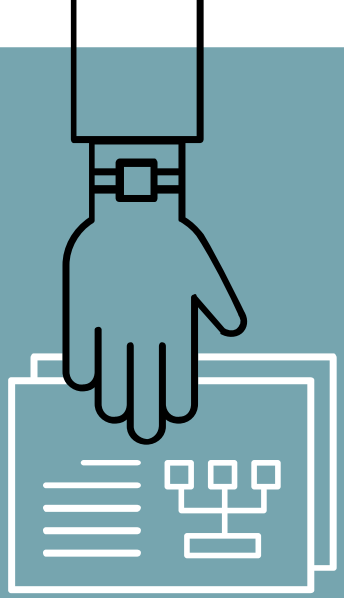
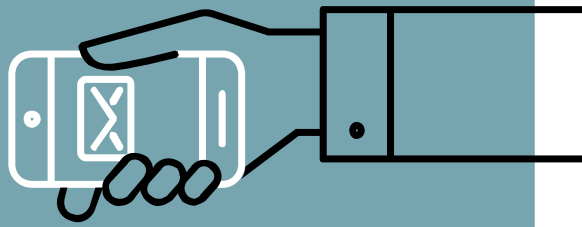
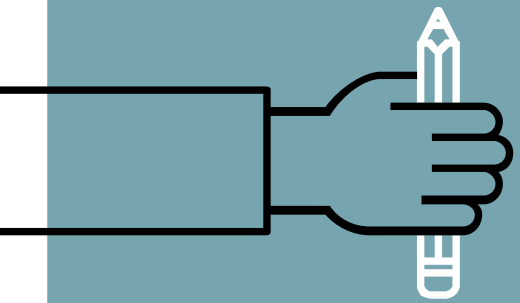
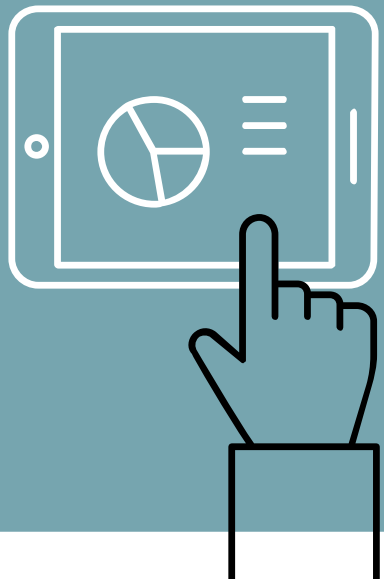




BST 270
Reproducible Data Science

Winter 2022
Session 6



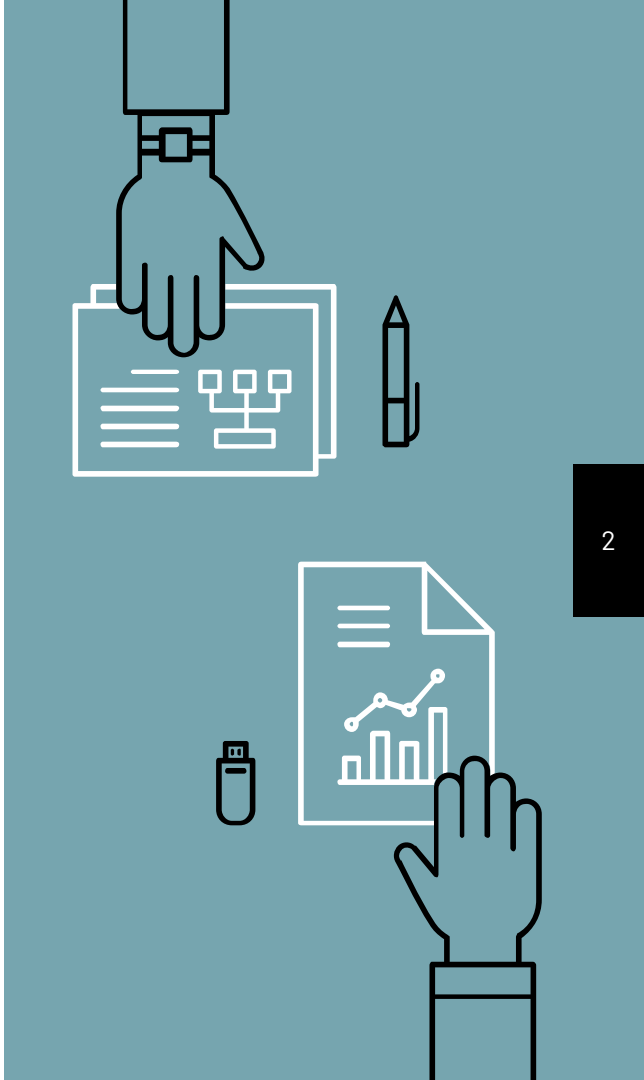


Module 6 Part 1 Comments

Data management seems by far one of the trickiest. This is because biomedical data comes in many different forms [e.g. EHR], and with privacy, what is shared is further complicated (maybe only summary statistics).

A centralized system that is secure and adherent to the data formats/standards in many fields would be amazing - privacy and policy seems like a couple of the biggest barriers to making this a reality. People create more and more tools to help overcome this, but honestly I feel like it will be way too far away until we reach, if we do.

It was interesting to hear about other IDEs for R. I've only heard of and used RStudio previously.



Module 6 Part 1 Comments

Our program can be very theoretical, so it was nice to go back to the basics of applied data science and solidify my understanding of the different metrics and ways of simulating data.

It was nice to see the metrics that we've learned before presented in a different notation like Lorenzo does. It was a little confusing at first but always good to learn things in different ways.

I always found it very interesting when I was applying for jobs that a lot of places would have Data Management positions in addition to Data Analyst positions. Those seem like difficult roles to separate, since, as Merce was saying, a very large component of the data analyst's job is determining what data management is required in order to effectively use the data.

Module 6 Part 1 Comments

I haven’t used Make before, but it sounds like such a powerful tool to automate the analysis pipeline!

I’ve only used the Git command line but didn’t know github had a desktop software that makes its usage Even more friendly.

