRESIDENT, NAT GOLDHABER - VICE PRESIDENT.

Instructions at the bottom of the card state: "To vote for a write-in candidate, follow the directions on the long stub of your ballot card."

hard for people



Considerations when choosing a dataset: comprehensive data

- Have access to all the data observations that exist, which is usually a lot
- Collected and digitized as part of generalized procedures of an institution

The New York Times

13 million articles



~500 million tweets per day

CONGRESS.GOV

100,000s votes per year

Considerations when choosing a dataset: sampled data

- When collecting individual data is relatively expensive
- Only a portion of the population is sampled
- Not just restricted to polling or surveys



1. **Clover Food Lab**

821 reviews

\$\$ • American (New),
Sandwiches, Cafes



nielsen
.....

Common biases in selecting the source of data

- **Omission:** Using only arguments from one side
- **Source selection:** Include more sources or more authoritative sources for one side over the other
- **Story selection:** Regularly including stories that agree or reinforce the arguments of one side
- **Placement:** Using the benefit of the perceived importance of position to highlight certain stories

Common biases in selecting the source of data

- **Labelling (two types):**
 - Using only arguments from one side
 - Labeling people on one side of the argument with labels and not the other
- **Spin:** Story provides only one interpretation of the events

Common biases in the data itself (i.e., sampled datasets)

- **A bias in sampled data occurs when a procedure causes the sample to overrepresent a subpopulation**
- Biases may not necessarily be intentional
- Even if you don't *think* your over-/ under-representation of a subpopulation will impact your results, it's still a bias
- Always strive to minimize any biases in your data collection procedures

Gallup Polls

- Randomly calls two groups of ~500 people a day by sampling among all possible phone numbers
- For landlines, asks for household member who has the next birthday
- Calls people living in all 50 states
- Tries to assure 70% cellphone, 30% landlines
- Weights data to reflect the demographics of the general population

IMDb Movie Ratings

- Registered users rate films 1-10 stars; they are an overrepresented subpopulation relative to the general population
- Registered users who rate movies in their free time further over represents a specific segment of the general population
- *“Men Are Sabotaging The Online Reviews Of TV Shows Aimed At Women¹”*
60% who rated Sex in the City were women. Women gave it a 8.1, men gave it 5.8.

