Welcome to Data Literacies!

Co-leads: Di Yoong & Leanne Fan

WORKSHOP PLAN.

- Stages of Data
- Team Activity
- Cleaning and Analyzing
- Visualizing
- Team Activity
- Wrap up
    - Preparing for tomorrow's workshop
- Curriculum Recap

THINKING ABOUT DATA.

What makes up research data? How do you define data?

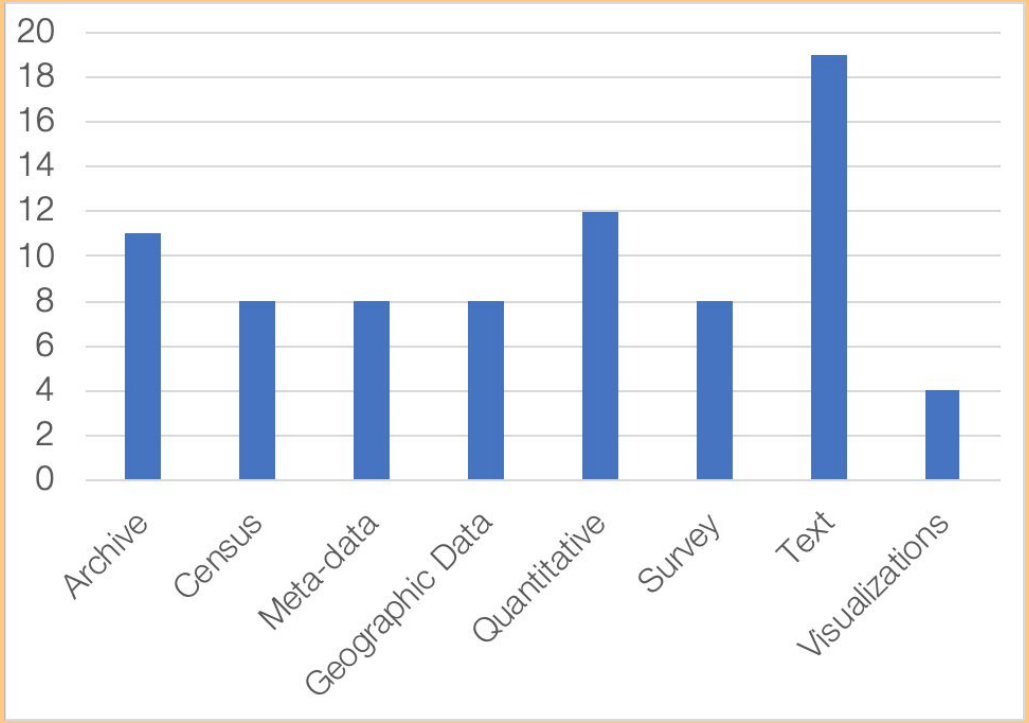
- What did Steve mention was the definition of data?

Non-digital text (field notes)
Digital text (Project Gutenberg)
Computer code

GIS and spatial data
Computer aided design (CAD)
Metadata & paradata

What kind of data is our group working with?

- Volunteer to share what you're working on!
- Structured versus unstructured data
- Email - structured or unstructured?

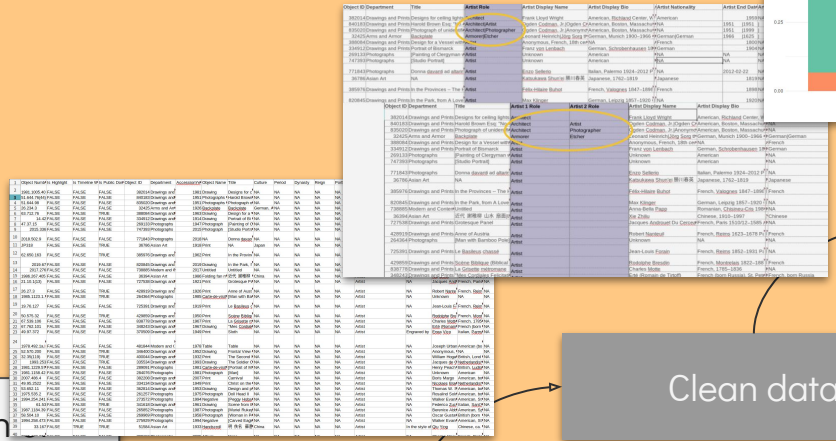


STAGES OF DATA.

