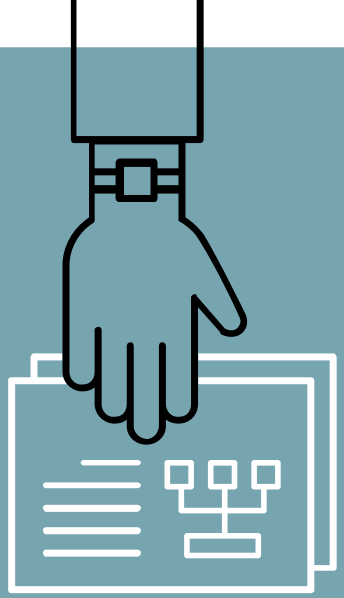
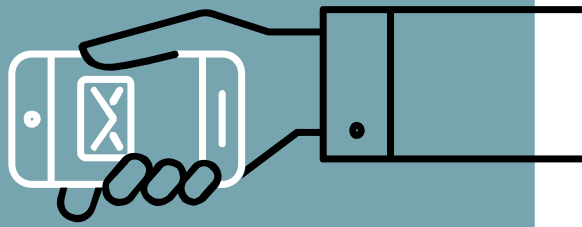
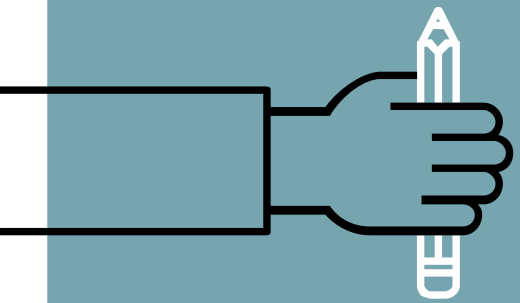
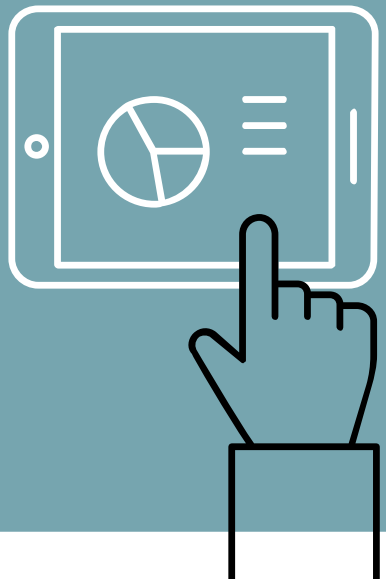




BST 270
Reproducible Data Science

Winter 2023
Session 2



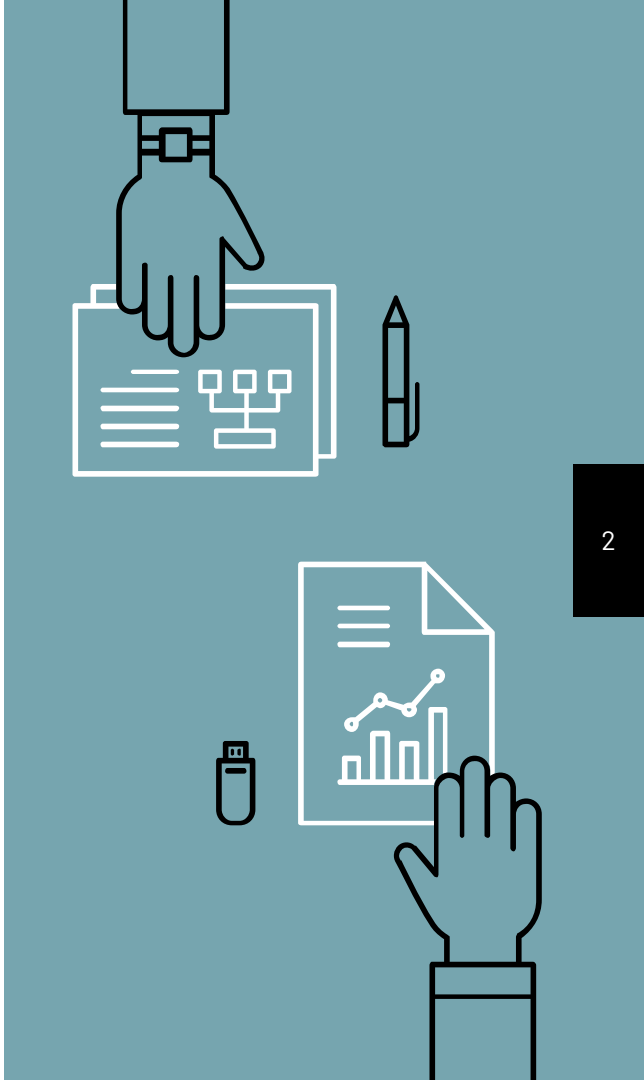


Discussion

- Regarding the definitions of **replicability** in Module 2.2.1, I'm curious if people think that for a different research team to "regenerate their own data" is really necessary for a study to be "replicable" -- e.g. could an analysis conducted regarding a one-time historical event, for which new data cannot be generated, not be considered "replicable"?
  To undermine my own question, I guess **I feel like the the distinction between reproducible and replicable is an important and useful one, but nailing down exactly what the notion of "replicability" does or does not include is probably not all that useful**. The point, as I see it, is not to deem individual studies replicable or not, but rather to develop an understanding of how to assess the extent to which studies ought to inform our own beliefs about the scientific question at hand.

