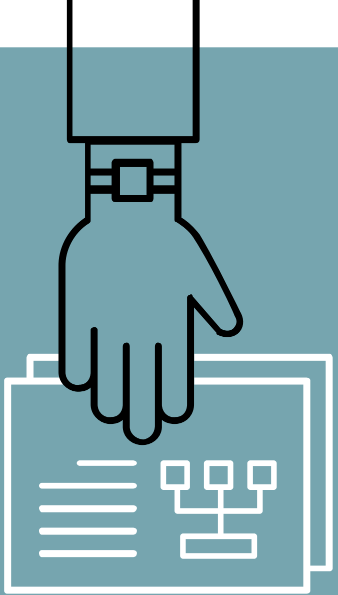
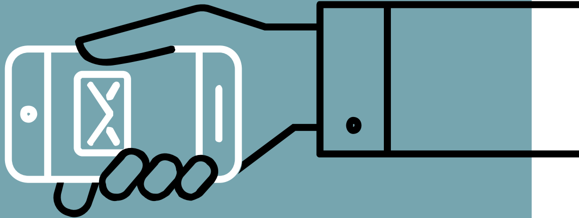


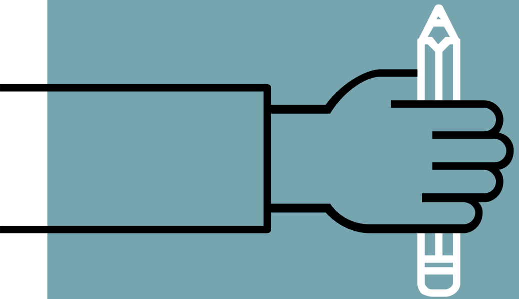


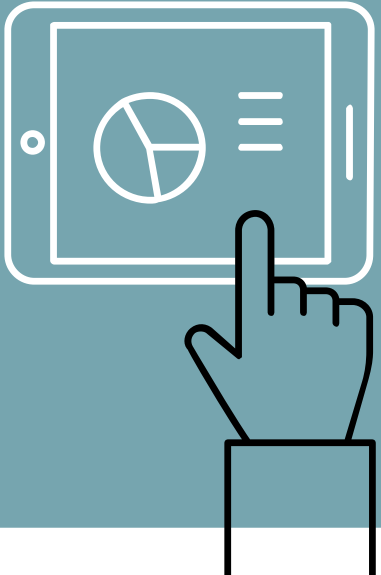
BST 270

Reproducible Data Science

Winter 2023

Session 1





Teaching Staff

Viola Fanfani

Email: vfanfani@hsph.harvard.edu

Office Hour: by appointment

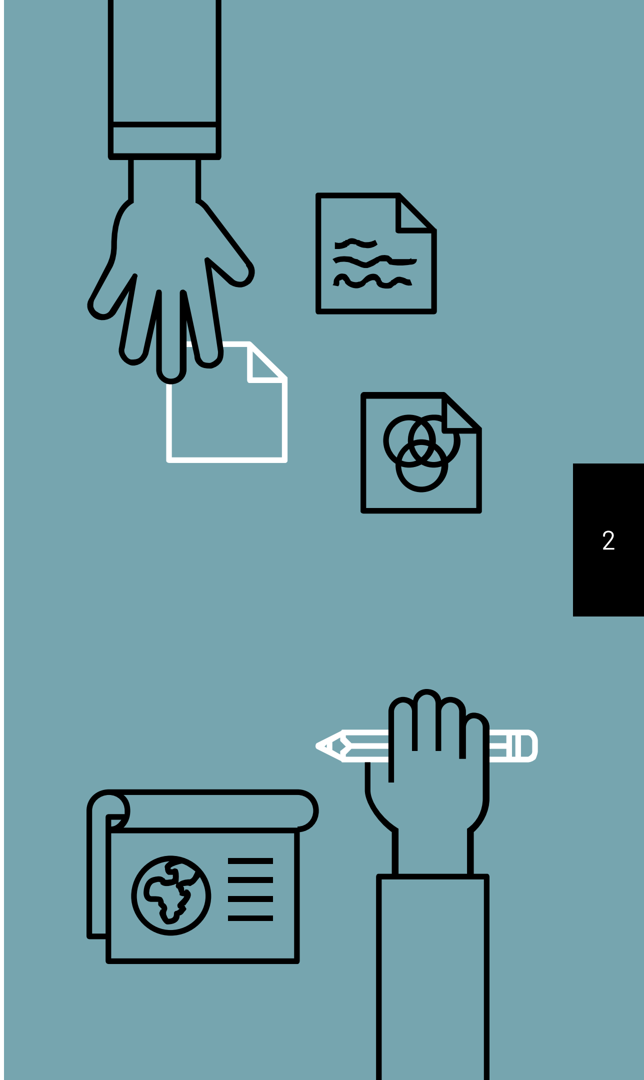
Beau Coker (TF)