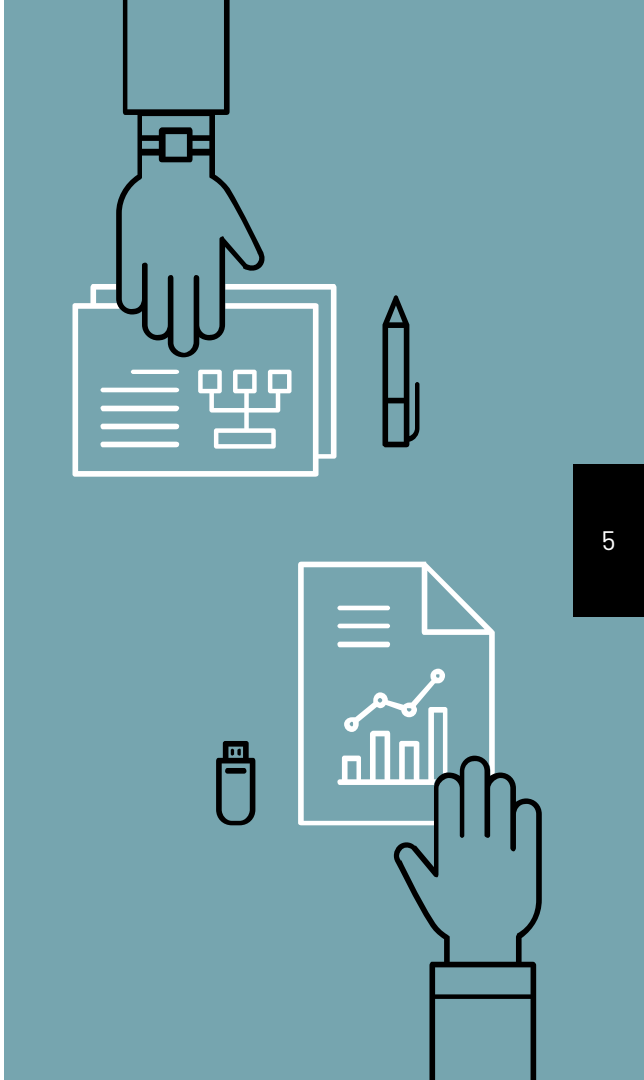
I have worked in an interdisciplinary team where researchers came from very different backgrounds. Most of the time, those with data science background would be very keen on using version controls tools such as git while others with a clinical background would sometimes be reluctant to learn/adapt to working with git. It is an interesting task to convince others in using a new tool and to balance between an interdisciplinary group of researchers.

Module 6 Part 1 Comments

A few more features of R which I think the videos do not talk about are list of variables (in global environment) on the top right, you can navigate to different folders using the files tab in bottom right, terminal in bottom left where one can run linux and git commands. Overall R studio is an all-encompassing environment for reproducible data science.

Dr. Crosas and colleagues suggests ten rules as a few things we can do as researchers to help improve our workflow and make the data more reusable, and easier to reproduce. I think these rules are very important for new researchers and can serve as guide for planning from the hypothesis generating to manuscript production.

I had very bad habit of naming and tracking my modifications on the code, and usually I end up spending hours on finding the latest version of the code. I will try using github to maintain my code in the future.

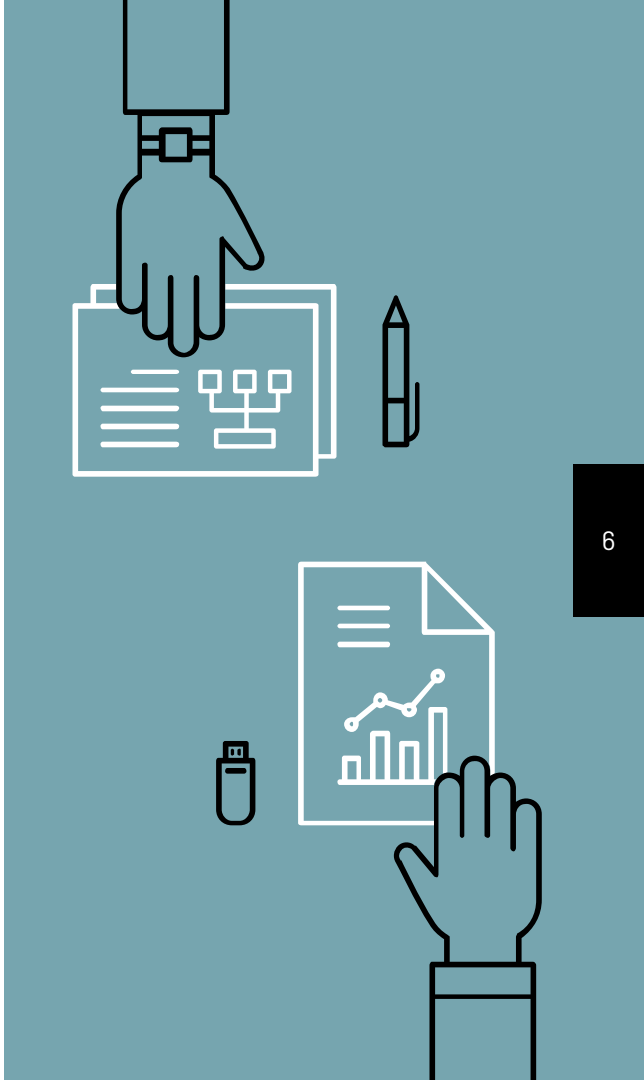


Module 6 Part 1 Comments

I hope data sharing could be made easier and easier in the future so that we could have more efficient collaborations with people who work on similar topics.

I'm again struck by the importance of statisticians being familiar with data management as well as being able to effectively communicate with data managers. I've worked on projects where less than ten hours of meetings between data managers and statisticians to set up a database with a mind to how the data would be used in analysis would have saved 100s hours spent cleaning and restructuring the data later.

