

COGS138: Neural Data Science

Lecture 10

C. Alex Simpkins, PhD

UCSD Dept. of Cognitive Science, Spring 2023

RDPRobotics, LLC

http://casimpkinsjr.radiantdolphinpress.com/pages/cogs138_sp23

rdprobotics@gmail.com | csimpkinsjr@ucsd.edu

Plan for today

- Announcements
- Previous project review
- Project overview
- Review - Last time
- Statistical data analysis, Part 2

Announcements

- A2 - don't forget - due **Friday 5/5**
- Reading 2 - Released on canvas and in web site password protected area, lecture quiz due next **Tues 5/9 R2 quiz**
- **Group formation** - check canvas for empty groups, please self-sign up
- Previous project review released when we get the groups together (this week)
- Podcasts added to webpage along with several links to readings
- New book and article references

Last time

Course links

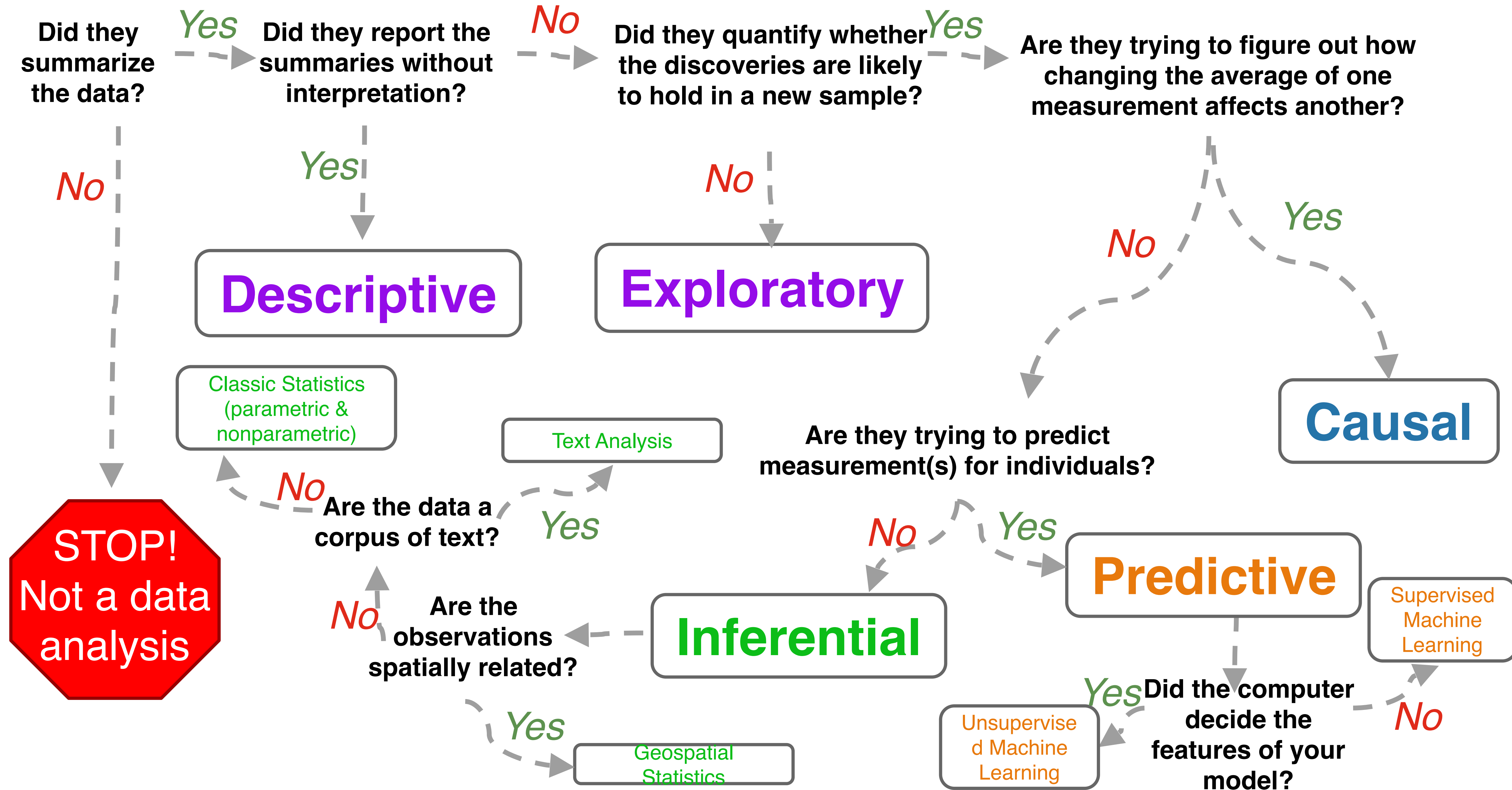
Website	http://casimpkinsjr.radiantdolphinpress.com/pages/cogs138_sp23	Main face of the course and everything will be linked from here. Lectures, Readings, Handouts, Files, links
GitHub	https://github.com/drsimpkins-teaching	files/data, additional materials & final projects
datahub	https://datahub.ucsd.edu	assignment submission
Piazza	https://piazza.com/ucsd/spring2023/cogs138_sp23_a00/home (course code on canvas home page)	questions, discussion, and regrade requests
Canvas	https://canvas.ucsd.edu/courses/44897	grades, lecture videos
Anonymous Feedback	Will be able to submit via google form	If I ever offend you, use an example you are uncomfortable with, or to provide general feedback. Please remain constructive and polite

“Data science is the process of formulating a quantitative question that can be answered with data, collecting and cleaning the data, analyzing the data, and communicating the answer to the question to a relevant audience.”

To do this, you have to *look at, describe, and explore* the data

Summary: Analytical Approaches

1. **Descriptive** (and **Exploratory**) Data Analysis are the first step(s)
2. **Inference** establishes relationships
 - a. Classic Statistics
 - b. Geospatial Analysis
 - c. Text Analysis
3. Machine Learning is for **prediction**
 - a. Supervised
 - b. Unsupervised
4. Experiments best way to establish the likelihood of **causality**
 - a. Remember you **cannot** establish causality with computational methods only correlations along with statistical beliefs



Statistical Data Analysis

- There are various definitions
- “Statistics” - the science of gathering data and discovering patterns
- “the science that deals with the **collection, classification, analysis, and interpretation of numerical facts or data**” [[dictionary.com](https://www.dictionary.com)]

What are the 2 types of statistics?

What are the 2 types of statistics?

- **Descriptive** - Summarizing the characteristics of data
- **Inferential** - Modeling, making 'inferences' from data

Descriptive statistics

- **Summarizing** the **characteristics** of data
 - Central tendency - (“center”) mean, median, mode
 - Variability - (“dispersion”) variance, standard deviation
 - Frequency distribution - (“occurrence within data”) counts
- Charts, plots, probability distribution shapes

Inferential statistics

- “Modeling” or making ‘inferences’ from the data
- Taking data from samples and making predictions about populations
- 2 types
 - *Estimating parameters*
 - *Hypothesis tests*

Estimating parameters

- Parametric data (data consisting of parameters)

Hypothesis testing

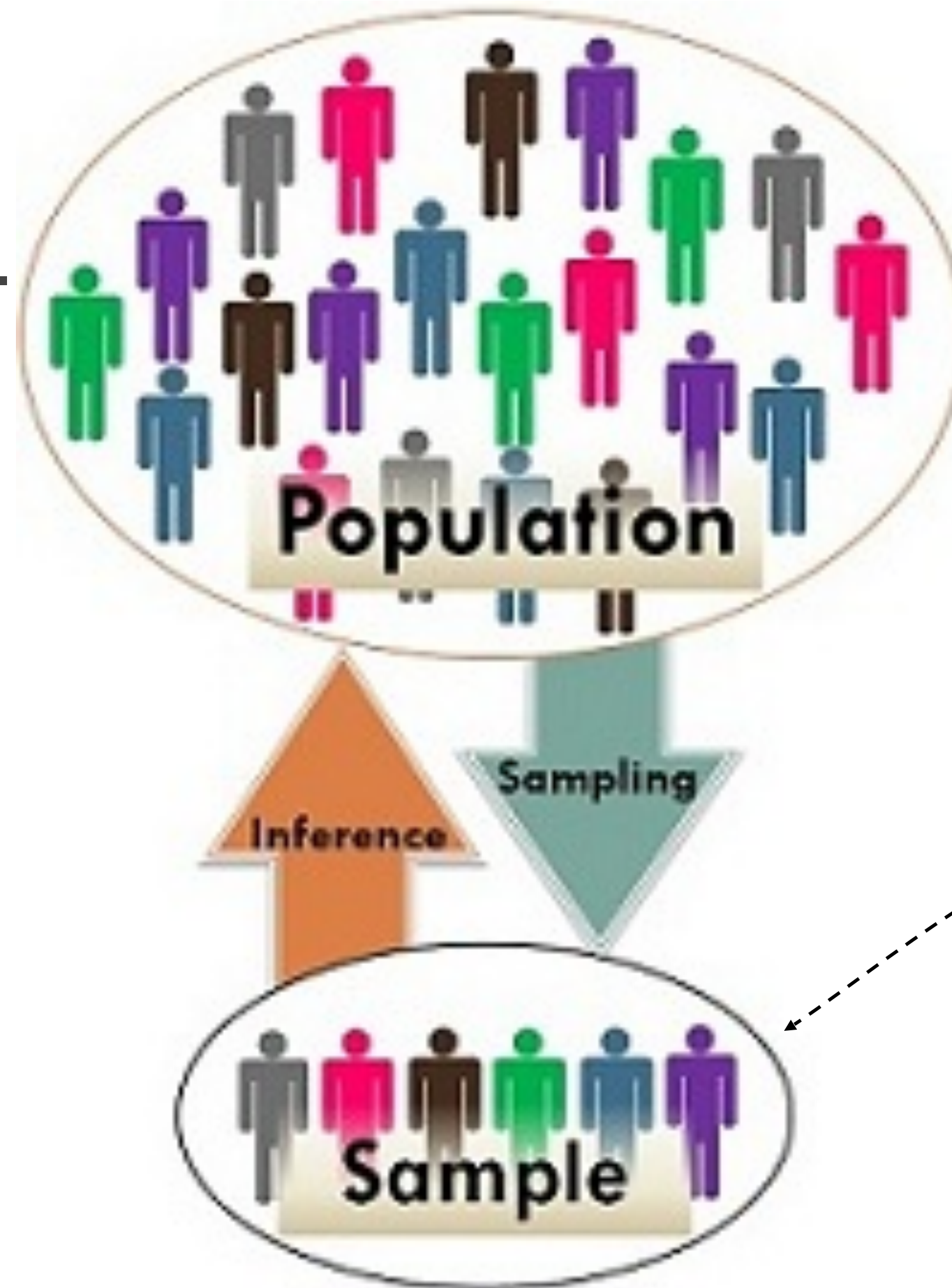
- Non-parametric data (no parameters)