What proportion of the artwork collected in the Met are by non cisgender men?

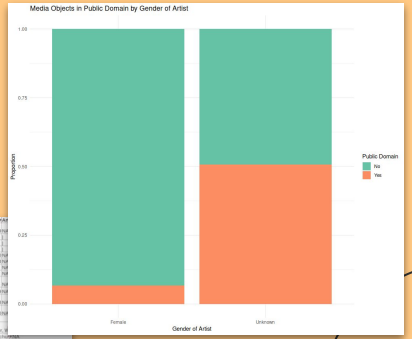
Formulate research question(s)

Collect data



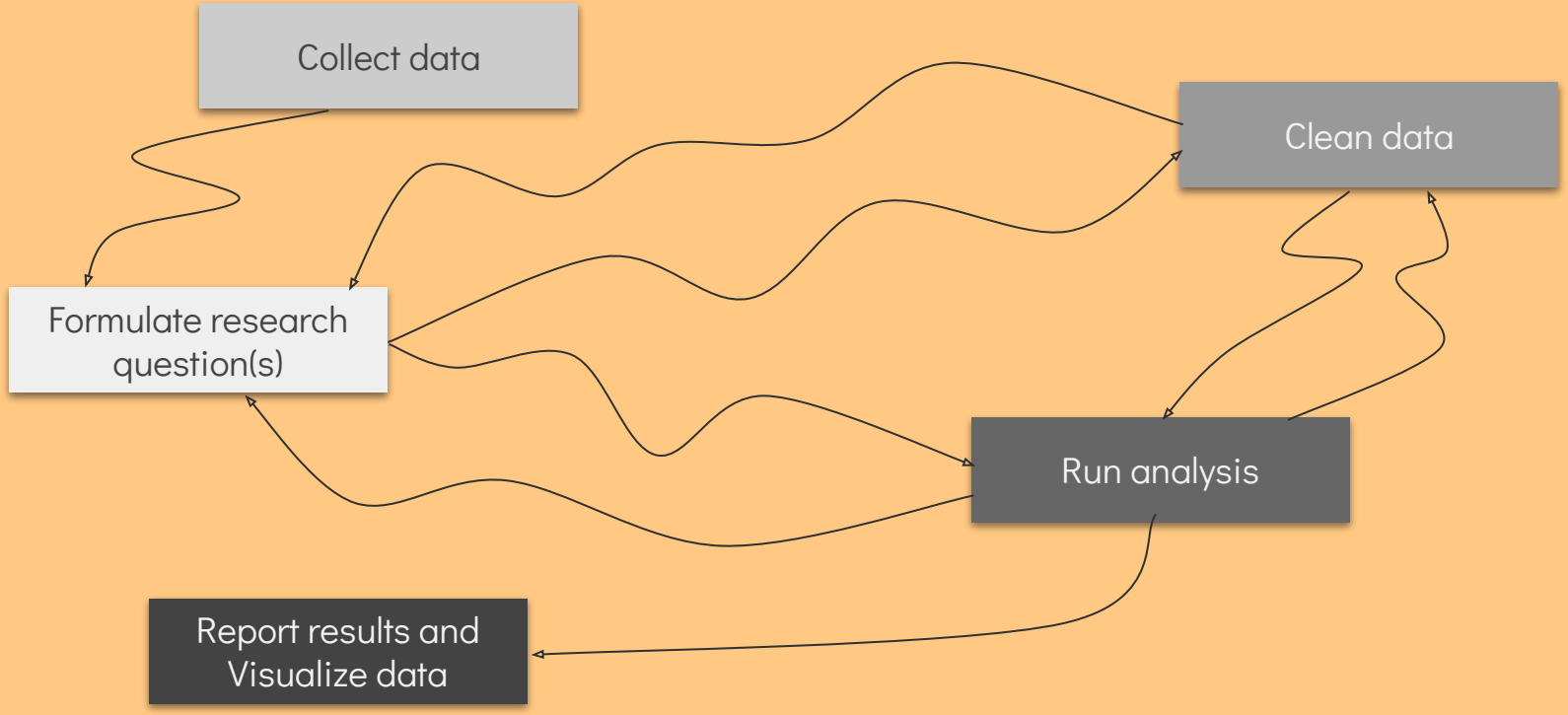
Clean data

Run analysis



Report results and Visualize data

STAGES OF DATA.