I felt mollified upon realizing that it seems even harder to make qualitative data / analyses reproducible than for most quantitative projects…

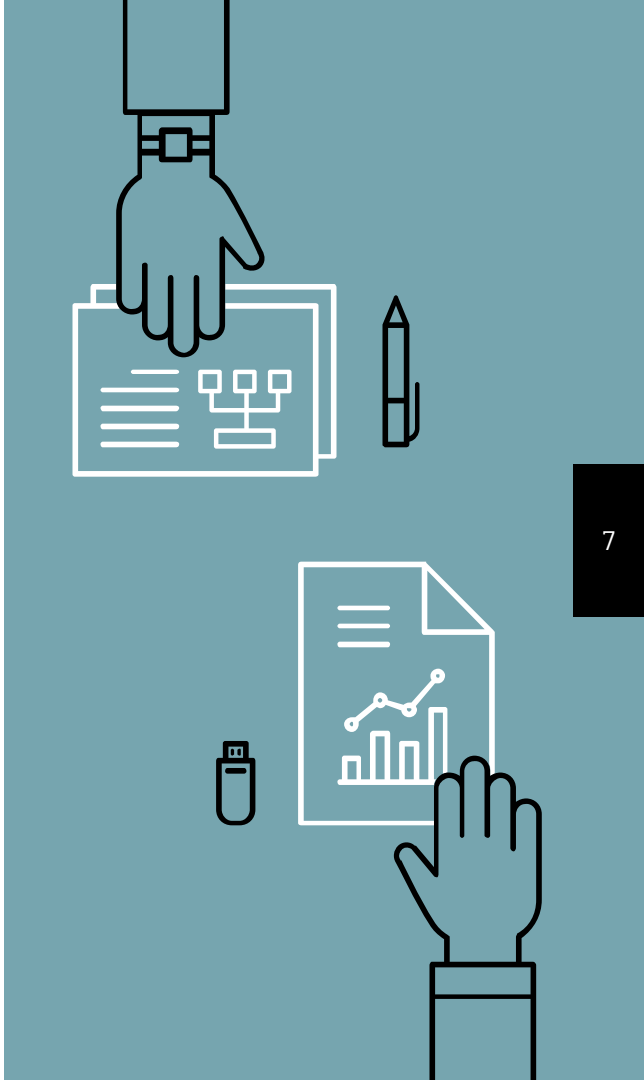


Module 6 Part 1 Comments

The 10 rules seem almost obvious but also very important to be written out like this. If everyone learns data science with these ten rules, it would make everyone's lives a whole lot easier.

For python projects with lots of separate files, test cases, classes, etc., I like PyCharm because it makes debugging, switching between files, and running tests much easier than JupyterLab. Also, Sublime was mentioned multiple times for R and Python, but I also like to use it for editing bash scripts, slurm job scripts, and readmes.

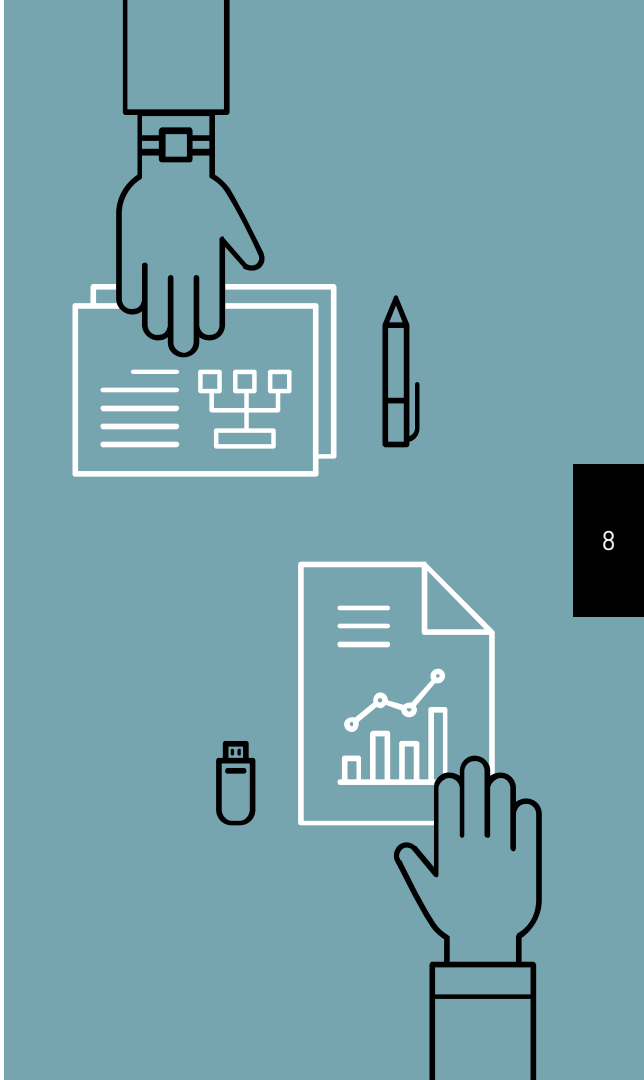
I really liked seeing how RStudio integrates with Git. Honestly, all of the Git videos were super informative and helpful since I've never really been taught Git but have been told to use it so it has always scared me.



Module 6 Part 1 Comments

Up to this module, the whole story of data reproducibility becomes very clear to me --- where we can find the data to use (data repository and data archive); how we deal with sensitive data (data privacy and 6 categories of data); how we maintain research data and publish our own data (data provenance and version control tools); if we want to use multiple data sources, how we integrate different data sources (statistical methods for data science and data pooling); how we publish our own results and validate others' results (data and research management and reproducible of research), etc.

The point that Merce Crosas made on the challenges of Big Data is interesting. She mentioned that data sharing is not easy for huge dataset. So what is needed to make that work is the combination of data repositories with computing computational resources and technologies. She mentioned that we can utilize the cloud computing environment to facilitate the computational workflow. One such platform I ever used is Redivis (https://redivis.com/), it collaborates with academic data repositories and allows researchers to create customized data portals using SQL language. I am curious if there are also such platforms for researchers at Harvard to use.



Module 6 Part 1 Discussion

Which is the python IDE that you would recommend, Atom or Jupyter Notebook? What are some of the differences?

What python IDE are you using?