Email: beaucoker@g.harvard.edu

Office Hour: by appointment

John Quackenbush
Professor of Computational Biology
and Bioinformatics

Email: johnq@hsph.harvard.edu



Course Details

- January 9-13, 18-20, 9am-12pm
  - We will be “in-class” for the first 1-2 hours
  - The rest of the time will be for watching the day’s videos any time after class and submitting discussion points

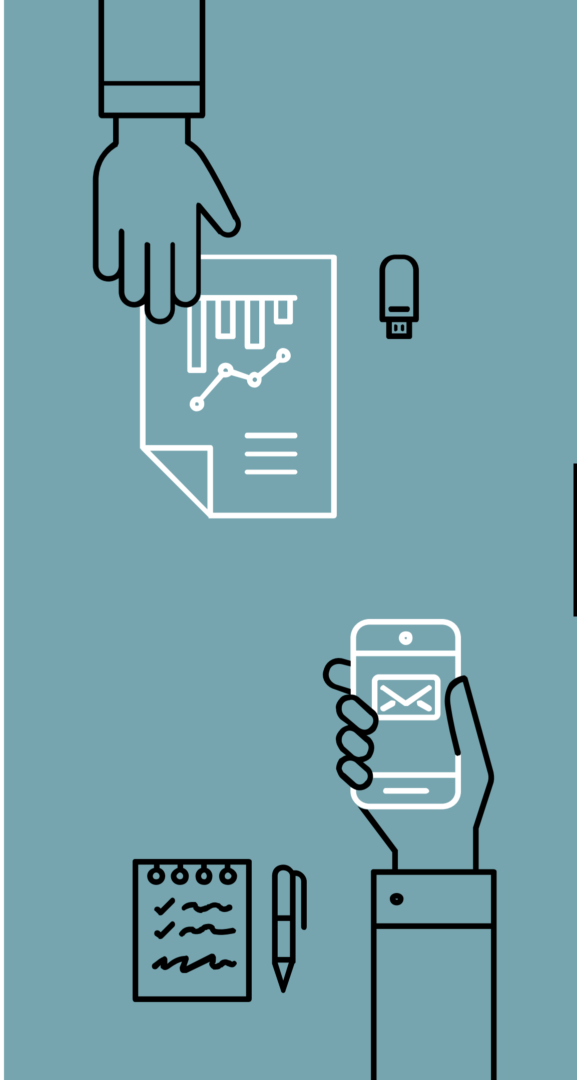
- 2.5 credits (Pass/Fail)
  - A minimum of 70% is needed for a Pass

- Grading
  - 60% homework, in-class participation and attendance
    - Submit discussion points after each day of videos