- How to handle situations where the data is not easily publicly available (e.g., EHR data)
- The videos talk about data availability being important for reproducible research. In my lab work, I use a lot of **health sensitive data** that cannot be publicly shared. Is there a best practice for what to do in this case?

Module 4: Data Provenance



Comments

- The module also discussed the availability of metadata. How do you most efficiently record/share relevant metadata for a data source/project? Reviewing the MIDUS, improved brevity would make these files much more decipherable, but all necessary information should be contained (and how does one identify the amount of ”minimally necessary” information?)

What is meta-data?

- Data about data
  - How the data was created
  - How the data was cleaned or transformed
  - Time and date of the creation/collection/transformation
  - Who collected/cleaned/transformed the data
  - Where the data is stored
  - Who has access to the data
  - File size(s)

What to put in the lab notebook?

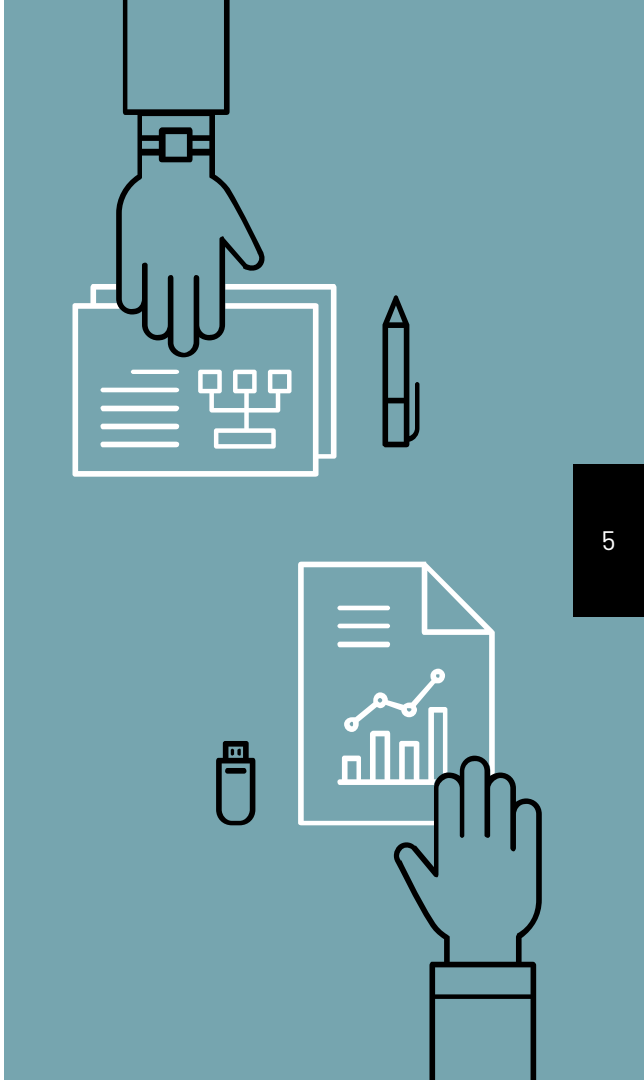
- Code, including comments that help another read and understand what the code is doing
  - State what each variable is, what the inputs and outputs of a function are, the type of object you expect to get after running the code (a list, array, data frame, dictionary, number, etc.)
- Text (not code comments) that guide the reader through your work
  - State the purpose of the project/code, why you are doing this and why you are doing it this way (why are you using that particular method?)

- Visualizations
- Interpretations (of plots, tables, output, etc.)
- If your project requires running several files with code, mention where this file fits in to the workflow
- Ideas for future work
- Any notes from your PI or collaborators

Discussion

- John & Curtis discuss their personal experience and backgrounds as relevant to reproducibility and how these affect their personal/lab workflows for reproducibility, but are there general or accepted **best practices for reproducibility** (or specific steps in the workflow)? And do these differ largely by different research areas (thinkin that workflows can greatly differ, say a large genetic consortium compared to localized or small sample clinical research)

- http://netbooks.networkmedicine.org/
- https://andersenlab.org/dry-guide/2022-03-09/
- http://www.stracquadaniolab.org/docs/