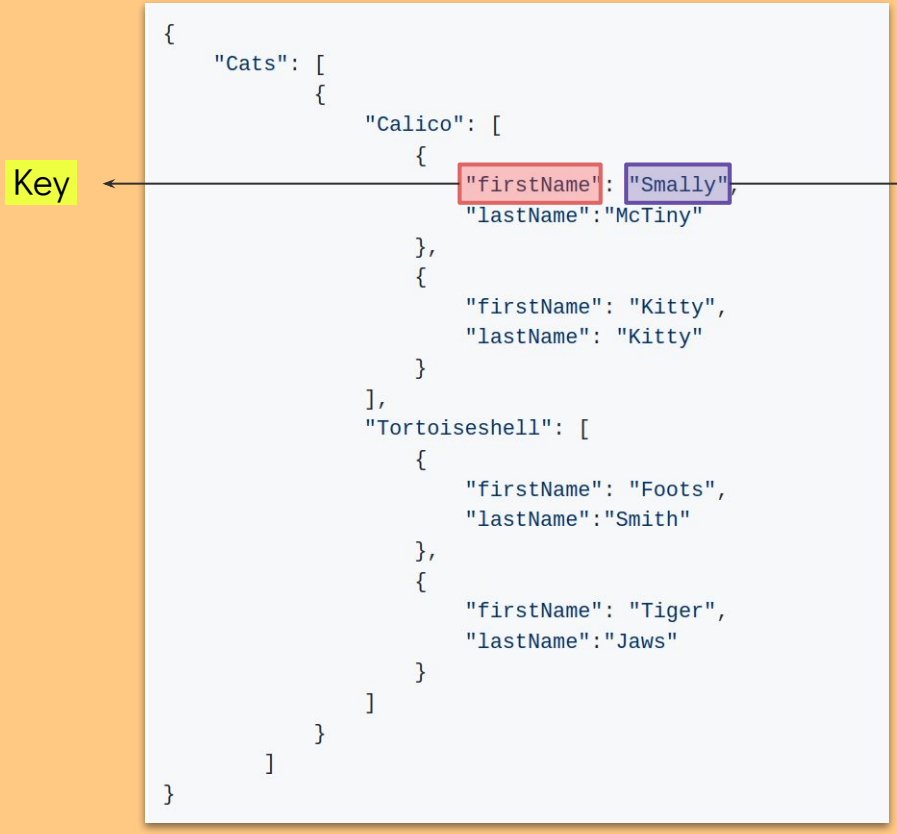


TIDY DATA STRUCTURE.

- Each variable is in a column
- Each observation is a row
- Each value is a cell

Object ID	Origin Location	Type
0001	Manhattan, NY	Sculptures
0002	San Diego, CA	Photographs
0003	Columbus, MS	Photographs
0004	Miami, FL	Oil Paintings

TIDY DATA STRUCTURE.



File type: JSON (.json)

TIDY DATA STRUCTURE.

How will you convert this data set to a tidy data structure?

Object ID	Origin Location	Type
0001	Manhattan, NY	Sculptures
0002	San Diego, CA	Photographs
0003	Columbus, MS	Photographs
0004	Miami, FL	Oil Paintings


```
{
    "Cats": [
            {
                "Calico": [
                    {
                        "firstName": "Smally",
                        "lastName":"McTiny"
                    },
                    {
                        "firstName": "Kitty",
                        "lastName": "Kitty"
                    }
                ],
                "Tortoiseshell": [
                    {
                        "firstName": "Foots",
                        "lastName":"Smith"
                    },
                    {
                        "firstName": "Tiger",
                        "lastName":"Jaws"
                    }
                ]
            }
        ]
}
```

TIDY DATA STRUCTURE.