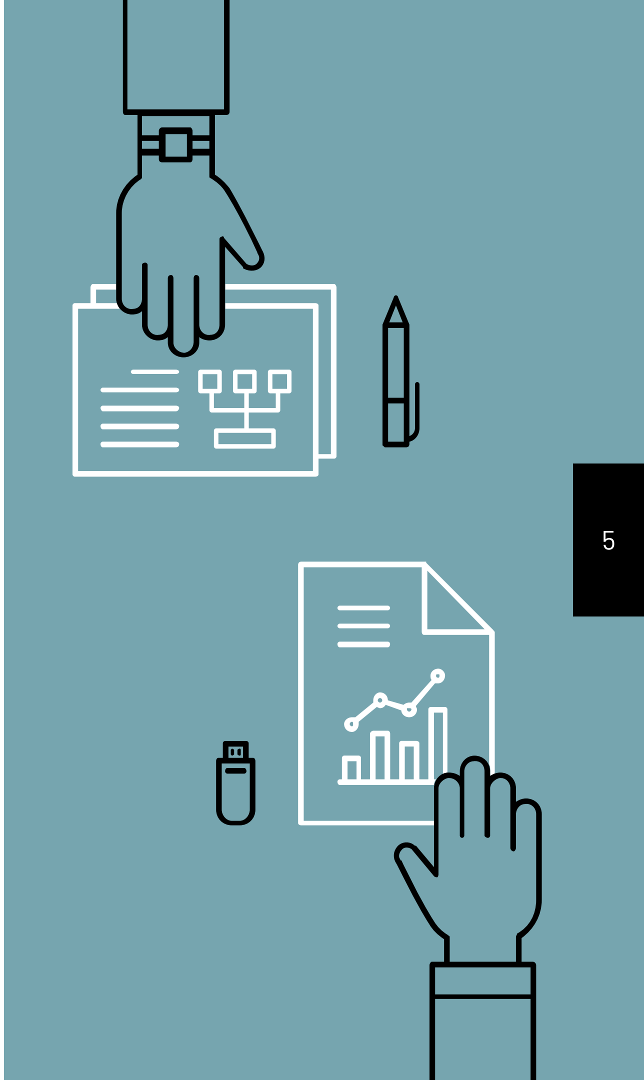
  - 40% individual project
    - Will be assigned next week

Course Details

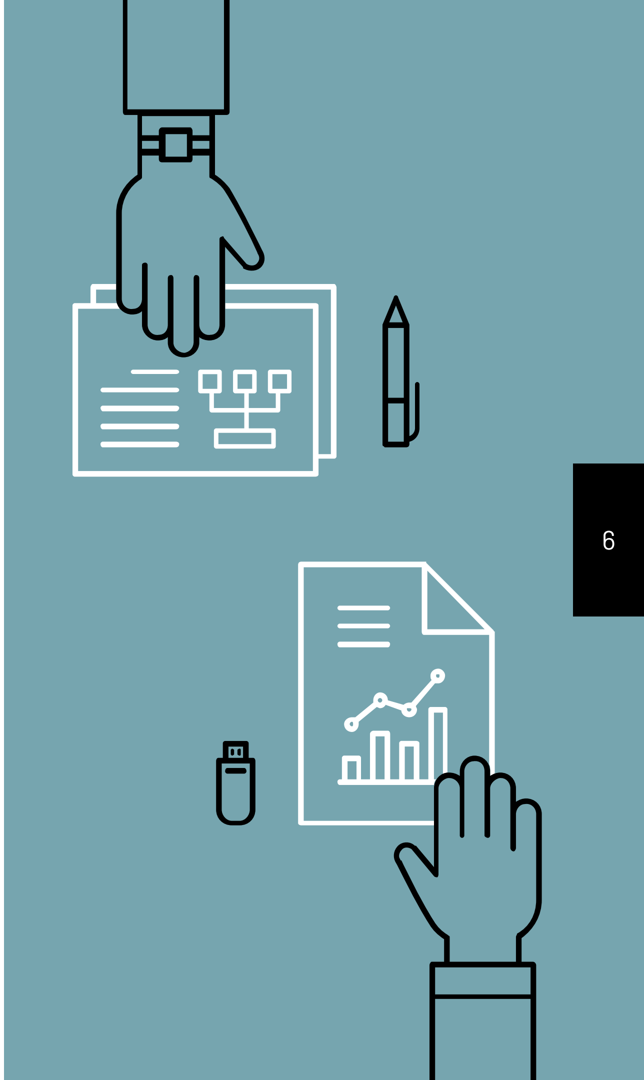
- Course videos
  - Principles, Statistical and Computational Tools for Reproducible Data Science (edX.org)
- Course GitHub
  - BST270-Winter2023
- Optional reading
  - Christopher Gandrud (2015), Reproducible Research with R and RStudio, 2nd Ed.
  - Kitzes, Turek, Deniz (2017), The Practice of Reproducible Research: Case Studies and Lessons from the Data-Intensive Sciences, 1st Ed.



What does *“reproducible data science”* mean to you?



What does *“replicable data science”* mean to you? Is it the same as *“reproducible”*?