Statistic

“A quantity computed from a sample”

Populations & Samples

We want to learn something about this..



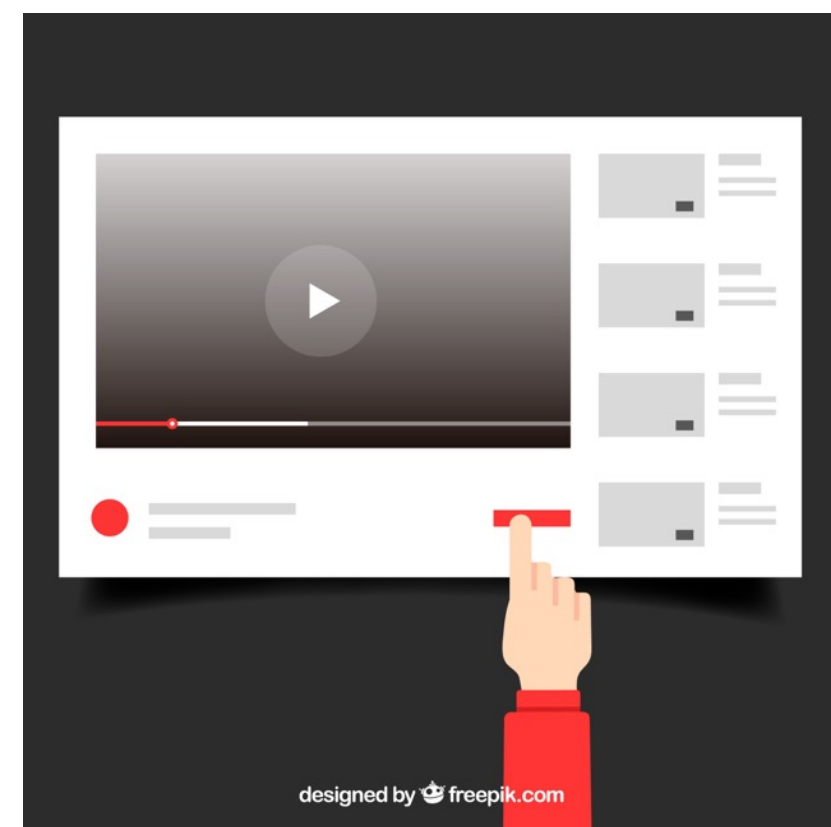
Our population: *all* Neurons in the motor cortex

Our sample: LFP ~ 1-10k neurons

....but we can only *actually* collect data from this

statistic

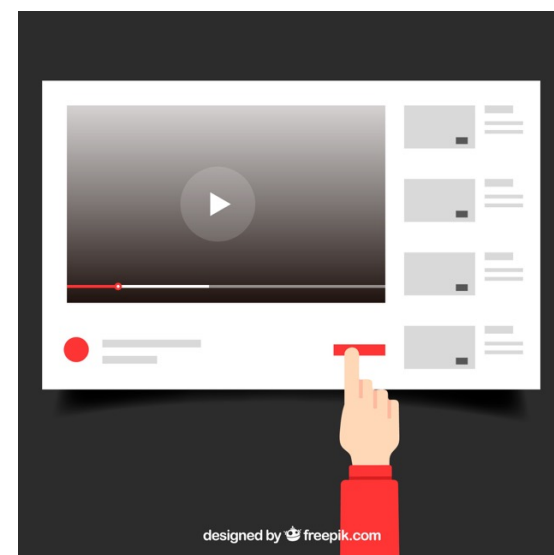
“*A quantity computed from a sample*”



For our YouTube analysis, we could take a random sample of comments from YouTube and calculate the following statistic: *the number of positive and the number of negative words in each review.*

Best sampling practices:

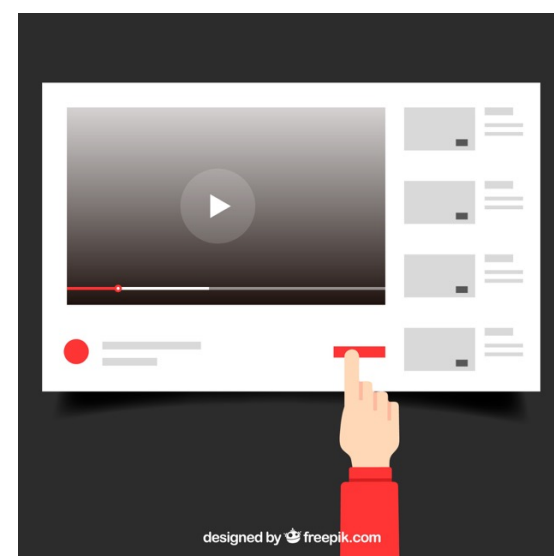
- Always think about what your population is
- Collect data from a sample that is representative of your population
- If you have no choice but to work with a dataset that is not collected randomly and is biased, be careful not to generalize your results to the entire population



You'd want to be sure you sample randomly across *all* YouTube comments, making sure not to get more comments from one genre over another, or one location over another, etc.

Examples of bad sampling:

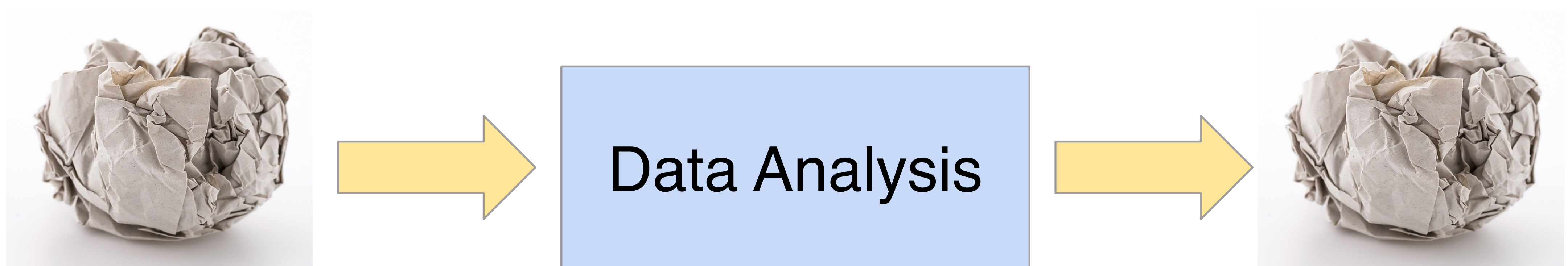
- Surveying subscribers of a Marvel movie magazine for research on Americans' attitudes toward DC movies
- Randomly sampling Facebook users for what TV shows people like



To understand *all* YouTube comments, you wouldn't just want to sample from one YouTube channel, or videos in a single language.

It's *always* worth spending time at the beginning of a project to determine whether or not the data you have are garbage. Be certain they are actually able to help you answer the question you're interested in.

GIGO : Garbage In. Garbage Out.