Type of Pet	Breed	First Name	Last Name
Cat	Calico	Smally	McTiny
Cat	Calico	Kitty	Kitty
Cat	Tortoiseshell	Foots	Smith
Cat	Tortoiseshell	Tiger	Jaws

TIDY STRUCTURE MAY SEEM TRIVIAL…

- What new variables were created?
- What does each of the three tables represent?

A

Untidy Data

species	habitat	weight	length	latitude/longitude	date
Alligator mississippiensis	swamp	431 lb	4 ft 2	29.531,-82.184	Sept 15, 2015
Puma concolor	forest	125 lb	2.2m	29.125,-81.682	08/10/2015
Ursus americanus	forest	88 kg	133 cm	N29°7'30"/W81°40'55.2"	07-13-2015

B

Tidy Data

meta-data

data

species_code	date	station_code	weight_kg	length_cm
TSN 551771	2015-09-15	1	196	127
TSN 55247	2015-08-10	2	57	220
TSN 180544	2015-07-13	2	88	133

station_code	habitat	latitude	longitude
1	swamp	29.531	-82.184
2	forest	29.125	-81.682

species_code	class	genus	species
TSN 551771	Reptilia	Alligator	mississippiensis
TSN 55247	Mammalia	Puma	concolor
TSN 180544	Mammalia	Ursus	americanus

Source: Ten Simple Rules for Digital Data Storage

BREAKOUT GROUPS!

DISCUSSION ACTIVITY: SCRAPING TWITTER

BREAKOUT GROUPS!

You will be going into your breakout groups for 15 mins to discuss considerations of a Twitter project

- Everyone will be discussing the same project and consideration
- Please *do* take notes on our collaborative google document.
- When we return, we will

TWITTER ACTIVITY.

You will be going into your breakout groups for 15 mins to discuss the questions. When we return, each group will have about 3 mins to share their main ideas.

- Please *do* take notes on our collaborative google document.

TWITTER ACTIVITY.

You are interested in looking at reactions to the presidential debates across time. You've decided that you would be using Twitter data for your project.

After collecting your data, you learned that your data has information from users who were later banned and also included some tweets that were removed/deleted from the site.

QUESTIONS TO GET YOU STARTED:

- When you first collect the data, would you anonymize users?
  - How would anonymity impact your decisions upon learning about the changes to your data?
- How would where you are at in your project (e.g. cleaning v. reporting results) affect your decisions?
  - What are considerations you might have when choosing to remove or not the impacted data?
- How would the number of Tweets and/or Twitter users impact your decision?

FURTHER EXPLORATIONS.

- If you were collecting and analyzing data on folx in power, such as the Tweets of Congress' project, would that change your answers to the previous questions?
- Current ethical guidelines from SAFE Lab at Columbia University have decided to alter the text of social media post to render it unsearchable. Why and when would you consider (or not) altering collected tweets for publication?

DIFFERENTIAL PRIVACY.

- Piecing back identifying information about a person no longer too challenging
  - Anonymity may no longer be sufficient to protect participants
- Differential privacy as an emerging statistical strategy
  - Adding random noise to your data
    - Balancing accuracy and privacy
  - Census 2020's approach

CLEANING AND ANALYZING DATA

CLEANING.

Generally, high quality data is measured in its **validity**, **accuracy**, **completeness**, **consistency**, and **uniformity**.

CLEANING: VALIDITY.