Discussion

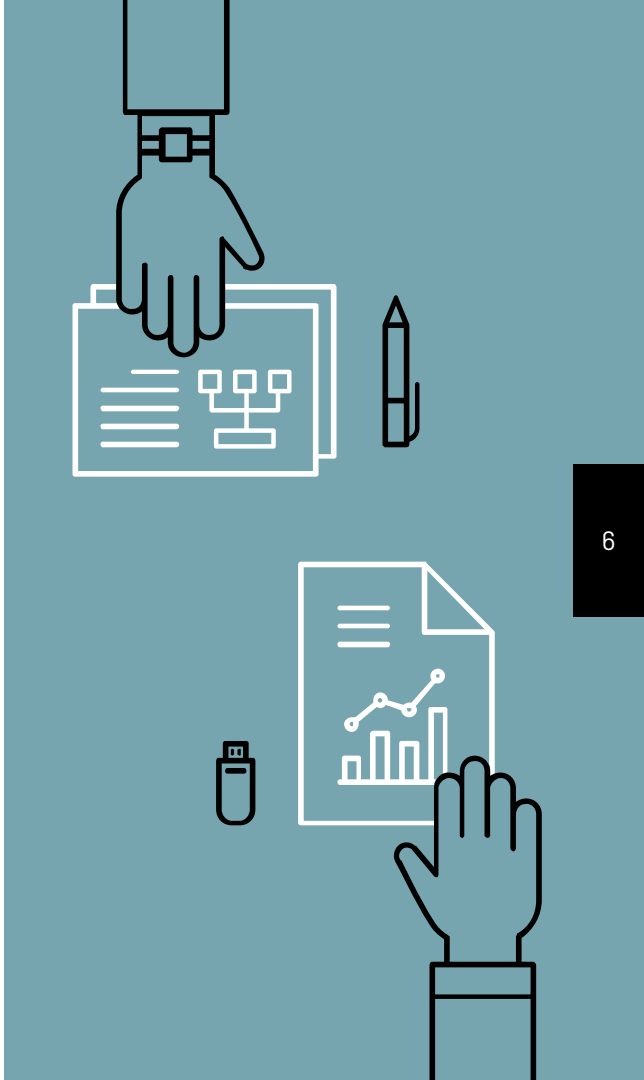
- In video 2.4.5, Dr. Huttenhower talked about making reproducible workflows in an automated way. It wasn't really clear to me from the video how to set this up and what types of software I can use to do this. Can we please discuss in class some approaches to set up an automated workflow to produce our results and rerun them whenever we need to?

- Can we talk more about the automated workflow tools (Doit, Snakemake, Taverna, Kepler, Knitr)? I was unfamiliar with these.

- How to provide reproducible examples when the code can only be run efficiently by taking advantage of parallelized cluster computing?

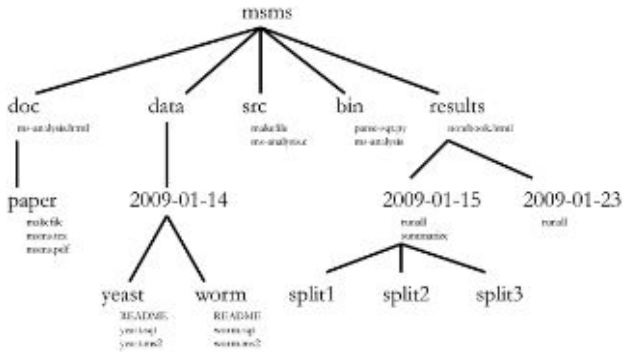
- I found Bill Noble's article on organizing computational biology projects to be quite relevant to how I conduct my research. I plan to try out Noble's file system as a way to keep my research materials organized and accessible for myself and others like collaborators and reviewers.

- I found the paper "A Quick Guide to Organizing Computational Biology Projects" very informative.



Discussion

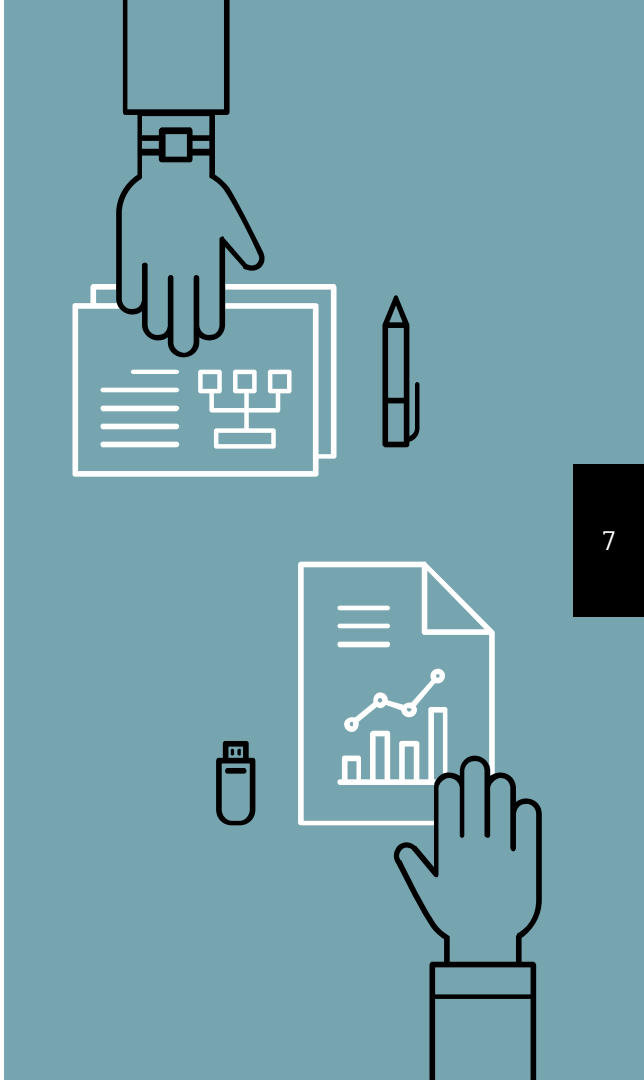
- **In Noble's paper, the author listed the top-level organization, but I'm still unsure of the distinction between source (src) and scripts (bin).**