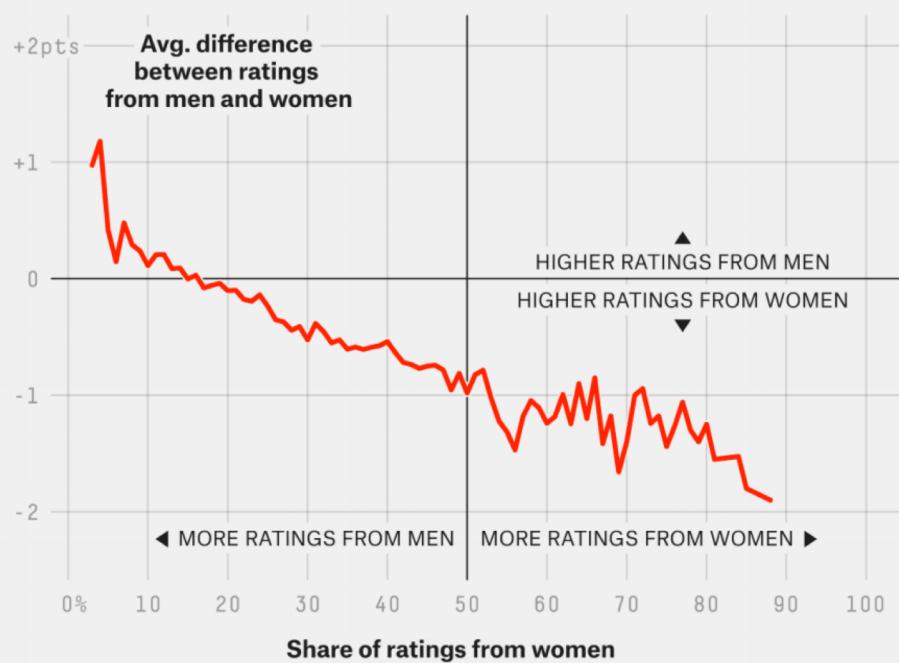
¹ fivethirtyeight.com

Considerations when choosing a dataset: biases

IMDb Movie Ratings

Men tank the ratings of shows aimed at women

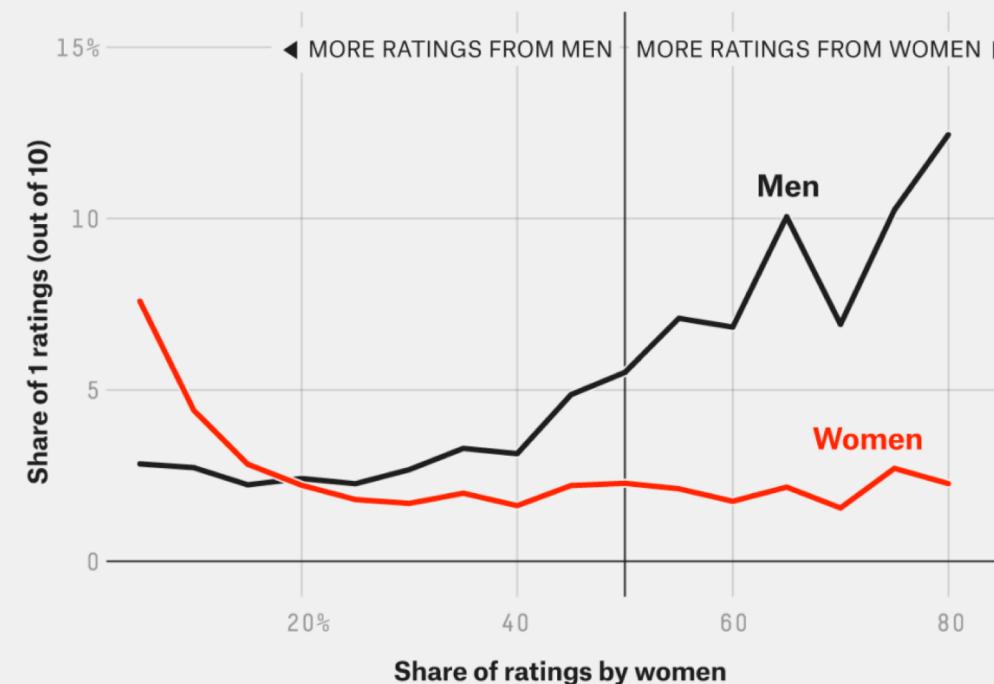
Average difference between IMDb ratings of TV shows from men and women by share of ratings from women



For English language shows with 1,000 or more ratings

Men are more likely to give the crappiest rating

Share of IMDb ratings of 1 (out of 10) for shows with at least 10,000 ratings by share of ratings from women*



*Rounded to nearest 5 percent

Yelp Reviews

- Registered users rate businesses on a 1-5 star scale
- Registered users tend to represent a certain subset of the population (those who are more social media inclined and opinionated)
- Customers with extreme experiences are more likely to voice their opinions

Yelp Reviews



6. Clover Food Lab



104 reviews

\$ · Sandwiches, Cafes,
American (New)



1. Clover Food Lab



821 reviews

\$\$ · American (New),
Sandwiches, Cafes

Yelp Reviews



6. Clover Food Lab



104 reviews

\$ · Sandwiches, Cafes,
American (New)



1. Clover Food Lab



821 reviews

\$\$ · American (New),
Sandwiches, Cafes

Longwood Medical

Harvard Square

Nearly all datasets involve a human in some way or another, and our world is far from being uniform and equal. **This is not an excuse but evidence** that your dataset probably has bias. The goal is to minimize it as much as possible.

When we learn about modelling, the same applies.

