COGS138: Neural Data Science

Lecture 19

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
- In class discussion about EDA checkpoint

Announcements

- ***<u>Deadlines upcoming this week:</u>***
- **Tuesday**:
- **Wednesday:**
- **Saturday**:
  - Reading Quiz 4 11:59pm
  - Lecture quiz
  - EDA checkpoint 11:59pm
- **Next Saturday** for A5 and EC assignment, missing quizzes, etc

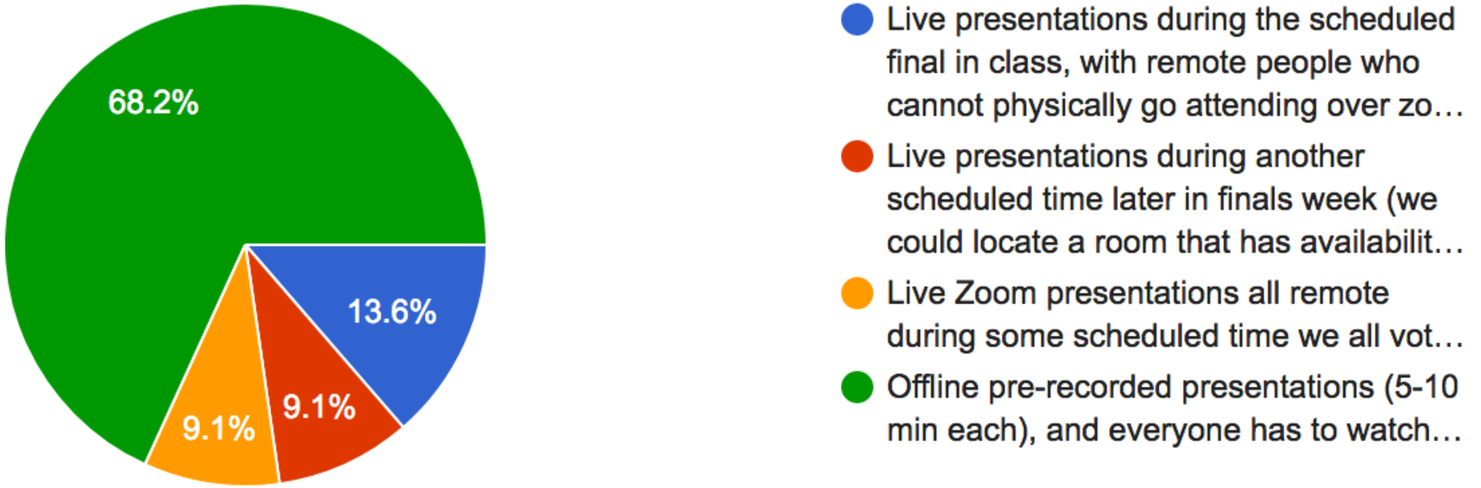
Announcements

- Extra office hours for Dr. Simpkins this week (Friday at 12-1pm)
    - Will be available all day either over zoom via quick appointment or piazza/email as always
- Project feedback for the Data checkpoint will be released by the end of the evening if not there yet

Preference for your **favorite** method of doing the final presentations and feedback for each other
(choose the top that you prefer, and you can vote on a secondary option that works for you, as well as comment)