For measurements to be valid they must conform to certain constraints

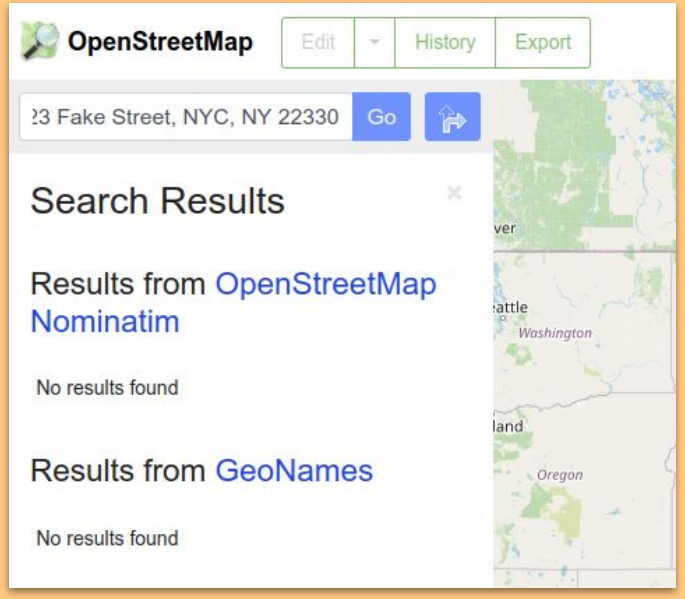
Object ID	Consent given?	Origin Location	Type
0001	1	Manhattan, NY	Sculptures
0002	1	San Diego, CA	Photographs
0003	1	Columbus, MS	Photographs
0004	1	Miami, FL	Oil Paintings

CLEANING: ACCURACY.

For measurements to be accurate they must represent the correct values

123 Fake Street, NYC, NY 22330

- Observations may be valid but inaccurate
- Cross-referencing external trusted sources can improve accuracy



CLEANING: COMPLETENESS.

For a measurement to be complete they must represent everything that might be known about the interested phenomenon

CLEANING: CONSISTENCY.

For measurements to be consistent different observations must not contradict each other

- A person cannot be represented as simultaneously dead and being born at the same time in your data set

CLEANING: UNIFORMITY.

For measurements to be uniform the same unit of measure must be used in all relevant measurements

- Person A's height in inches and Person B's in centimeters

CLEANING.

Generally, high quality data is measured in its **validity**, **accuracy**, **completeness**, **consistency**, and **uniformity**.

- Removal and transformation of “raw“ data

ANALYZING.

Analysis can take many form but they generally fall under:

- Descriptive
  - Geared towards a ***description*** and ***summary*** of a data set
- Inferential
  - Geared towards making ***predictions*** and ***hypothesis testing***
- Qualitative
  - Geared towards ***understanding*** a phenomenon

ANALYZING.

Research question: Why are cis-gender men more likely to be represented in museum collection?

- Data:
  - Interviews with curators and collectors
- Category of analysis:
  - Qualitative

TEXT ANALYSIS.

A quick note on text analysis:

- Libraries and dictionaries are **constructed** and **created**
  - Clean and convenient but can be limited

