

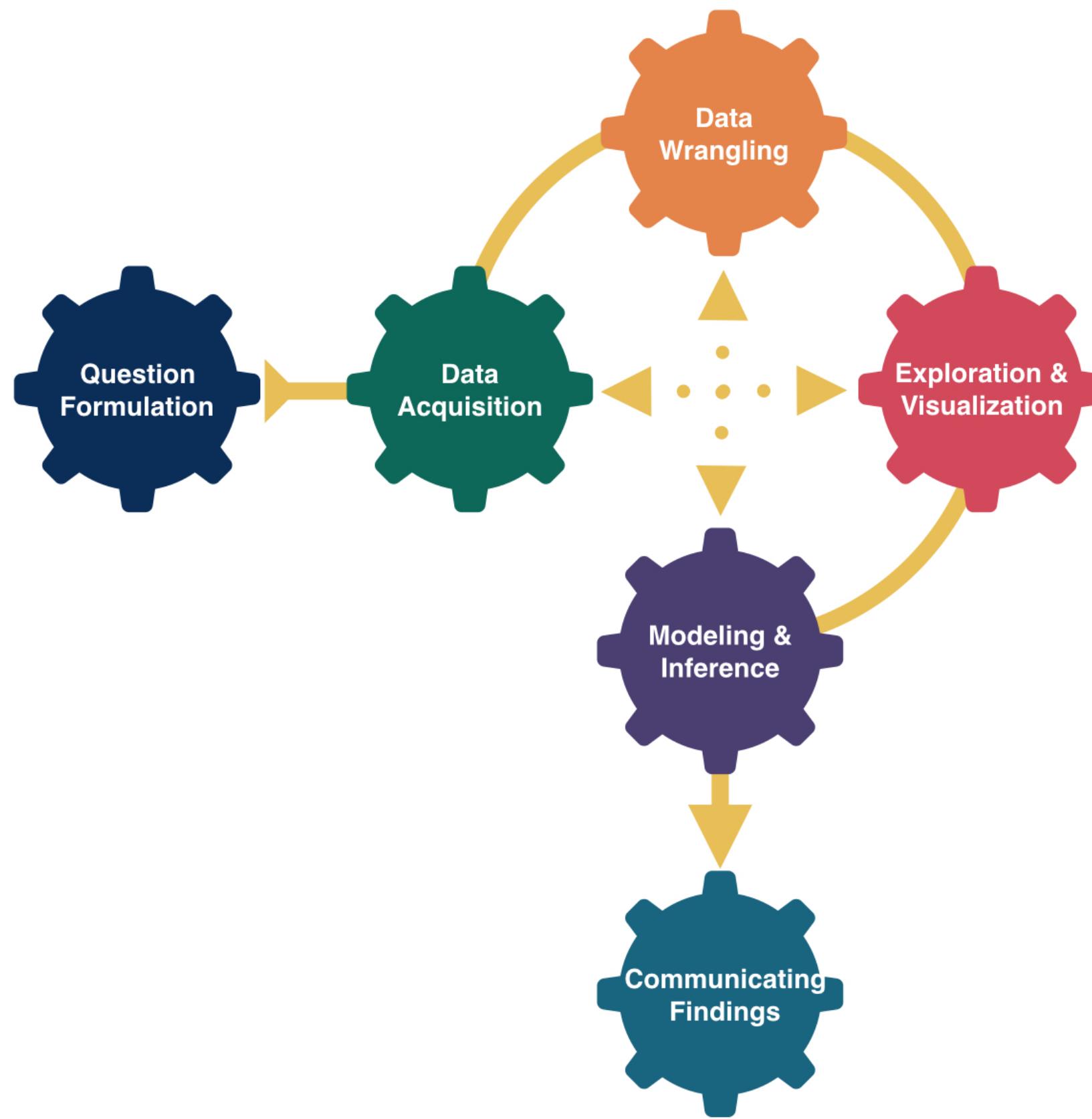
What words or phrases do you think of when you hear the word “Harvard”?

This being a **data** class, I’d like to collect some data related to “statistical thinking.”

Go to [bit.ly/stat-100-think](http://bit.ly/stat-100-think) to provide the words or phrases you think of when you hear “statistical thinking.”



# Statistical Thinking



Kelly McConville  
Stat 100  
Week 1 | Fall 2023

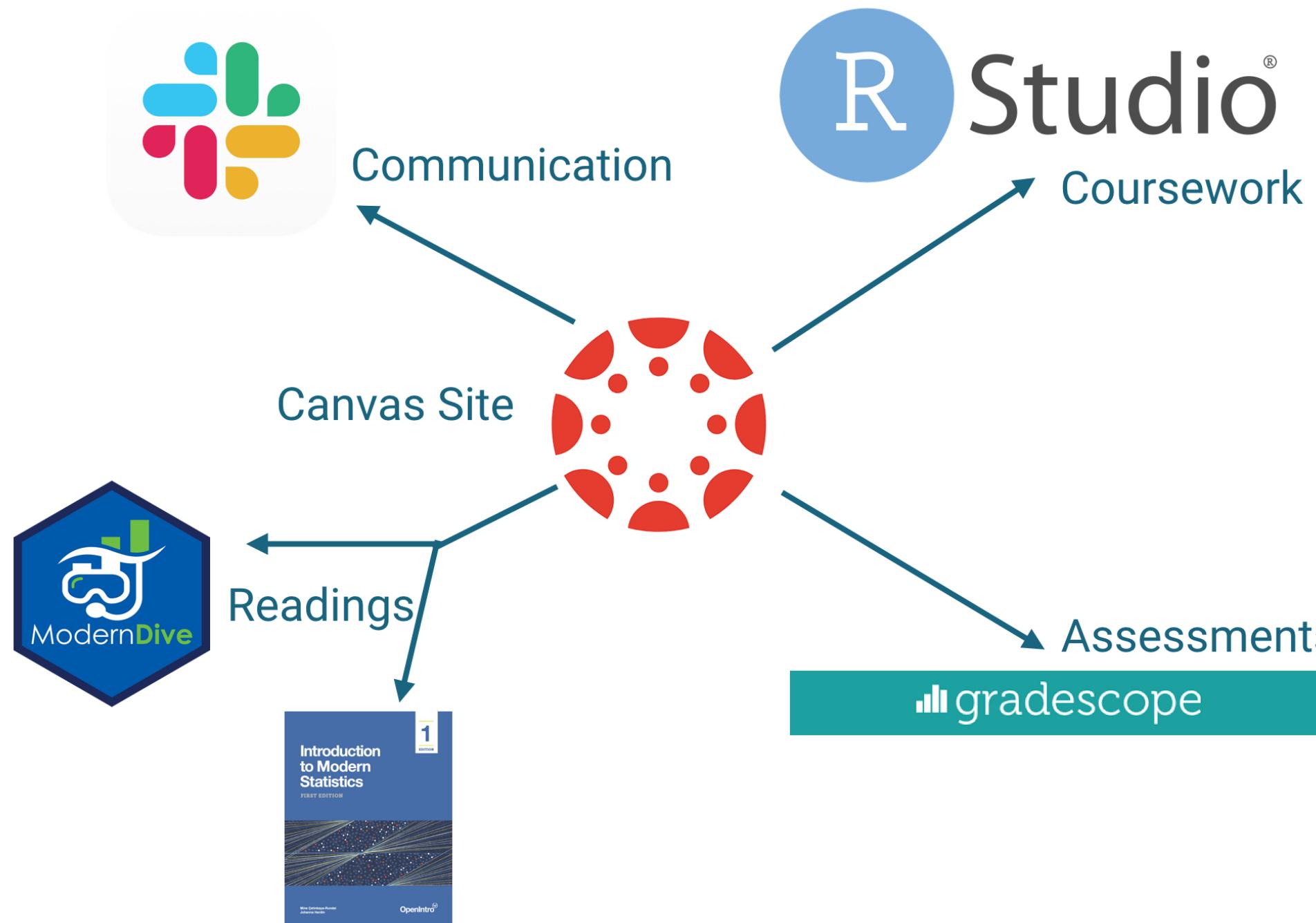
# Getting Started in Stat 100

## Step 1: Getting Started Module in Canvas

### ▼ Getting Started

-  **Read over the course syllabus**
-  **Check out the office hours schedule! (The standard schedule will begin in Week 2.)** 
-  **Join the class Slack workspace**

# Stat 100 Tech & Materials





# Announcements

- Lecture slide decks will always be posted and linked to a Canvas Module the day before lecture.
  - Will also bring printed versions for those who prefer paper copies.
- No section and no lecture quiz this week.
  - But be on the look-out for section preference form from my.harvard.
- Only I will be running office hours this week at the following time:
  - Today 1:30 - 3:00 pm in Science Center 316 (This week only)
- The regular office hour schedule will be posted later this week and will start next week.
- **If able, please bring a laptop or tablet to Mondays's lecture.**