Copy

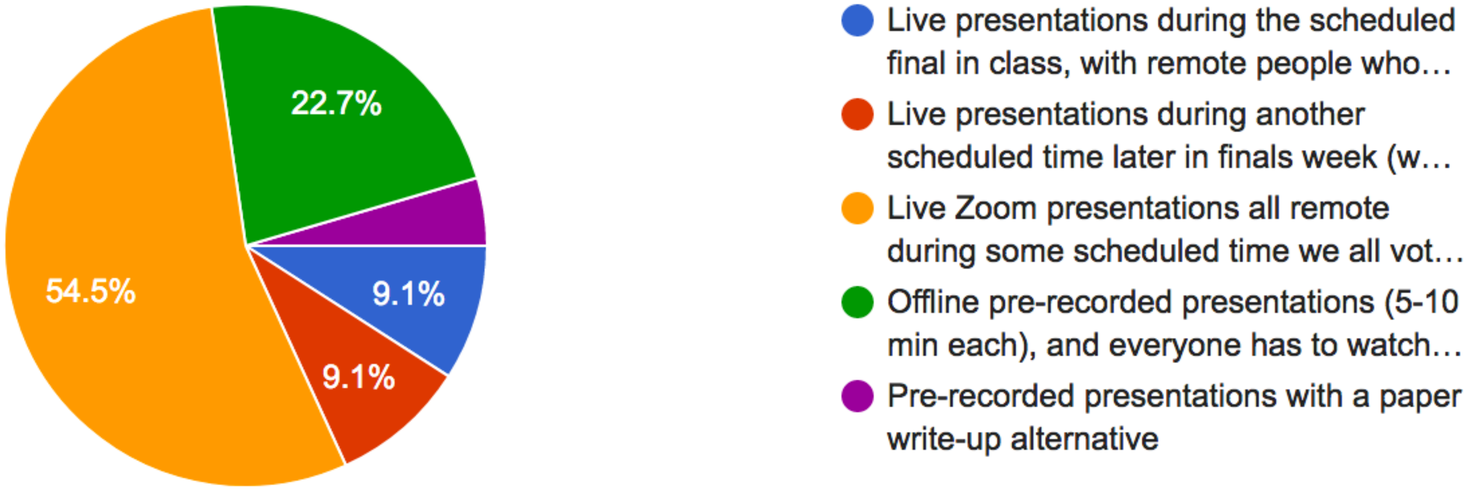
22 responses



Preference for your **second favorite** method of doing the final presentations and feedback for each other
(choose the second favorite that you prefer, and feel free to add detail in the comments area)

Copy

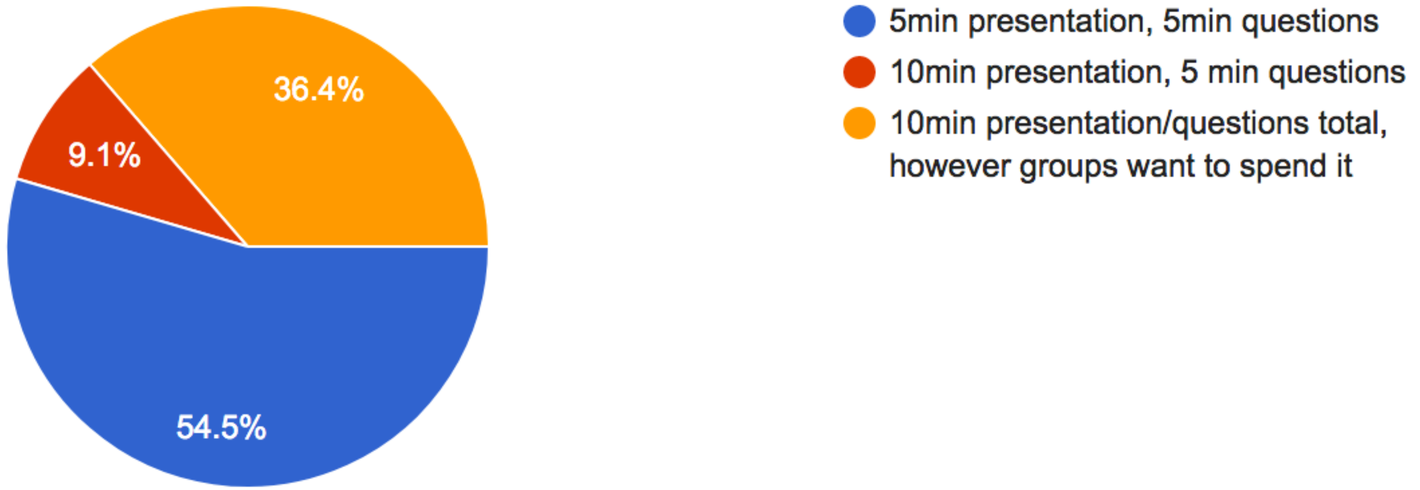
22 responses



What length do you prefer? 5min presentation with potentially 5min questions is short to get it all across, but takes less overall time, 10min is longer especially if we do 5min questions, or we could set 10min total however the groups want to spend it.

