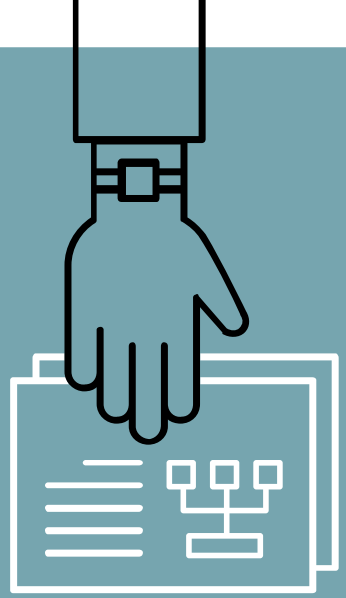
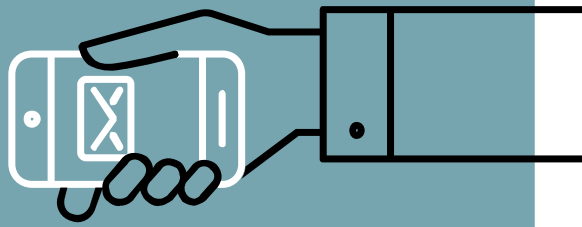
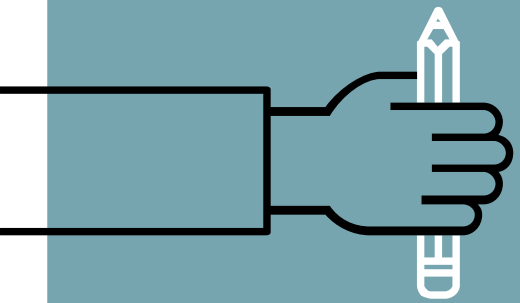


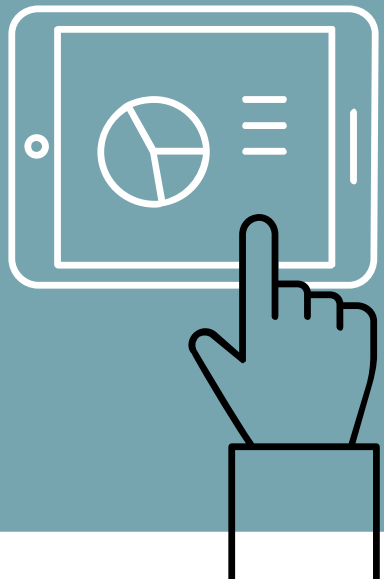


# BST 270

## Reproducible Data Science

Spring 2020





# Teaching Staff


Heather Mattie

Instructor of Data Science

Email: hemattie@hsph.harvard.edu

Office Hour: Mondays 5:15-6:15pm or by appointment

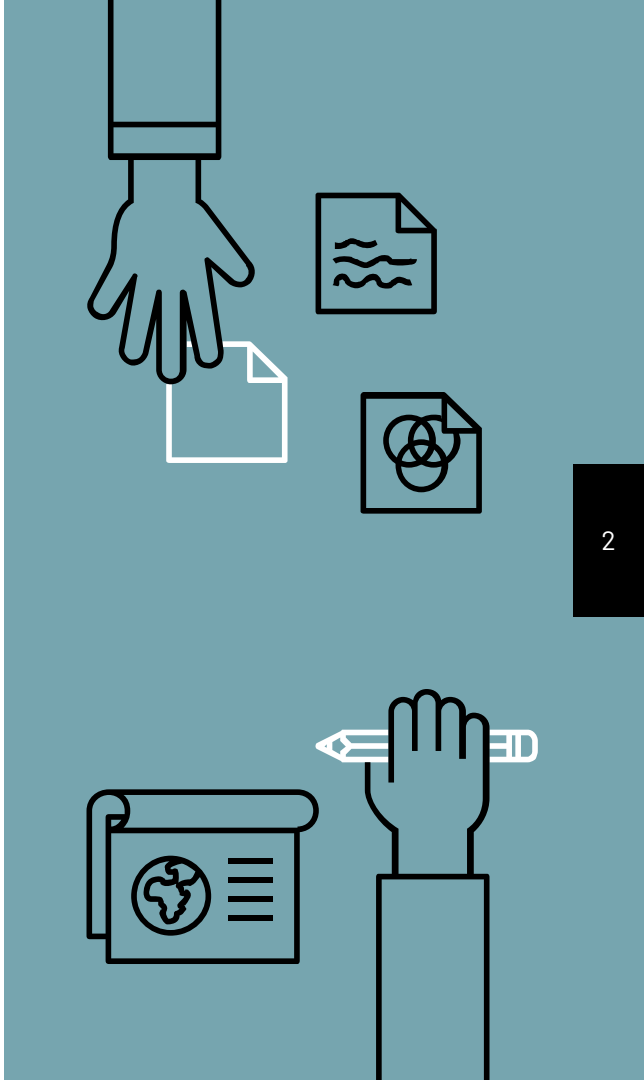
Office: Building 1, 4th floor, room 421A