# Day 1 Goals

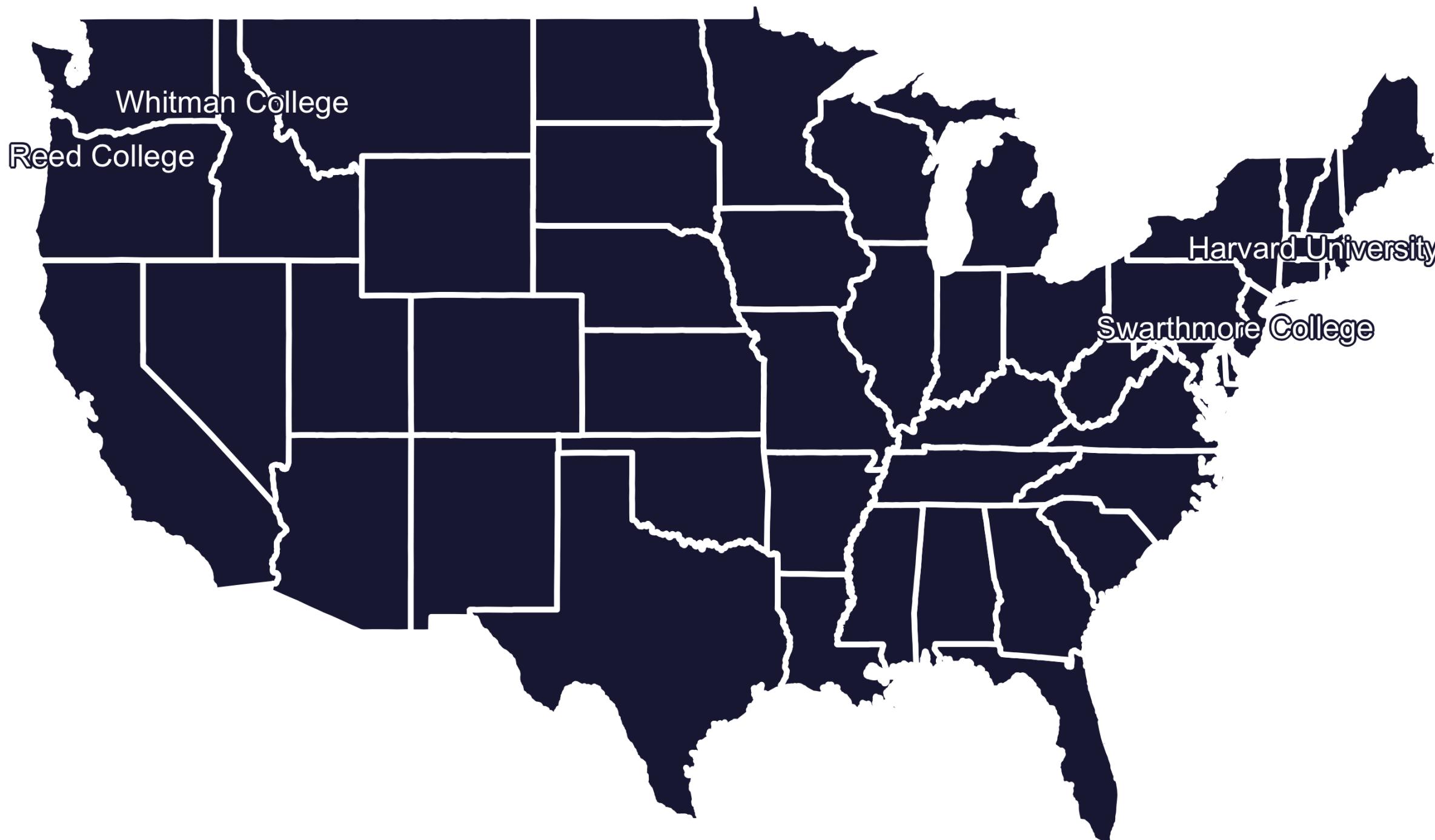
- Start engaging in statistical thinking
- Introduce data
- Consider hand-drawn visualizations as a way to tell stories with data
- Hop into the RStudio Server using Posit Cloud

# Looking Ahead to Day 2...

- Discuss course structure (lecture, section, wrap-ups, office hours, assessments...)
- Present important course policies (engagement, code of conduct, chatGPT, ...)
- Get started in **RStudio** and with **Quarto** documents

But first, let me quickly introduce  
myself...

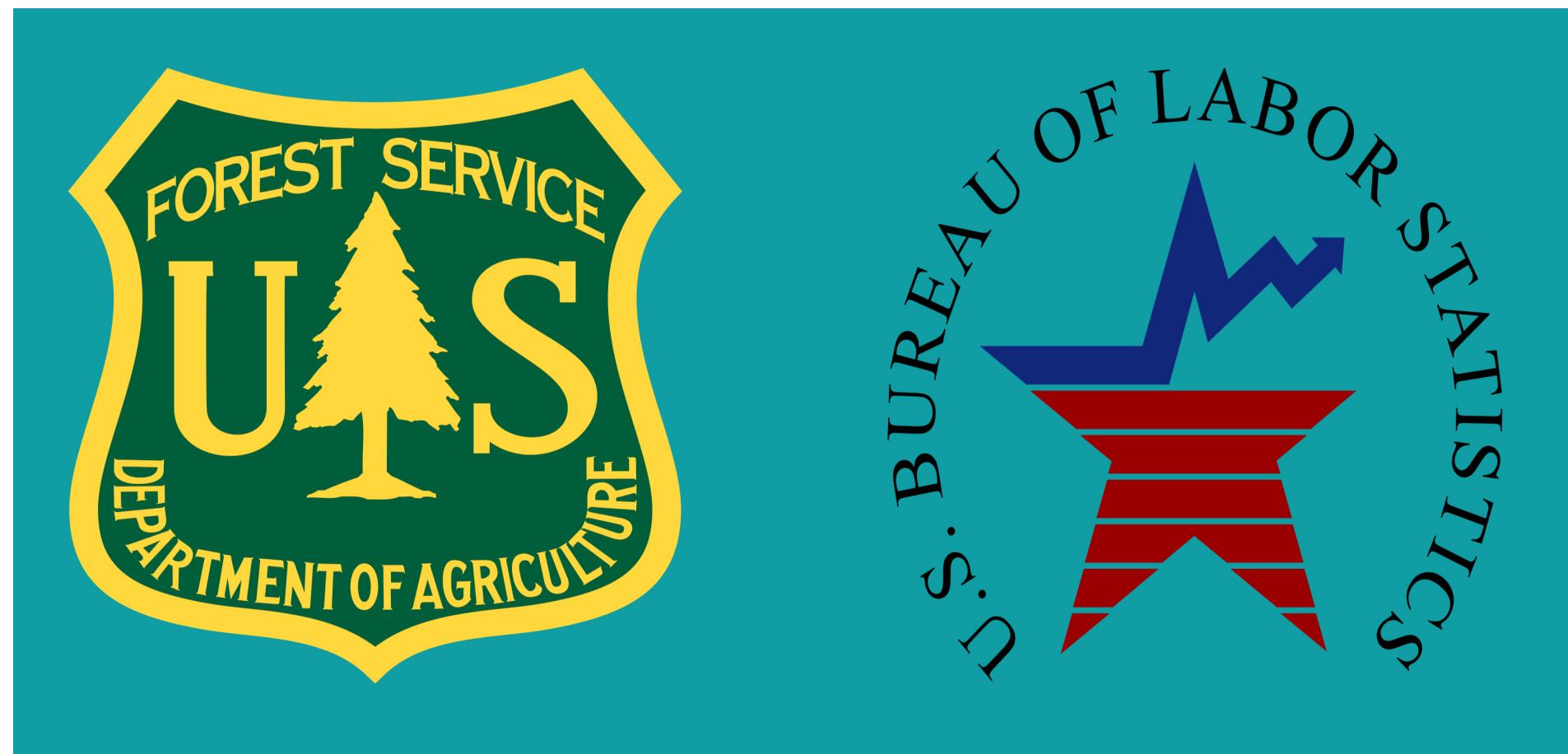
# Let's start with my path to Harvard...





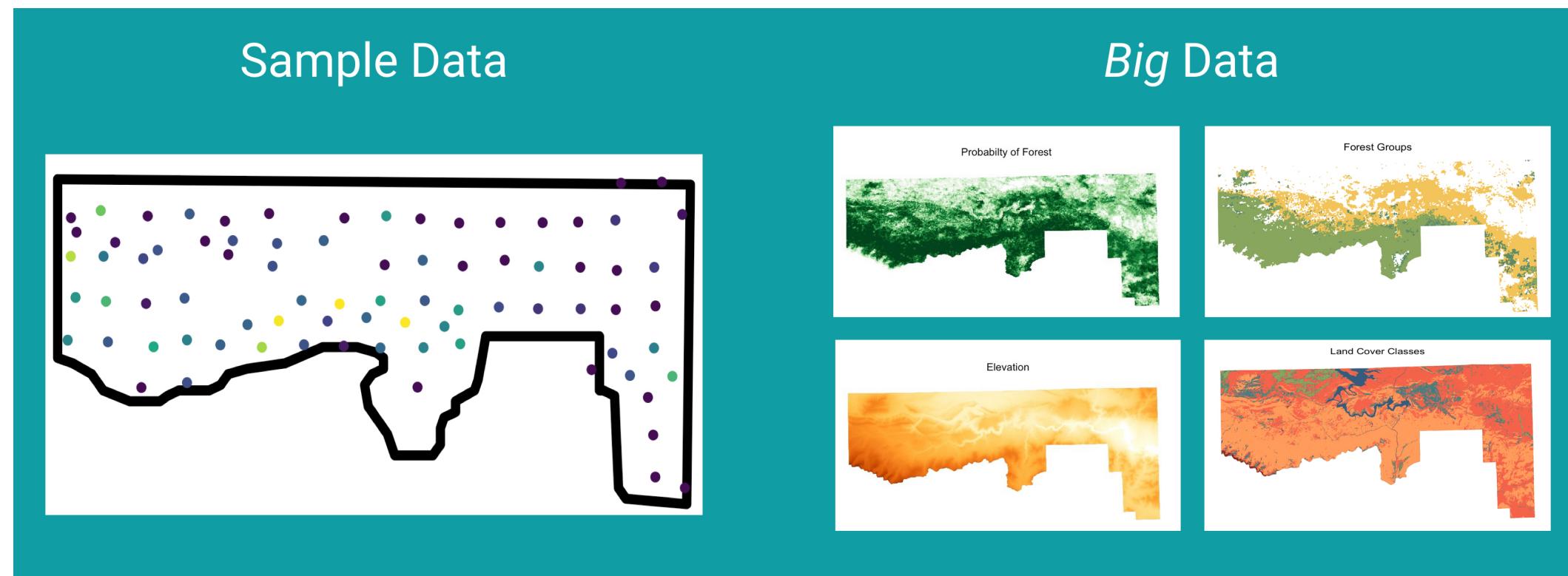
# Research Interests

## Survey statistics and collaborate with



# Research Interests

## Where survey statistics meets data science



# Advising Undergraduate Forestry Data Science Research







- I **love** teaching stats and coding.
- But, learning stats and coding is **hard**.
- With the **right scaffolding, good strategies, and sustained effort**, you can excel at both!
- And mistakes are part of the learning process. They don't imply that you are bad at stats.