While computers are getting better at
'understanding' photos and videos, **text and numbers are much easier.**

Further, **structured** data (e.g., spreadsheet formatted data)
is much easier than
unstructured data (e.g., free-flowing essays)

Plain Text

- Ends in .txt (generally)
- No formatting, font type, font size, color, etc.
- Text position is provided by whitespace characters (space, tab, return)

ALICE'S ADVENTURES IN WONDERLAND

Lewis Carroll

THE MILLENNIUM FULCRUM EDITION 3.0

CHAPTER I. Down the Rabbit-Hole

Alice was beginning to get very tired of sitting by her sister on the bank, and of having nothing to do: once or twice she had peeped into the book her sister was reading, but it had no pictures or conversations in it, 'and what is the use of a book,' thought Alice 'without pictures or conversations?'



- CSV (.csv)
- Tab-separated (.tsv)
- **Delimiter:** The character that separates each value

Plain Text

sample.csv	
Bill #,	Jack Reed (RI), Elizabeth Warren (MA)
Bill 27,	Yay, Yay
Bill 28,	Yay, Nay
Bill 30,,	Nay
Bill 47,	Nay, Nay
Bill 91,	Nay, Nay
Bill 105,	Yay,Yay

- XML (.xml)
- These colors —>
aren't actually stored
in the file, the editor
just adds them on
your screen to help
make it look prettier

XML

```
<roll_call_vote>
<congress>115</congress>
<session>1</session>
...
<members>
  <member>
    <member_full>Alexander (R-TN)</member_full>
    <last_name>Alexander</last_name>
    <first_name>Lamar</first_name>
    <party>R</party>
    <state>TN</state>
    <vote_cast>Yea</vote_cast>
  ...
  </member>
</members>
</roll_call_vote>
```

JSON

- JSON (.json)
- **JavaScript Object Notation**
- Like XML, data is annotated
- A nesting of key-value pairs
- When whitespace is removed, can be more space efficient than XML

```
{  
    "firstName": "John",  
    "lastName": "Smith",  
    "isAlive": true,  
    "age": 27,  
    "address": {  
        "streetAddress": "21 2nd Street",  
        "city": "New York",  
        "state": "NY",  
        "postalCode": "10021-3100"  
    },  
    "phoneNumbers": [  
        {  
            "type": "home",  
            "number": "212 555-1234"  
        },  
        {  
            "type": "office",  
            "number": "646 555-4567"  
        },  
        {  
            "type": "mobile",  
            "number": "123 456-7890"  
        }  
}
```

Plain Text vs XML vs JSON

- They can all express the same content
- Plain Text doesn't have structure, but is universally robust
- XML is the most verbose, harder to parse
- JSON doesn't have </stuff_here> end tags
- JSON is more succinct than XML (easier to parse)

It's important to re-evaluate your previous steps to ensure you're on the right track

Ask an interesting question

Get the Data

Explore the Data

Model the Data

Communicate/Visualize the Results

Agenda

- What is data?
- Aspects of data: formats, scope, biases, etc
- Asking precise questions
- Parsing data with Regular Expressions

Asking the right questions

It's crucial for your question of inquiry to have precise, defined terms
that can be proven true or false

"Voting is at an all-time high"

It's crucial for your question of inquiry to have precise, defined terms
that can be proven true or false

"Voting is at an all-time high"

- Where? USA? World-wide?
- What type of voting? Presidential races, local elections?
- What is our metric? Number of total votes. Percentage of the population?
- What's our actual time scale? "all-time"?

Imagine the following assertion as the thesis of one's project:

"People never vote against party lines anymore"

They then collect some data, run experiments, and conclude by saying they proved their hypothesis.

Take a guess as to what they were trying to investigate, form a more precise question. If time allows, imagine what type of data you'd use and how you'd go about answering it.

The more specific your questions, the more meaningful your results can be.

I urge you all to be aware of biases (both in your data and in your modelling) as much as you can. Doing so will ensure you are providing results that accurately represent reality, leading to more equitable interpretations and uses of your work.

This is immensely important, for Data Science will only continue to play an increasingly powerful and influential role in our society and world at large.

Agenda

- What is data?
- Aspects of data: formats, scope, biases, etc
- Asking precise questions
- Parsing data with Regular Expressions

Regular Ex*

* Not a Taylor Swift song

Many datasets are laboriously curated, and are said to be “**cleaned**”

A **cleaned** dataset is one that has been taken from its original raw form, and has been modified (e.g., errors fixed, missing data amended, extraneous information removed) to a form that is easy for processing.

