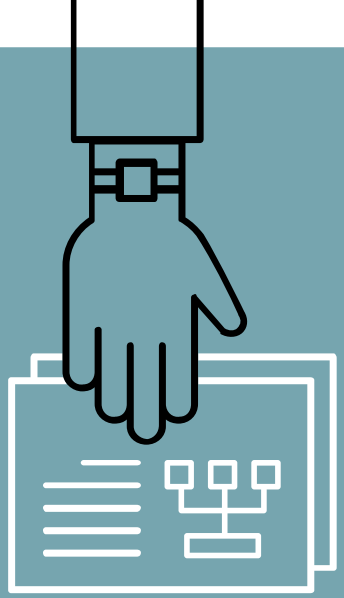
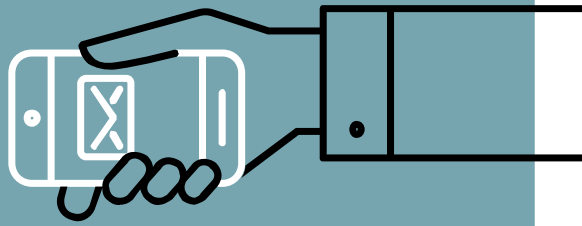
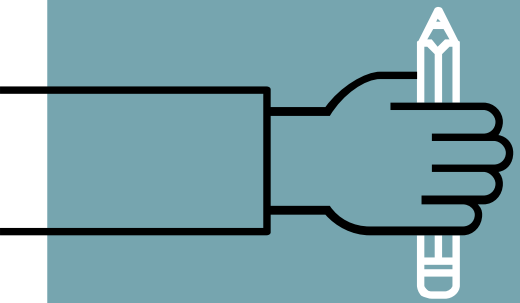
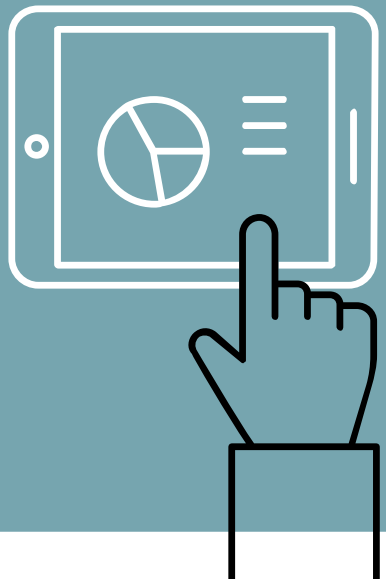




BST 270
Reproducible Data Science

Winter 2021
Session 2

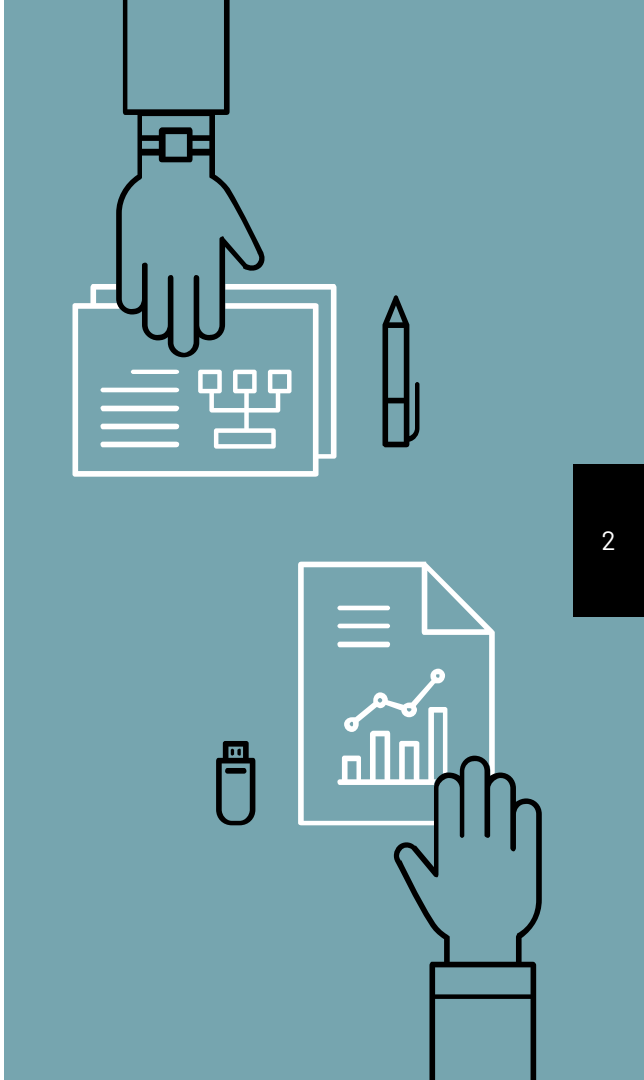




Module 2 Discussion

Questions

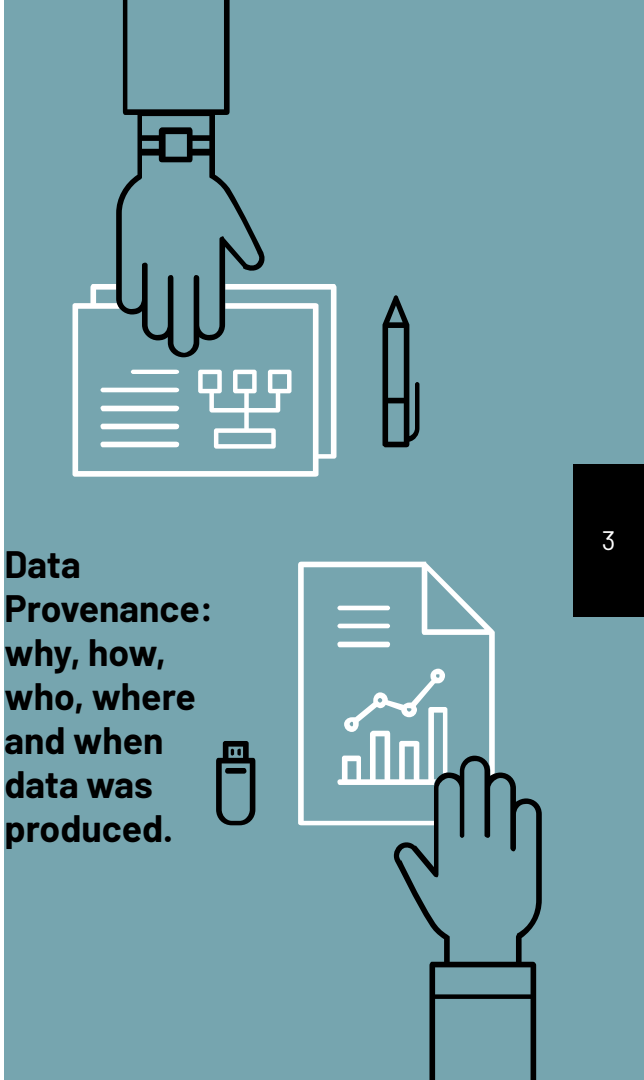
- **What is meta-data?**
  - Data about data
    - How the data was created
    - How the data was cleaned or transformed
    - Time and date of the creation/collection/transformation
    - Who collected/cleaned/transformed the data
    - Where the data is stored
    - Who has access to the data
   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