**Also, the Stat 100 Teaching Team  
are so excited to support your  
learning!**

**Stat 100 is about developing our statistical thinking skills.**

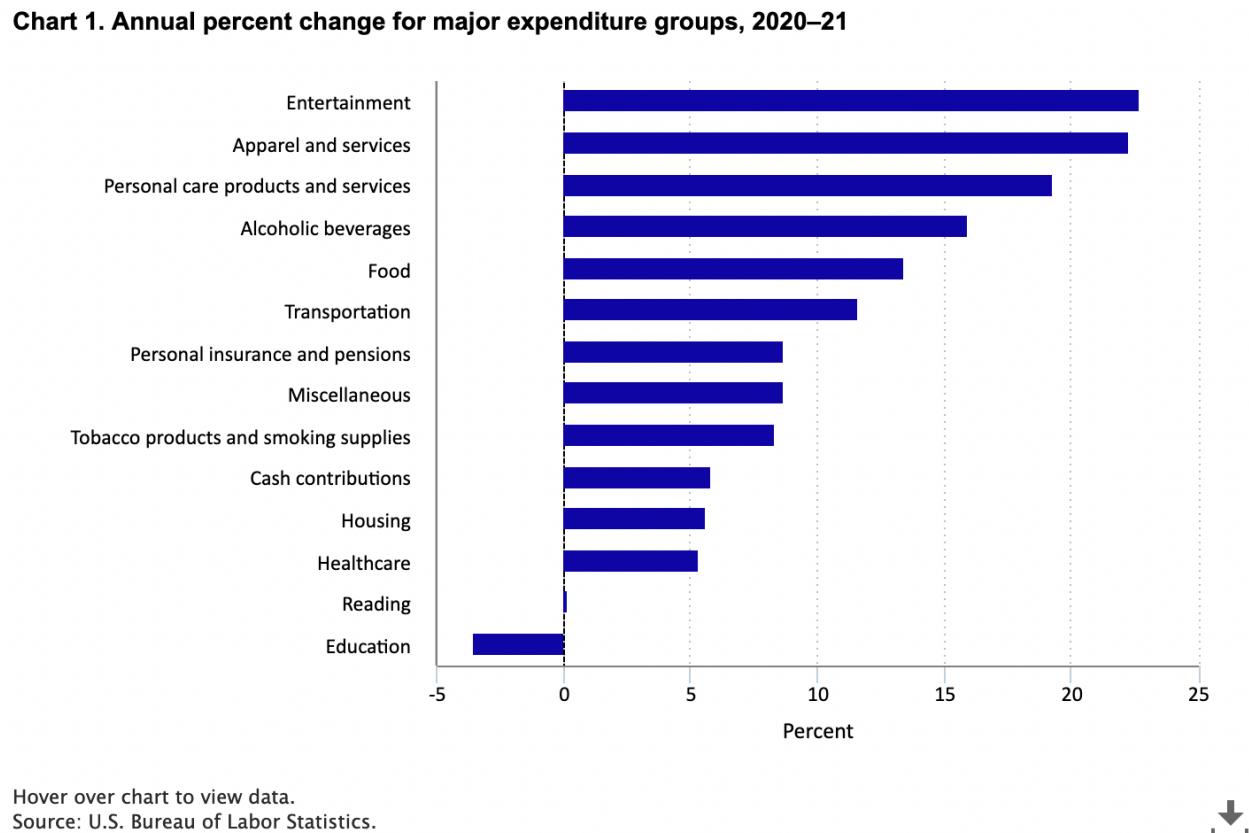
What is **statistical thinking**?

It is not the same as mathematical thinking.

Let's discover what **statistical thinking** is through some examples.

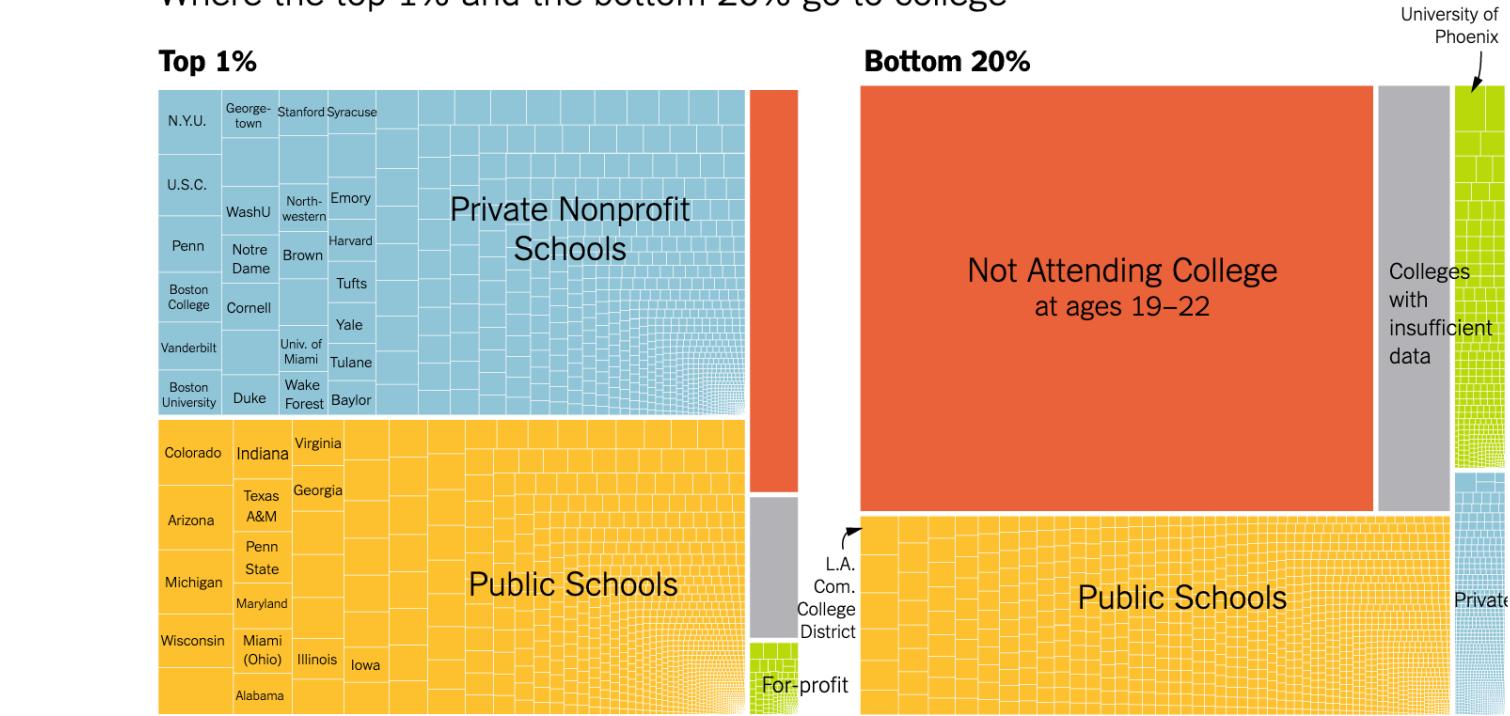
# Data in Stat 100

Will use a wide-range of **real** and **relevant** data examples

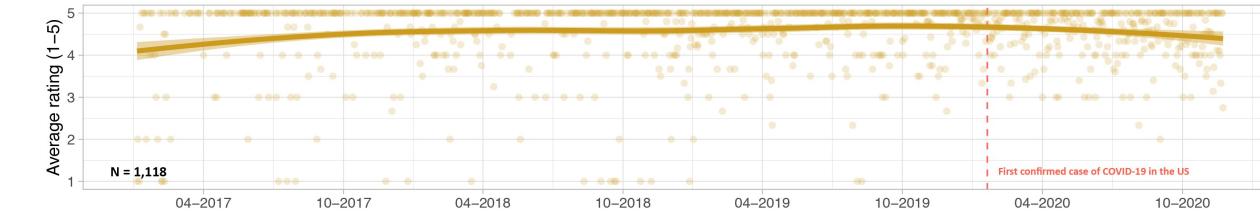


# Data in Stat 100

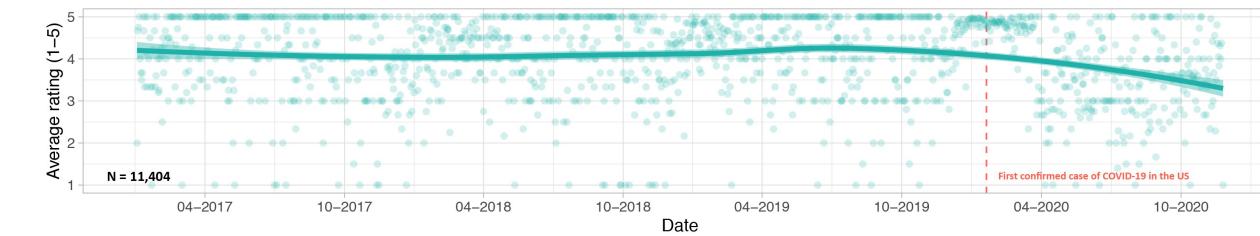
Where the top 1% and the bottom 20% go to college