Terminology

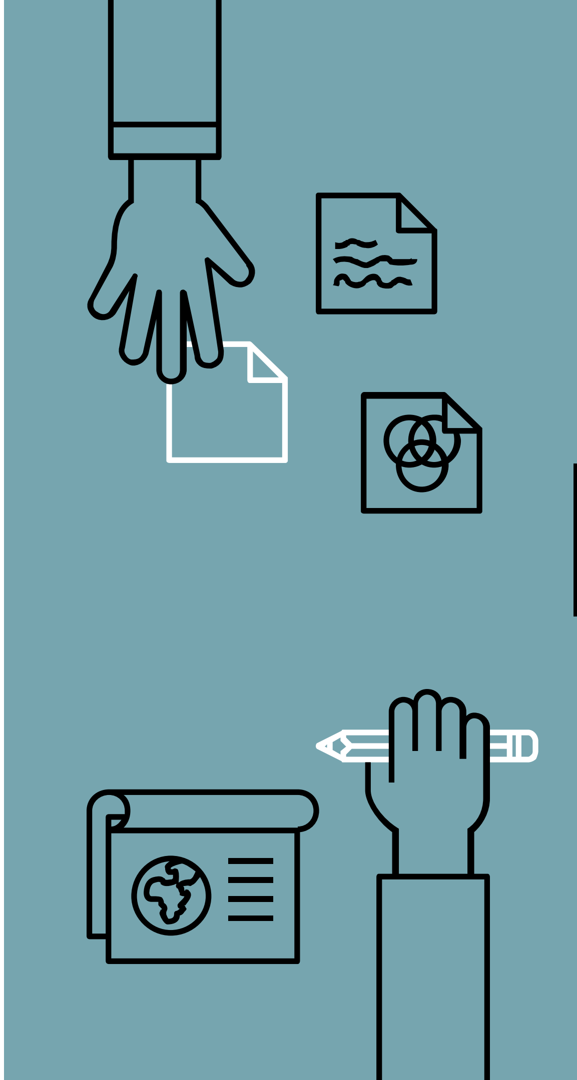
- **Reproducible**: A range of best practices for quantitative research including management and sharing of data and computational methods. More formally, an experiment can be considered "reproducible" if a different research team can obtain its input data and computational tools, and rerun the same methods to obtain the same result.

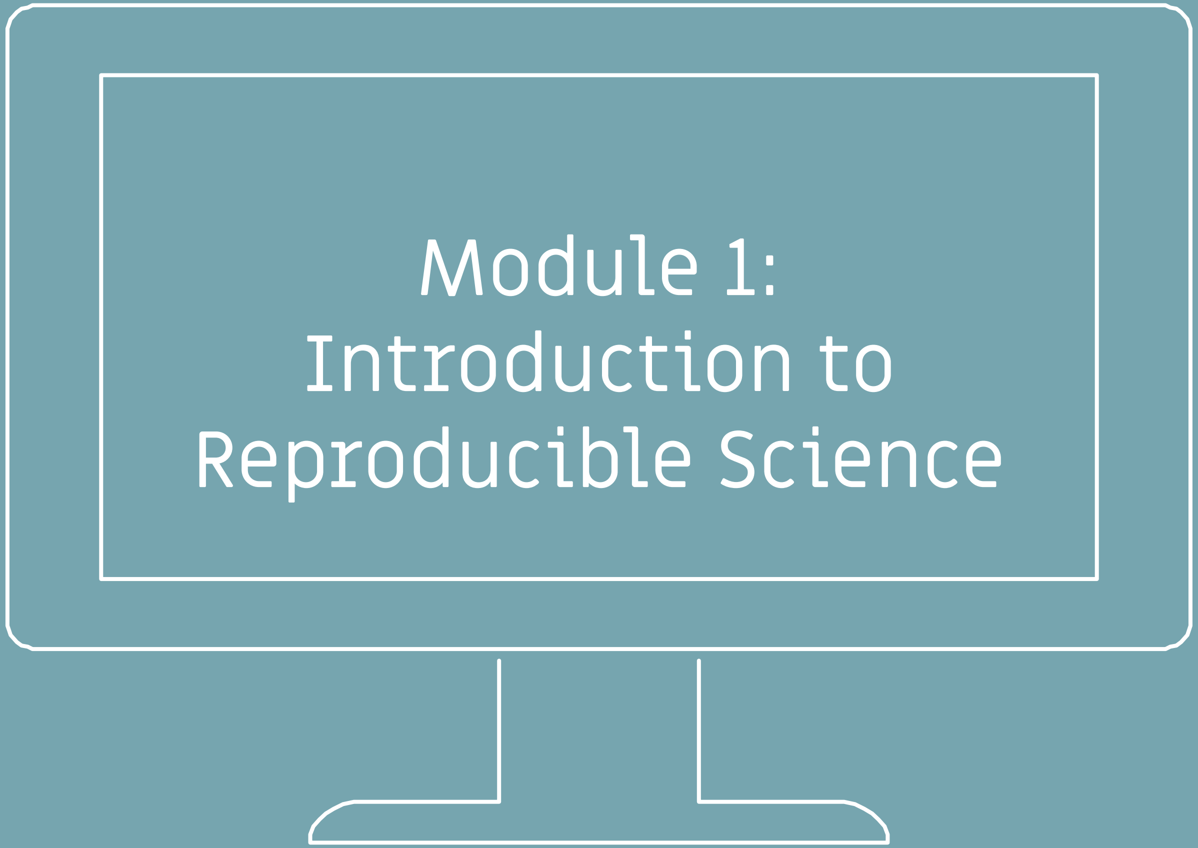
- **Replicable**: A prior study being duplicated using the same procedures or concept, but with new data.