The source code `src/ms-analysis.c`is compiled to create `bin/ms-analysis`and is documented in `doc/ms-analysis.html`

- **How can you make sure that merging code from two people won't result in conflict assuming they worked on it at the same time?**

- **What's metadata database?**

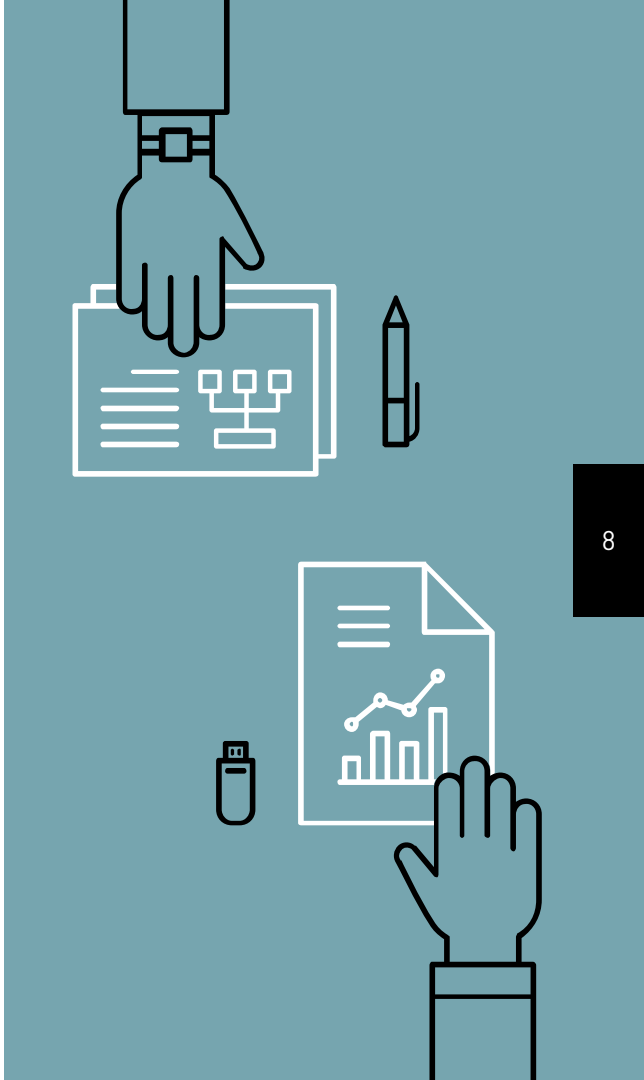


Discussion

- Responding to Module 2.3, I hope that there will eventually be more statistical work on these types of data/model criticism -- **i.e. iteratively "sanity-checking" your analysis as you conduct it -- in a principled way.** This type of iterative checking is good practice in many senses, as the video suggests. However, it also introduces data-dependence into your analysis which is generally unaccounted for in statistical procedures.

- I strongly agree with the statement that one of the key elements to assuring our work is reproducible is assuring that we have good experimental design.

- I found the discussion on including positive and negative controls to be quite interesting. Going forward, I plan to be more intentional about incorporating tests that involve situations where a known result is expected and where no results are expected.