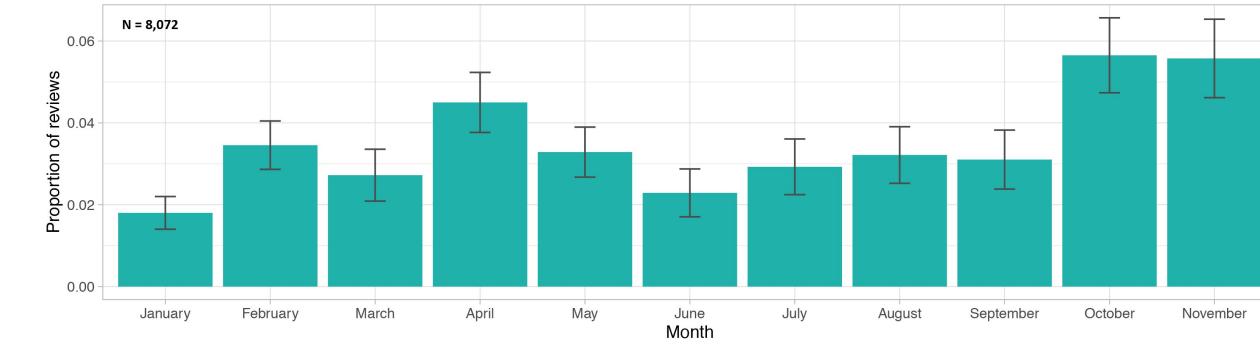
Top 3 unscented candles Amazon reviews 2017–2020



Top 3 scented candles Amazon reviews 2017–2020



Top 5 scented candles on Amazon:  
Proportion of reviews mentioning lack of scent by month 2020

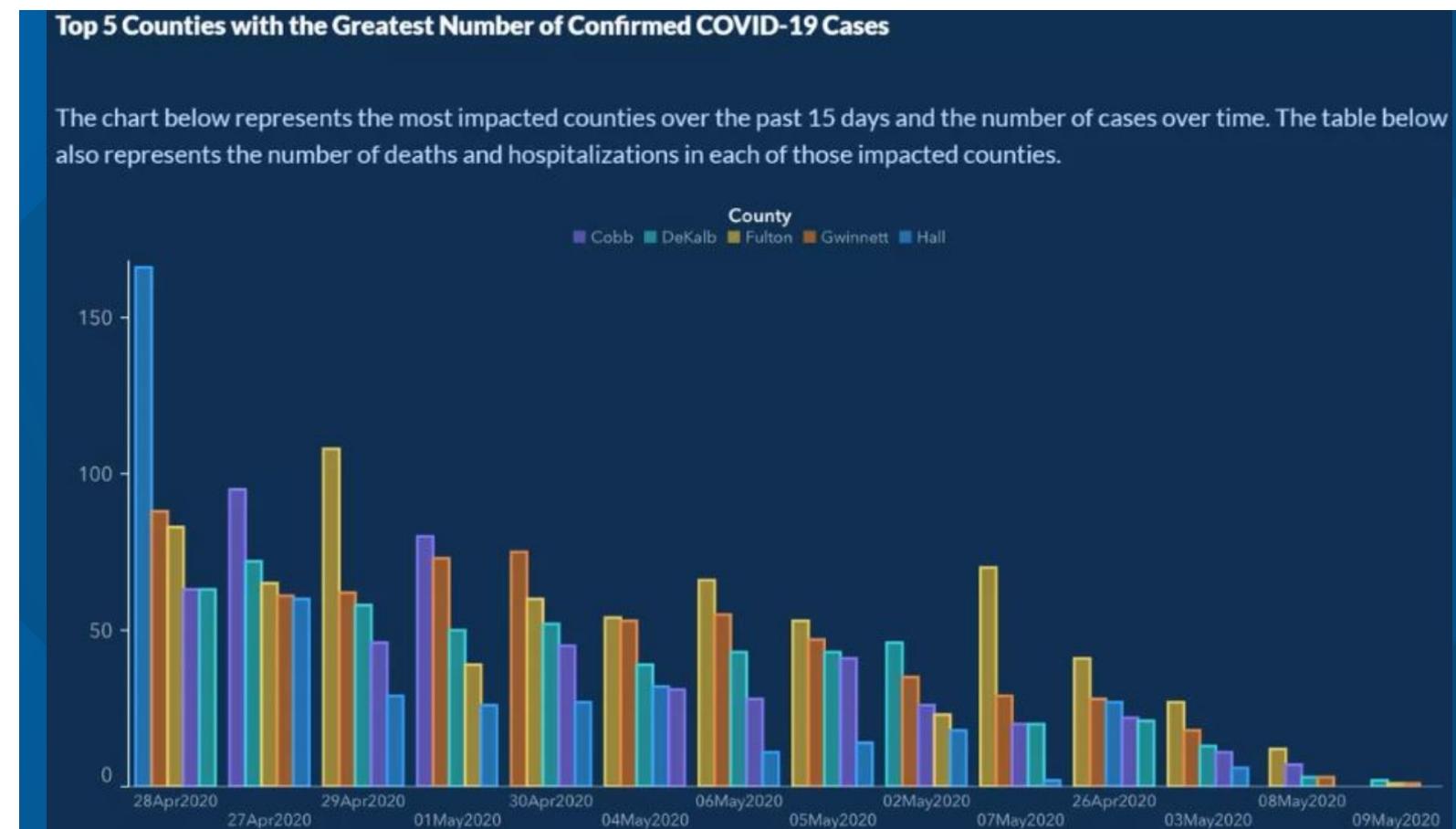


- I understand that some of these topics have likely had profound impacts on your lives.
- We will focus class time on the key course objectives but will use these current topics to empower ourselves and to see how we can productively participate with data.

# Example: Visualizing COVID Prevalence

# Example: Visualizing COVID Prevalence

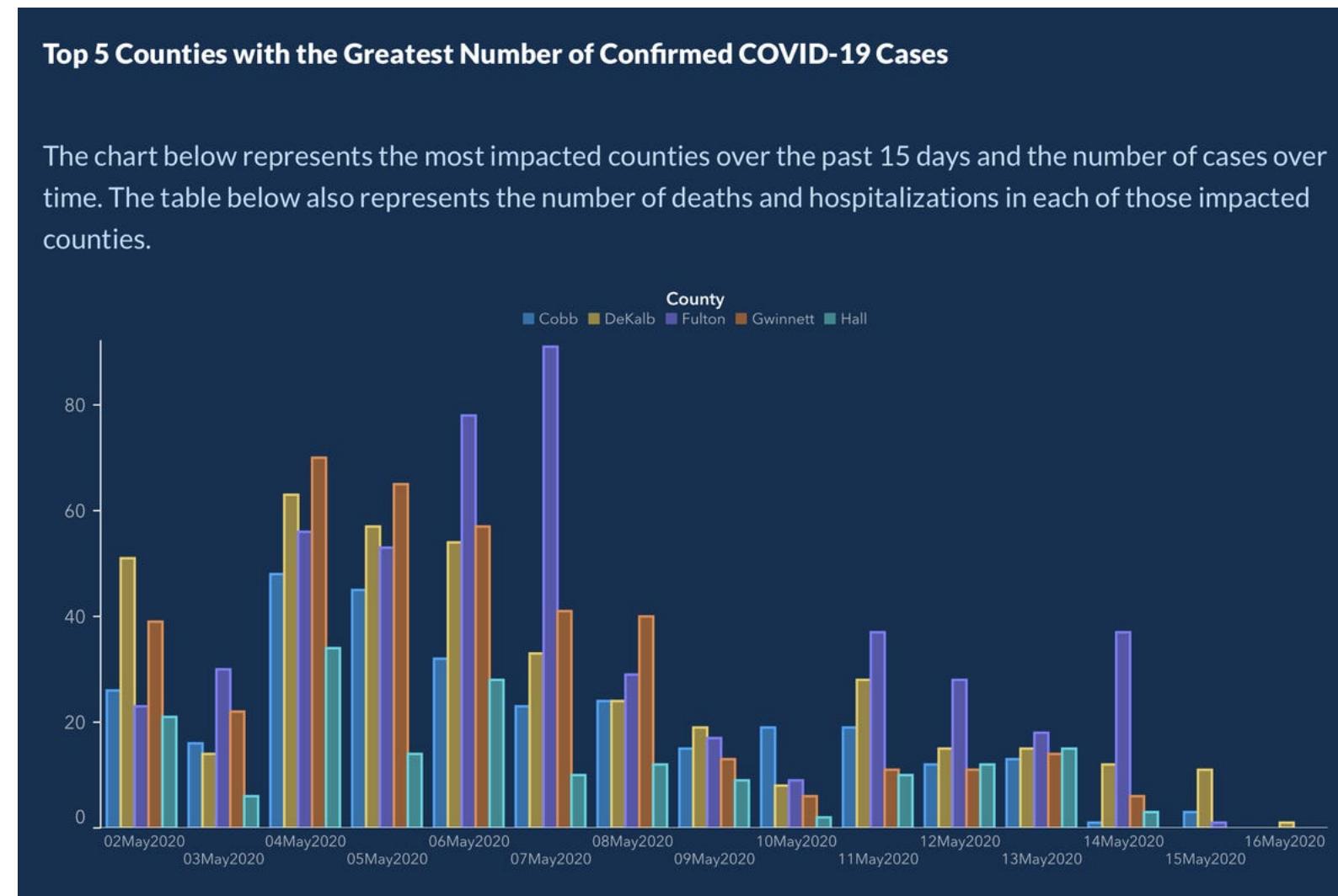
- In May of 2020, the Georgia Department of Public Health posted the following graph:



- At a quick first glance, what story does the Georgia Department of Public Health graph appear to be telling?
- What is misleading about the Georgia Department of Public Health graph? How could we fix this issue?

