EVERYDAY REPRODUCIBILTY

REPRODUCIBILITY AND ITS GOALS

OUTLINE:

- What is "reproducibility"?
- Goals of reproducibility
- Discussion

REPRODUCIBILITY…

… and other words that start with "R"

- reproducibility
- replicability
- repeatability
- robustness
- rigor

Our focus: computational reproducibility

REPRODUCIBILITY: GOALS

What are we really trying to achieve?

- **Reproducible:** Entire analysis fully and exactly reproducible
- **User friendly:** Easy to access, install, run
- **Transparent:** Easy to inspect, understand, modify
- **Reusable:** Others may build upon the project
- **Version controlled**
- **Archived**

FULLY AND EXACTLY REPRODUCIBLE

- The "most original" data should be available
- Include all code necessary to get from the original data to the final results
- The code should directly produce the plots / tables / numbers in the paper
- All software dependencies should be specified and ideally included with the code
- Random seeds specified
- etc.

USER FRIENDLY

- Code easy to access and inspect, ideally even without downloading
- Should require minimal effort for a user to install and run
- Should cause minimal disruption to a user's resources (e.g., not install unwanted software on their system)
- etc.

TRANSPARENT

- Code should be well-organized and documented, ideally in a notebook format.
- Highlight key analytical choices, such as statistical tuning parameters.
- Use interactive elements (e.g., widgets) to help users explore different options.
- Enable results caching for quick re-runs after minor modifications.
- Provide both raw and cleaned/reformatted data when necessary for clarity.

REUSABLE

- Code should be portable across platforms
- Code should be modular to facilitate re-use in other project
- Depending on the project, creating a new software package may be helpful
- etc.

PERMANENTLY ARCHIVED

- In a (file) format suitable for long-term preservation
- in a (physical) format suitable for long-term preservation

VERSION CONTROLLED

- This aids transparency
- Ultimately, most valuable for **you**

GOALS

1. Exactly reproducible
2. User friendly
3. Transparent
4. Reusable
5. Archived
6. Version controlled

These are distinct goals

They pose distinct challenges

FOCUS: *EVERYDAY REPRODUCIBILITY*

- Most of our goals are readily achievable for "everyday" projects
- The hard(er) challenges:
  - restricted access datasets
  - massive datasets
  - proprietary software
  - highly computationally intensive code

DISCUSSION

- Which reproducibility goals have you attempted on a project? Which do you regularly strive for?
- What challenges have you run into?
- Any interesting / notable experiences?
- What lessons have you learned?
- What tools do you find helpful?

OUR GOALS

We are going to talk about some tools that can aid computational reproducibility.

Reproducibility isn't binary

- You don't need to use all these tools all the time.
- More strategies improve reproducibility, but everyone must decide what's the line.

You can't learn this all in one day

- We're just aiming to dip a toe in.
- **Goal:** Try using one new tool in your next (or current) project.