Why is reproducibility important?

- The scientific method requires that work can be falsified or verified by others
  - If something new is discovered, it should be true at any point in time and in anybody's hands

- Your own analysis requires reproducibility
  - Checking for errors
  - Preparing a manuscript
  - Having to tweak something after manuscript reviewers propose changes or have questions
  - Having someone else use your work for future work

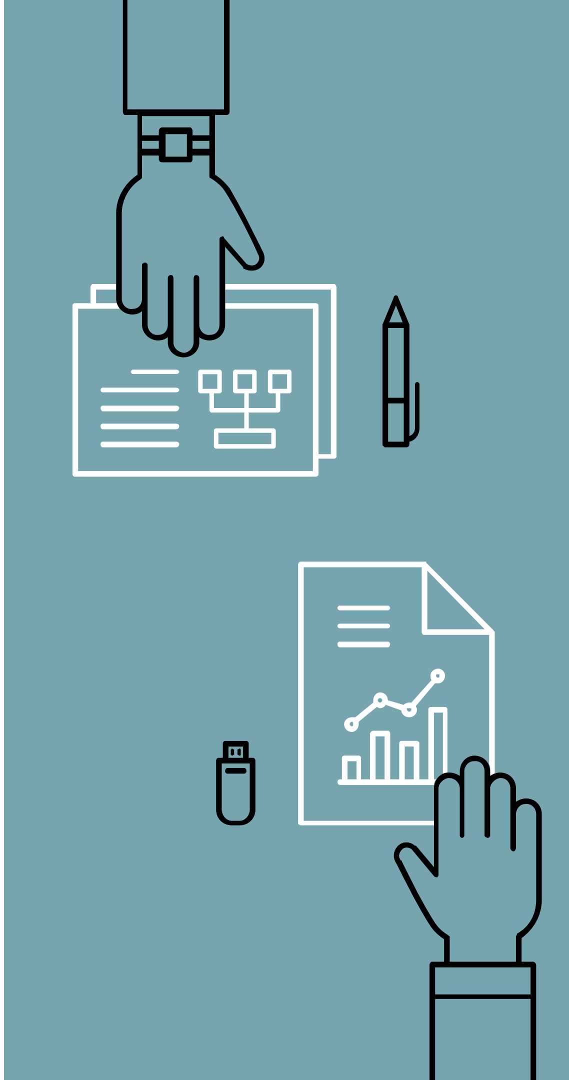




Module 1: Introduction to Reproducible Science

Videos

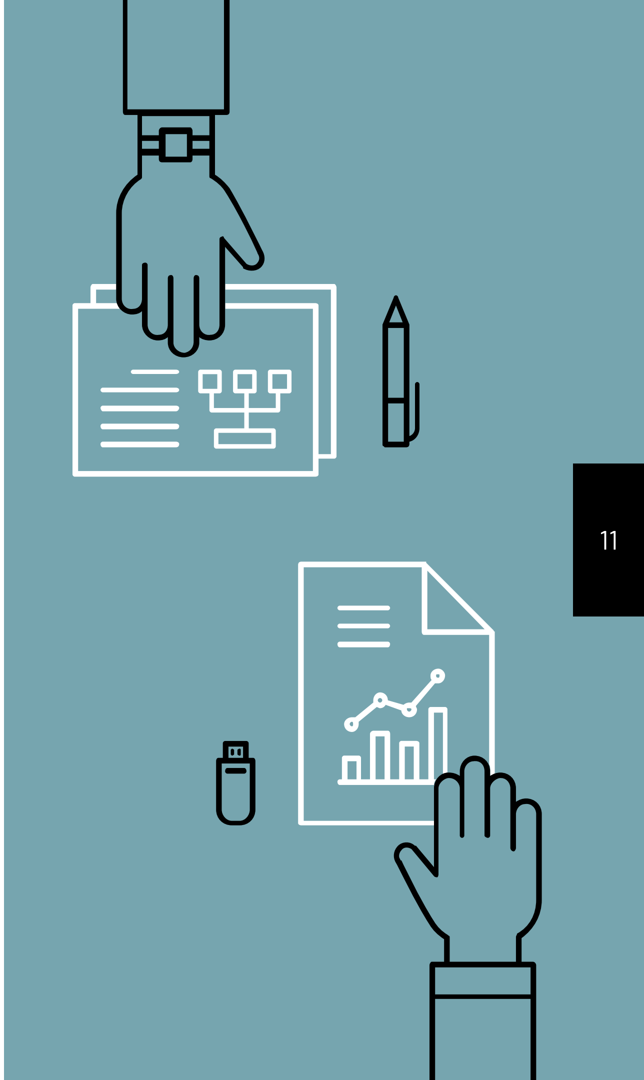
- **Intro to reproducible science**
- Module 2: **Fundamentals of reproducible science** (concepts and language)