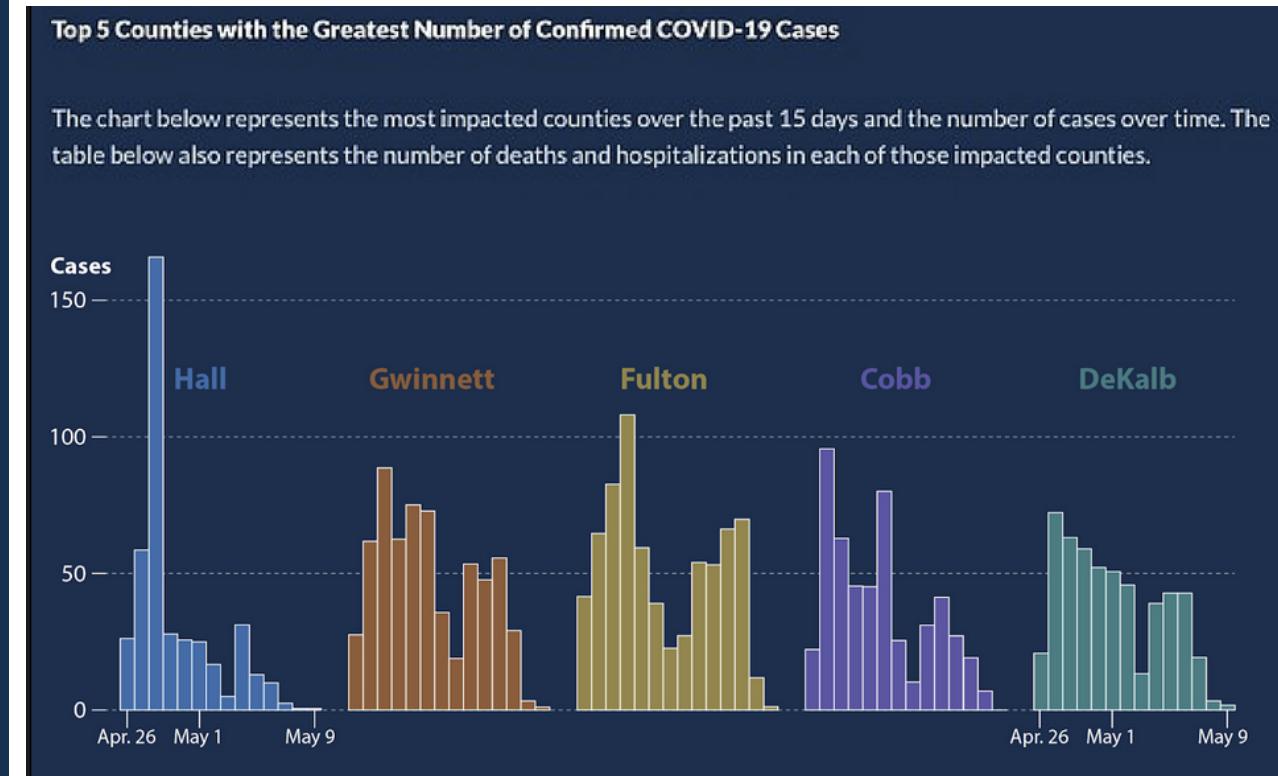
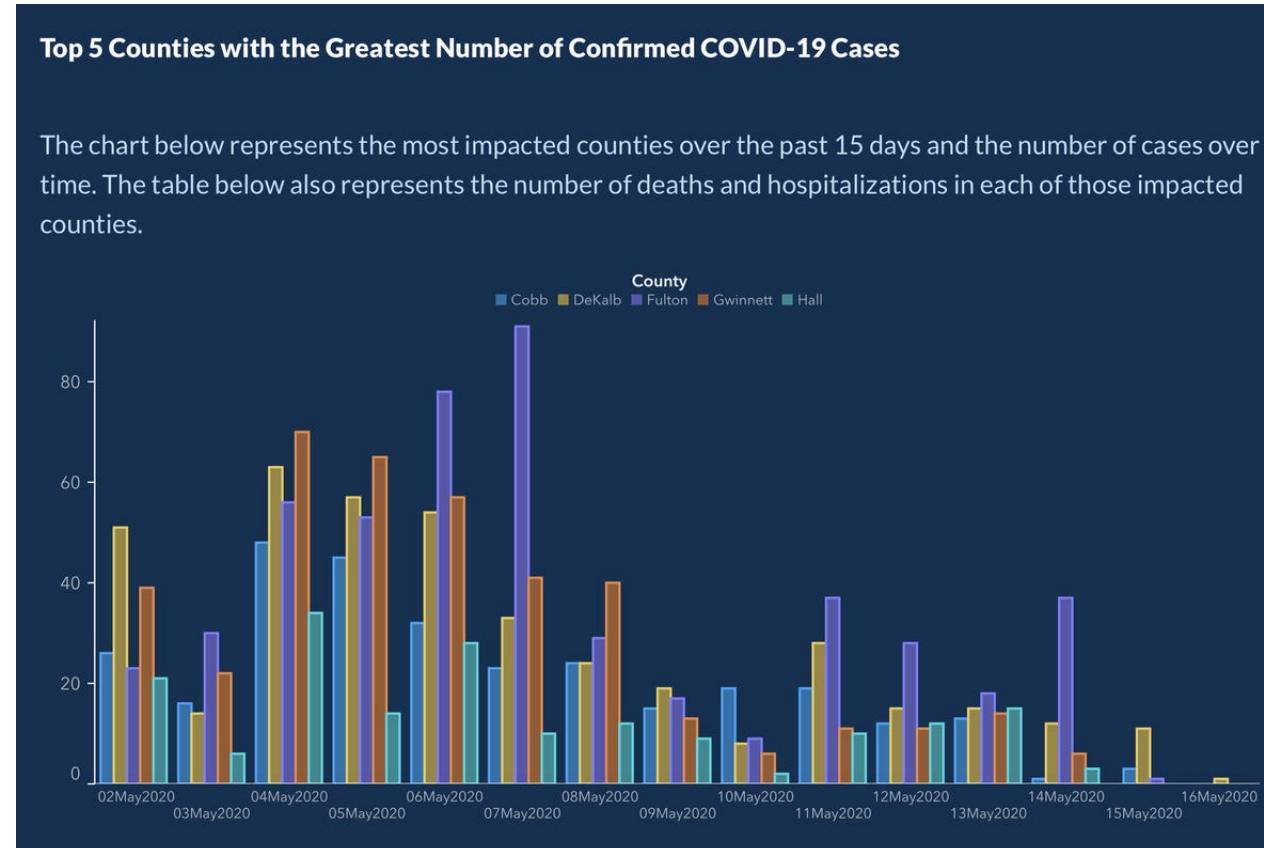
# Example: Visualizing COVID Prevalence

- After public outcry, the Georgia Department of Public Health said they made a mistake and posted the following updated graph:



- How do your conclusions about COVID-19 cases in Georgia change when now interpreting this new graph?