Aaron Sonabend

PhD Candidate, Biostatistics

Email: asonabend@g.harvard.edu

Office hour: TBD

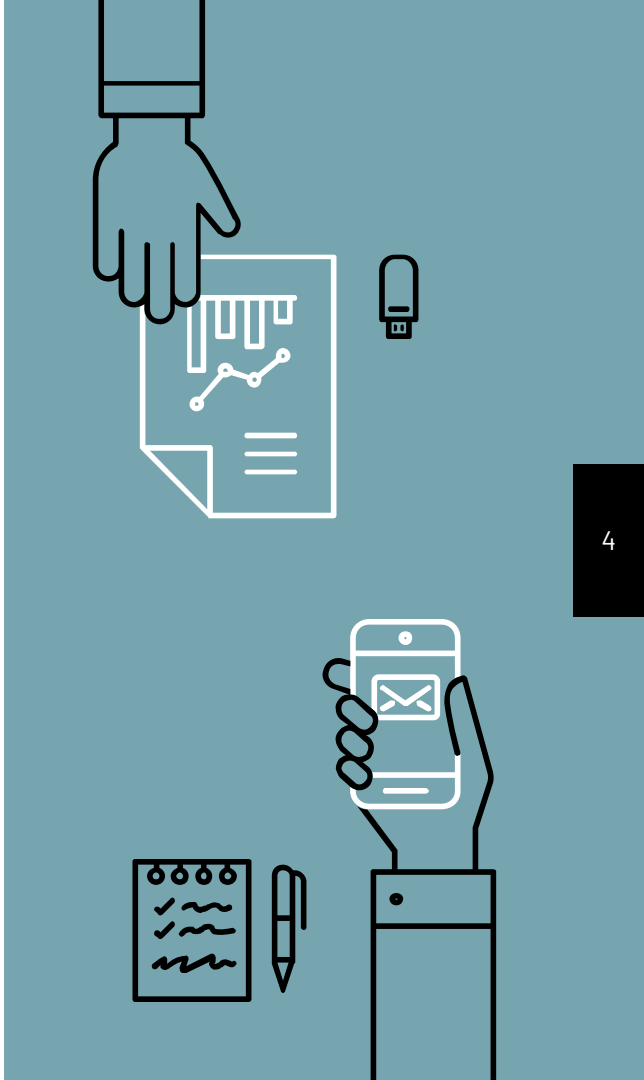



# Course Details

- Mondays 3:45-5:15pm
- Kresge 205
- 2.5 credits (Pass/Fail)
  - A minimum of 70% is needed for a Pass
- Grading
  - 40% homework, in-class participation and attendance
  - 20% case study
  - 40% group project
- Holidays (no class)
  - February 17th
  - March 16th

# Course Details

- Course videos
  - edx.org
- Course GitHub
  - BST270-Spring2020
- 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?