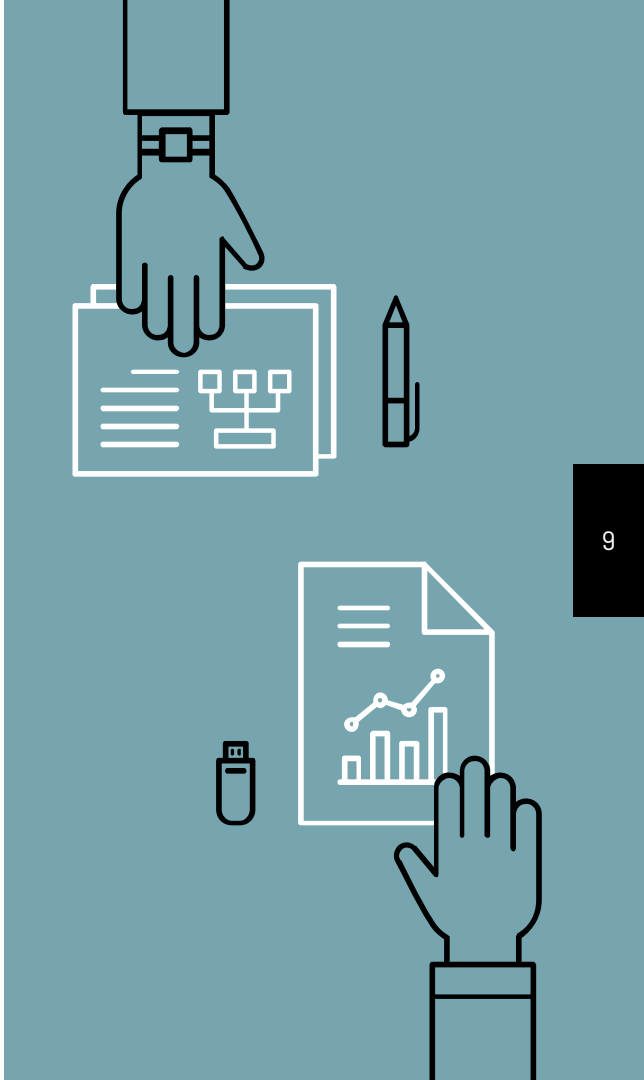
Video 6.2.3 talks about different editors for Python - what has worked best for you, and what would you suggest?

Are there better IDEs for Python than Jupyter? One thing I love about RStudio is the environment tab, which is missing in something like Jupyter notebooks. In Jupyter, I feel it's so much easier to run things out of order.

- I've never used Atom but now that I've checked it out I really want to use it!
- Jupyter Notebook and Jupyter Lab via Anaconda Navigator have worked the best for me. I know a few colleagues who prefer Spyder and VS Code.

Are there a few examples of situations where one would prefer an editor that's not IDE?

- I think it has to do with preference. I know some people who learned how to code with plain scripts and don't like IDEs. I personally really hate coding without one. Sometimes you will need to save your code to a plain script in order to execute it. I've had to do this when using the School's computing cluster. But I always use an IDE to write and test the original code.



Module 6 Part 1 Discussion

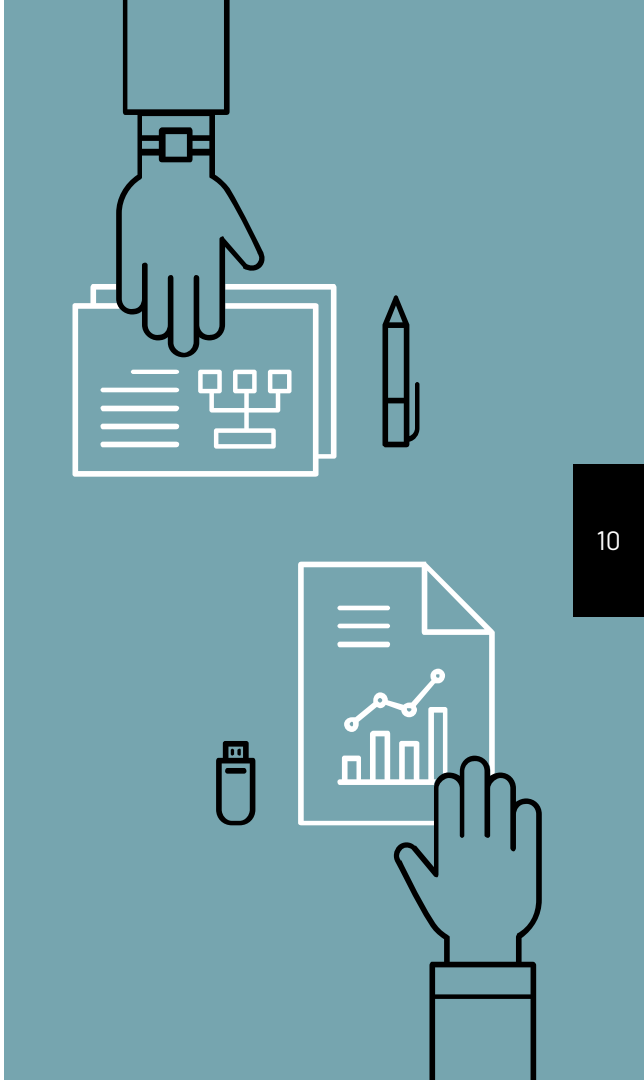
Can you use Git to do version control for datasets that contain sensitive information?

When talking about using services like Github and bitbucket, how do you navigate the privacy settings if you store data on those sites? Are there different tiers that would allow for different levels of sensitive data to be stored, or is it best practice not to store any data there?

- Yes. GitLab is recommended for this. Google Cloud Platform can also be made HIPAA compliant and host Git repositories with version control.

I understand the master and development branches in Git, but what is the use of the bug branch? I assume this is where you can keep documentation of past bugs in your code, but does it contain actual code as well? Why is this branch necessary if you can just add a new version to your master branch that fixes the bug?

- Bug branches are used to explicitly track the difference between bug development and feature development. Bug branches will be created when there is a bug on the live site that should be fixed and merged into the next deployment.