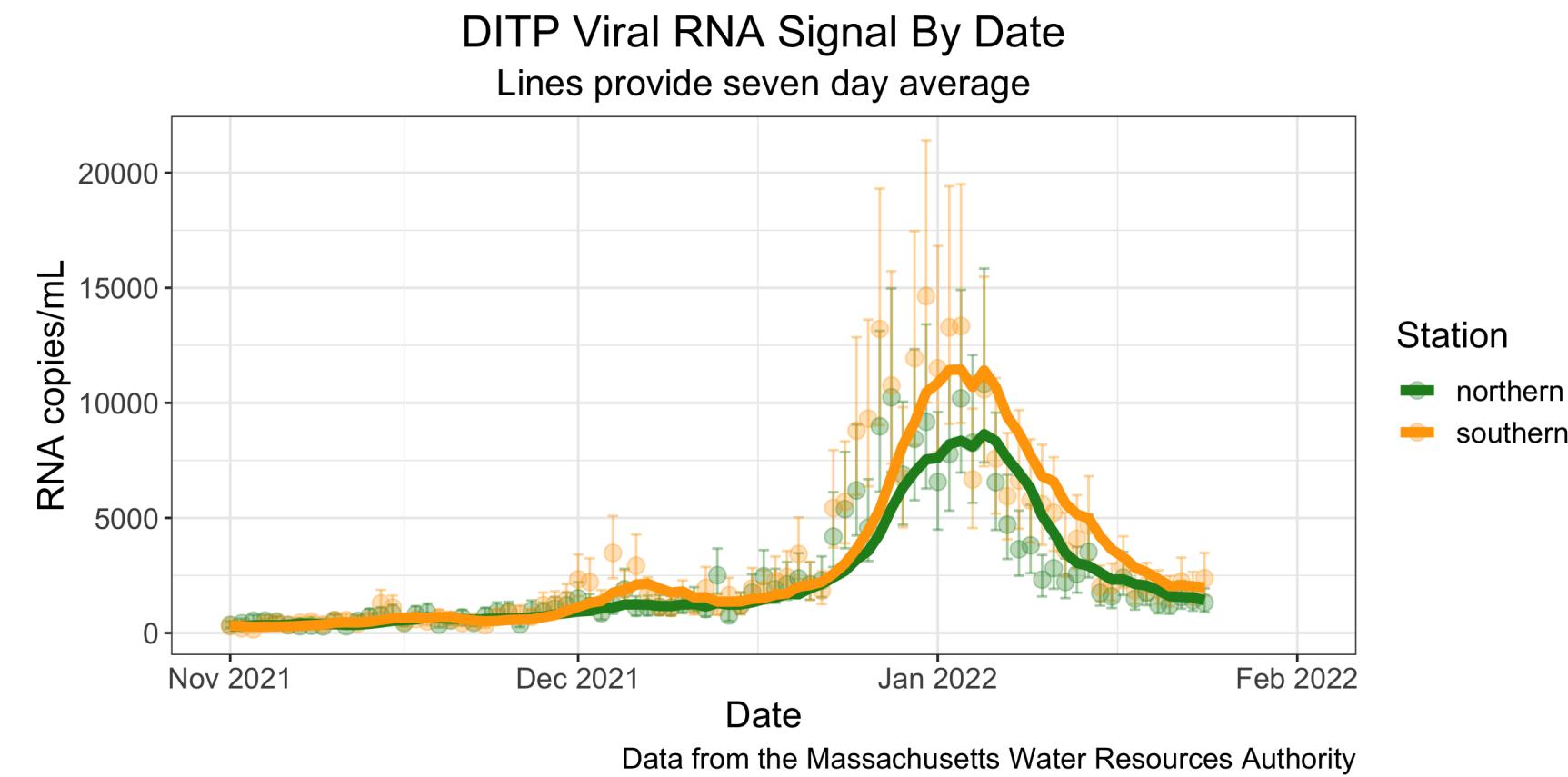
Alberto Cairo, a journalist and designer, created the second graph of the Georgia COVID-19 data:



- A key principle of data visualization is to “**help the viewer make meaningful comparisons**”.
- What comparisons are made easy by the lefthand graph? What about by the righthand graph?
- From these graphs, can we get an accurate estimate of the COVID prevalence in these Georgian counties over this two week period?

# Example: Visualizing COVID Prevalence

- The Massachusetts Water Resources Authority (MWRA) graph tracks the presence of COVID-19 in the Boston-area wastewater.



- What are the pros of using wastewater over nasal swabs to assess COVID prevalence? What are the cons?
- One more note: The graph also incorporates **uncertainty measures**, a key statistical thinking idea that we will learn more about later in the semester!



# What is “Statistical Thinking?”

# Statistical Thinking

- Understanding the importance of **context**.
  - Context explains the Monday jumps in the COVID counts.
- How we **encode** information in a graph should be driven by our research question.
  - **Design choices** impact the conclusions the viewer draws.
- How the data are **collected** impacts the conclusions we can draw.
  - Voluntary COVID test results don't likely provide good estimates of COVID prevalence.
- Often we are using a **sample** of data to say something about a larger group. In this case, we should measure how certain our estimates are!
  - We will learn to **compute** and **interpret** certainty estimates (like those in the wastewater graph) later in the course!

# Statistical Thinking

- About developing **reasoning** (not just learning definitions and formulae).
- Developing our statistical thinking skills will allow us to soundly **extract knowledge from data!**
- Statistical thinking requires **judgment** that takes time to develop.
  - Will see **examples** and **practice** applying statistical thinking throughout the course.

# What are/is Data?

| “*Raw data*’ is an oxymoron.” – Lisa Gitelman

| “*Data ... is information made tractable.*” – Catherine D’Ignazio and Lauren Klein

# Data Frames

Data in **spreadsheet**-like format where:

- Rows = Observations/cases
- Columns = Variables

ID	kind	.pred_AI	.pred_class	detector	native	name	model
1	Human	0.9999942	AI	Sapling	No	Real TOEFL	Human
2	Human	0.8281448	AI	Crossplag	No	Real TOEFL	Human
3	Human	0.0002137	Human	Crossplag	Yes	Real College Essays	Human
4	AI	0.0000000	Human	ZeroGPT	NA	Fake CS224N - GPT3	GPT3
5	AI	0.0017841	Human	OriginalityAI	NA	Fake CS224N - GPT3, PE	GPT4
6	Human	0.0001783	Human	HFOpenAI	Yes	Real CS224N	Human

- Data from **GPT Detectors Are Biased Against Non-Native English Writers.** *Weixin Liang, Mert Yuksekgonul, Yining Mao, Eric Wu, James Zou.* [CellPress Patterns](#) and available in the **R** package **detectors**.

# Data Frames

ID	kind	.pred_AI	.pred_class	detector	native	name	model
1	Human	0.9999942	AI	Sapling	No	Real TOEFL	Human
2	Human	0.8281448	AI	Crossplag	No	Real TOEFL	Human
3	Human	0.0002137	Human	Crossplag	Yes	Real College Essays	Human
4	AI	0.0000000	Human	ZeroGPT	NA	Fake CS224N - GPT3	GPT3
5	AI	0.0017841	Human	OriginalityAI	NA	Fake CS224N - GPT3, PE	GPT4
6	Human	0.0001783	Human	HFOpenAI	Yes	Real CS224N	Human

Rows = Observations/cases

What are the cases? What does each row represent?

# Data Frames

ID	kind	.pred_AI	.pred_class	detector	native	name	model
1	Human	0.9999942	AI	Sapling	No	Real TOEFL	Human
2	Human	0.8281448	AI	Crossplag	No	Real TOEFL	Human
3	Human	0.0002137	Human	Crossplag	Yes	Real College Essays	Human
4	AI	0.0000000	Human	ZeroGPT	NA	Fake CS224N - GPT3	GPT3
5	AI	0.0017841	Human	OriginalityAI	NA	Fake CS224N - GPT3, PE	GPT4
6	Human	0.0001783	Human	HFOpenAI	Yes	Real CS224N	Human

Columns = Variables

**Variables:** Describe characteristics of the observations

- **Quantitative:** Numerical in nature
- **Categorical:** Values are categories
- **Identification:** Uniquely identify each case

ID	kind	.pred_AI	.pred_class	detector	native	name	model
1	Human	0.9999942	AI	Sapling	No	Real TOEFL	Human
2	Human	0.8281448	AI	Crossplag	No	Real TOEFL	Human
3	Human	0.0002137	Human	Crossplag	Yes	Real College Essays	Human
4	AI	0.0000000	Human	ZeroGPT	NA	Fake CS224N - GPT3	GPT3
5	AI	0.0017841	Human	OriginalityAI	NA	Fake CS224N - GPT3, PE	GPT4
6	Human	0.0001783	Human	HFOpenAI	Yes	Real CS224N	Human

Every time you get a new dataset, spend time exploring the variables.

Example questions:

- Is the variable capturing what I want?
- For categorical variables, what are the categories? Do those categories adequately represent the data represented by that variable?
- For quantitative variables, what values are possible? Were the data rounded or binned? Are those values actually encoding categories? What are the units of measurement?