Copy

22 responses



Announcements II

- Final presentations
  - Vote was mainly for offline, about 1/3 online
  - Backup was for zoom - technical issues
- We will do offline - simpler
  - Every group will record a 5-10min presentation, share a link (you can do however it works but we must all be able to access it)
  - Each individual will review 3 at minimum, fill out google form, then each additional one you review will give you an additional 0.5% on the group review project grade portion
  - We will encourage a piazza discussion for each project
  - End of quarter optional additional hour discussion

Project schedule

Task due	Date due	Description
Previous project review	5/23/2023 at 11:59pm (Tuesday)	Select 2 of the 3 available, review as individuals and then come together as a group to submit your responses to the questions after a discussion. This will orient you to the class project
Project proposal	5/26/2023 at 11:59pm (Friday wk8)	Generate your question, hypothesis, initial data sets you'll be working with, etc., describe your plan, schedule, who is doing what, potential issues, suggested analysis and how it will answer your question
Data checkpoint	6/2/2023 at 11:59pm (Friday wk9)	Builds on the proposal by taking the feedback from PP above and actually getting, loading, describing your data,
EDA checkpoint	*6/10/2023 at 11:59pm (Saturday wk10)*	*Builds on the previous checkpoint, essentially most of your analysis should be done by this point*
Final report	6/15/2023 at 11:59pm (Thursday Fin wk)	Due Thursday of finals week so we can grade before the Tuesday deadline, otherwise your grade may be delayed
Group evaluations	6/15/2023 at 11:59pm (Thursday Fin wk)	You will evaluate each other based on participation and performance, this will contribute to your overall final project grade 5%)

EDA checkpoint

- Link to EDA checkpoint:
  - https://github.com/drsimpkins-teaching/cogs138/blob/main/main_project/EDACheckpoint_groupXXX.ipynb
- One additional question to add - what do you think given the exploration you have done that your biggest challenges are and how will you address them?
- Link to outline of what to include:
  - https://github.com/drsimpkins-teaching/cogs138/tree/main/main_project

Group issues?

- Communicate with us for assistance working things out
- Group strategies, clear communication to avoid misunderstandings, regular updates, be open with ideas (no shooting down approach)

Remaining assignments schedule

- A5 wk10, A4 extra credit
- Lecture quizzes - will be released and you complete by the end of finals week
- Final course survey
- Otherwise just project

A4 - getting pysurfer working is a task…

- Working with UCSD IT on the dependencies since the user install version does not appear to work well - complex paths to update
- This will be an optional assignment or for your edification

A5 - Mouse v. Human cells

- Less dependent on complex dependencies
- https://allensdk.readthedocs.io/en/latest/install.html
- installation
- path

Last time…

Refining the hypotheses

A hypothesis should be

- Narrow
- Very specific
- ***Not*** include a conclusion or interpretation
- Consist of a research and null hypothesis
- Remember we are trying to reject or fail to reject the null, which basically says we either
  - ***'didn't find anything' or***
  - ***'failed to not find anything'***

Developing a hypothesis - overview readings to review

- https://www.scribbr.com/statistics/hypothesis-testing/
- https://opentext.wsu.edu/carriecuttler/chapter/developing-a-hypothesis/#:~:text=A%20researcher%20begins%20with%20a,prediction%20is%20called%20a%20hypothesis.
- https://www.skillsyouneed.com/num/hypotheses-testing.html
- https://www.nedarc.org/statisticalhelp/advancedstatisticaltopics/hypothesisTesting.html
- https://www.youtube.com/watch?v=joNb67F1UbY