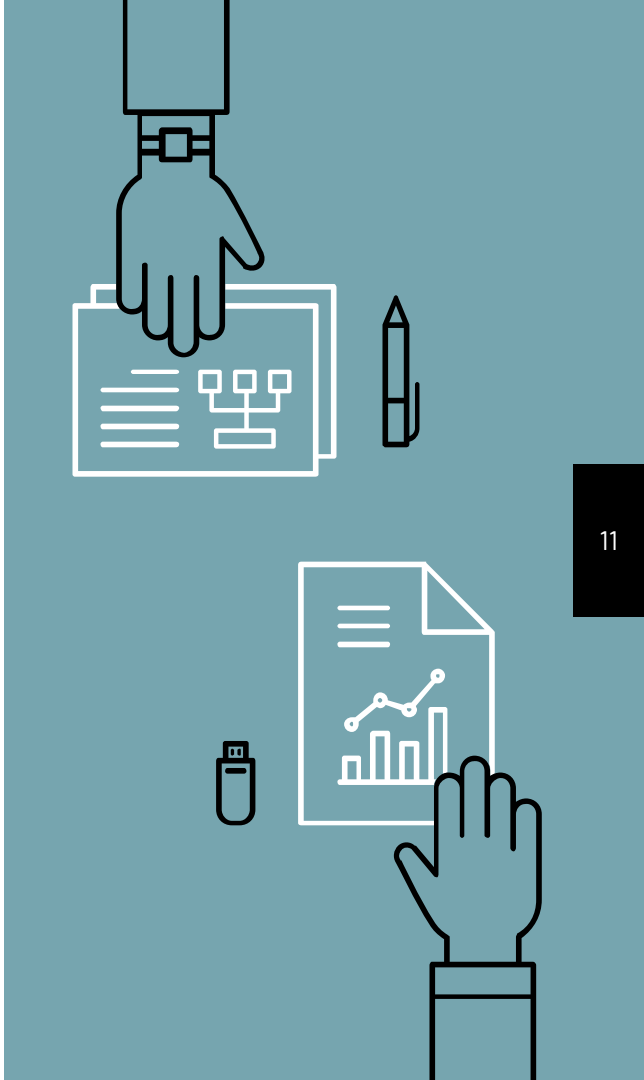
Module 6 Part 1 Discussion

I found the Github modules very interesting as I had never used Github before prior to this class. What are some of the best practices / key considerations for Github to be successful in larger teams?

- I recommend GitLab for larger teams. And usually, the data science and/or engineering teams will have specific protocols when using GitLab. Like a process/rules for creating/naming repositories, merging branches, checking code, etc.

I'm a bit confused as to what gitignore actually does; what happens if you don't use it? Why exactly do you not want your R workspace to be pushed to Git?

- gitignore is used when you don’t want to track certain files, meaning the different versions aren’t subject to version control. This can be for files with sensitive information, or files that aren’t necessary to your analysis that you don’t want to track. I use gitignore for my intro to data science course so students can copy my lecture files and take notes without disrupting the repository. See details at the bottom of the repo’s README file.



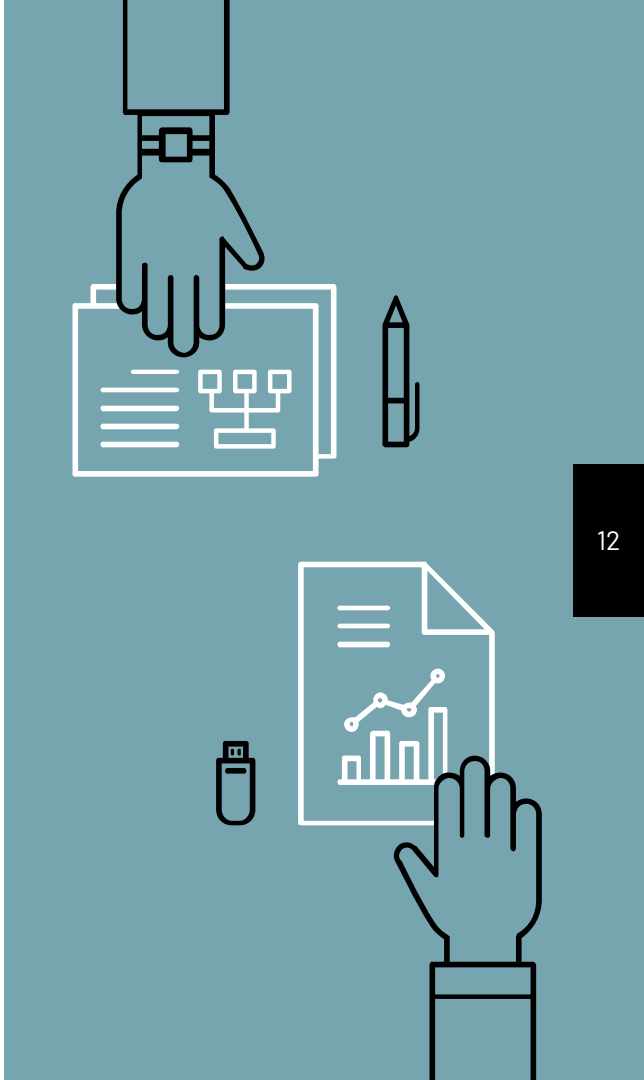
Module 6 Part 1 Discussion

What about GitHub that makes it unique compared to other resources of project sharing resources?

- Two big things that come to mind are version control (although it's not the only version control resource out there) and the community. The community of GitHub users is huge and they are usually very helpful if you have a bug or issue with anything.

Are there ways for two people to collaborate on a project in different languages? For example, can Git or any other version control software convert python to R?

- Git can't because it is for version control. But you can add python code chunks and run them in an Rmarkdown file and you can add R code chunks and run them in a Jupyter Notebook. So, there are ways to collaborate in multiple languages. It can make things more difficult at times but it's doable.



Module 6 Part 1 Discussion

What does a license do inside a Github repository? Can we create one without it?

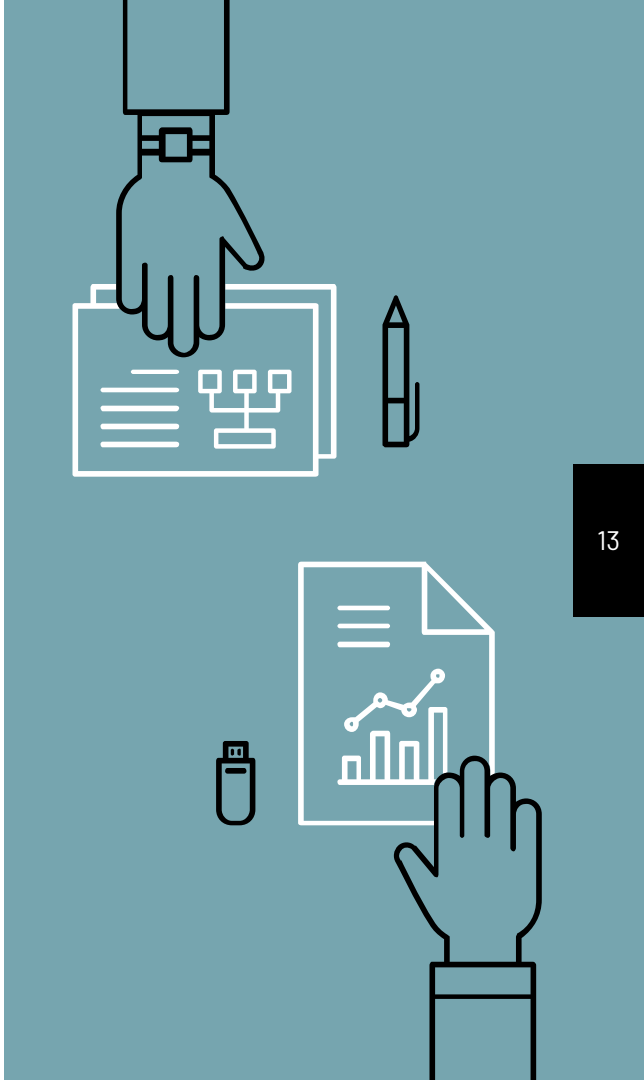
- A license ensures others are free to use, change, and distribute the software. The MIT License permits users to do anything with a given project as long as they credit the developer and don't hold him or her liable for the project's use.
- You can create a repository without one, but if you want to make your code and/or data open source you'll need one so that people who use it will have to cite you/the repository. You can check out how to create one here.