- Module 3 : **Case Studies** (good and bad examples)
- Module 4: **Data Provenance** (reporting requirements, best practices recording data analysis operations, recording and communicating data manipulation)
- Module 5: **Statistical Methods** (study design, +/- controls, FPR/FNR, cross-validation)
- Module 6: **computational tools for reproducible science** (tools for data annotation and tracking, workflows, literate programming, documentation, Python/R)

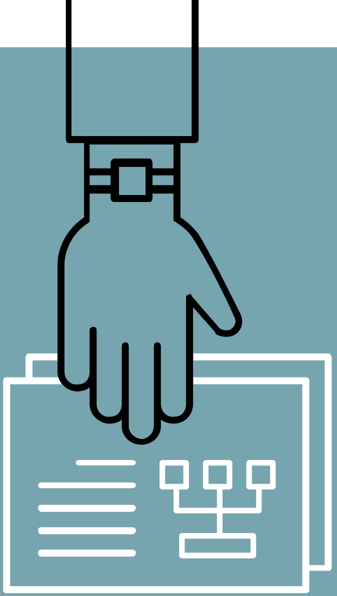


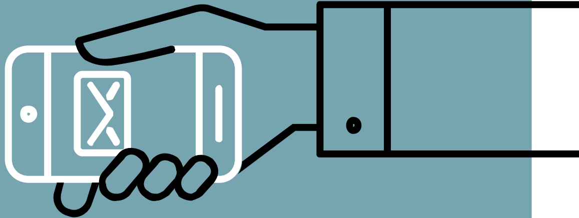
Module 1 Videos

- 1.1 Welcome to reproducible science
- 1.2 Intro to the teaching team
- 1.3 Intro to the modules

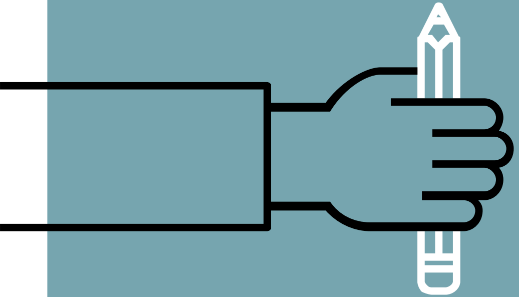
Discussion points: https://forms.gle/qS4WbREVKuYpVrzA6