Module 2 Discussion

Questions

- **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)



Data Provenance: why, how, who, where and when data was produced.

Module 2 Discussion

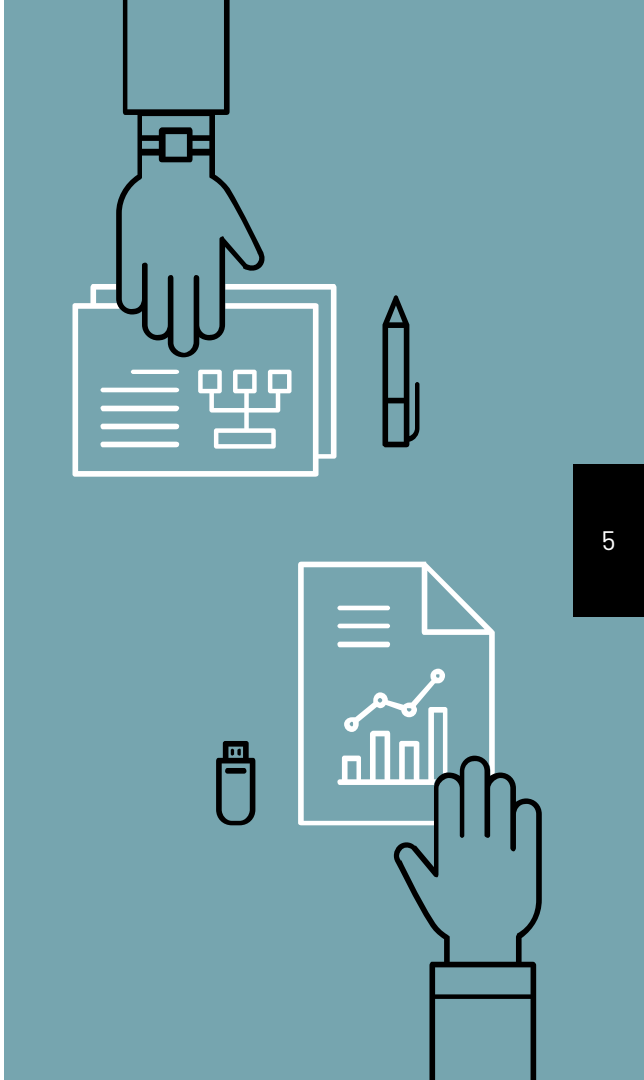
Questions

- **Problems from lack of metadata?**
  - You don't know the "history" of the data
    - This can lead to a misunderstanding/ambiguity of what variables represent or how they were transformed into a particular type which can lead to incorrect/invalid results and conclusions
    - Could lead to a privacy breach
    - Can mask any biases in data collection or cleaning

Module 2 Discussion

Questions

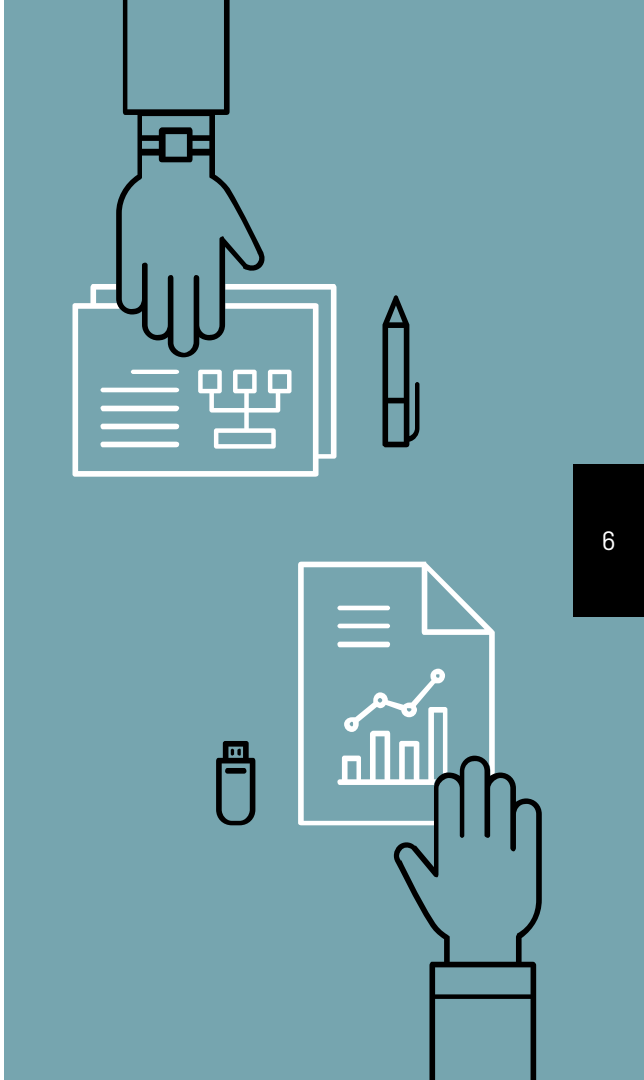
- **How much should we care about how reproducible the collection of the actual data was? Are there any procedures to evaluate how valid the data itself is?**
  - We care about this a lot, but in practice it can be difficult to validate data collection methods
  - Usually if a data set is used a lot in published papers or was collected by a reputable institution, we tend to trust it
  - Different data collection biases can be revealed when conducting analyses after the fact
    - Selection bias for example