**Goal: Start collecting data from  
your life so that you can visualize  
it on P-Set 1.**

# Hand-Drawn Data Viz

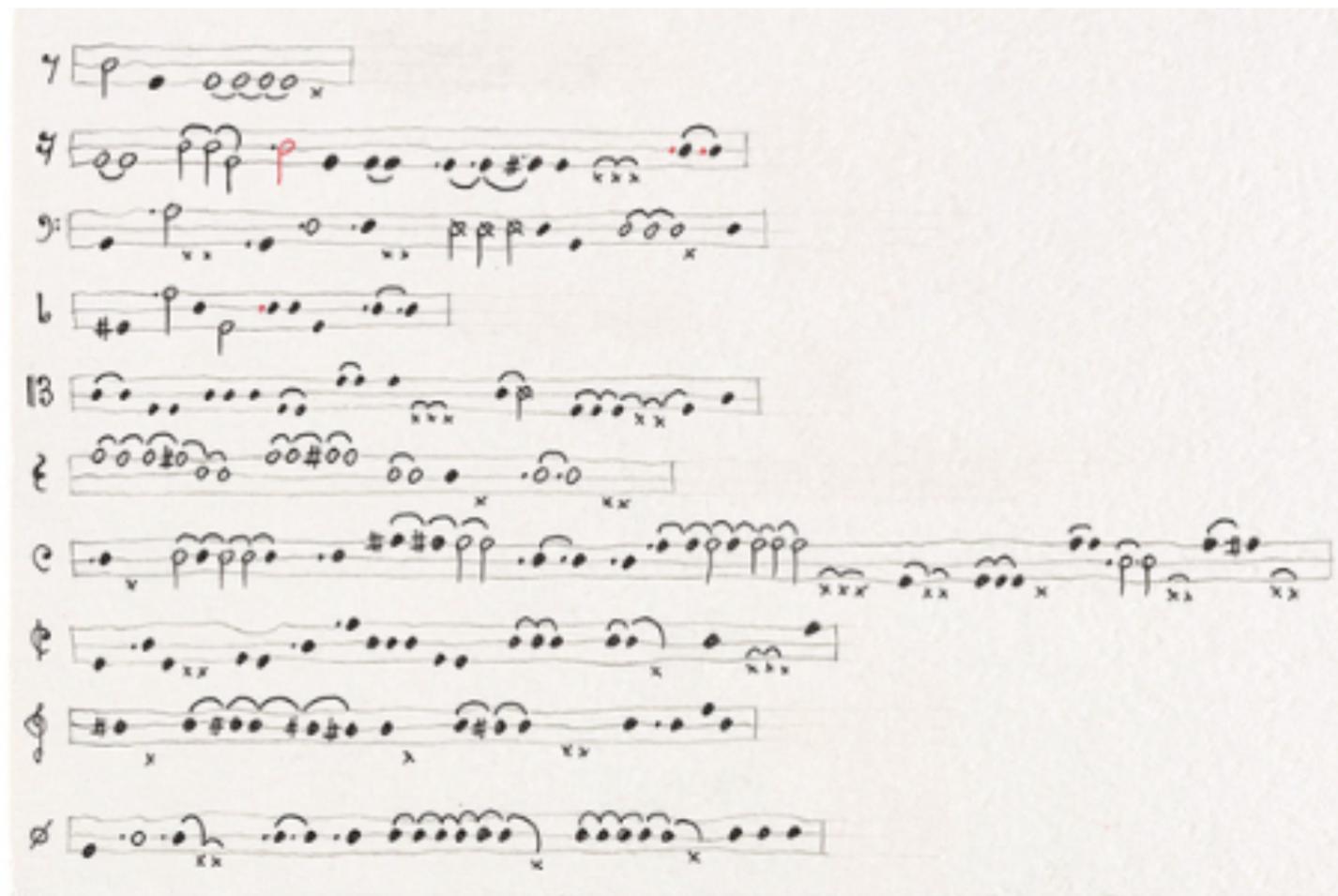
- Once we have collected data, a common next step is to visualize it.
- Two key aspects of data visualization:
  - Determining how you want to display the data.
  - Figuring out how to tell the computer to do that mapping.
- Hand-drawn data visualizations allow us to focus on the first part with full control over the **creative** process!

# Hand-Drawn Data Viz Examples

## Dear Data

*“Each week, and for a year, we collected and measured a particular type of data about our lives, used this data to make a drawing on a postcard-sized sheet of paper, and then dropped the postcard in an English “postbox” (Stefanie) or an American “mailbox” (Giorgia)!“*

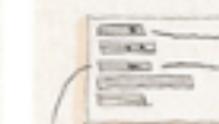
# Dear Data Examples



## 66 DEAR DATA

### WEEK 07: MUSICAL COMPLAINTS

#### HOW TO READ IT:



Each "note" is a single complaint I said.  
(i.e. every single time I expressed dissatisfaction or annoyance about a situation or particular thing)

Each "Score" represents a typology of things I complained about, featuring complaints in chronological order.

#### SCORES:

- 4 - ME AS A PERSON (e.g. "I am so... ugly / obnoxious...")
- 5 - ME AT WORK (e.g. "I should've done...")
- 6 - WORK (e.g. "this project isn't going well!")
- 7 - TECHNOLOGY (e.g. "the computer is not working!")
- 13 - SERVICE/FOOD (e.g. "the waiter is so slow!")
- 8 - SOMEBODY (e.g. "He's really a jerk...")
- 9 - COLD (e.g. "I am freezing! The A.C. is crazy!")
- 4 - HOW I FEEL (e.g. "so tired!!", "so bored!!")
- 9 - BOYFRIEND (e.g. "you're cheating! you haven't...")
- 15 - OTHER (e.g. "I spent 1 hour waiting for...")

#### POSITIONS OF NOTES:

- |   |       |   |
|---|-------|---|
| 1 | •     | → ACTUAL need to complain   |
| 2 | •     | → average " " "   |
| 3 | •     | → MOREAL " " "  |
| 4 | -->-- | → MISSED COMPLAINTS:<br>Thought of complaining,<br>but didn't do! |

#### ATTRIBUTES

- to boyfriend
- to friendly family
- to stranger
- qq → in english  
(all my others were in ITA)
- - - - -
- via txt/email  
(digital life)
- #○ adding EMPHASIS
- oo o close in time  
(same situation)
- p → to Stefanie
- about s.thing related  
to DEAR DATA

FROM:  
GIORGIA LUPI  
NYC  
NEW YORK  
NY - USA



#### SEND TO:

STEFANIE POSAVEC

~~NEW YORK~~ ENGLAND

LONDON ENGLAND

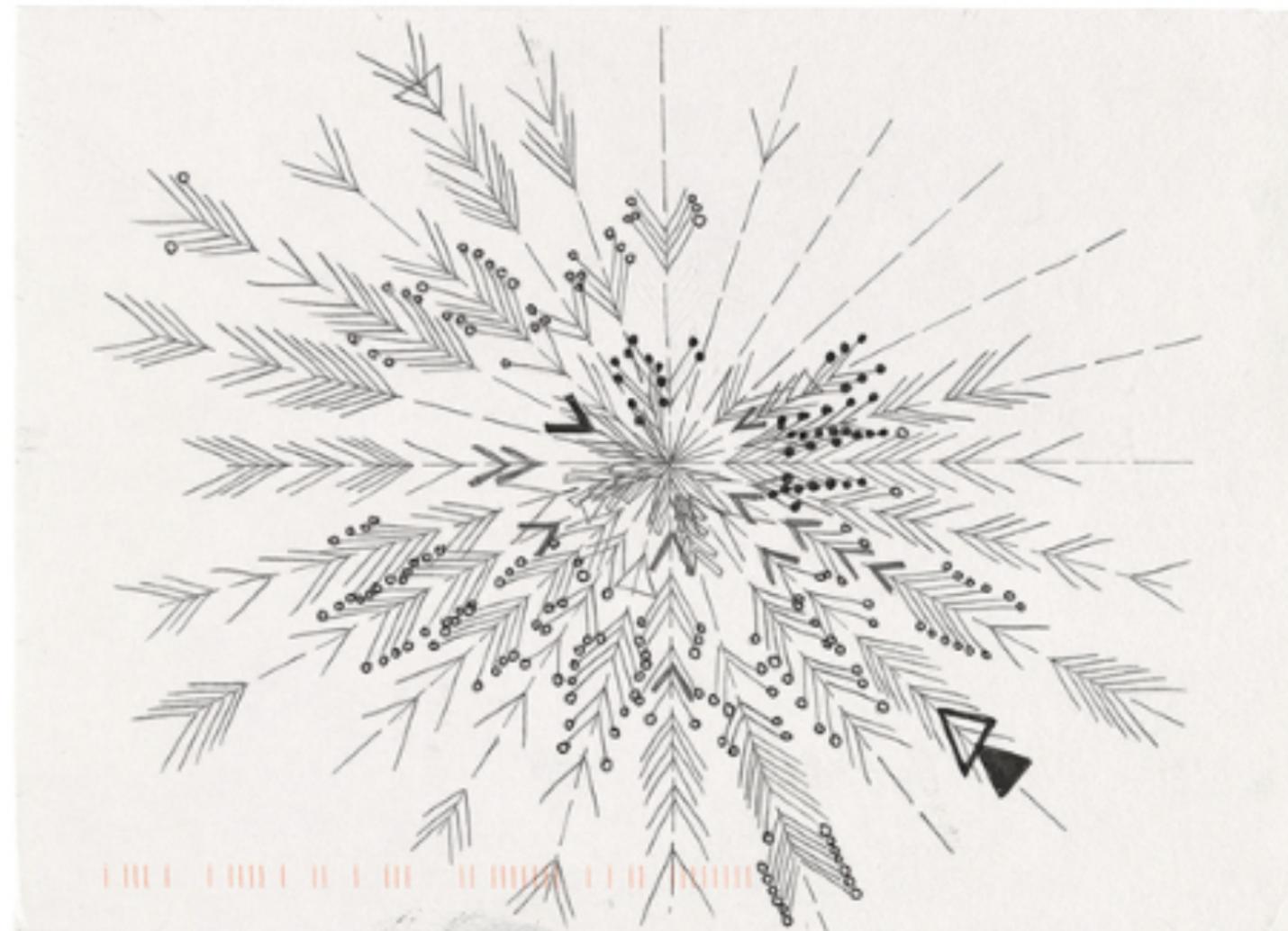
- UK -

ENGLAND

DELIVERED BY  
HAND (SPECIAL NYC  
DELIVERY!)