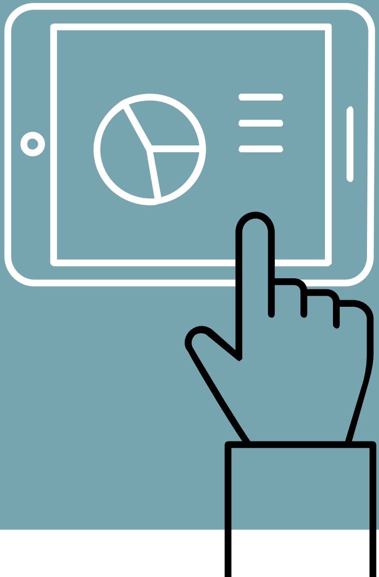






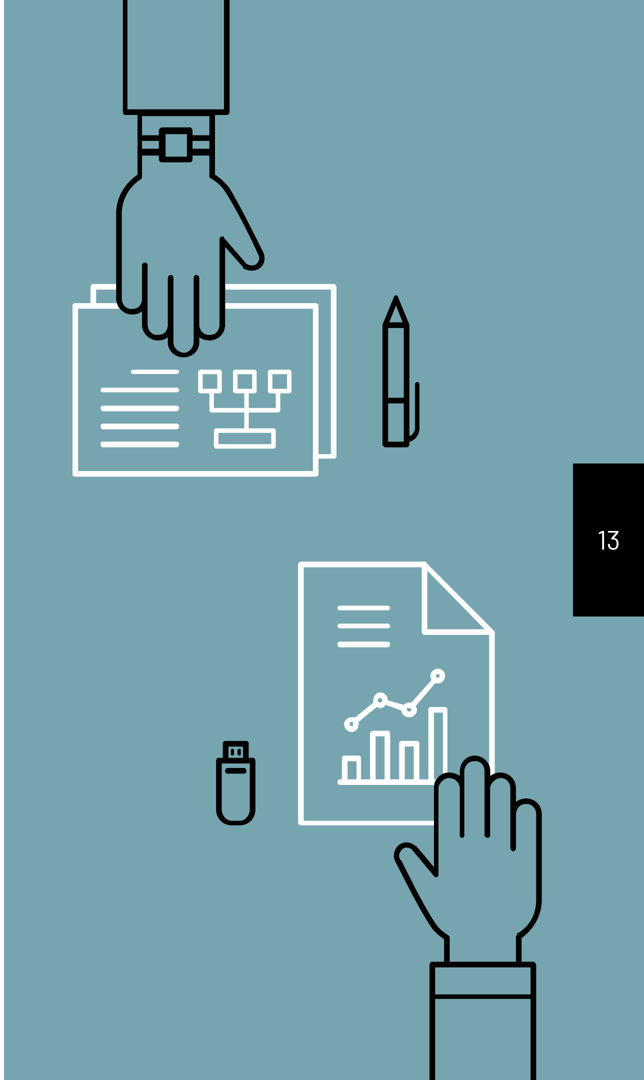
Tools for Reproducible Research

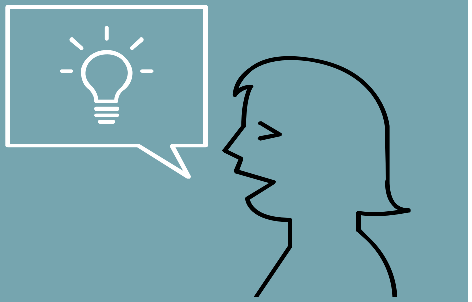




Tools

- RMarkdown
- Jupyter Notebook
- git/GitHub
  - Tutorials
  - Happy Git and GitHub for the useR
- Overleaf