Hypothesis : Simplicity, narrowness

- KISS principle
- Boiled down to the essence of the relationship you are testing
- Research/Alternative and Null are opposites

Hypothesis testing

-Cannot prove hypothesis

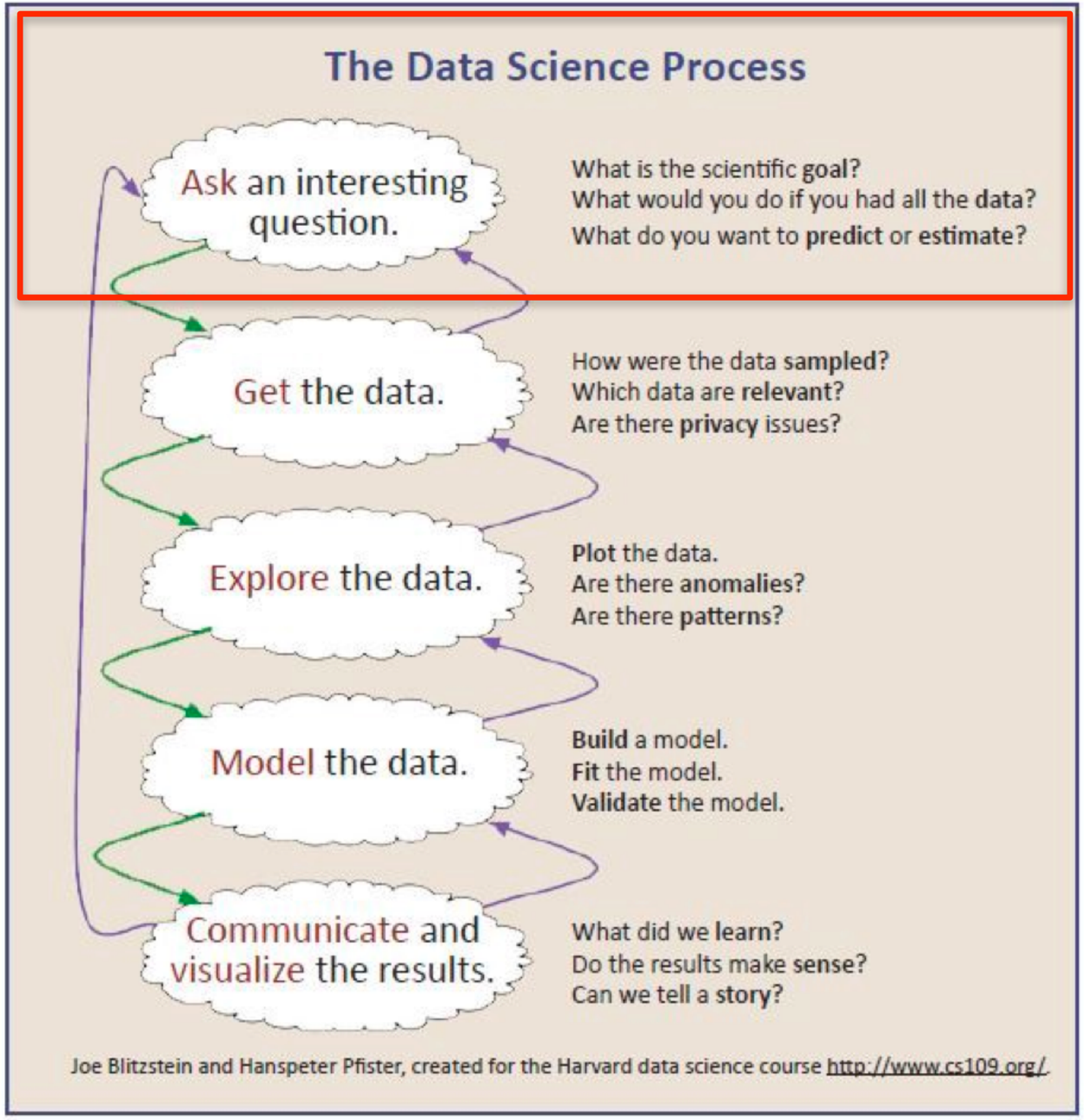
-Can only reject or fail to reject null hypothesis

-Why?