Module 2 Discussion

Questions

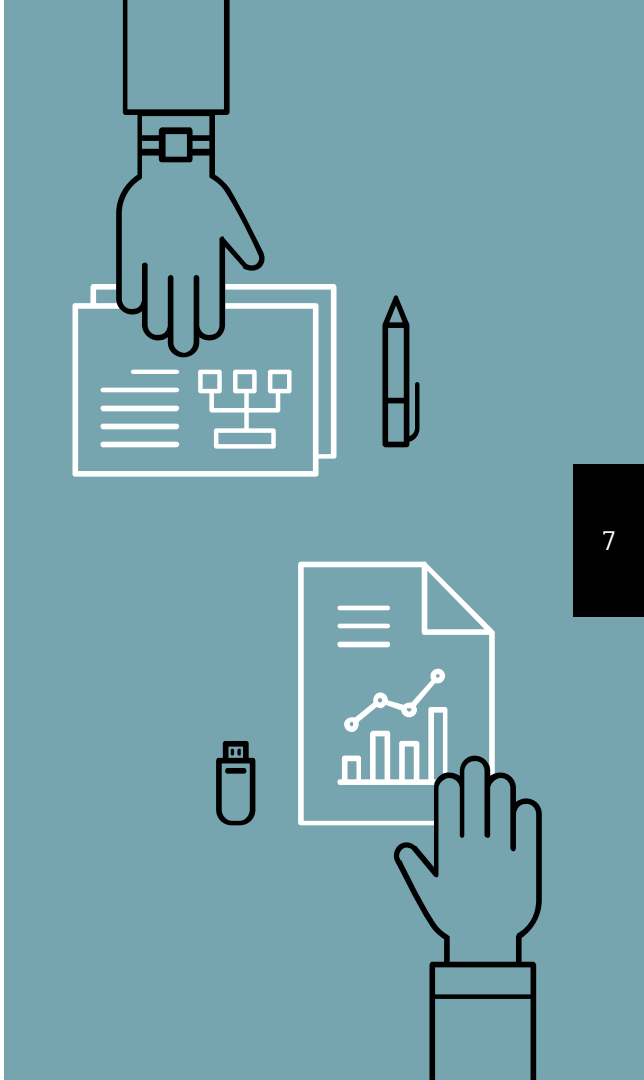
- **What should a data scientist put in the lab notebook?**
  - Code, including comments that help another read and understand what the code is doing
    - State what each variable is, what the inputs and outputs of a function are, the type of object you expect to get after running the code (a list, 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?



Module 2 Discussion

Questions

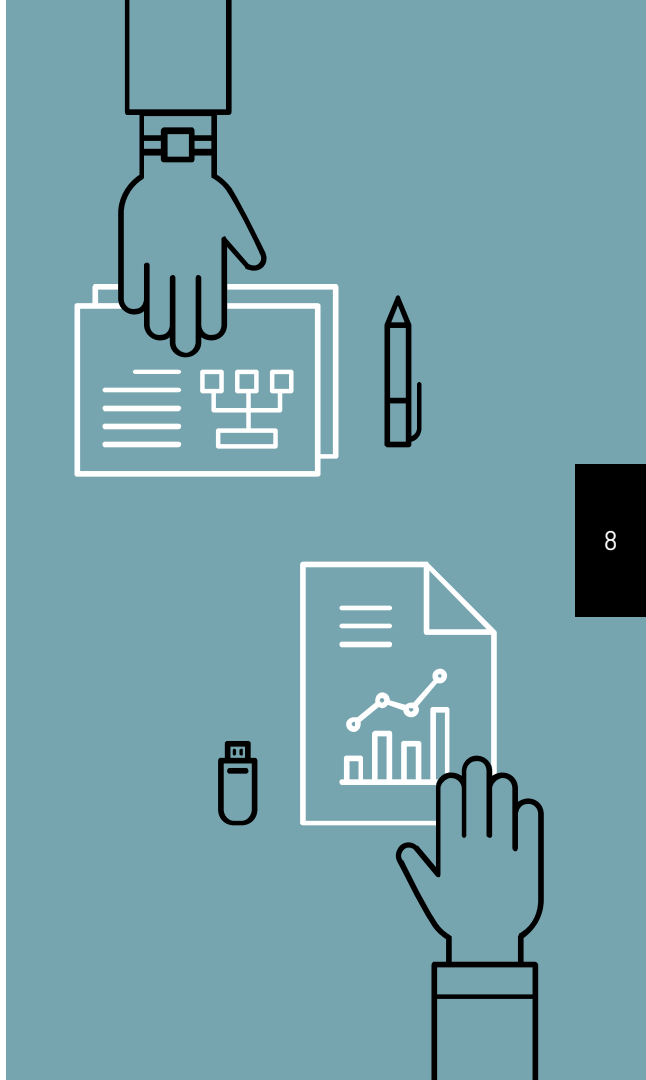
- **What should a data scientist put in the lab notebook?**
  - Visualizations
  - Interpretations (of plots, tables, output, etc.)
  - If your project requires running several files with code, mention where this file fits in to the workflow
  - Ideas for future work
  - Any notes from your PI or collaborators



Module 2 Discussion

Questions

- **Is there a standard for well-written code?**
  - Short answer: yes!
  - Long answer: depends on your advisor/PI/team
  - Examples
    - Google's Python Style Guide
    - Google's R Style Guide
    - R Style Guide
    - The tidyverse style guide