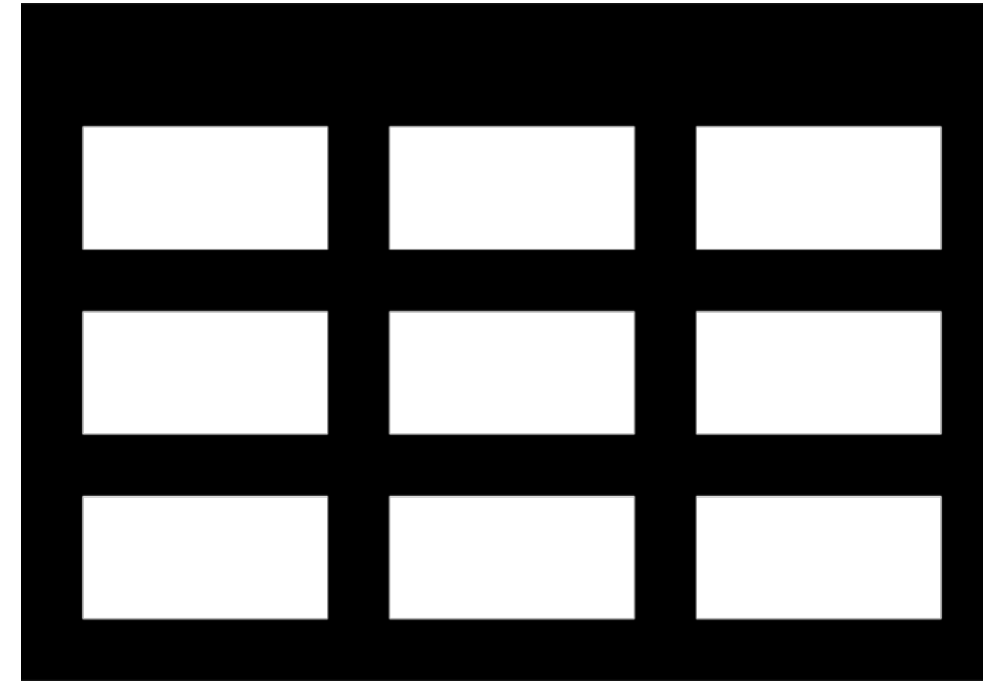


For survey data I collect from you all, which of the following best describes the population I could generalize findings back to.

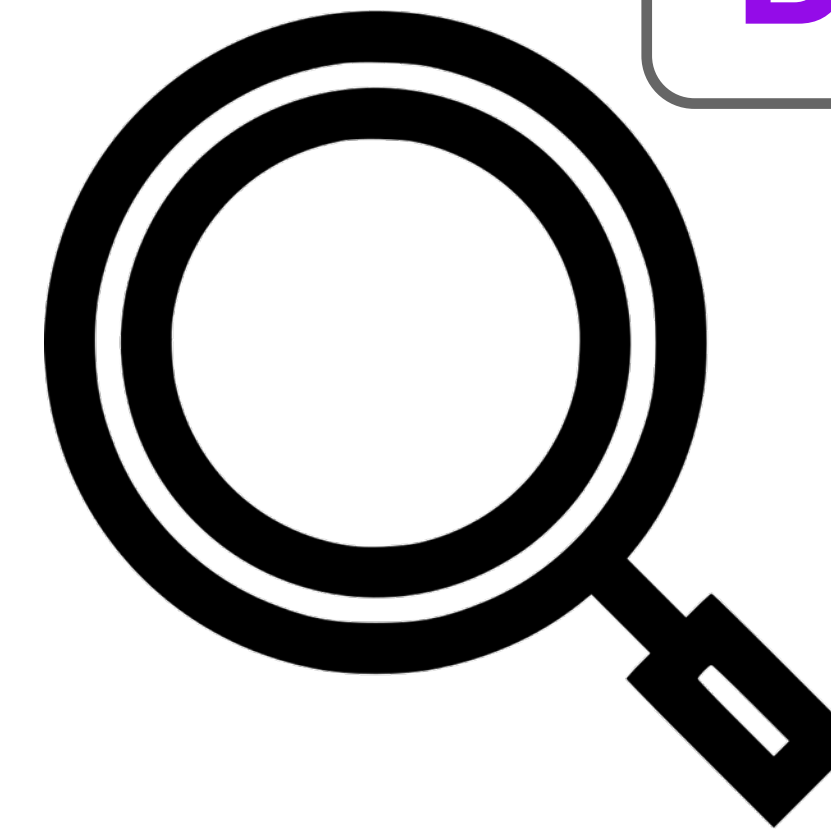
- A** Undergraduates
- B** Undergraduates in the US
- C** Undergraduates at UCSD
- D** Students aged 18-25
- E** UCSD COGS138 students