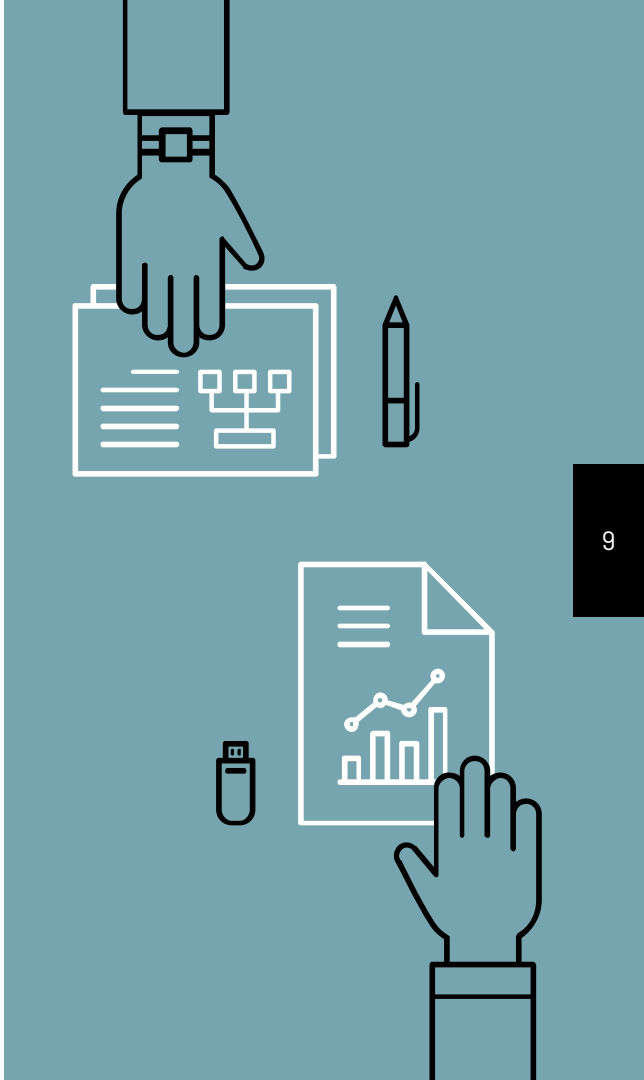


Discussion

- Is there a good example of a biomedical research paper incorporating all of these design suggestions that we can reference for future projects?

  Reproducibility standards for machine learning in the life sciences

- What is the difference between ”reproduce” and ”cross-validate” the result?



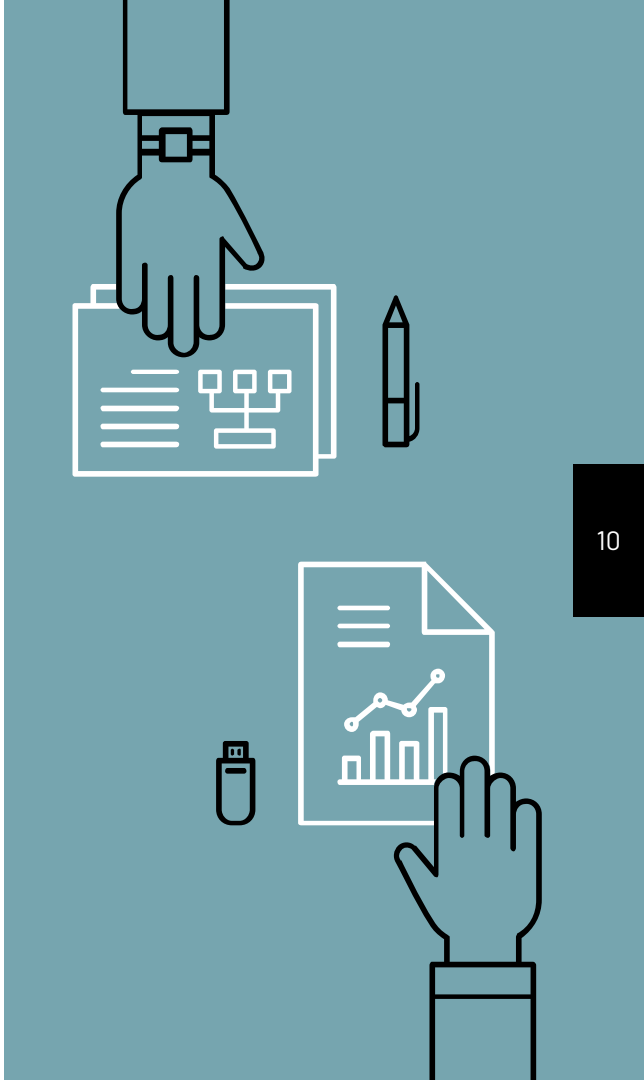
Discussion

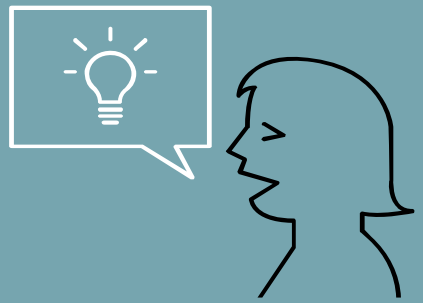
- Responding to Module 2.3, I hope that there will eventually be more statistical work on these types of data/model criticism -- **i.e. iteratively "sanity-checking" your analysis as you conduct it -- in a principled way.** This type of iterative checking is good practice in many senses, as the video suggests. However, it also introduces data-dependence into your analysis which is generally unaccounted for in statistical procedures.

- I strongly agree with the statement that one of the key elements to assuring our work is reproducible is assuring that we have good experimental design.

- I found the discussion on including positive and negative controls to be quite interesting. Going forward, I plan to be more intentional about incorporating tests that involve situations where a known result is expected and where no results are expected.

- When I worked as a software engineer, a lot of attention was put on code reviews: every change in the code was checked by somebody else. In research, this doesn't seem to be done very often. It would be interesting to assess whether translating this practice into an habit in a research environment would be beneficial (easier to find typos, precise comments about what the colleague is doing...) or whether it would slow down the whole process.