Start the presentation to see live content. Still no live content? Install the app or get help at **PollEv.com/app**

# Are "reproducible" and "replicable" equivalent?

Yes

No

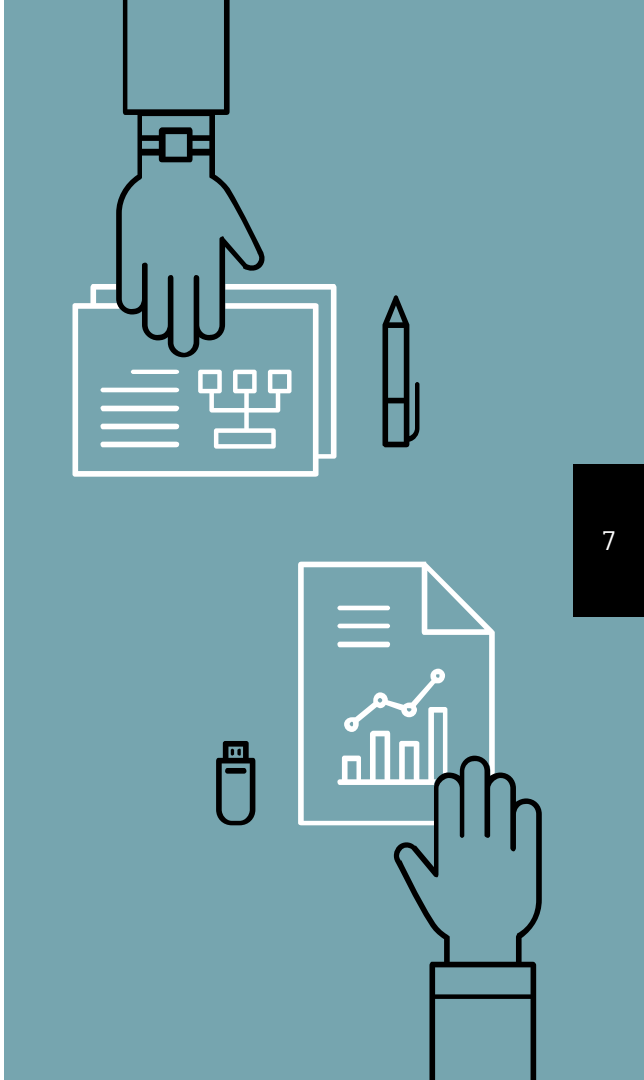
The way you pose the question makes me think yes...

Start the presentation to see live content. Still no live content? Install the app or get help at **PollEv.com/app**

# 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.

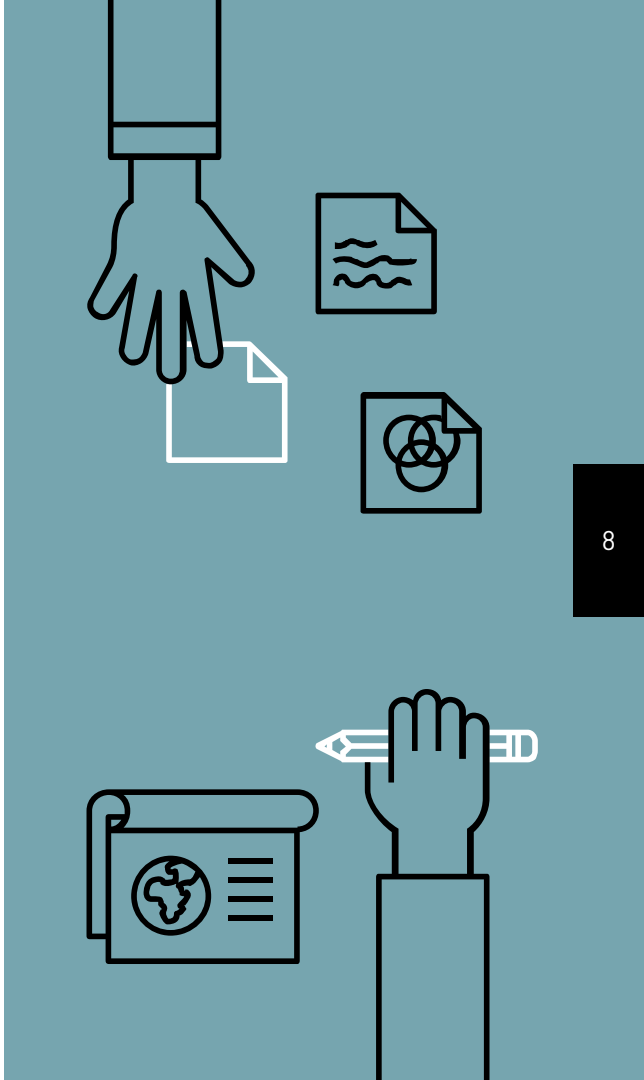
- **Replicable**: A prior study being duplicated using the same procedures or concept, but with new data.