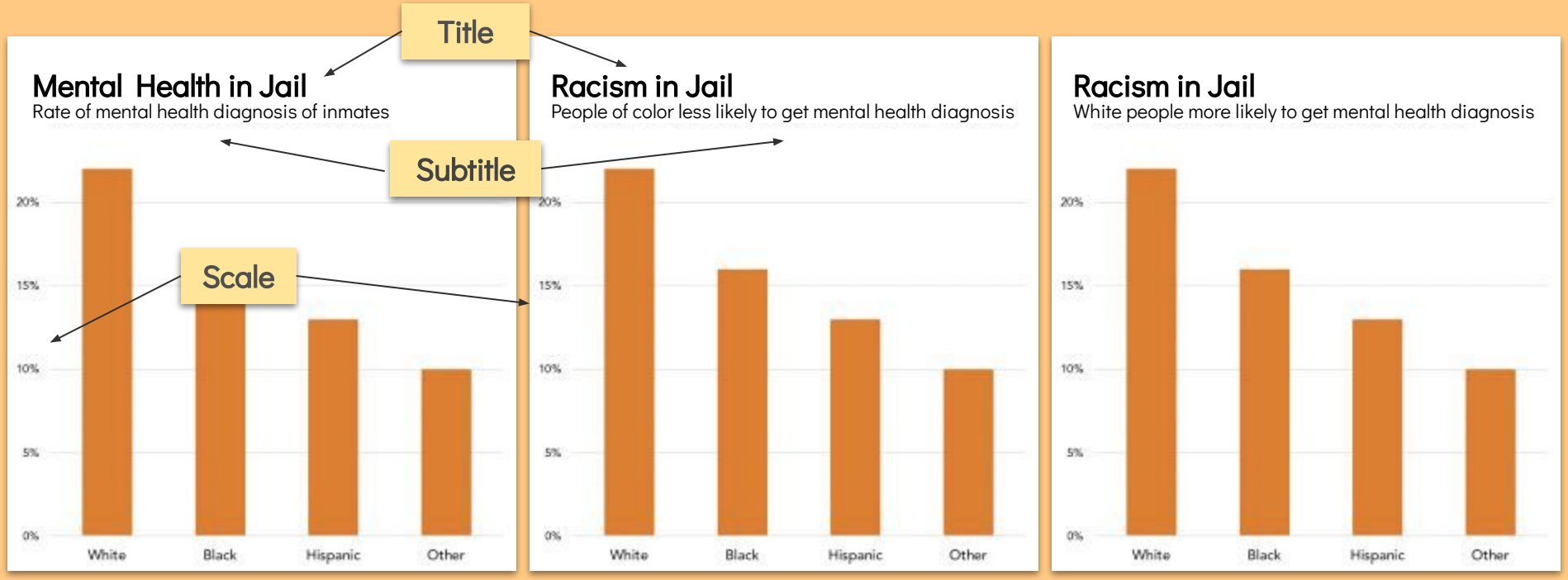
VISUALIZING

VISUALIZING.

Visualizing your data helps you tell a story and construct a narrative that guides your audience in understanding your interpretation of a collected, cleaned, and analyzed dataset.

- From trying to understand how a chart work to *understanding what the data is saying*
- Can you suggest some common assumptions we make in visualizing data?

TELLING A STORY.



Graphics from The Numbers Don't Speak for Themselves, *Data Feminism*

BEFORE WE GO INTO DISCUSSION~

BREAKOUT GROUPS!

BREAKOUT GROUPS!

You will be going into your breakout groups for 15 mins to discuss the question. When we return , each group will have about 3 mins to share their main ideas.

- Groups 1-4: cleaning and analyzing; Groups 5-8: visualizing
- Please *do* take notes on our collaborative google document.
- When we return, the person in your group whose birthday is closest to today's date will prepare to share their group's main ideas. 😊 in your pair groups

LET'S DISCUSS: PAIR & SHARE

Let's team up to pair and share

- Room 1: Team 1 and Team 8
- Room 2: Team 4 and Team 7
- Room 3: Team 2 and Team 6
- Room 4: Team 3 and Team 5

LET'S DISCUSS!

Team 1 & 2:

- Looking at the Met Museum data set, if you are interested in answering the question, what is the gender breakdown for art work in the Met collection, for the variable "Artist Gender," suggest at least **2 decisions** you might make in cleaning the responses for this variable.
  - E.g. How would you address NAs and empty fields?

LET'S DISCUSS!

Team 3 & 4:

- It is standard practice to share quotes from long transcripts or provide example Tweets when describing qualitative results. As results are shared and circulated, these quotes can be taken out of context.
  - Can you suggest at least **2 possible** safeguards you might place on unintended secondary (mis)uses of our data?

LET'S DISCUSS!

Team 5 to 6

In the following data visualizations, compare between the left and right visualization:

- What narrative are we telling with the visualization on the left v. the right,
- What information is missing,
- What might be misleading, and
- How would you change them?

Original visualizations from Nathan Yau, *How to Spot Visualization Lies*.