Descriptive

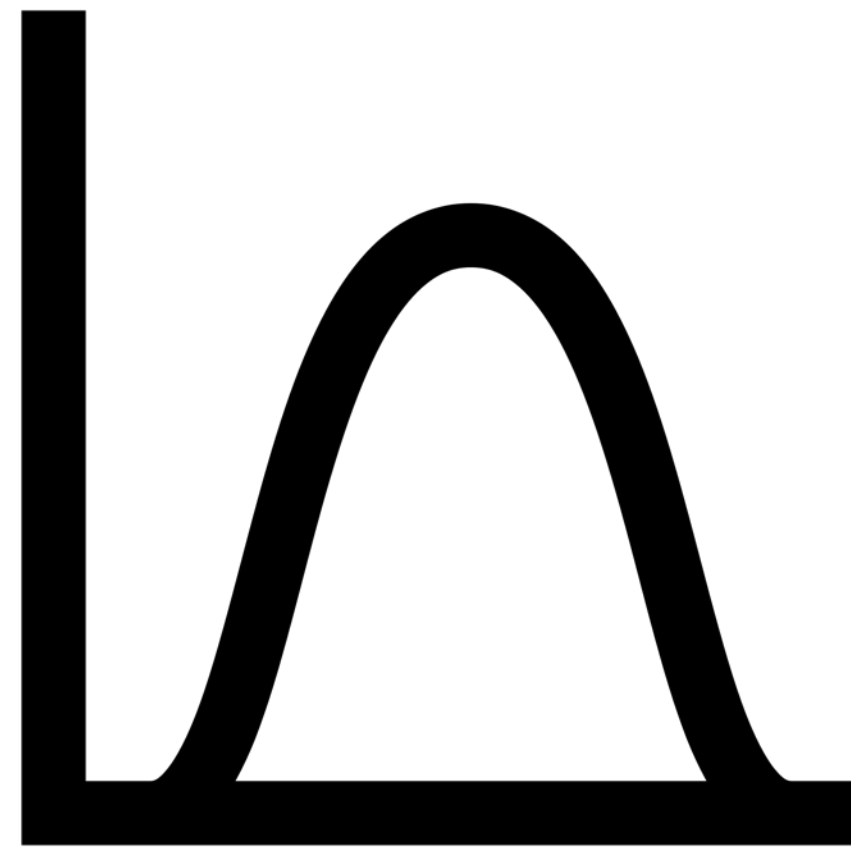
Descriptive Analysis



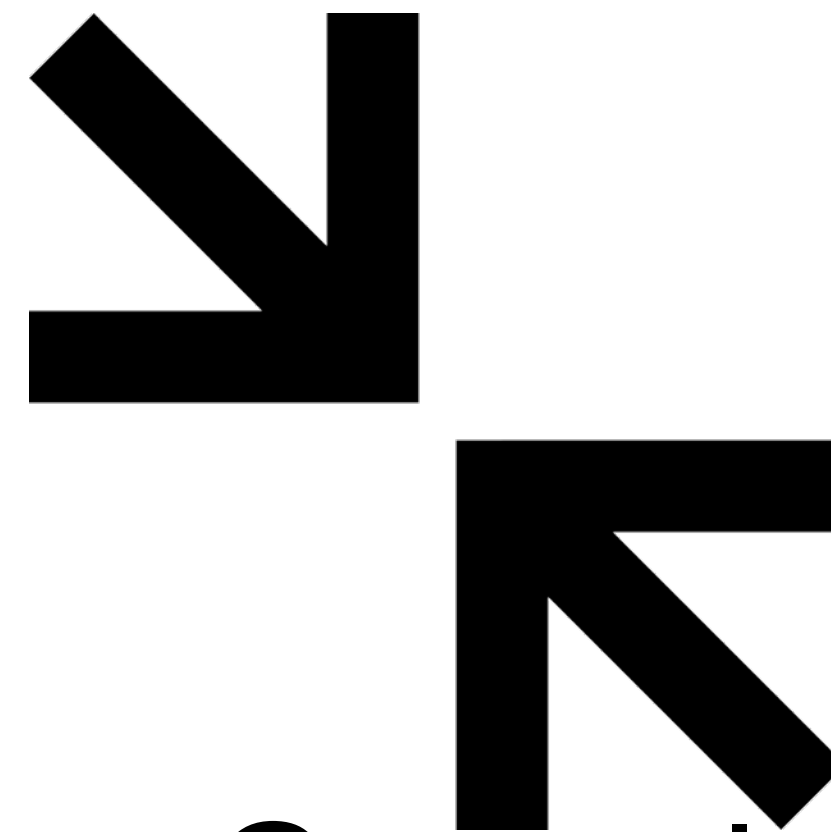
Size



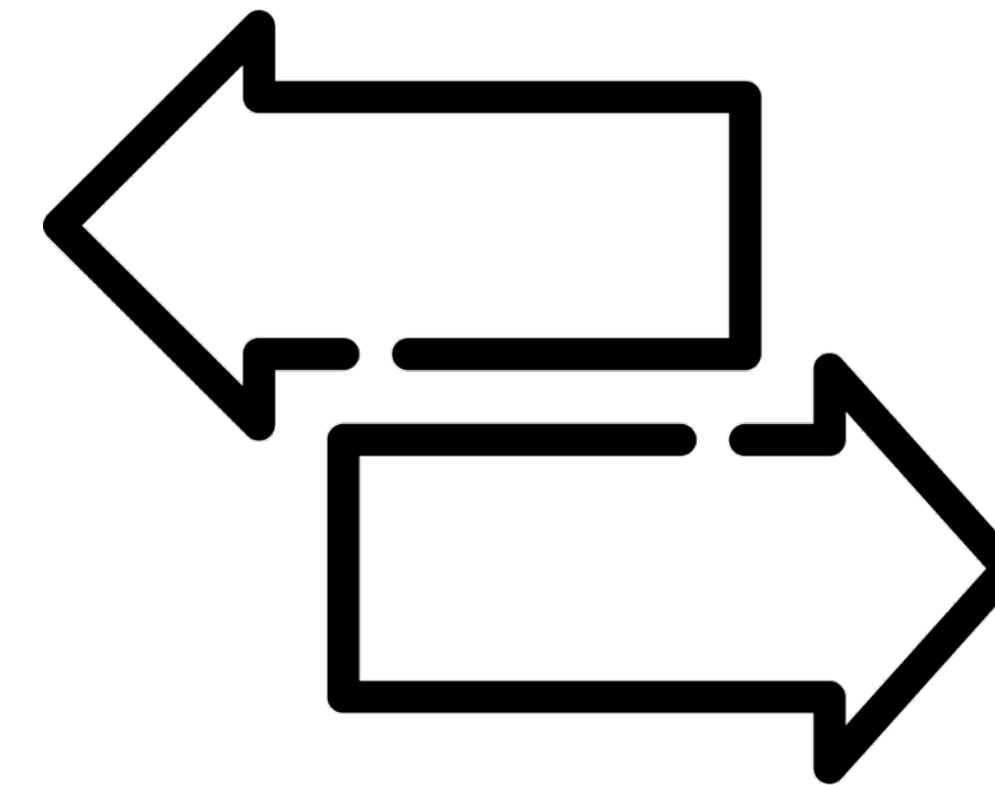
Missingness



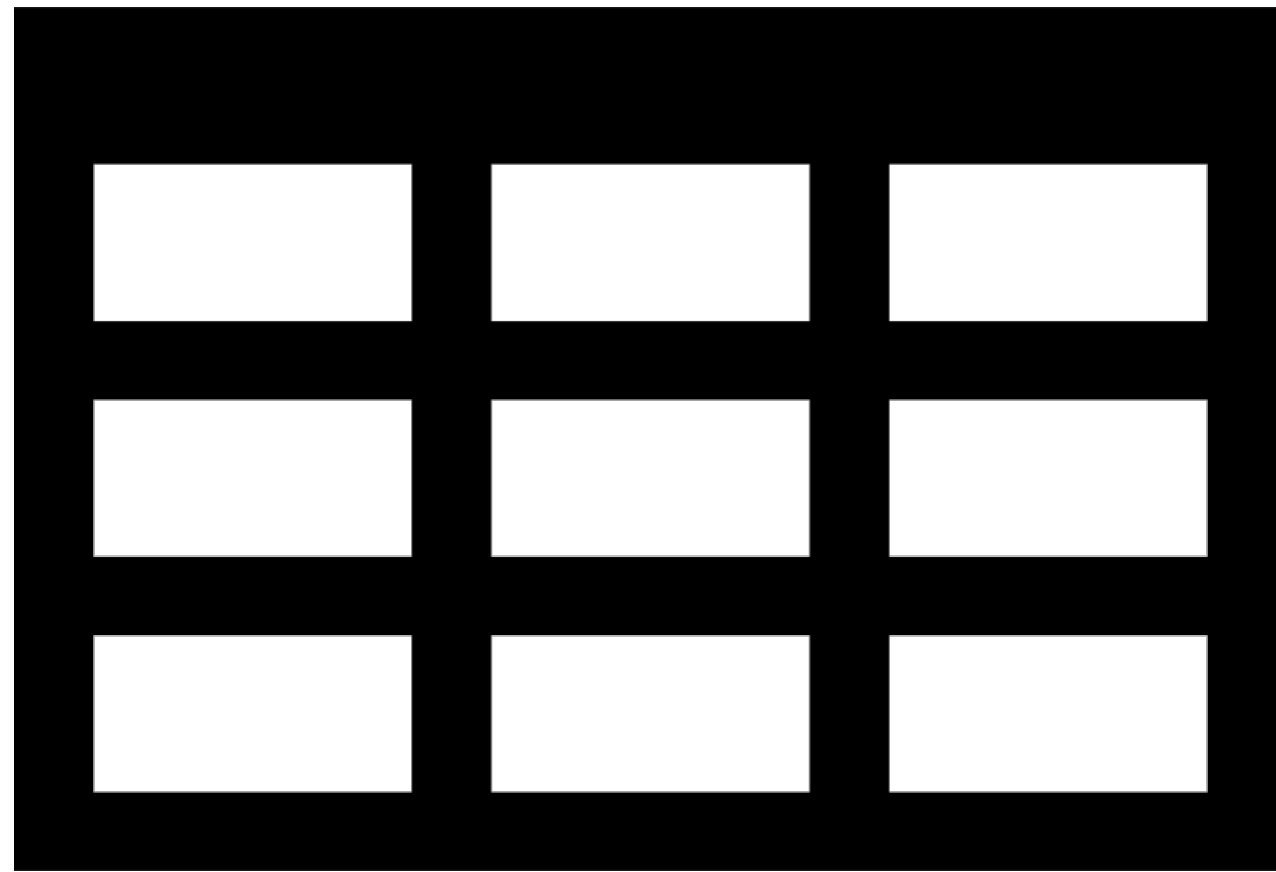
Shape



Central
Tendency



Variability



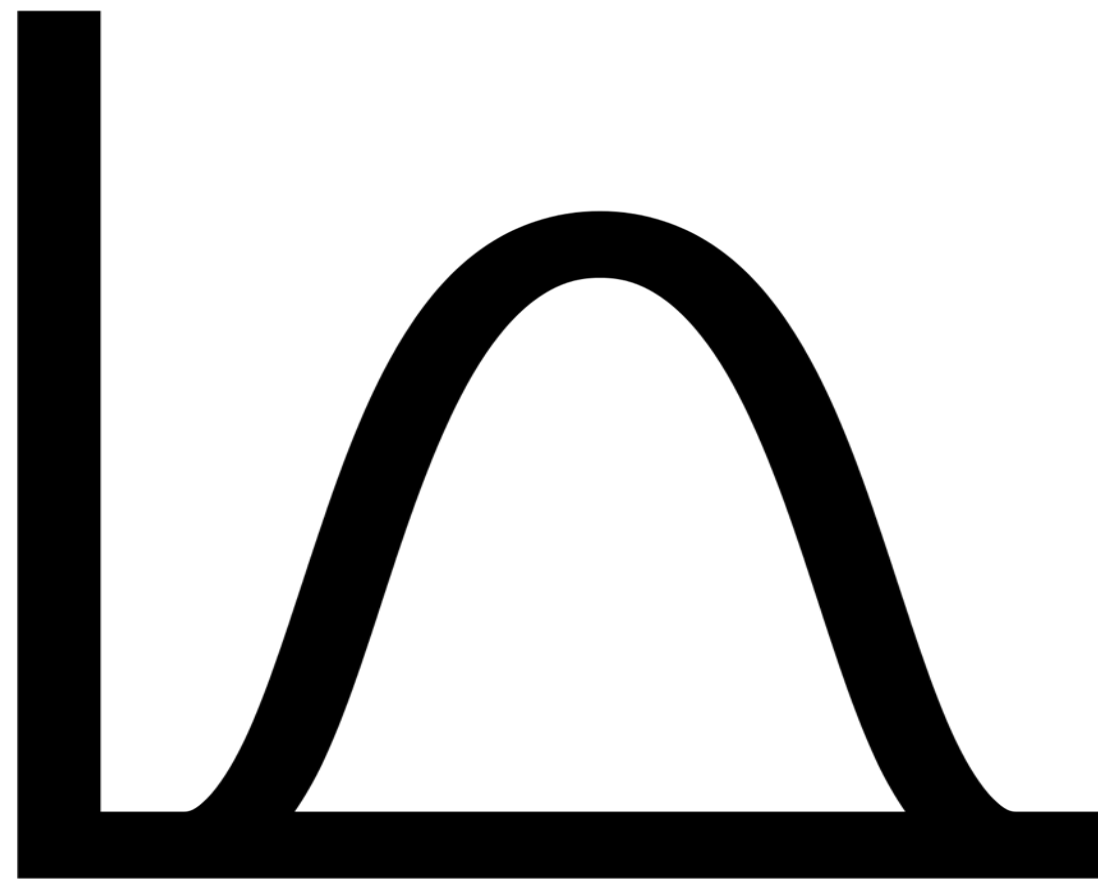
Size

How many observations (rows) and variables (columns) you have is an important first step. You should always be aware of the **size** of your dataset .