It is surprising to me that Github is not the official service from the git team. Is there any alternative to Github that still keeps the great features of git?

- Here's a good resource about alternatives. A lot of the alternatives are actually more flexible than GitHub and more popular - depends on your team and type of project.

StackOverflow is always telling me to --force push in response to any git error. How bad of an idea is this?

- Depends on how many people are working on the same project/files(s). It can be really bad if you overwrite someone else's code. There is a better option: –force-with-lease. You can read more about it here.



Module 6 Part 1 Discussion

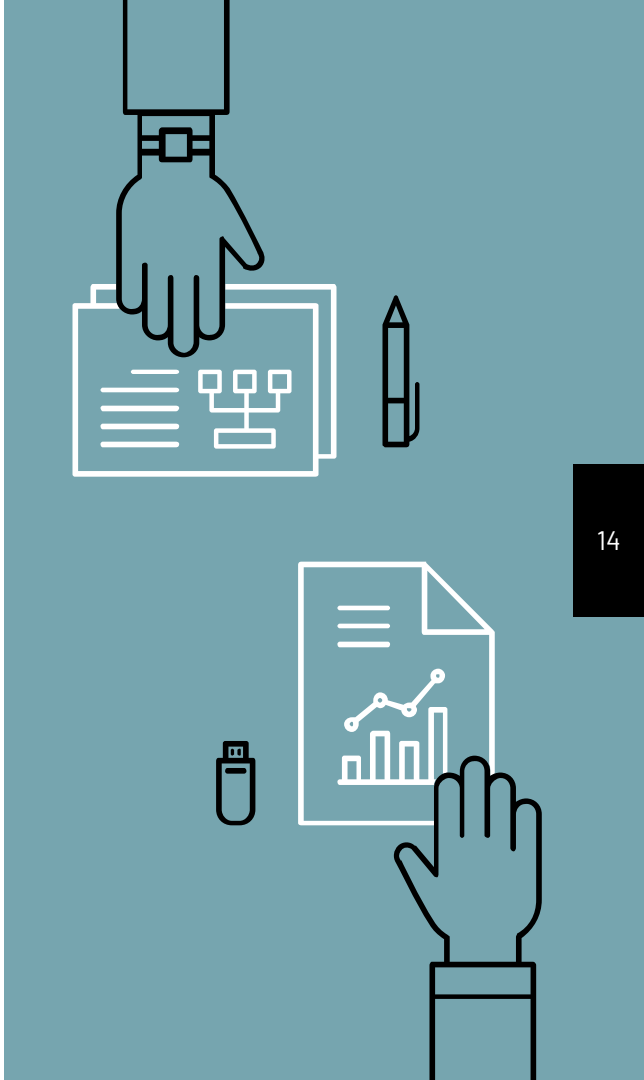
What are the most common commands used in Git while collaborating on a project?

I feel like a lot of people I've worked with have familiarity with the main git commands (add, commit, push, pull) but things can go off the rails quickly if things fall outside of this realm (especially with merges). What tips do you suggest to get better practices w/ merges.

- Once you start working with collaborators using git, it can get complicated quickly. You can get up and running with git/GitHub/GitLab basics really quickly, but learning how to merge and perform more advanced commands takes training. I would recommend having everyone on the team receive the same training, and discuss/decide on a set of best practices/rules/SOP for working together with git.

- The most common commands are add, commit, push, pull, clone, branch, merge, checkout, fork

- Resources for tutorials: GitHub collaboration documentation, GitHub collaboration tutorial, merging branches,

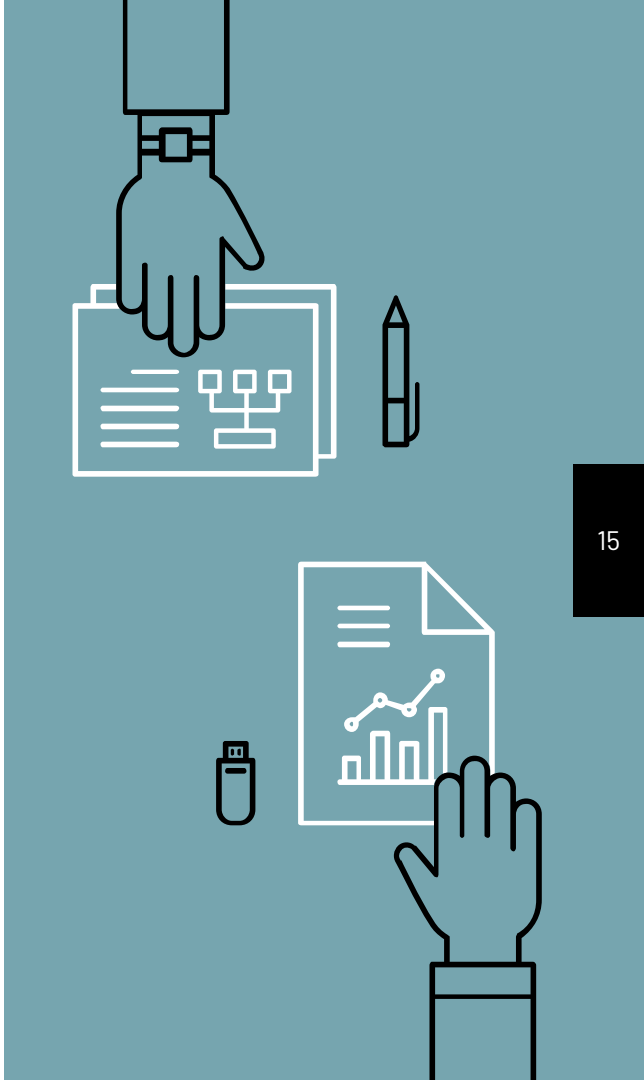


Module 6 Part 1 Discussion

I'm still a bit confused about what "git pull --rebase" actually does. Do you (or anyone else here) have insight into when we would want to use this instead of a regular "git pull" or branch?