# 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

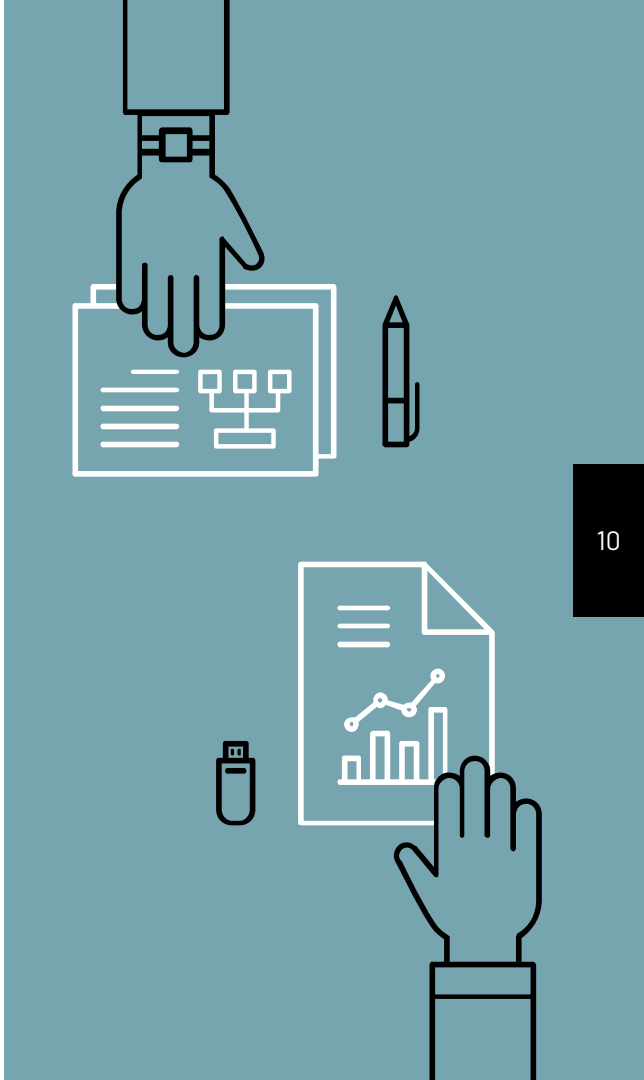
- Your own analysis requires reproducibility
  - Checking for errors
  - Preparing a manuscript
  - Having to tweak something after manuscript reviewers propose changes or have questions
  - Having someone else use your work for future work