Data Science questions should...

- Be specific
- Be answerable with data
- Specify what's being measured

What makes a question a good question?


The Data Science Process
Ask an interesting question.
What is the scientific goal?
What would you do if you had all the data?
What do you want to predict or estimate?
Get the data.
How were the data sampled?
Which data are relevant?
Are there privacy issues?
Explore the data.
Plot the data.
Are there anomalies?
Are there patterns?
Model the data.
Build a model.
Fit the model.
Validate the model.
Communicate and visualize the results.
What did we learn?
Do the results make sense?
Can we tell a story?
Joe Blitzstein and Hanspeter Pfister, created for the Harvard data science course http://www.cs109.org/

Working toward a strong data science question

Vague: How does the brain change when you have a brain injury?

Better: What neurological changes are there after a stroke?

Even better: What neurological and behavioral changes can be measured with EEG and motion capture between an average normal subject and a stroke patient who had a recent stroke that impaired motor function?

Best?

Group A

- Question: What is the relationship between EEG measurements from the Emotiv EEG Neuroheadset and the eye state (open or closed), and can these measurements be used to accurately predict the eye state?

- Hypothesis: It should be possible to, with reasonable accuracy, predict whether a datapoint was recorded with eyes open or eyes closed given the EEG data. Alpha waves manifest when eyes are closed in a relaxed state, and we predict that our classifier will be able to use this to differentiate between the two states.

Group A

- Data:

Hypothesis prompt

- Include your team's hypothesis
- Ensure that this hypothesis is clear to readers
- Explain why you think this will be the outcome (what was your thinking?)

- What is your main hypothesis/predictions about what the answer to your question is? Briefly explain your thinking. (2-3 sentences)
- Include the refined 1-3 sentence research and null hypothesis statement below the above description

Group B

- Question: Are there specific genetic risk factors and electrophysiological signatures for Autistic Spectrum Disorders which could be correlated with the behavioral abnomalities observed in children with ASD?

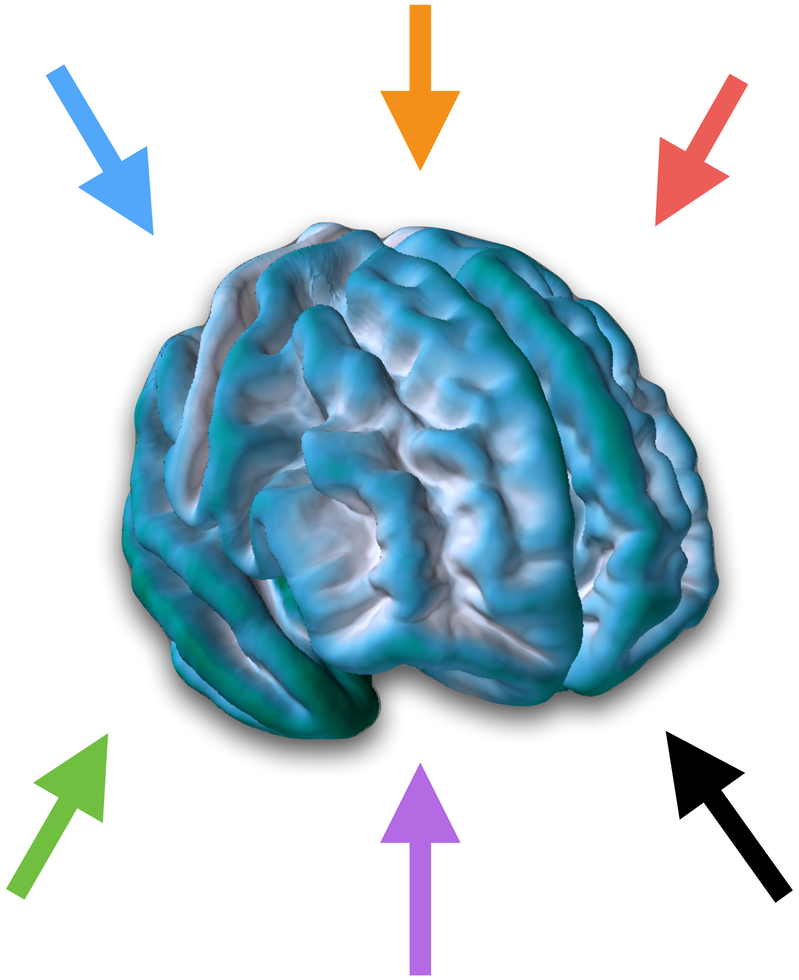
- Hypothesis: Our hypothesis is : there are several genetic risk factors that can be targeted which has impact on the dysfunctioning of lower visual regions which explains the abnormalities in visual detection tasks found in behavioral studies and the unique electrophysiological signatures measured by EEG techinque. We think it is true because previous studies have found that patients with ASD demonstrate genetic variations as well as sensory abnormalties which we think should be interconnected.