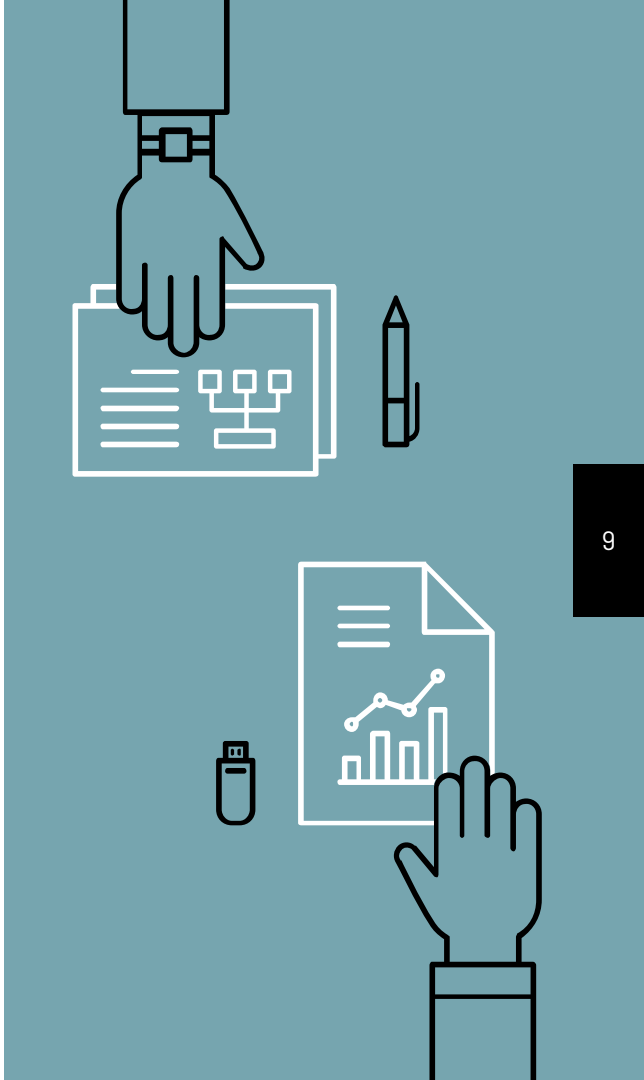


Module 2 Discussion

Questions

- **Is there a way to automatically backup code to GitHub?**
  - Yes! But $$$
  - You can also use Google Cloud Platform or AWS (also $$$)
  - Or this code (but I don't know if this works or not)

- **Automated data/results sharing sounds really useful but probably requires a lot of careful planning out. Is it likely to accidentally share the wrong information?**
  - There is always that risk, which is why there are so many training modules to complete before touching biomedical data
  - There are (or should be) several checks made by multiple individuals before releasing anything to the public or other collaborators/team members.

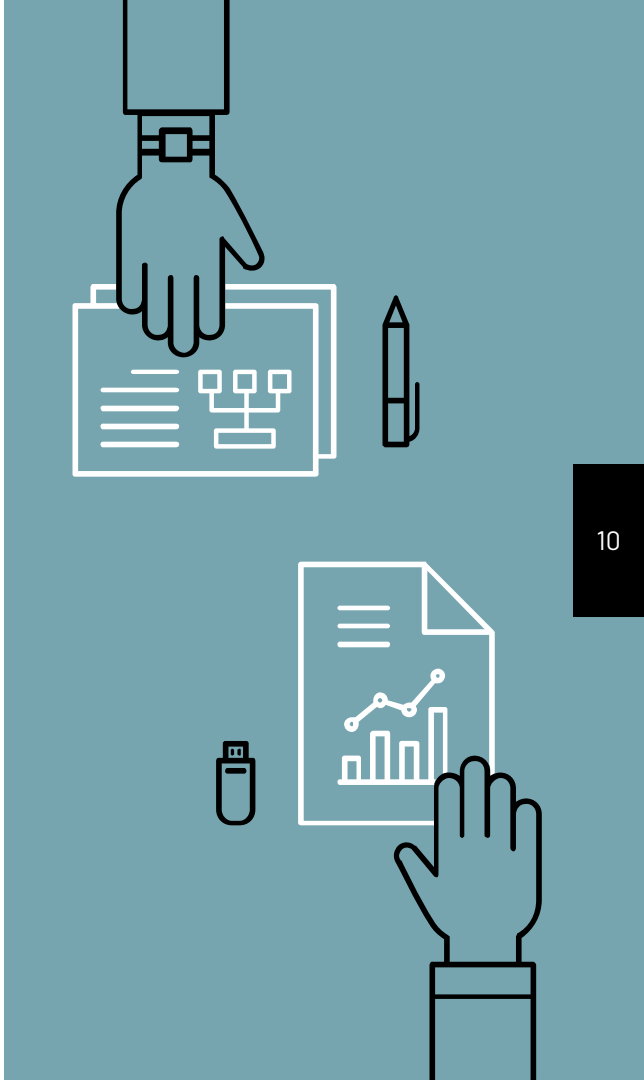


Module 2 Discussion

Questions

- **How do you organize your data/results/papers?**
  - Google Drive, Dropbox, Overleaf
  - A document that explains in detail where all of the pieces are (i.e., the code, the manuscript, the lit review, etc.)
  - I personally keep detailed notes of meetings and any thoughts or issues I experience during the project

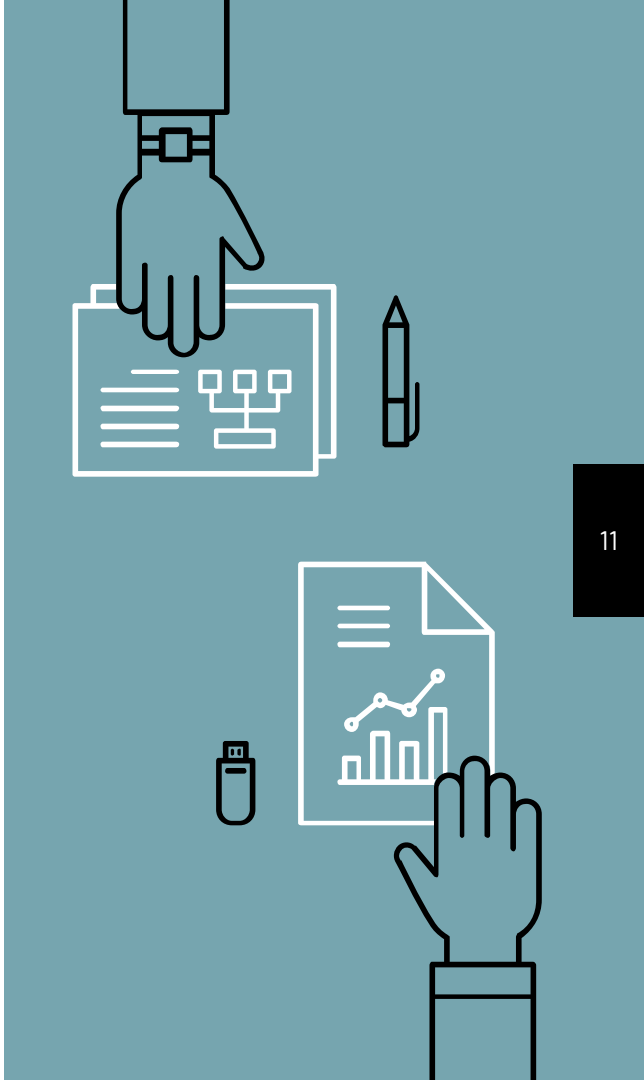
- **Are there ever problems that arise when researchers use proprietary software for analysis, where the actual code is not available to the public?**
  - There can be - you'll see this in Module 3
  - This does happen a lot and journals can require example code for synthetic data or a toy example
  - Sometimes you can request to see the code - but this is difficult