Missingness

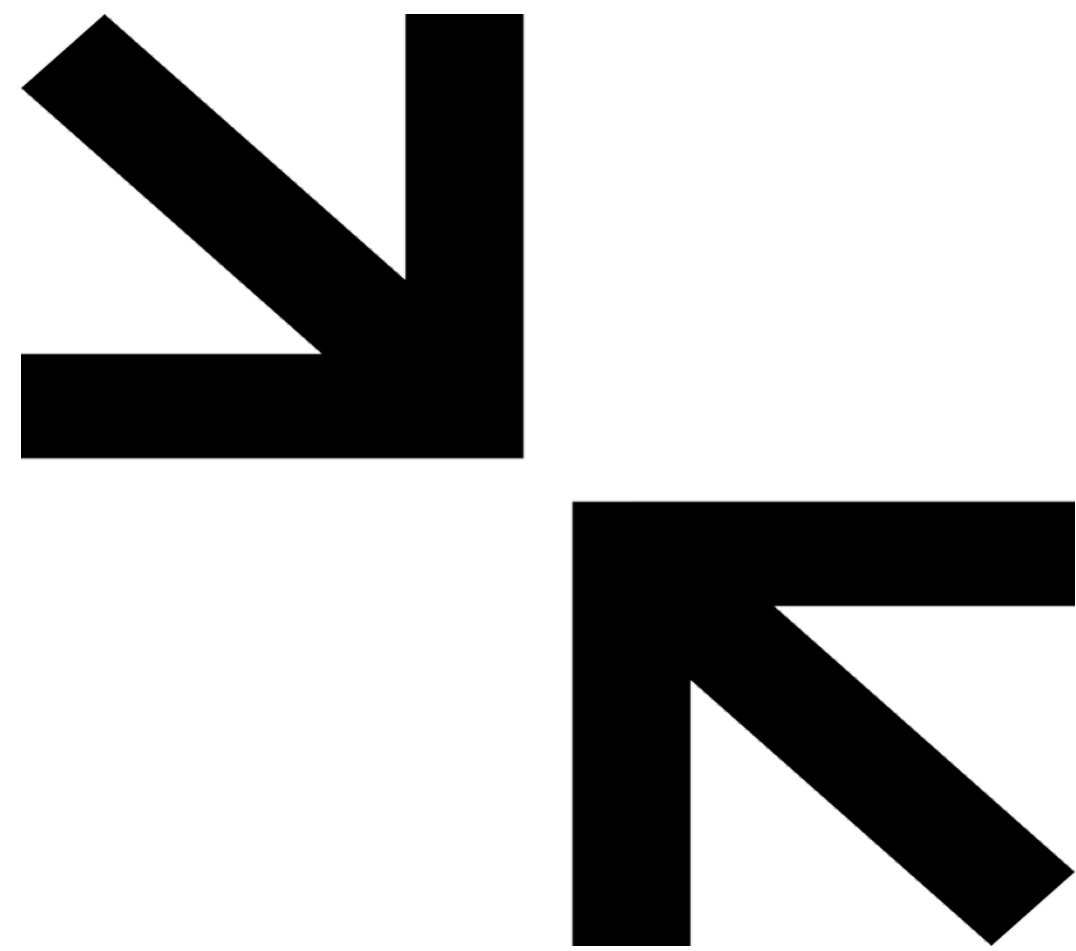
It's critical to know **how many observations have missing data** for variables of interest in your data. Knowing *why* their missing is also important.



Shape

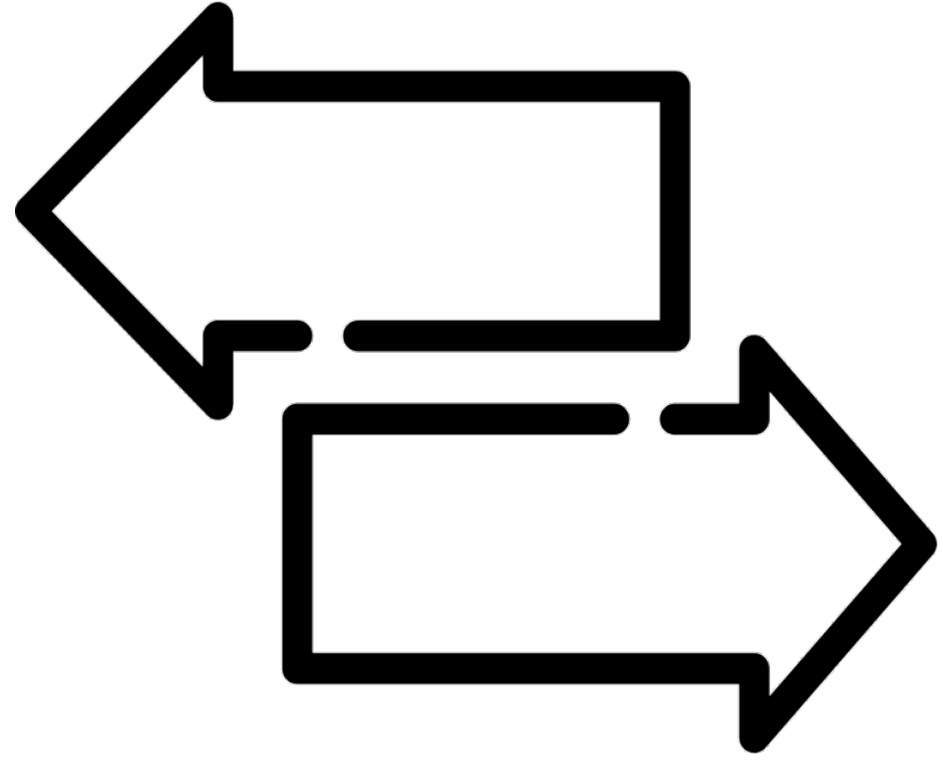
It's critical to know **the distribution of the variables** in your dataset. Certain statistical approaches can only be used with certain distributions.

Descriptive



Central Tendency

Knowing the **mean, median, and/or mode** can help you get an idea of what a typical value is for your variable(s) of interest



Variability

The central tendency tells you part of the story. The **variability in the values** in your observation helps fill in the rest.



Which of the following is NOT something accomplished by a descriptive analysis?

A Describes typical values in your dataset

B Determines the size of your dataset

C Establishes causal relationships between variables

D Identifies missing data

E Determines how variable values in your dataset are

Descriptive Statistics & Summary

- We look for the summary, the relevant features of the data to the study
- Use statistics to express them

Descriptive

Descriptive Analyses are often included as “Table 1” in academic publications

Table 1. Baseline Characteristics of the Patients.*

Characteristic	Ranibizumab Monthly (N=301)	Bevacizumab Monthly (N=286)	Ranibizumab as Needed (N=298)	Bevacizumab as Needed (N=300)
Age — no. (%)				
50–59 yr	2 (0.7)	1 (0.3)	6 (2.0)	2 (0.7)
60–69 yr	33 (11.0)	28 (9.8)	31 (10.4)	34 (11.3)
70–79 yr	102 (33.9)	84 (29.4)	115 (38.6)	103 (34.3)
80–89 yr	142 (47.2)	150 (52.4)	126 (42.3)	142 (47.3)
≥90 yr	22 (7.3)	23 (8.0)	20 (6.7)	19 (6.3)
Mean — yr	79.2±7.4	80.1±7.3	78.4±7.8	79.3±7.6
Sex — no. (%)				
Female	183 (60.8)	180 (62.9)	185 (62.1)	184 (61.3)
Male	118 (39.2)	106 (37.1)	113 (37.9)	116 (38.7)
Race — no. (%)†				
White	297 (98.7)	281 (98.3)	296 (99.3)	294 (98.0)
Other	4 (1.3)	5 (1.7)	2 (0.7)	6 (2.0)
History of myocardial infarction — no. (%)				
	34 (11.3)	40 (14.0)	30 (10.1)	36 (12.0)
History of stroke — no. (%)				
	14 (4.7)	18 (6.3)	22 (7.4)	16 (5.3)
History of transient ischemic attack — no. (%)				
	12 (4.0)	25 (8.7)	12 (4.0)	19 (6.3)
Blood pressure — mm Hg				
Systolic	134±18	135±19	136±17	135±17
Diastolic	75±10	75±10	76±9	75±10
Visual-acuity score and Snellen equivalent				
68–82 letters, 20/25–40 — no. (%)	111 (36.9)	94 (32.9)	116 (38.9)	103 (34.3)
53–67 letters, 20/50–80 — no. (%)	98 (32.6)	118 (41.3)	108 (36.2)	119 (39.7)
38–52 letters, 20/100–160 — no. (%)	67 (22.3)	53 (18.5)	58 (19.5)	58 (19.3)
23–37 letters, 20/200–320 — no. (%)	25 (8.3)	21 (7.3)	16 (5.4)	20 (6.7)
Mean score	60.1±14.3	60.2±13.1	61.5±13.2	60.4±13.4
Total thickness at fovea — μm‡				
	458±184	463±196	458±193	461±175
Retinal thickness plus subfoveal-fluid thickness at fovea — μm				
	251±122	254±121	247±122	252±115
Foveal center involvement — no. (%)				
Choroidal neovascularization	176 (58.5)	153 (53.5)	176 (59.1)	183 (61.0)
Fluid	85 (28.2)	81 (28.3)	77 (25.8)	72 (24.0)
Hemorrhage	20 (6.6)	24 (8.4)	24 (8.1)	25 (8.3)
Other	18 (6.0)	20 (7.0)	15 (5.0)	18 (6.0)
No choroidal neovascularization or not possible to grade	2 (0.7)	8 (2.8)	6 (2.0)	2 (0.7)

* Plus-minus values are means ±SD.

† Race was self-reported.

‡ Total thickness at the fovea includes the retina, subretinal fluid, choroidal neovascularization, and retinal pigment epithelial elevation.