On to today…

Practical challenges in neural data science

Neural Data Science

- New way of putting together disparate methods
- Integrate many perspectives to build a better picture of brain, behavior, cognition
- We have explored many different approaches, discussed their integration, discussed how to think in all the ways you need to in order to implement techniques in a single study/groups of studies



What are the biggest challenges of making neural data science work?

What are the biggest challenges of making neural data science work?

- Many different libraries, many different dependencies
  - Have to get good at or build a team with the skill set to make complicated practical analyses, measurements possible
  - New attitude
- Many different modalities means you need to be aware of the issues associated with each in order to avoid spurious conclusions
  - Need skills or to know how to develop the skills for each - where do you look, how to ask questions (technical and content)

What are the biggest challenges of making neural data science work?

- A lot of data requires strategies to deal with it - you can’t just load terabytes into RAM (at this point)
- Changing landscape of what is available
- Increased heterogeneity of teams, and interdisciplinary has challenges as well

What are the biggest challenges of making neural data science work?

- How do you learn about, for example motion capture? Or how do you learn about eye tracking?
- How do you review all the literature?
    - Reading unfamiliar literature?
- Where do you get the data?
- How do you know if the data is good?

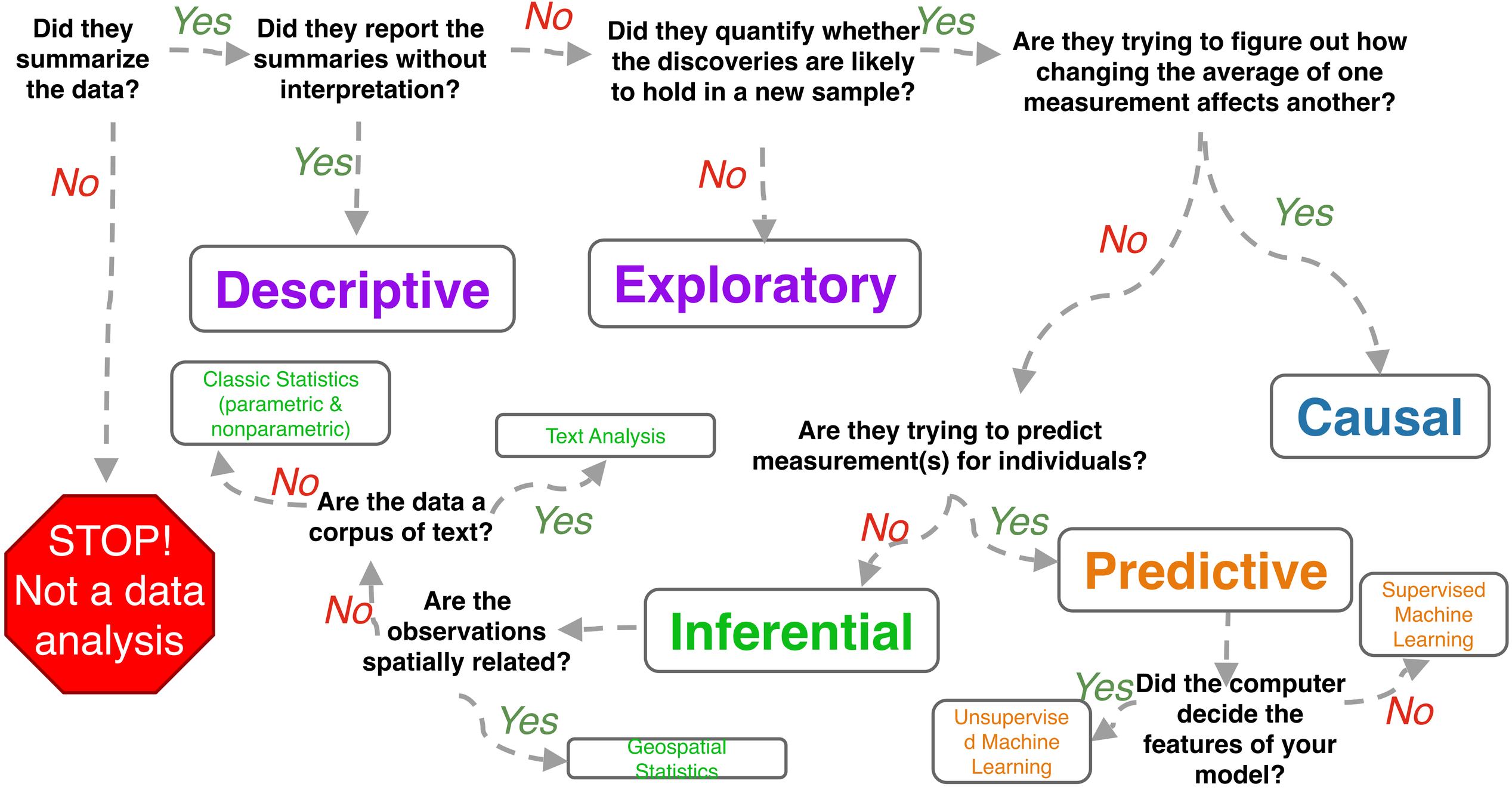
What are the biggest challenges of making neural data science work?

- Free and open does not mean good, easy to use, understand, or relevant