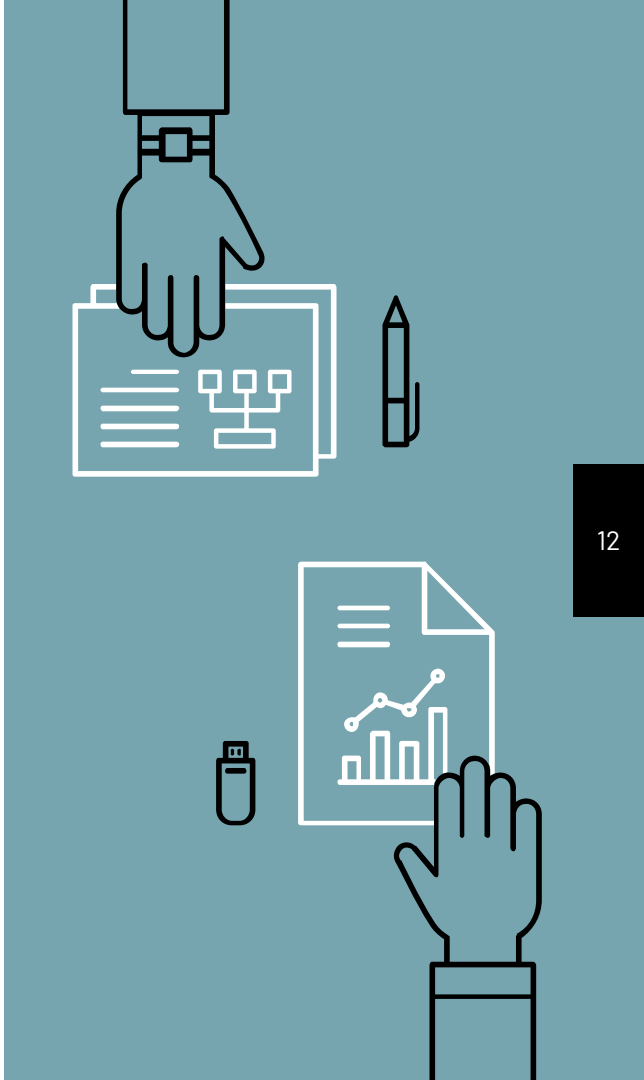


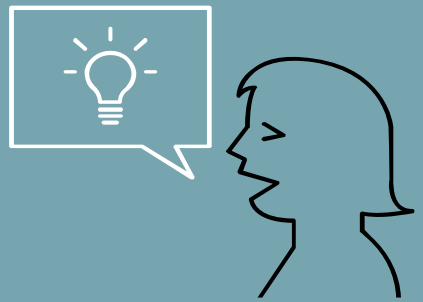
Individual Project

1) Reproduce Covid-19 Figures from the NYT

2) Reproduce 2 tables/figures from a FiveThirtyEight article

Write “reproducible” workflow on how you did it, why, where the data is…





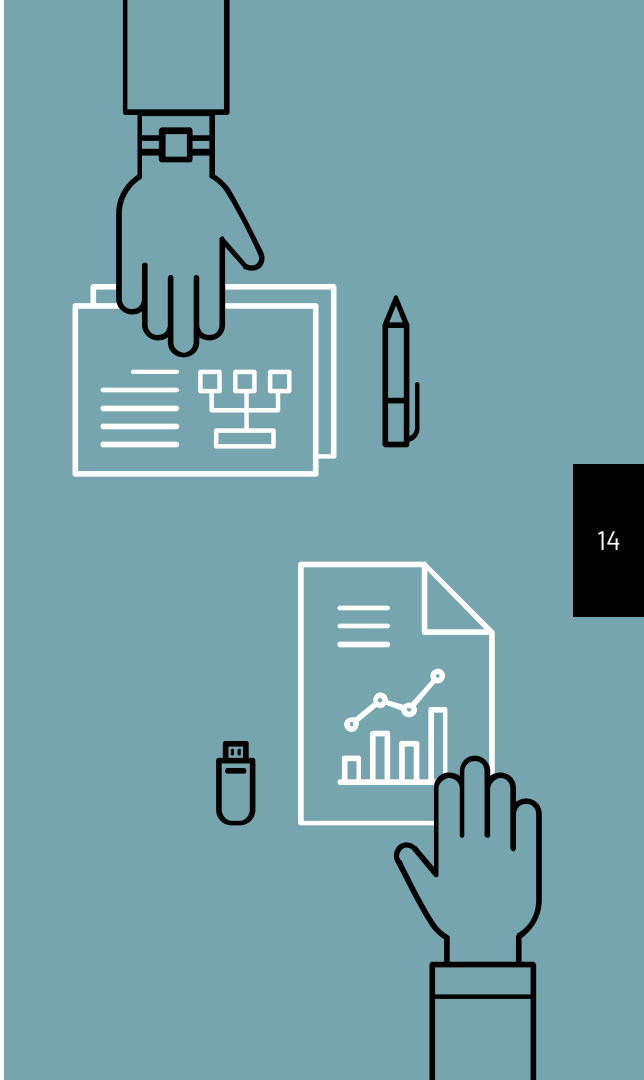
In-Class Project

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

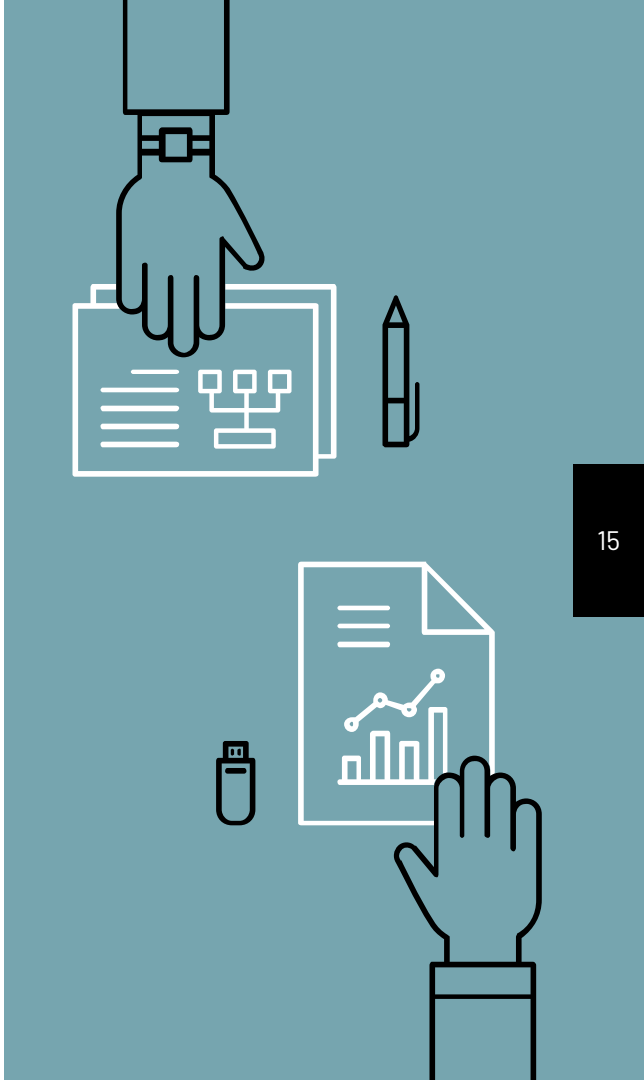


We will attempt to reproduce the results from and critique the reproducibility of:

[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

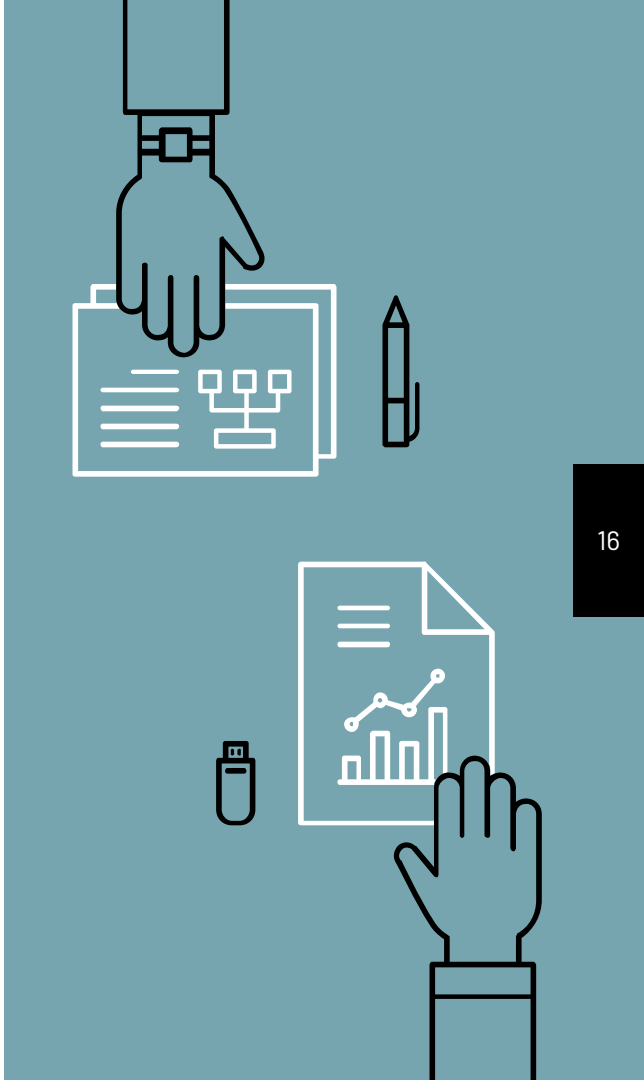


MIDUS II Data Sets

1. Data and supporting codebook and other documents

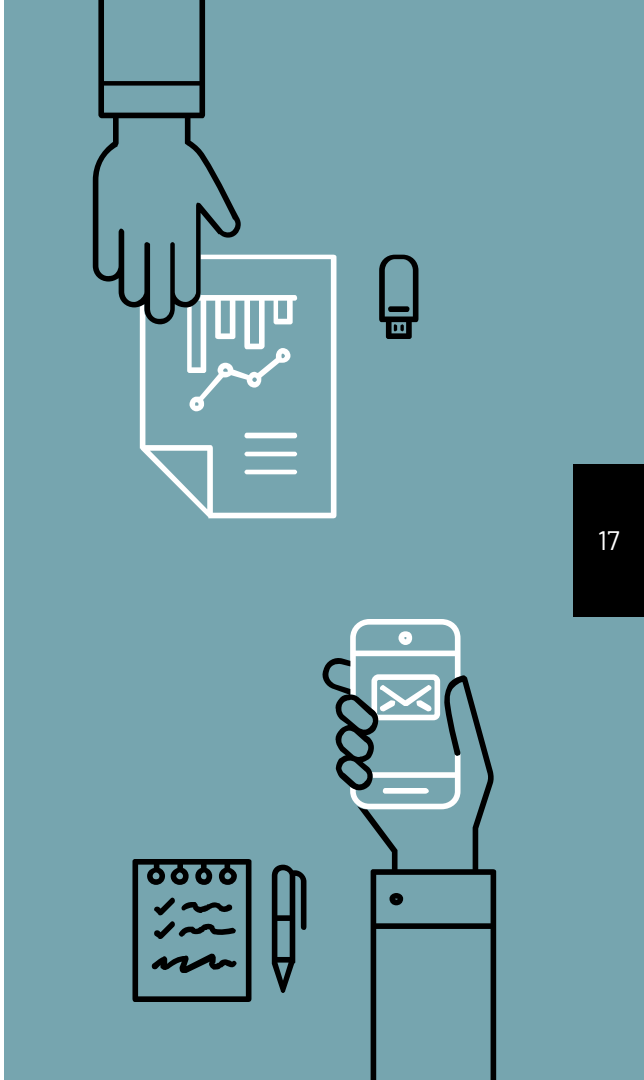
2. Biomarker data

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.



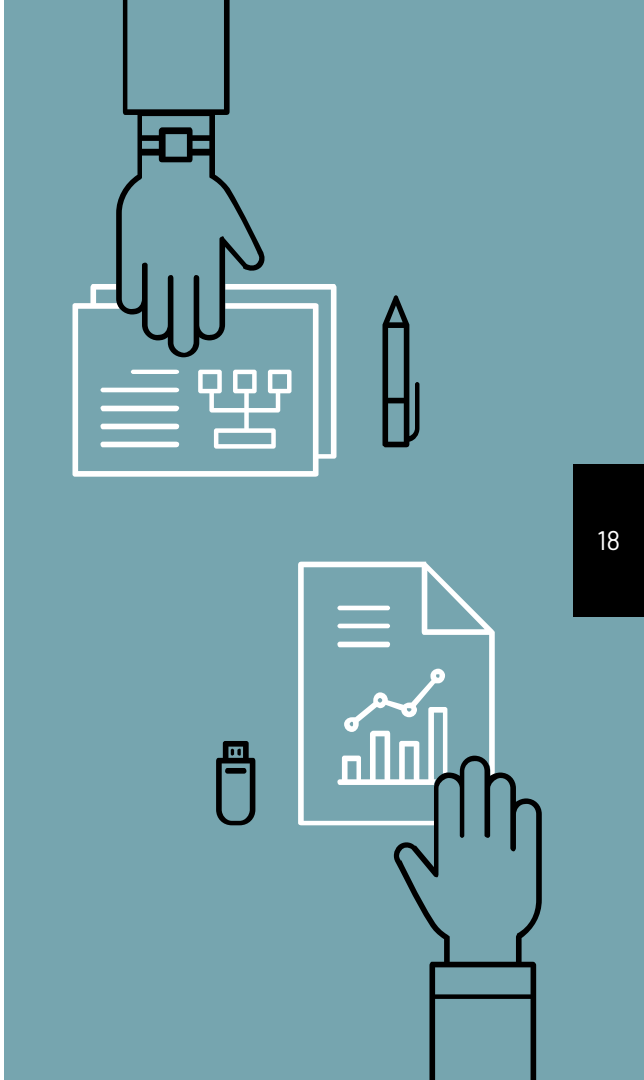
Data Dictionary

- We will be working in teams to create a data dictionary for this project
- All variables used for the analysis should be posted here with any notes you think are informative/necessary
- Assignments
  - Group 1: Salvador, Dylan, Randy, Chen
  - Group 2: Zhuoran (Joanne), Jodeci, Christian, Lee
  - Group 3: Dominic, Yi, Lauren, Khondoker,



Data Wrangling

- Once we have the variable names organized, we need to wrangle the data and determine our sample
- In the same groups, write code to deal with any missing values, weird values, recoding, etc., according to the paper
- We'll discuss all code tomorrow



Homework

- Watch Module 3 videos
- Submit Module 3 discussion points