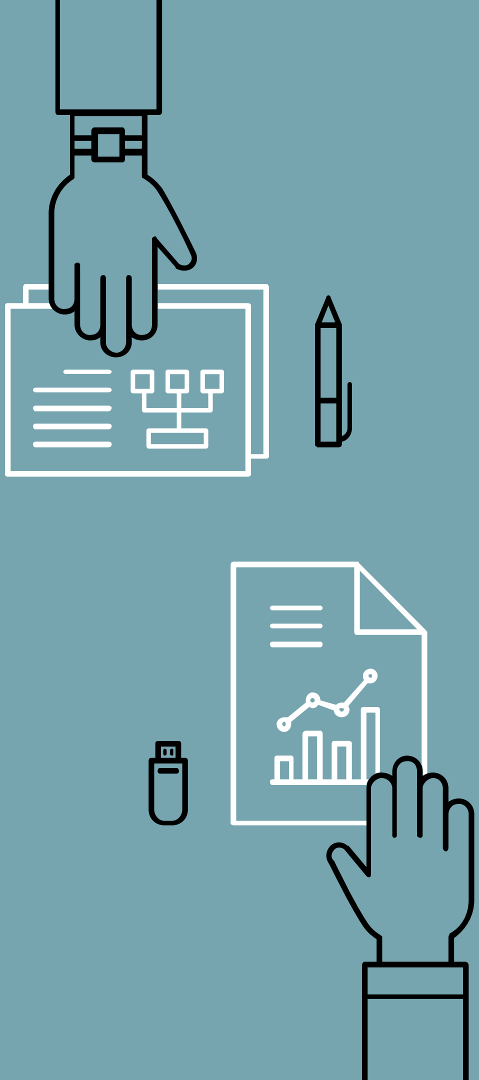
In-Class Project

We will attempt to reproduce the results from and critique the reproducibility of:

[1] Boehm, J. K., Williams, D. R., Rimm, E. B., Ryff, C., \& Kubzansky, L. D. (2013). Relation between Optimism and Lipids in Midlife. The American Journal of Cardiology, 111(10), 1425-1431.

Specific Tasks:

- Create a data dictionary
- Wrangle data
- Recreate Figure 1
- Recreate Tables 1-5
- Critique reproducibility

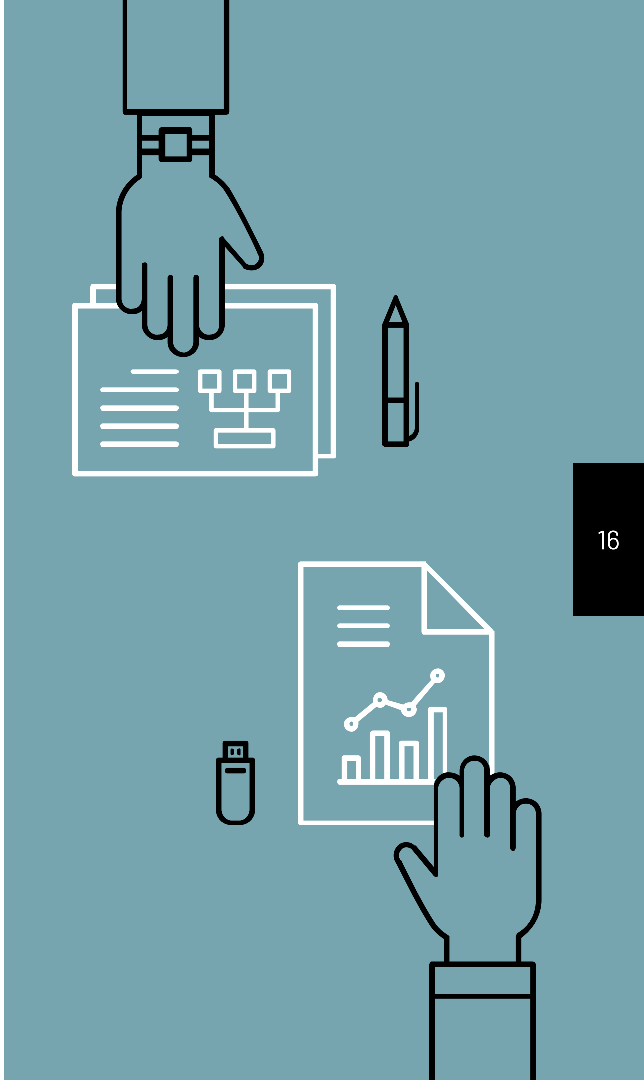


MIDUS II Data Sets

1. Data and supporting codebook and other documents

2. Biomarker data

This particular article focuses only on MIDUS II data, including biomarker data, and investigates the relationship between optimism and lipids. You can download the data in multiple formats. We will be using the **R files in class** and performing all data cleaning and analyses in R and an RMarkdown file.



Homework

- Create GitHub account (if you don’t already have one)
  - Send Viola your GitHub username
- Clone class repository
- Read [1]
- Download data to reproduce [1]
- Watch Module 2 videos
- Submit Module 2 discussion points