Module 2 Discussion

Questions

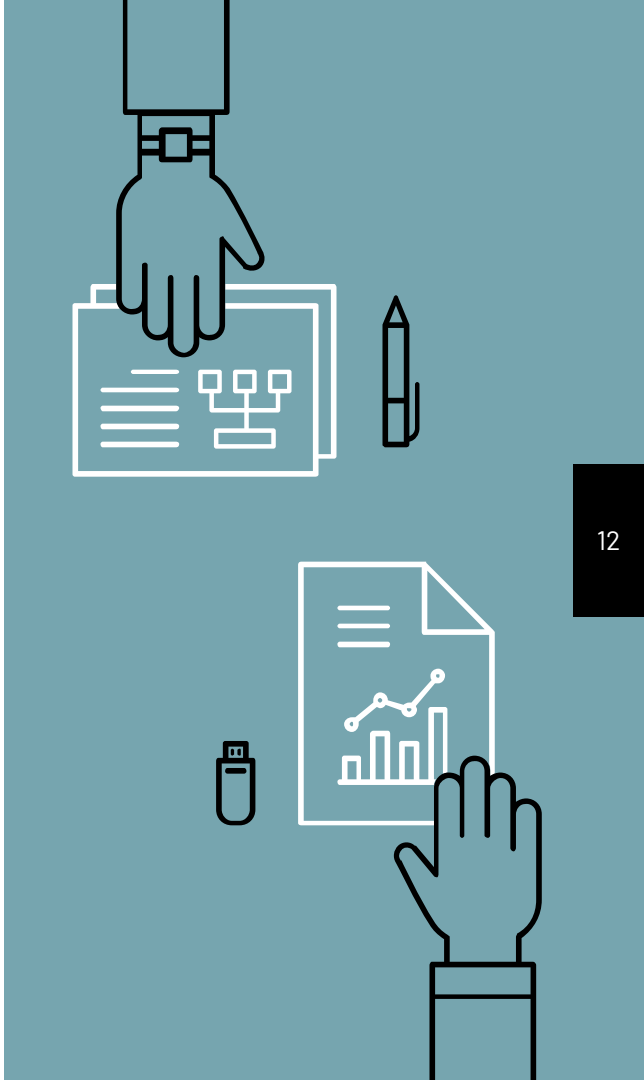
- **Video 2.3 mentions doing visual representations and tests to check the data, and I wonder what that looks like for high dimensional data or other data that makes visualization complicated.**
  - Break it down
    - Don't try to plot everything together
    - Univariate/bivariate visualizations
      - Example: plot the age distribution. If you have someone who is 300 years old, you have a data entry error
  - I like this Towards Data Science post



Module 2 Discussion

Questions

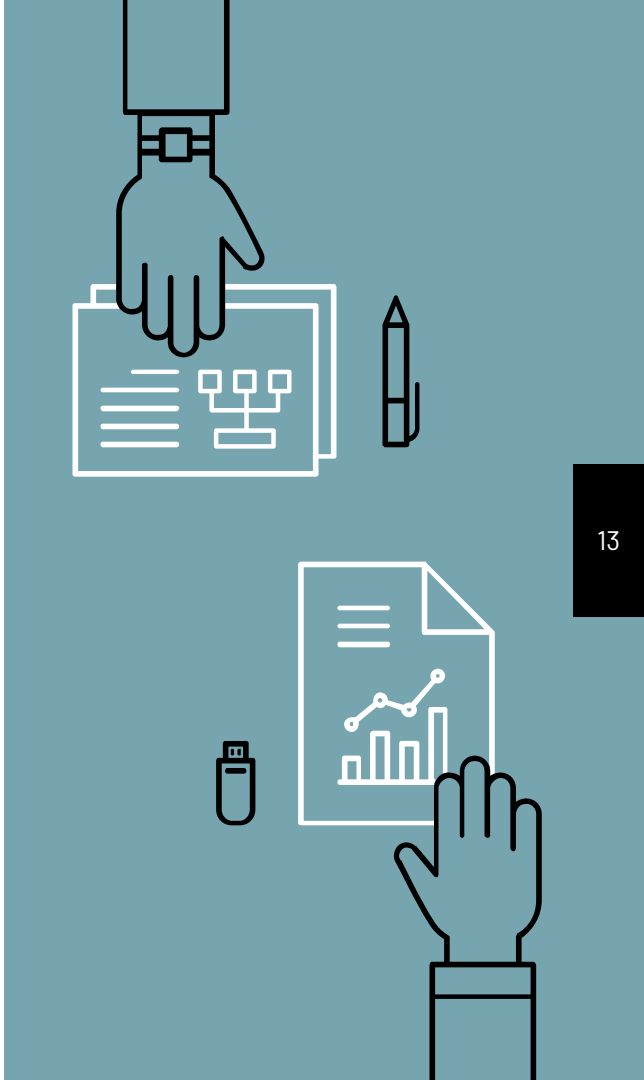
- **When submitting to a journal, are the reviewers allowed or encouraged to reproduce the results?**
  - Short answer: I don't know for sure.
  - Long answer: I *think* reviewers can do this if they really want to and have access to the data. I haven't ever experienced a reviewer request access to the code or other materials needed for the analysis. I imagine the overwhelming majority of reviewers don't have the time to do this (they are also in academia and there are (usually) deadlines for reviews). If it's really impactful research like the case study presented in Module 3, I would hope someone would attempt to reproduce the results.
  - New Experimental Results journal