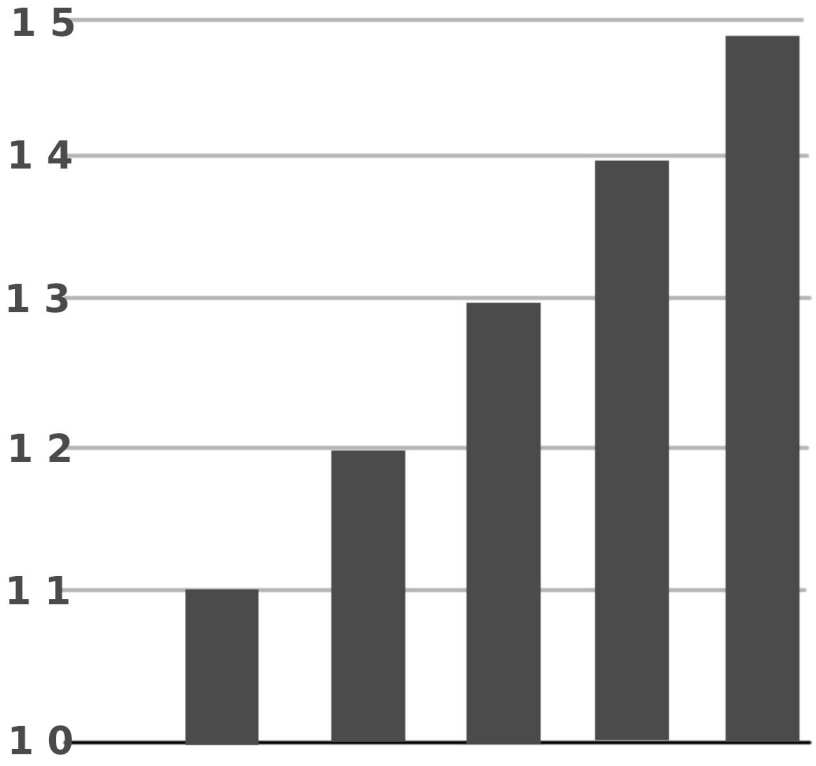
LET'S DISCUSS!

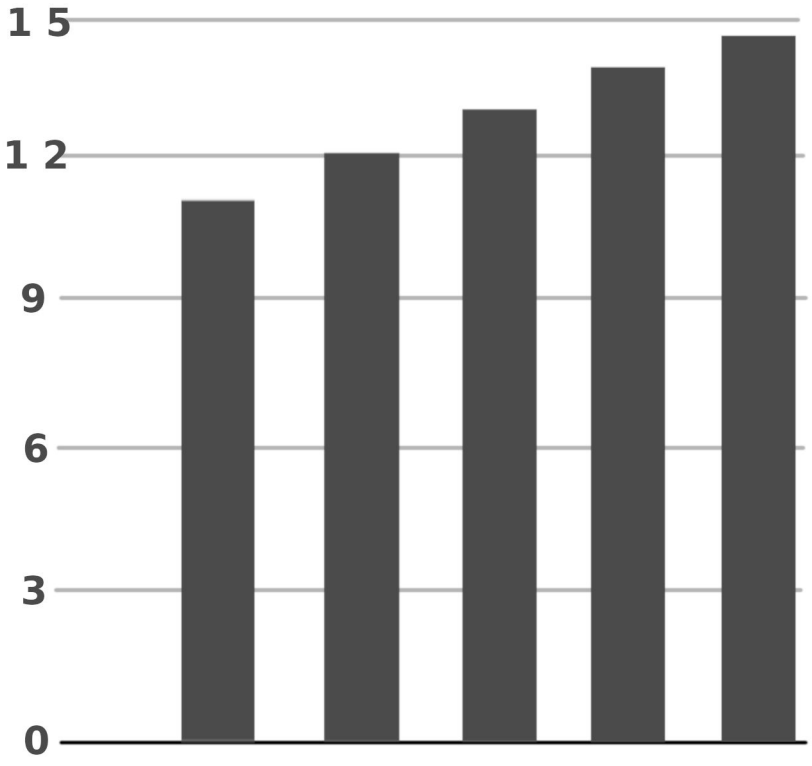
Team 7 to 8

Look through the WTF visualizations on this tumblr (viz.wtf)