Descriptive

Size

Zooming in on this we see variables stratified by Age, Sex, and Race

Table 1. Baseline Characteristics of the Patients.*

Characteristic	Ranibizumab Monthly (N=301)	Bevacizumab Monthly (N=286)	Ranibizumab as Needed (N=298)	Bevacizumab as Needed (N=300)
Age — no. (%)				
50–59 yr	2 (0.7)	1 (0.3)	6 (2.0)	2 (0.7)
60–69 yr	33 (11.0)	28 (9.8)	31 (10.4)	34 (11.3)
70–79 yr	102 (33.9)	84 (29.4)	115 (38.6)	103 (34.3)
80–89 yr	142 (47.2)	150 (52.4)	126 (42.3)	142 (47.3)
≥90 yr	22 (7.3)	23 (8.0)	20 (6.7)	19 (6.3)
Mean — yr	79.2±7.4	80.1±7.1	78.4±7.8	79.3±7.6
Sex — no. (%)				
Female	183 (60.8)	180 (62.9)	185 (62.1)	184 (61.3)
Male	118 (39.2)	106 (37.1)	113 (37.9)	116 (38.7)
Race — no. (%)†				
White	297 (98.7)	281 (98.3)	296 (99.3)	294 (98.0)
Other	4 (1.3)	5 (1.7)	2 (0.7)	6 (2.0)

* Plus-minus values are means ±SD.

† Race was self-reported.

‡ Total thickness at the fovea includes the retina, subretinal fluid, choroidal neovascularization, and retinal pigment epithelial elevation.

Shape

Central variability

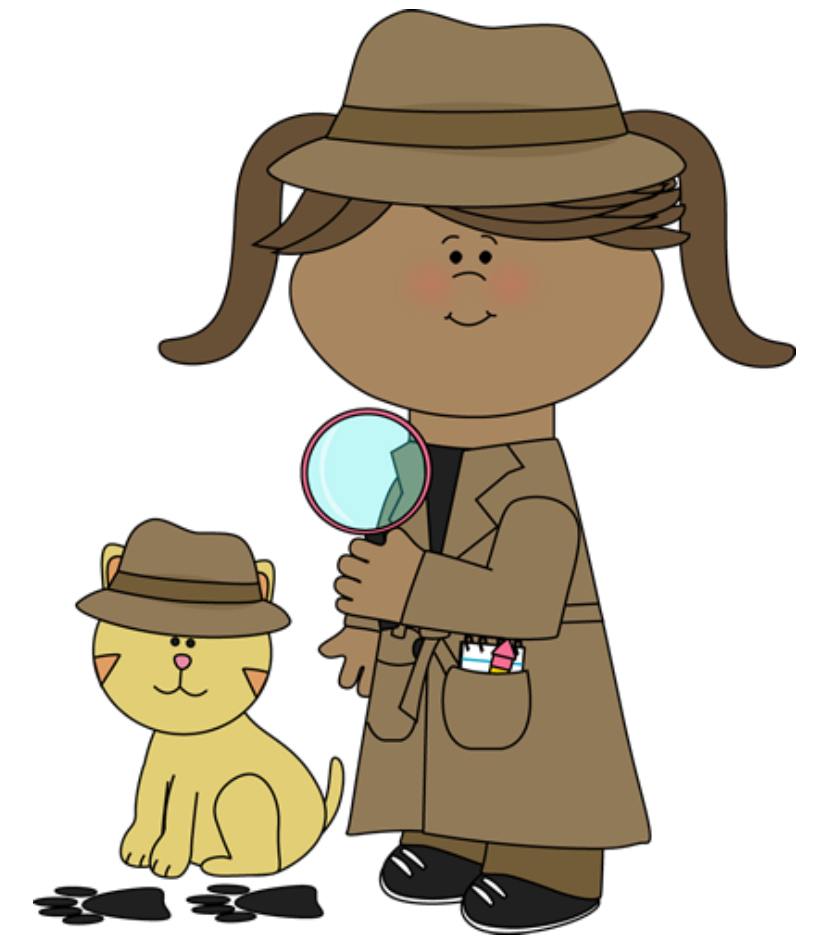
tendency

Descriptive Statistics & Summary

Calculating descriptive statistics, understanding what they tell you about your data, and reporting them are critical steps in every analysis.

Exploratory: The goal is to find unknown relationships between the variables you have measured in your data set. Exploratory analysis is open ended and designed to verify expected or find unexpected relationships between measurements.

Exploratory



Exploratory Data Analysis (EDA)
detective work answering the question:
“What can the data tell us?”

Why EDA?

Exploratory

- Understand data properties
- Discover Patterns
- Generate & Frame Hypothesis
- Suggest modeling strategies
- Check assumptions (sanity checks)
- Communicate results (present the data)

.....and if you don't, you'll regret it

You must always explore your data

衆瞽
摸象之圖



The dataset

You



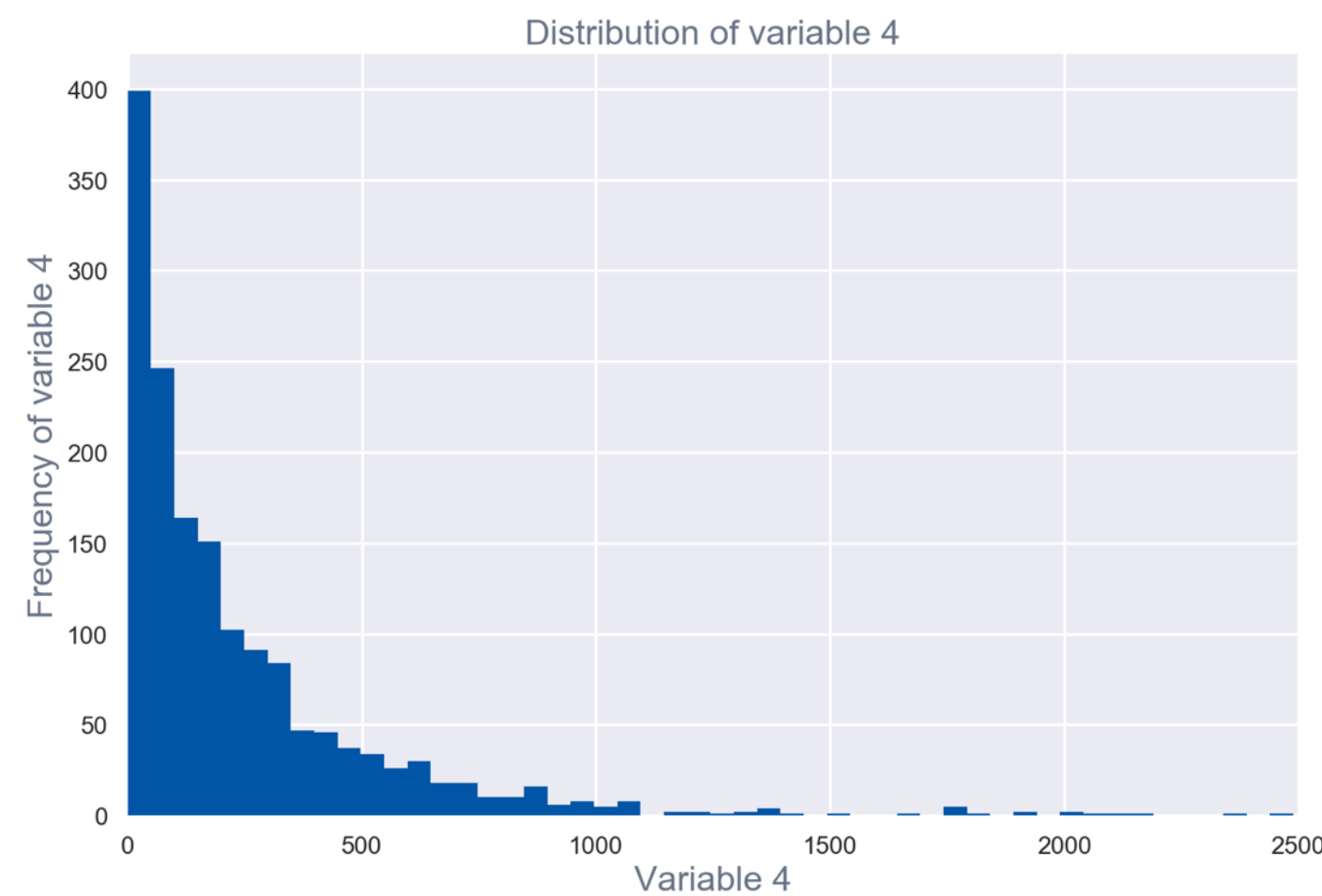
The general principles of exploratory analysis:

- Look for missing values
- Look for outlier values
- Calculate numerical summaries
- Generate plots to explore relationships
- Use tables to explore relationships
- If necessary, transform variables

EDA Approaches to “Get a Feel for the Data”