Module 2 Discussion

Key Takeaways

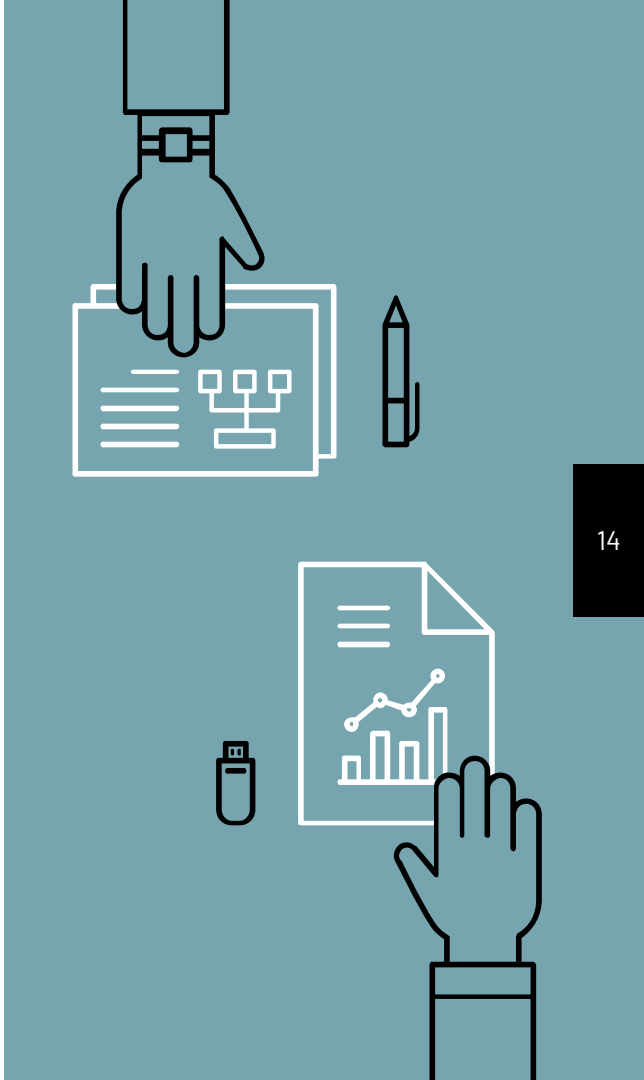
- The structure of file storage is important
- Writing code that detects its own potential errors is important
- Metadata (data provenance) is essential
- Variable naming
  - Should be as informative/intuitive as possible



Module 2 Discussion

Key Takeaways

- **Reproducibility: Data vs Analyses**
  - Reproducibility of data: more about *files* – where you put them and how you manage them
    - Example: databases with notations for data
  - Reproducibility of analyses: more about *programs*, *code* and *workflows*
    - Example: literate programming or revision control repositories for code that carries out analyses

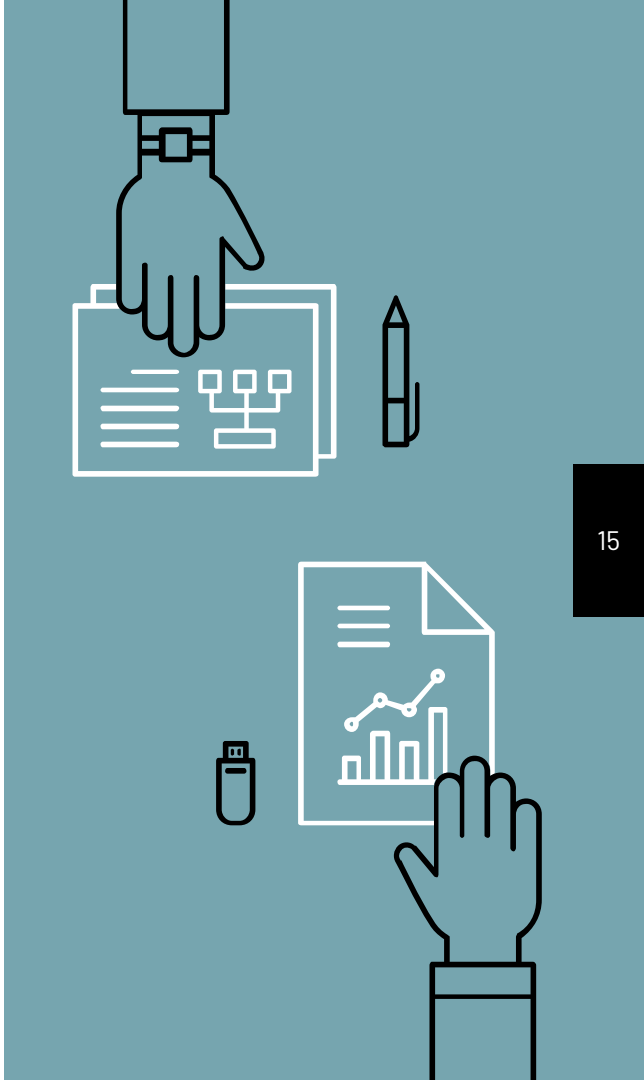


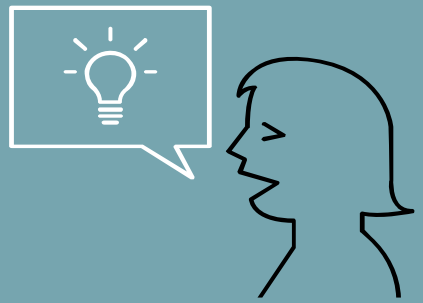
Module 2 Discussion

Key Takeaways

- **Reproducibility: Electronic vs Protocol**
- Electronic reproducibility: activities that a computer can carry out for you.
  - Checking assertions, control results, or unit tests
  - Storing documentation or data in a particular repository
  - Managing revision control history for an analysis workflow.

- Protocol reproducibility: activities and best practices you must do yourself.
  - Using a revision control repository or a public database
  - Designing your experiment to support convenient reproducibility by others
  - Picking a consistent naming scheme for your files and folders
  - Writing enough documentation for another user or your future self to read and understand