In video 5.2.4, Erik gives advice about using git-pull rebase. In my limited experience of working with version control (git), I never heard that we should do a 'git-pull rebase' before we push any changes. Is this common practice in industry/academia? Although it's sound advice and makes sense to do that before pushing every time. I wonder if it is common practice.

- git pull --rebase is used to combine your local unpublished changes with the published remote changes. Your local changes are reapplied on top of the remote changes. git pull pulls the remote version onto your computer but does not apply your local changes.
- I'm not sure how common this is. In my intro to data science course I haven't used it in an attempt to keep the coding as simple as possible, but it would be really helpful for students who forget to pull before they commit and push new code - they get an error message and can't push. git pull --rebase would solve this problem. I'm sure it's used a lot more in industry and bigger teams.



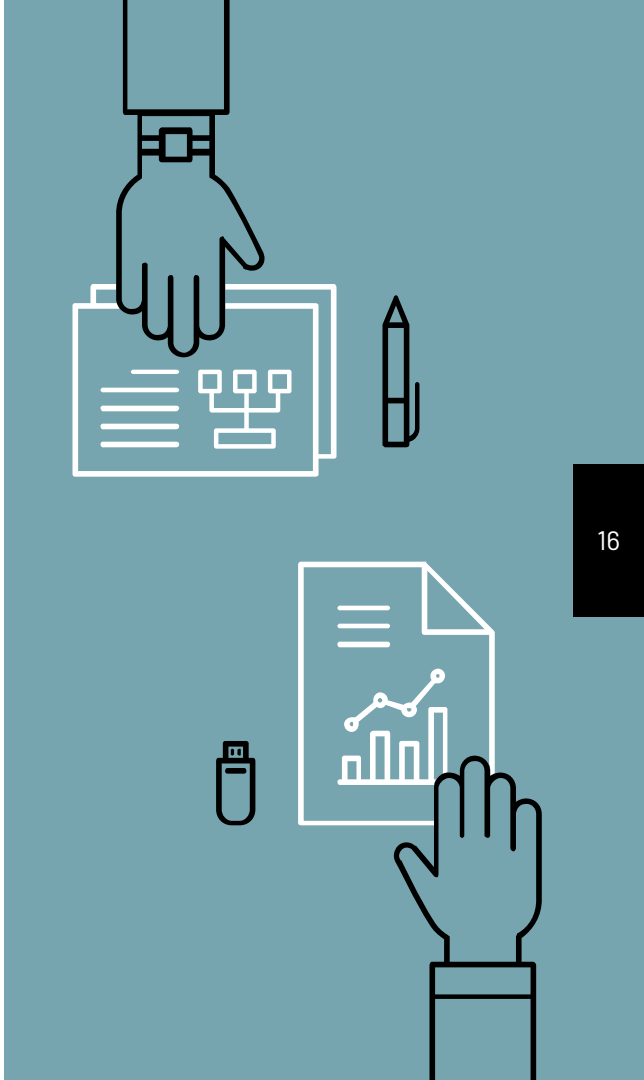
Module 6 Part 1 Discussion

Are journals moving towards allowing GitHub repositories in publications, and if not, is that something that should be encouraged?

- I think they're moving towards any kind of open access or limited access repositories, not just GitHub. More and more journals are requiring authors to report/explain your plan for data access when submitting a manuscript for publication.

Are there ever issues where data-sharing is hindered due to a paywall? For example, do organizations that collect data ever force users to pay some kind of monetary fee to access the data? If so, does this have significant effects on what must be done to make research reproducible?

- Yep. One example is Twitter data. You can pay less (or nothing in some cases) for a limited number of tweets. You can pay hundreds of thousands of dollars for the Twitter "firehose" - all of the Tweets matching your keyword criteria in real time. Data that is paid for can make analyses very difficult to reproduce, but since Twitter data is technically public, I believe you can share the data you received. If the data is sensitive and can't be shared, I think a researcher would have to gain access from the paper author(s) to access the data. Twitter did launch other (less expensive) tiers of data access in 2017.



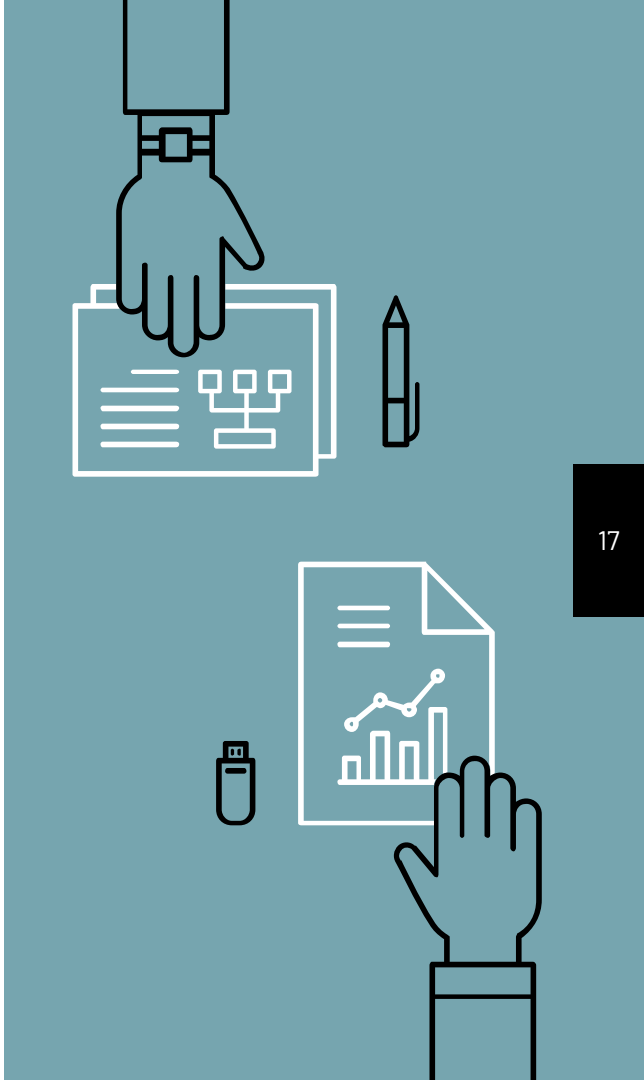
Module 6 Part 1 Discussion

Suppose you ran a study and obtained data that looks very promising. Is there a rule of thumb for how many papers/publications you should attempt to obtain before it is shared with other researchers (friends, colleagues, doppelgängers)?