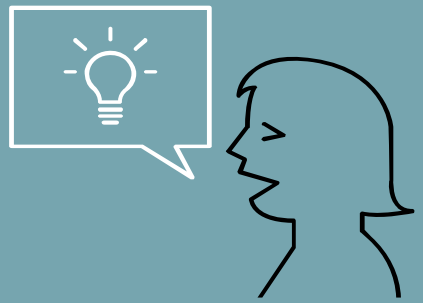


# Module 1: Introduction to Reproducible Science

# Module 1 Videos

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





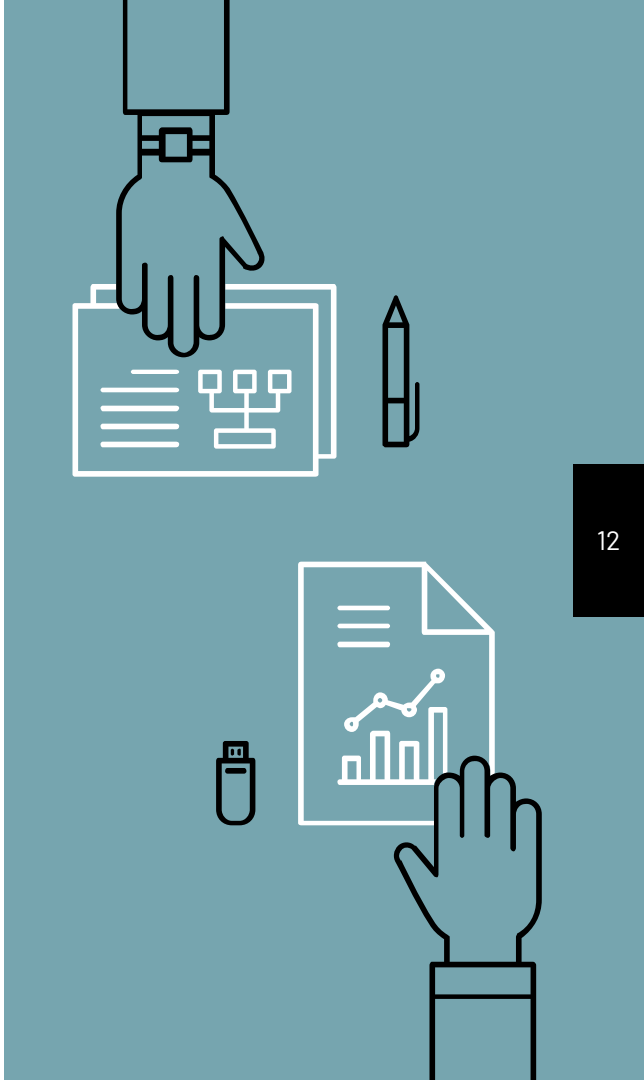
# In-Class Project

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:

- Figure 1
- Tables 1-5
- Critique reproducibility

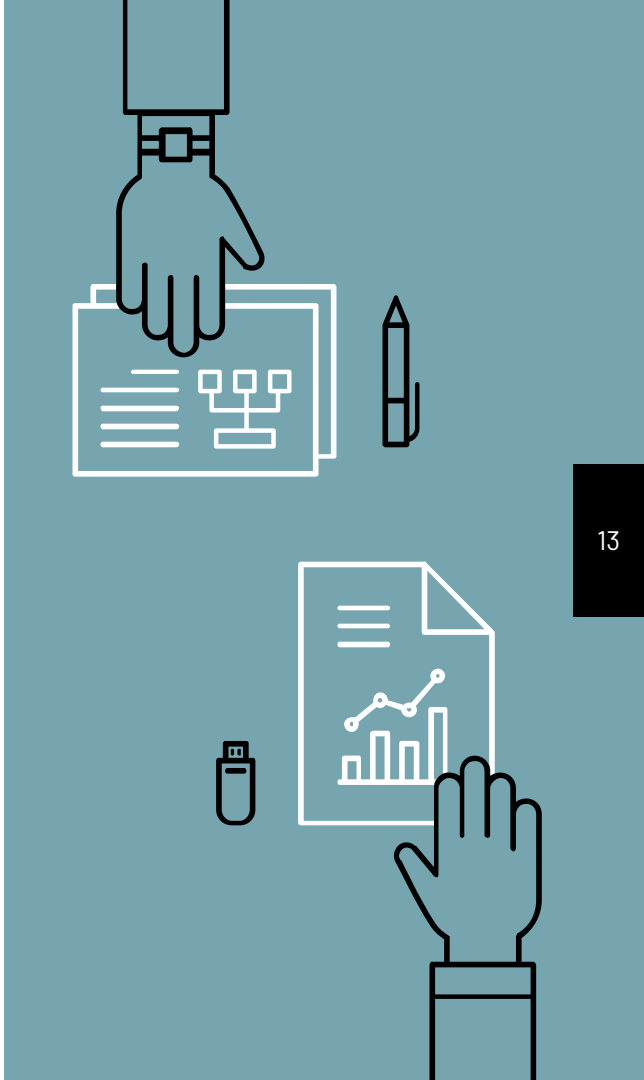


# MIDUS II Data Sets

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
- Clone class repository
- Read [1]
- Watch Module 2 videos