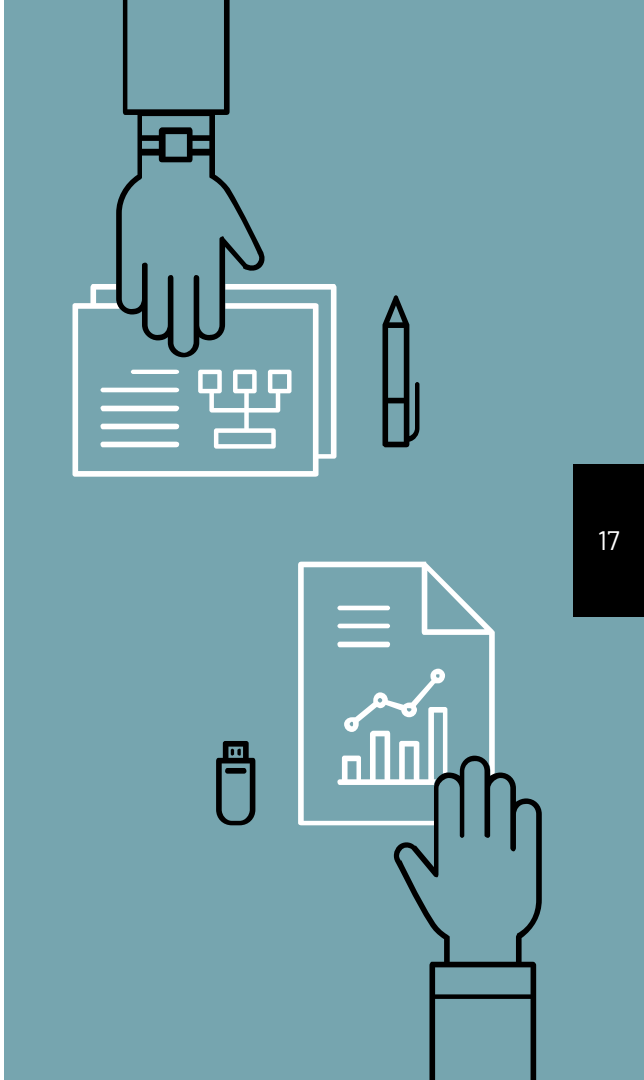
In-Class Project

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

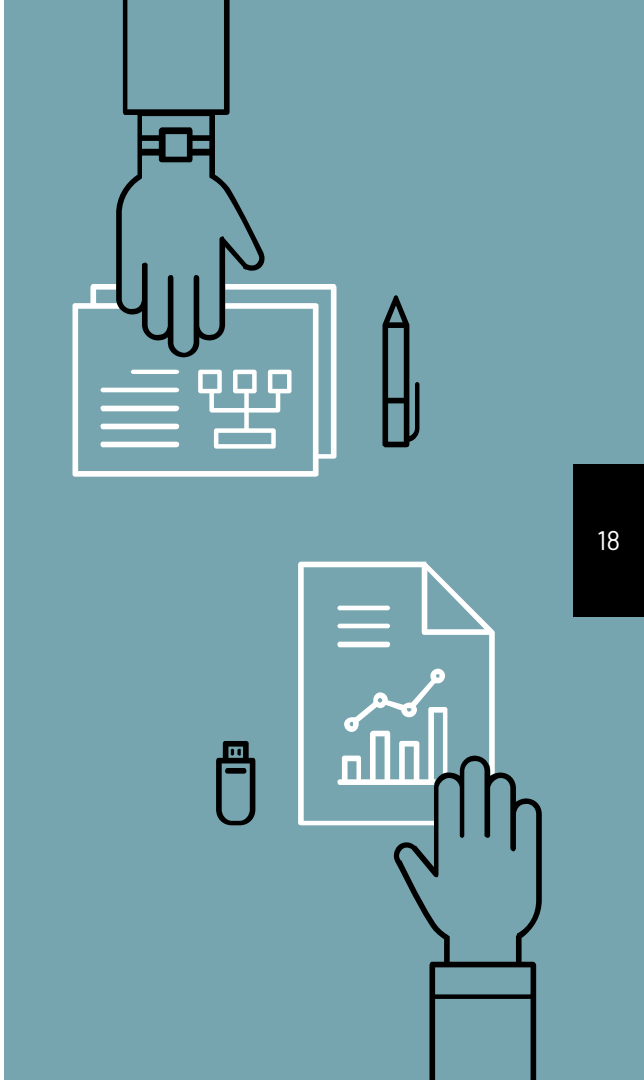


We will attempt to reproduce the results from and critique the reproducibility of:

[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

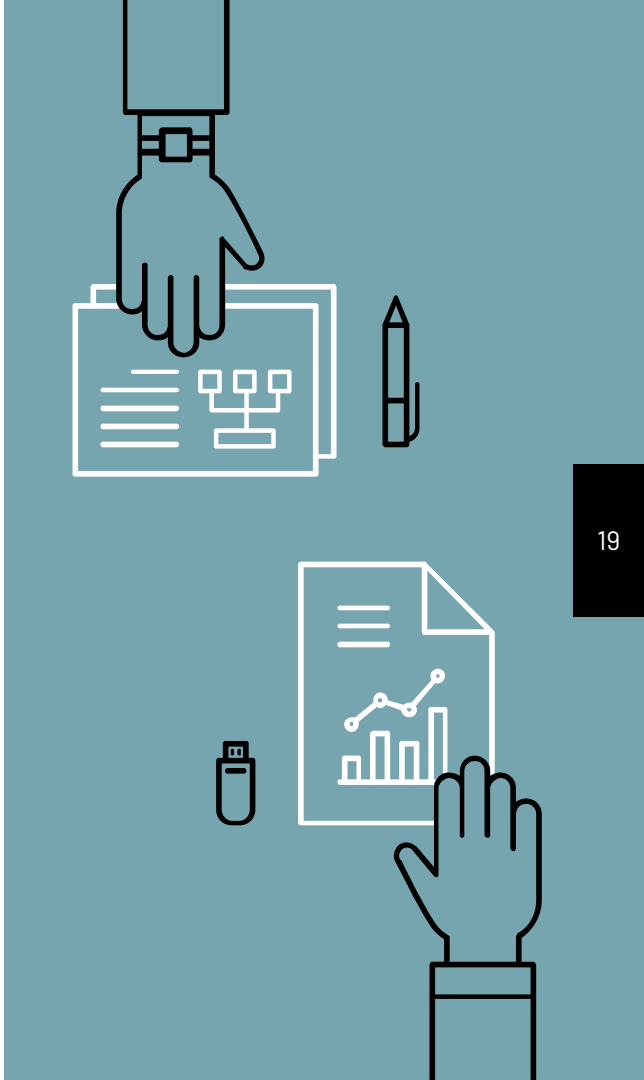


MIDUS II Data Sets

1. Data and supporting codebook and other documents

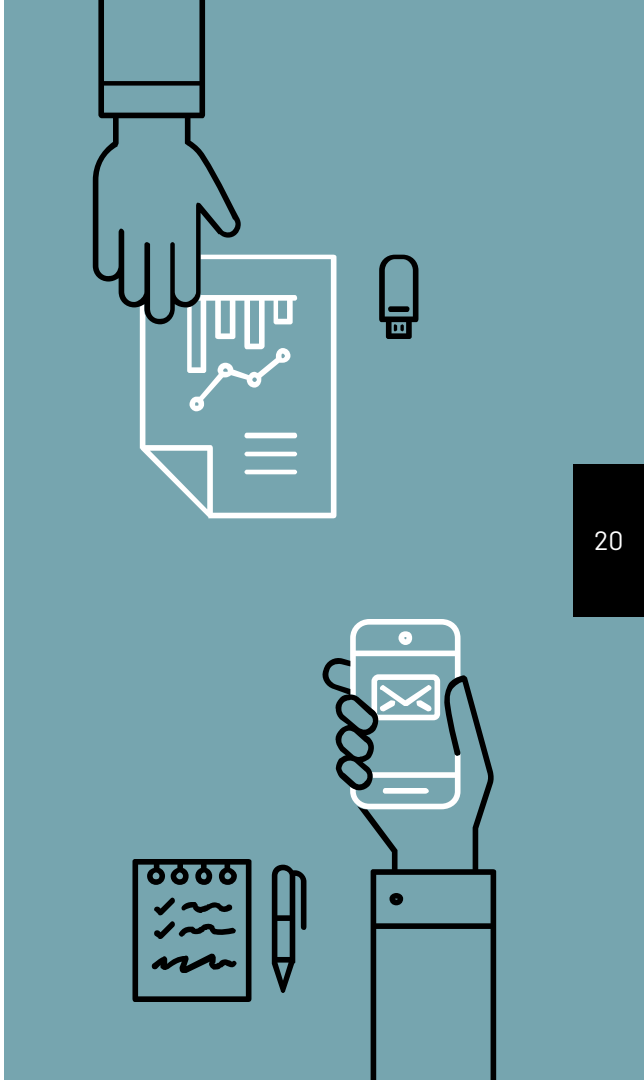
2. Biomarker data

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.



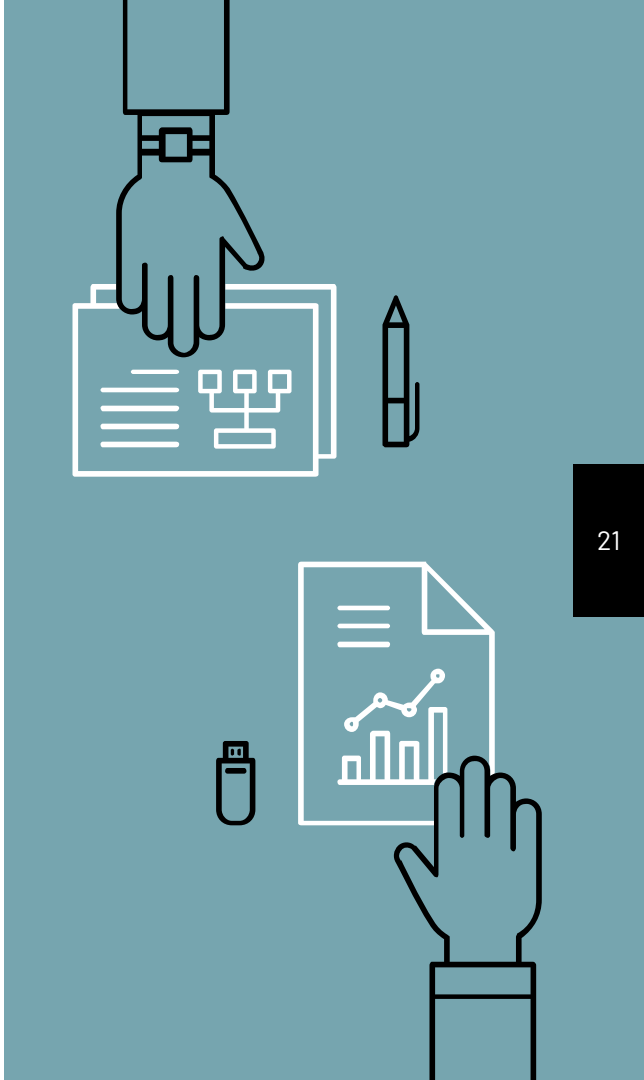
Data Dictionary

- 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: Kat, Evan, Corri
  - Group 2: Jeanette, Tony, Chidimma
  - Group 3: Joe, Gabrielle, Arpan
  - Group 4: Yu, Zebin, Daniel



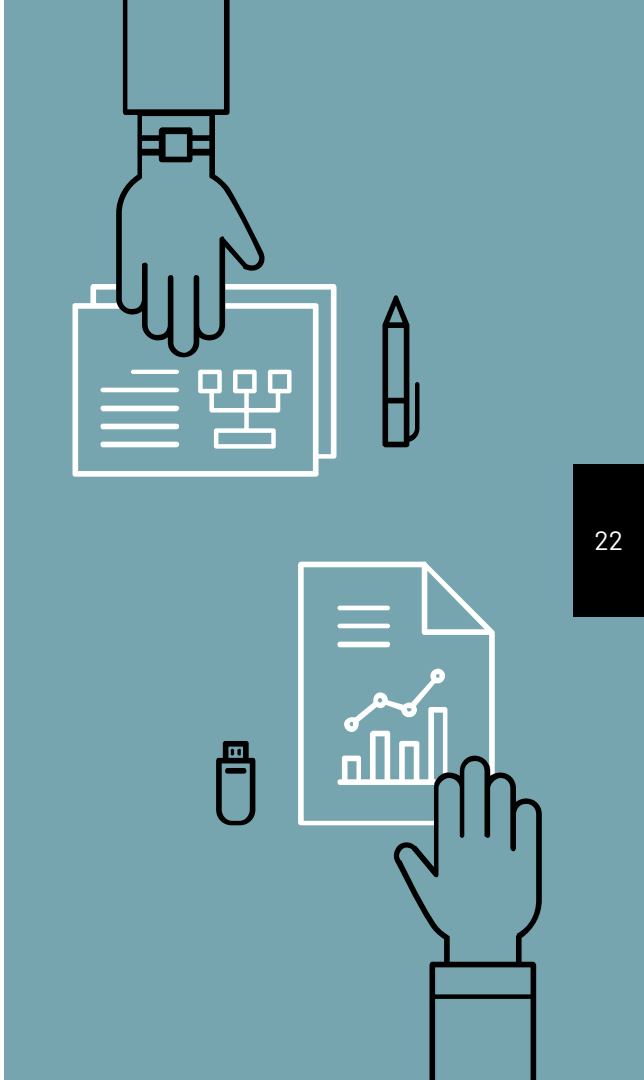
Data Wrangling

- 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 after



Sample Size

- What did we get for a sample size? Does it match the paper?
- Why?



Homework

- Watch Module 3 videos
- Submit Module 3 discussion points