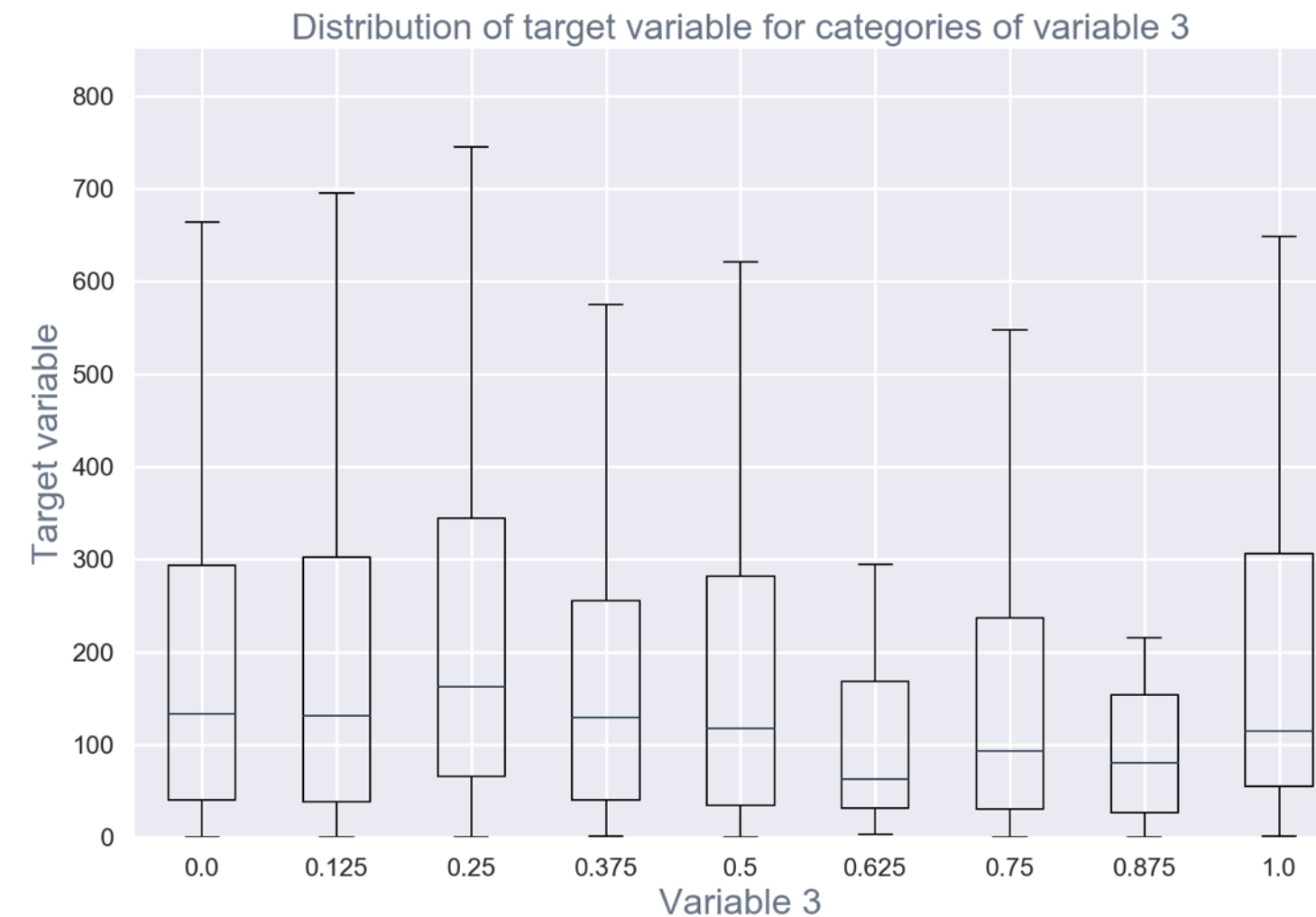
Understanding the relationship between variables in your dataset

Exploratory



Univariate

understanding a single variable
i.e.: histogram, densityplot, barplot



Bivariate

understanding relationship between 2 variables
i.e.: boxplot, scatterplot, grouped barplot, boxplot



Dimensionality Reduction

projecting high-D data into a lower-D space
i.e.: PCA, ICA, Clustering

On to today...

About the final projects

Finding the project files

- Blank starter document for report, handouts and info to start with (draft): https://github.com/drsimpkins-teaching/cogs138/tree/main/main_project
- Links to old projects: <https://github.com/NeuralDataScience/Projects>

Where will you turn in, work from?

- Github - we will create a repository for each group with the files in previous slides as a starter directory
- You'll add your data, notebook file etc there
- This will be the final turning location

Final Project Overview

1. Identify **questions** that you can answer by using publicly available datasets
- 1. Integrate** different datasets
1. Implement **technical skills** to answer your scientific questions
1. Work effectively with a **team**

What will you be turning in??

Proposal

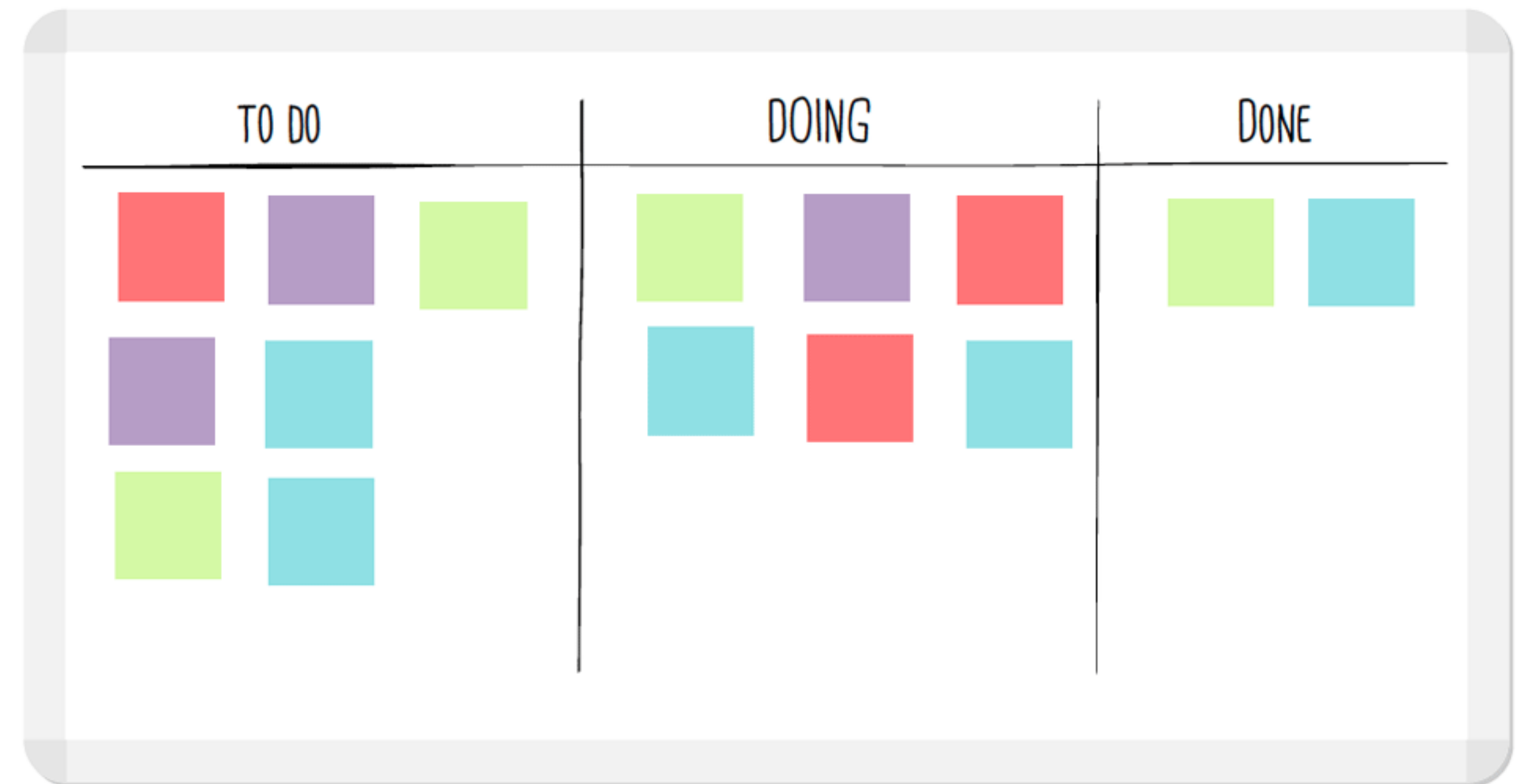
- 1 page document
- Pitching your question & approach

Final Project

- Jupyter notebook
- Steps of analysis that answers your main question
- Background and discussion sections