For example, often we want to remove or replace characters from the original text to simplify the grouping of words or sentences

Often with text data, you are interested in finding words or phrases that match a pattern (e.g., a bunch of letters together followed by a comma)

If the pattern is found, then you often want to either

- **replace** that pattern (e.g., remove the comma) and/or
- **return** the contents that matched the pattern

Finding all Matched Patterns — Manually!



BEYONCÉ  @Beyonce · 19 Aug 2013

RT @jleicole For #WHD2013, I ran 5.312 @CharityMiles to help @GirlUp educate girls in the developing world. #EveryMileMatters

#BeyGood



8.5K



22K



47K



How would you extract the hashtag from this tweet?

Regular Expressions

Finding all Matched Patterns — Manually!

```
tweet = "RT @jleicole For #WHD2013, I ran 5.312 @CharityMiles to help @Girl  
current_hashtag = ""  
hashtags = []  
is_in_hashtag = False  
for i in range(len(tweet)):  
    if tweet[i] == " ": # found a space, so we've possibly ended a hashtag  
        if current_hashtag != "":  
            hashtags.append(current_hashtag)  
            current_hashtag = ""  
        is_in_hashtag = False  
    else:  
        if tweet[i] == "#": # the start of a hashtag  
            is_in_hashtag = True  
        if is_in_hashtag == True:  
            current_hashtag += tweet[i]  
if current_hashtag != "":  
    hashtags.append(current_hashtag)
```

REGULAR EXPRESSIONS TO THE RESCUE!

We can import Python's Regular Expression library via:

```
import re
```

Finding all Matched Patterns — with Regex!

findall() returns a list of all substrings that match the pattern

Finding all Matched Patterns — with Regex!



BEYONCÉ  @Beyonce · 19 Aug 2013

RT @jleicole For #WHD2013, I ran 5.312 @CharityMiles to help @GirlUp educate girls in the developing world. #EveryMileMatters
#BeyGood



8.5K



22K



47K



```
tweet = "RT @jleicole For #WHD2013, I ran ..."  
pattern = "#[^, ]+"  
hashtags = re.findall(pattern, tweet)
```

Replacing Text

```
sentence = "Ms. Smith, are you okay?!? Please talk to me! Oh dear ..."
```

Imagine we want to replace all end-punctuation with a period, so that our text looks like:

```
sentence = "Ms. Smith, are you okay. Please talk to me. Oh dear."
```

Replacing Text

```
sentence = "Ms. Smith, are you okay?!? Please talk to me! Oh dear ..."
```

This would normally be annoyingly tedious to write code for.

```
pattern = "[ ?! ]+| \s*\.\+"  
sentence = re.sub(pattern, '.', sentence)
```

Replacing Text with Regex!

```
re.sub(pattern, replacement, text)
```

sub() replaces all matches in text with the replacement text

Patterns work by matching on:

- specific characters (e.g., 'z') or
- large categories of characters (e.g., all lowercased letters or all digits)

WORKED EXAMPLE:

“Code didn't work, no idea why...”

Specific Characters

```
text = "Code didn't work, no idea why..."  
pattern = 'a'  
re.findall(pattern, text)
```

Output: a

Specific Characters

```
text = "Code didn't work, no idea why..."  
pattern = '[aeiouy]'  
re.findall(pattern, text)
```

The [] brackets denote “any of these characters”

Output: ['o', 'e', 'i', 'o', 'o', 'i', 'e', 'a', 'y']

Specific Characters

```
text = "Code didn't work, no idea why..."  
pattern = '[ a-z ]'  
re.findall(pattern, text)
```

The [] brackets denote “any of these characters”

Output: ['o', 'd', 'e', 'd', 'i', 'd', 'n', 't', 'w', 'o', 'r', 'k', 'n',
'o', 'i', 'd', 'e', 'a', 'w', 'h', 'y']

Specific Characters

```
text = "Code didn't work, no idea why..."  
pattern = '[ a-zA-Z ]'  
re.findall(pattern, text)
```

The [] brackets denote “any of these characters”