# Dear Data Examples

a week of clocks

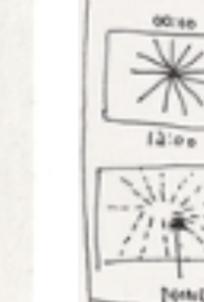


Stefanie

DEAR DATA : WEEK 01:

## A WEEK OF CLOCKS

### LEGEND



→ EACH LINE = ONE HOUR OF THE DAY, MOVING CLOCKWISE  
SEGMENT  
Each line now HOUR LINE =  
ONE DAY. ~~WEEK~~ WEEK BEGINS IN CENTRE + MOVES OUTWARD  
MONDAY SUNDAY

AN INSTANCE OF CLOCK-WATCHING IS INDICATED BY A SYMBOL:

PHONE      LAPTOP      TABLET      HUSBAND'S PHONE      WATCH

SYMBOL	TOTAL Instances	CAR	SHIRT	TELEVISION
PHONE	151	MICROWAVE	20	1
LAPTOP	84	FRIEND'S OVEN	1	1
TABLET	10	CHURCH CLOCK	1	1
HUSBAND'S PHONE	3			
WATCH	11			

1124995282

This week Giorgia and Stefanie tried gathering data in small notebooks (tedious), but soon switched to making notes on their phones (much easier). Stefanie's favourite clock to capture: a bell tolling the time in a town in Devon.

From:  
S. POSAVEC  
[REDACTED]  
LONDON  
UK  
[REDACTED]  
SCOTLAND 2014

Royal Mail  
Jubilee  
1st Class  
15-09-2014  
54007997



GIORGIA LUPI

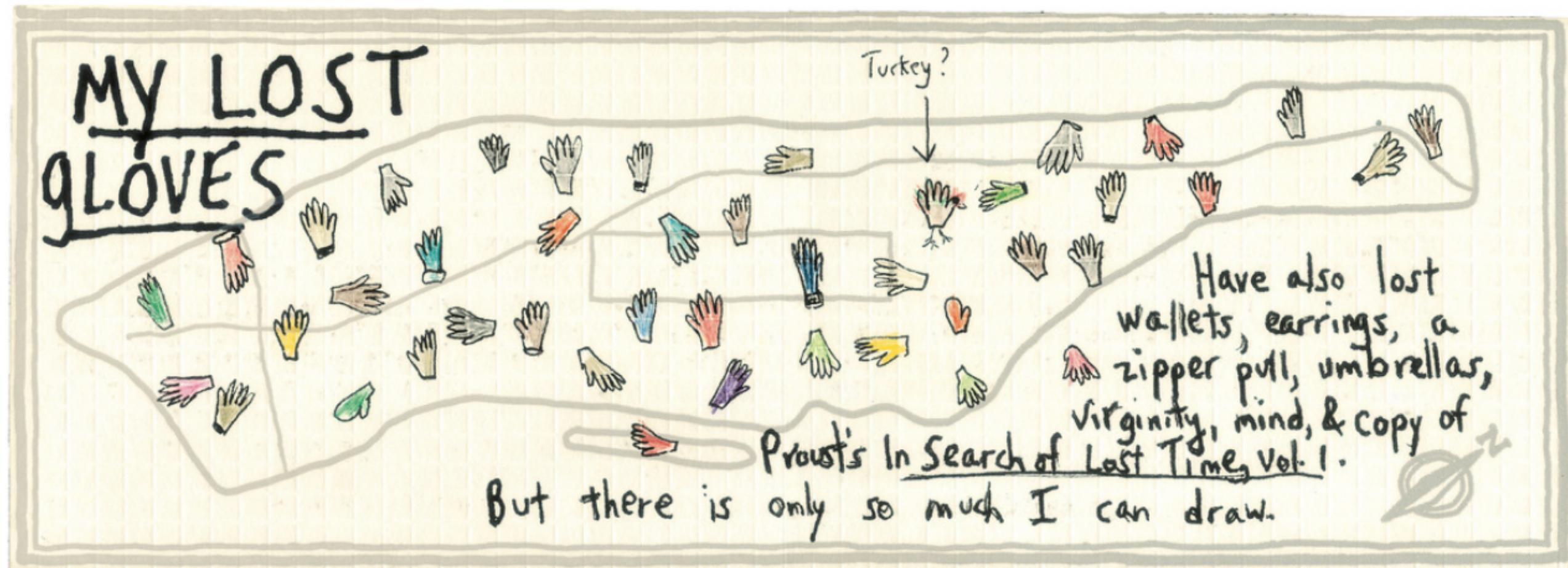
[REDACTED], [REDACTED]  
BROOKLYN, NY  
USA

BY AIR MAIL  
par avion

Royal Mail

## More Dear Data Examples

- Becky Cooper handed out hand-drawn maps of Manhattan to strangers and asked them to “map their Manhattan.”



Map drawn by New Yorker staff writer Patricia Marx

- What would the data frame for this visualization look like?

# More Dear Data Examples



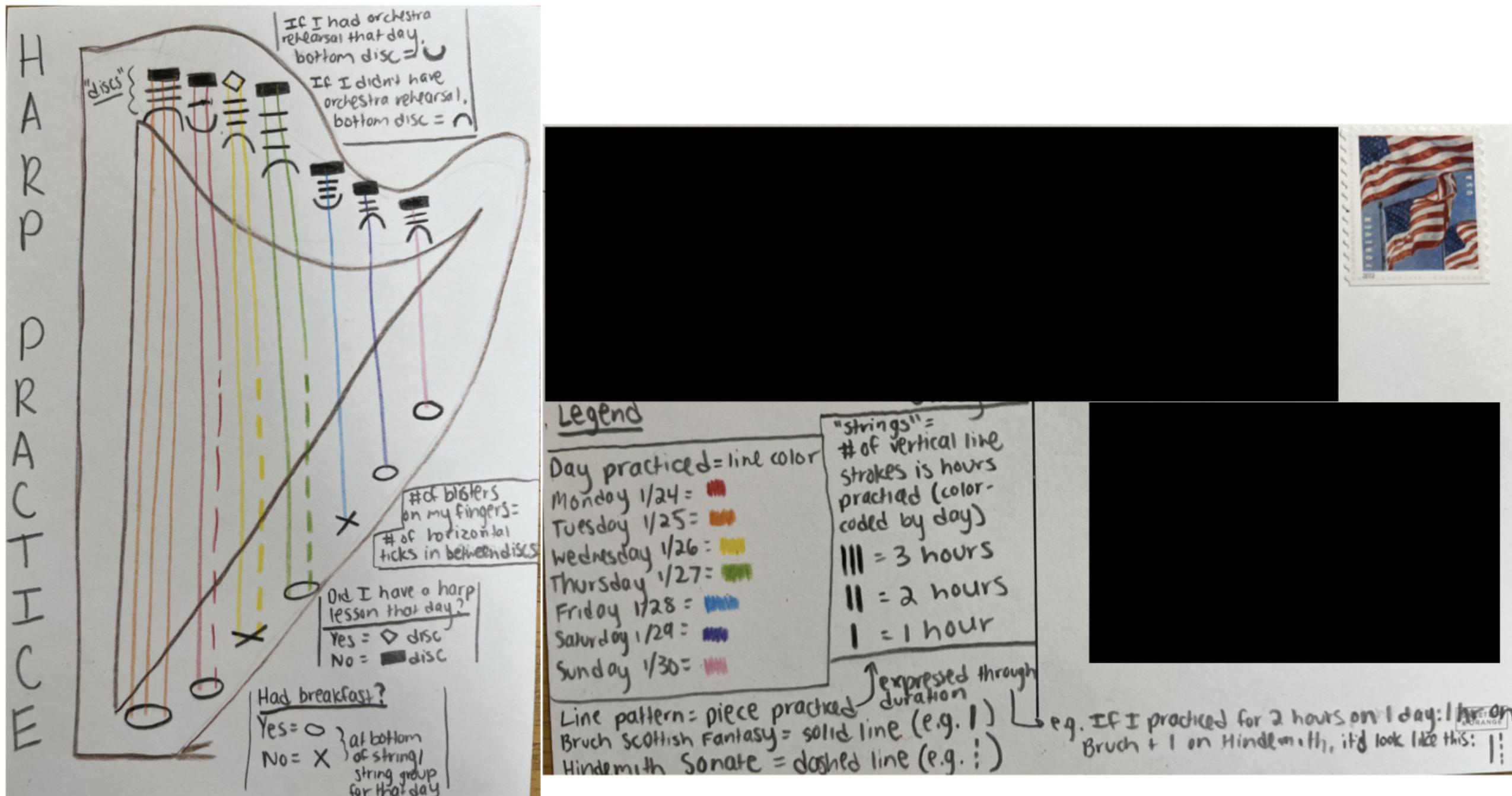
## Happiness and Calories at Annenberg

The front includes a data visualization on the number of calories in lunches taken at Annenberg in the past week and the self-reported happiness levels before and after each lunch.



- What would the data frame for this visualization look like?

# More Dear Data Examples



- What would the data frame for this visualization look like?

# Goal: By next Wed, collect data from your life so that you can visualize it on P-Set 1.

## Recommendations

- Store the data in your favorite spreadsheet program (Google Sheets, Numbers, Excel).
- Determine what your cases/observations will be.
- Collect data on **more** variables than you will likely visualize. It is hard to know beforehand what the interesting relationships will be.

## Next Week

- Will get a blank postcard and further guidance on the visualization with P-Set 1.

# Demo of accessing the RStudio Server on Posit Cloud

Try to access the RStudio Server between now and next lecture.  
Come back to the recording if need help with the steps.



# Reminders

- If able, please bring a laptop or tablet to Mondays's lecture.
- No section, no wrap-ups, and no lecture quiz this week.
- Make sure to go through the syllabus, which can be found on Canvas.
  - Will discuss assessments and course policies on Monday.
- Only I will be running office hours this week at the following time:
  - Today 1:30 - 3:00 pm in Science Center 316 (This week only)
- The regular office hour schedule will be posted later this week and will start next week.
- Be on the look-out for the section preference form.