Proposal

1. Group member names
1. Experimental question
1. Background
1. Approach



Final Project

1. Intro:

a. Overview, Question, Background, Hypothesis

2. Data Analysis:

a. Wrangling, Viz, Results

3. Conclusion:

a. Discussion, Limitation, Future Steps

Final Project

1. Intro:

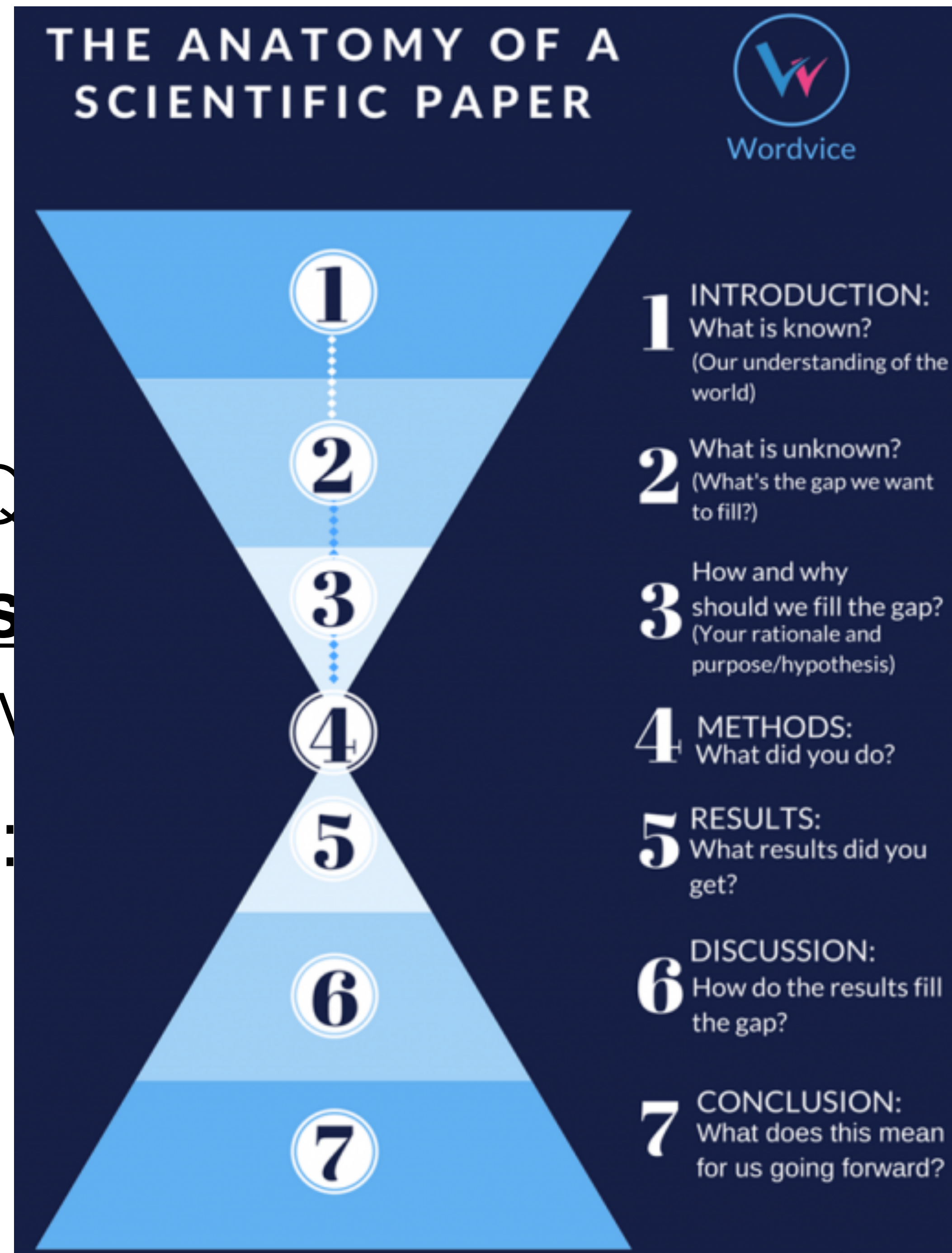
a. Overview, Q

2. Data Analysis

a. Wrangling, V

3. Conclusion:

a. Discussion,



hypothesis

Final Project

Overview

In this work, we aim to develop new approaches to automated hypothesis generation by utilizing the vast neuroscience literature. Building on prior work in semi-automated hypothesis generation, we present a hypothesis-first algorithm for computing a hypothesis "attractiveness" metric which is more

Data Collection

We based our data collection on the same 591 neuroscience terms that were analyzed in Voytek[1], to provide a reasonable benchmark to compare our models against. Using these terms, we utilized the NCBI E-Utils API to fetch intersection and union counts for the ~174k hypotheses (term-pairs) for the years 1700-2000 and 2000-2021 that were based on a term's presence in either the title and/or abstract of a given paper. For example, given a term pair (t1, t2), we would query all papers in a given time period whose titles and/or abstracts contain both t1 AND t2, and compute the sum of t1 NOT t2 and t2 NOT t1. From here, we constructed three matrices where the entries in each represented the intersection, union, and intersection over union (connectivity).

To make the fetching of multiple time periods reasonable we designed a scraper, called MicroLisc (*full source: ./data_collection/microlisc.py*), that utilizes multi-threading to make requests with a throughput of ~200 requests/sec. Additionally, we were able to vastly reduce the number of requests via caching since $|X \text{ NOT } Y| + |Y \text{ NOT } X| = |X| + |Y|$, when there are no intersections between X AND Y. Hence, by checking for intersection first, we could then retrieve the cardinalities of X and Y from a cache, leaving most hypotheses only requiring only 1 HTTP request (not intersecting), and at max 3 HTTP requests (if intersecting). The combination of these two optimizations allowed us to collect the complete set of hypothesis data for an entire time period in under 1.5 hours. We believe this decrease in data collection time will facilitate further analysis of multiple time periods.

Hypothesis

Our algorithm's method of computing the "attractiveness" metric allows for the hypotheses to be compared on an ordinal scale. In doing so, our hypotheses will be more robust to spurious correlations and have greater specificity. We conjecture that these features of the resulting hypotheses make them more meaningful in terms of their utility in predicting future trends in neuroscience.

computing percent increase in citations between the two periods.

<https://github.com/NeuralDataScience/Projects/blob/main/Wi2021/FinalNotebookGroup-TextMining.ipynb>

Final Project

Good commenting and following PEP guidelines (<https://www.python.org/dev/peps/pep-0008/>)

Data Wrangling

We configure our data directories and nest terms with their synonyms, producing a list of terms. We initialize our scraper(`data_collection/microlisc.py`) with this specific term set and run a job with a specified conjunction, intersection, and connectivity counts.

Configuration and Brain Term Initialization

```
[2]: # set up directory & file hierarchy
proj_dir = os.getcwd()
data_dir = os.path.join(proj_dir, 'data')
old_data_dir = os.path.join(proj_dir, 'old_data')
new_data_dir = os.path.join(proj_dir, 'new_data')
neighbor_data_dir = os.path.join(new_data_dir, 'neighbors_of_neighbors')
hypothesis_data_dir = os.path.join(new_data_dir, 'hypothesis_first')

# load in the 591 terms from the original analysis
df = pd.read_csv(os.path.join(data_dir, 'brain_terms_new.csv'), names=None)

# extract useful information from the dataframe
term_types = df.domain.unique()
mapping = dict(df[['term', 'domain']].values)
mapping_lower = {k.lower(): v for k, v in mapping.items()}

# Nest all unique terms from the terms dataframe
lis = [[f'"{j}"' for j in list(set(list(df[df.term == i]['synonyms']) \
                                + list(df[df.term == i]['term'])))] for i in df.term.unique()]

assert len(lis) == 588
```

PEP 8 -- Style Guide for Python Code

Final Project

Discussion and Further Directions

Based on the approaches implemented so far and the success we have seen, there is a range of different directions we could choose to further develop our architecture for automated hypothesis generation.

A more descriptive dataset: Currently, our architecture is limited by the scope of terms we are investigating and the data we are capturing related to those terms. To be specific the current architecture leans heavily on the scraped term-pair counts of just PubMed published literature. Whilst this is sufficient as a proof of concept, it isn't comprehensive and there are many ways we can further develop this stage of the project. One idea was to explore the citations of the papers we are scraping and the h-index of the authors publishing said papers. Based on how the current architecture works, terms can co-occur frequently in a range of papers, however, these papers can be largely insignificant, rarely being cited or reviewed. By utilizing the citations of each paper and how citations are growing over time we can add an extra dimension of scrutiny to our produced hypotheses, potentially boosting the utility of the output.

Addition of NLP to add extra insight into scraped hypotheses: One proposed area for further investigation is the addition of NLP-driven insights into the data collection process. Currently, the models derive insight from data relating to term occurrence. The addition of sentiment analysis could add a level of scrutiny to the number and type of occurring matches, possibly resulting in more accurate hypotheses. This is because simply capturing term occurrence fails to account for how the terms are occurring together. For example, take into account these two passages:

- Alzheimer's disease is related to the deterioration of the prefrontal cortex.
- Alzheimer's disease has no relation to the deterioration of the prefrontal cortex.

Although these statements communicate contradictory information, our current algorithm counts them the same. By building functionality for understanding sentiment, our model would be better able to make better decisions relating to the hypothesis it produces.

Construction of a research tool: Following the testing and finalization of the hypothesis generation framework, the team is interested in developing a Python Package Index API for public use. This API would function as a streamlined way for prospective researchers to query for hypotheses based on our refined models, allowing for deployment in the wider scientific community.

<https://github.com/NeuralDataScience/Projects/blob/main/Wi2021/FinalNotebookGroup-TextMining.ipynb>

Groups

Distributions

- <http://localhost:8888/notebooks/Documents/teaching/cogs138/old/Tutorials-master/Distributions.ipynb>
- <http://localhost:8888/notebooks/Documents/teaching/cogs138/old/Tutorials-master/Central%20Limit%20Theorem.ipynb>
- <http://localhost:8888/notebooks/Documents/teaching/cogs138/old/Tutorials-master/Correlation%20resampling.ipynb>