Output: ['C', 'o', 'd', 'e', 'd', 'i', 'd', 'n', 't', 'w', 'o', 'r', 'k',
'n', 'o', 'i', 'd', 'e', 'a', 'w', 'h', 'y']

Repeated Characters

```
text = "Code didn't work, no idea why..."  
pattern = '[ a-zA-Z ]+'  
re.findall(pattern, text)
```



The + sign means 1 or more occurrences must appear
(greedy approach of matching)

Output: ['Code', 'didn', 't', 'work', 'no', 'idea', 'why']

Repeated Characters

```
text = "Code didn't work, no idea why..."  
pattern = '[ a-zA-Z ]*'  
re.findall(pattern, text)
```



The * sign means 0 or more occurrences must appear
(greedy approach of matching)

Output: ['Code', '', 'didn', '', 't', '', 'work', '', '', 'no', '',
'idea', '', 'why', "", "", "", "", ""]

Repeated Characters

Instead of matching on 0 or more or 1 or more occurrences, you can also specify an exact number of occurrences N with {N}

Regular Expressions

```
text = "555-123-1234, 33-555-123-5678"  
pattern = '\d{3}-\d{3}-\d{4}'  
re.findall(pattern, text)
```

$\backslash d\{3\}$ means exactly 3 single-digits in a row

Output: ['555-123-1234', '555-123-5678']

```
text = "555-123-1234, 33-555-123-5678"  
pattern = '\d{1,3}-\d{3}-\d{3}-\d{4}'  
re.findall(pattern, text)
```

What do you think this matches?

Regular Expressions

```
text = "555-123-1234, 33-555-123-5678"  
pattern = '\d{1,3}-\d{3}-\d{3}-\d{4}'  
re.findall(pattern, text)
```

Output: ['555-123-1234', '555-123-5678']

RegEx Syntax

- **\w** Any alphanumeric character and underscore,
equivalent to [a-zA-Z0-9_]
- **\s** Matches any whitespace (spaces, tabs, line breaks)
- **\d** Matches any digit character, equivalent to [0-9]

RegEx Syntax

Regular Expression Character Classes

[ab-d] One character of: a, b, c, d

[^ab-d] One character except: a, b, c, d

\b Backspace character

\d One digit

\D One non-digit

\s One whitespace

\S One non-whitespace

\w One word character

\W One non-word character

Special characters

\	escape special characters
.	matches any character
^	matches beginning of string
\$	matches end of string
[5b-d]	matches any chars '5', 'b', 'c' or 'd'
[^a-c6]	matches any char except 'a', 'b', 'c' or '6'
R S	matches either regex R or regex S
()	creates a capture group and indicates precedence

Quantifiers

*	0 or more (append ? for non-greedy)
+	1 or more (append ? for non-greedy)
?	0 or 1 (append ? for non-greedy)
{m}	exactly m m occurrences
{m, n}	from m to n . m defaults to 0, n to infinity
{m, n}?	from m to n , as few as possible

Special sequences

\A	start of string
\b	matches empty string at word boundary (between \w and \w)
\B	matches empty string not at word boundary
\d	digit
\D	non-digit
\s	whitespace: [\t\n\r\f\v]
\S	non-whitespace
\w	alphanumeric: [0-9a-zA-Z_]
\W	non-alphanumeric
\z	end of string
\g<id>	matches a previously defined group

Special sequences

(?iLmsux)	matches empty string, sets re.X flags
(?:...)	non-capturing version of regular parentheses
(?P...)	matches whatever matched previously named group
(?P=)	digit
(?#...)	a comment; ignored
(?=...)	lookahead assertion: matches without consuming
(?!...)	negative lookahead assertion
(?<=...)	lookbehind assertion: matches if preceded
(?<!...)	negative lookbehind assertion
(?)	match 'yes' if group 'id' matched, else 'no'
(id)yes no	

Learning Objectives

- Understand different types and formats of data
- Be able to soundly select appropriate data
- Have awareness of biases that exist
- Be able to refine questions to suite your true inquiry
- Understand how to parse text with regular expressions

BREAK-OUT ROOM TIME!

ALL THE