    - Need to use judgment, sometimes ask questions, consider peer reviewed, should be well documented
- Integration of different datasets that may not have been meant to be combined
    - We discussed integration of heterogeneous sets
- Reducing dimensionality
    - Care not to cut out important information early
    - Sometimes reducing data can lead to focusing on richer content
    - Other times the rich content is not at all obvious and we must operate on it all to determine that

EDA review/discussion

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


Did they summarize the data?
Yes
Did they report the summaries without interpretation?
No
Did they quantify whether the discoveries are likely to hold in a new sample?
Yes
Are they trying to figure out how changing the average of one measurement affects another?
No
Yes
Descriptive
No
Exploratory
No
Yes
Causal
Classic Statistics (parametric & nonparametric)
No
Text Analysis
Are the data a corpus of text?
Yes
Are they trying to predict measurement(s) for individuals?
STOP! Not a data analysis
No
Yes
Predictive
No
Are the observations spatially related?
Inferential
Supervised Machine Learning
Yes
Yes
Did the computer decide the features of your model?
No
Unsupervised Machine Learning
Yes
Geospatial Statistics

Descriptive: The goal of descriptive analysis is to understand the components of a data set, describe what they are, and explain that description to others who might want to understand the data.

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.

Statistics

*"the science that deals with the **collection, classification, analysis, and interpretation of numerical facts or data**"*

Statistic - "*A quantity computed from a sample*"

Source: dictionary.com

Some of the lectures to review - a pointer

- Lecture 9, 10, 11, etc