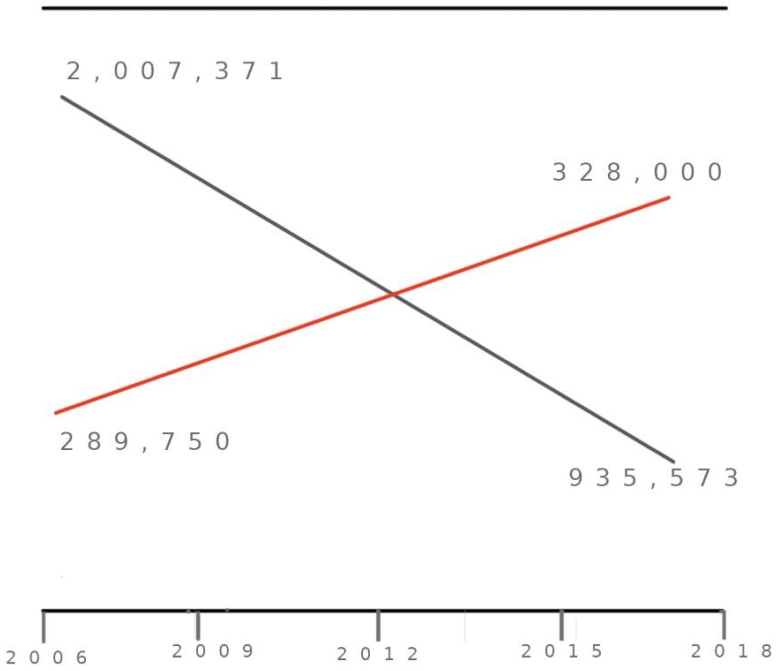
- What do these "bad" visualizations have in common?
- Making an assumption about the data source, choose one to "improve"
- Pick one best "bad" visualization for pair share

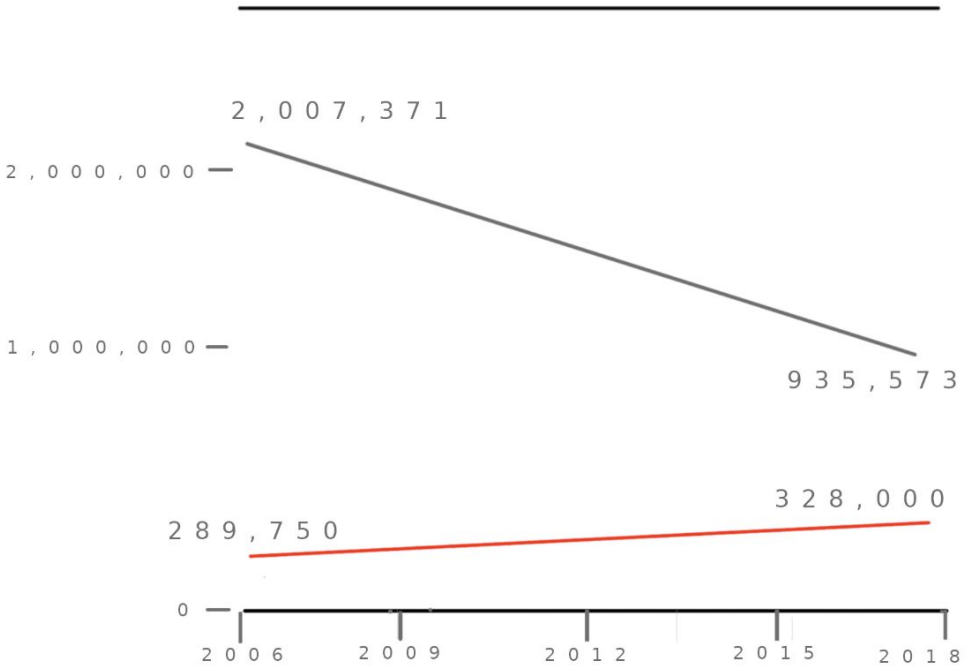
Group 5





Group 6





Data from Planned Parenthood

CURRICULUM RECAP

WRAP-UP & ADDITIONAL RESOURCES

WRAPPING UP.

Data is *people*...

...and there isn't a risk-free approach

ADDITIONAL RESOURCES:

- What to Consider When Planning a Digital Project
  - List of questions to consider through the lifespan of a project
- Feminist Data Manifest-no
  - Collection of digital projects that work through feminist frameworks and considerations
- Digital Precarity Manifesto
  - Precarity Lab's critical approach to digital work and data