- I don't think so. It's always nice to have as many as possible, but you also (usually) want to share your data as soon as possible. Sharing it sooner means more publications and citations and science.

One quote I once heard was ".pdfs are where data goes to die". Is this because of the difficulty that comes from getting the values for analysis, or is there another reason?

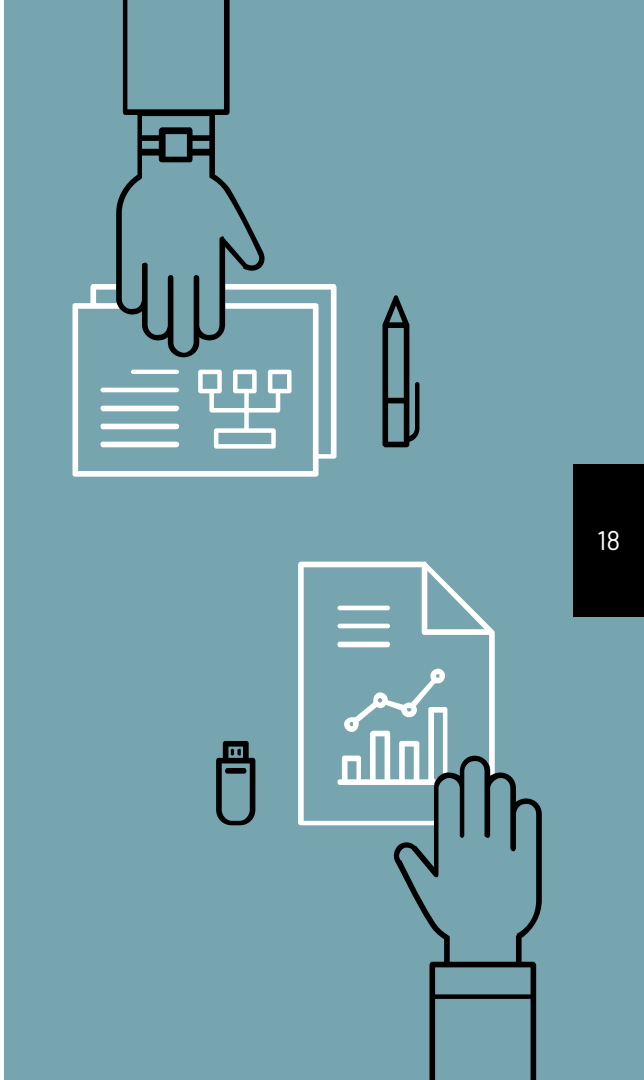
- This is because PDFs are not data formatted to be machine-readable and it takes a lot of tools to extract data from them.



Module 6 Part 1 Discussion

One thing that I've wondered for a while but never seen explicitly addressed is the question of when you should cite a package in a paper versus just having it in your code (which is then shared on GitHub, etc.). For example, I assume it would be overkill to cite ggplot2 or dplyr in your paper, but if one uses a package more directly in the analysis pipeline, many people cite the package. What do you think is the cutoff for citability?

- Great question. I've never seen it explicitly addressed either. I personally would report the language version and then any packages that are used for actual analysis, especially if they are very specialized and not commonly used.



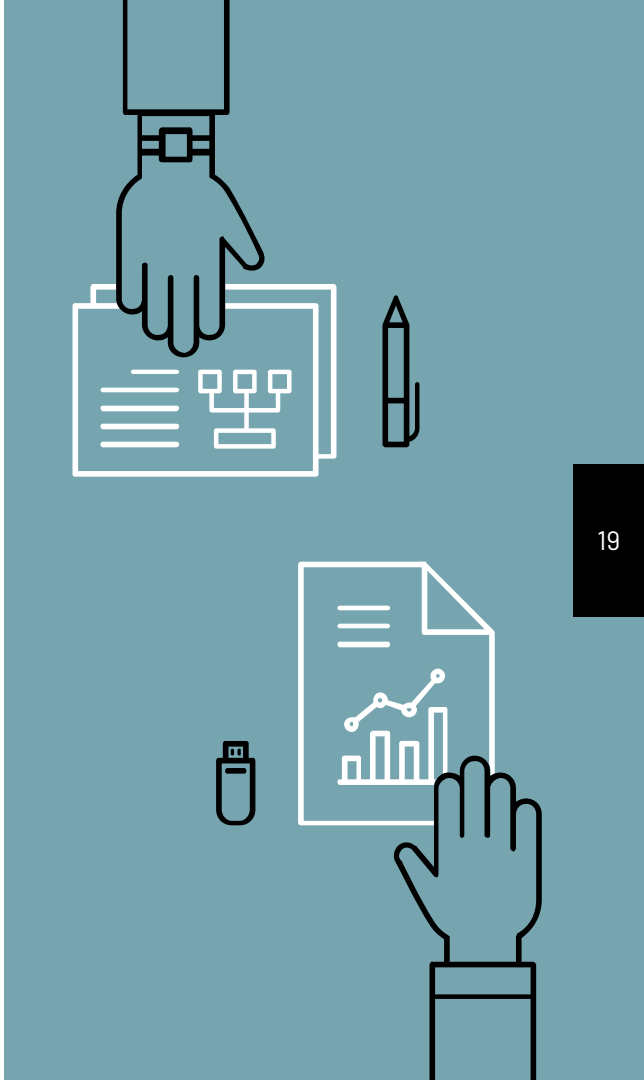
Module 6 Part 1 Discussion

I liked the idea of codifying a set of rules for data management. Particularly, the waiver related to how you want to get credit and the Creative Commons 0 license is fairly necessary. I wonder if there is a watching-eye effect that makes for better data processing since your name would be attached to it.

- I would think/hope so. It would work for me!

As Dr. Crosas mentions with the example of physics or astrophysics, the way that data is used for a given project may be completely novel. What would be a reasonable threshold for the specific data usage to be transformative enough for documentation?

- If I'm interpreting this question correctly, I would say anything that you do to/with the data needs to be documented. It may not end up in the Methods or Supplementary section of a paper, but you would still need to document it.



Module 6 Part 1 Discussion

Is there a movement to make pharmaceutical companies also make their data/results more transparent and reproducible?

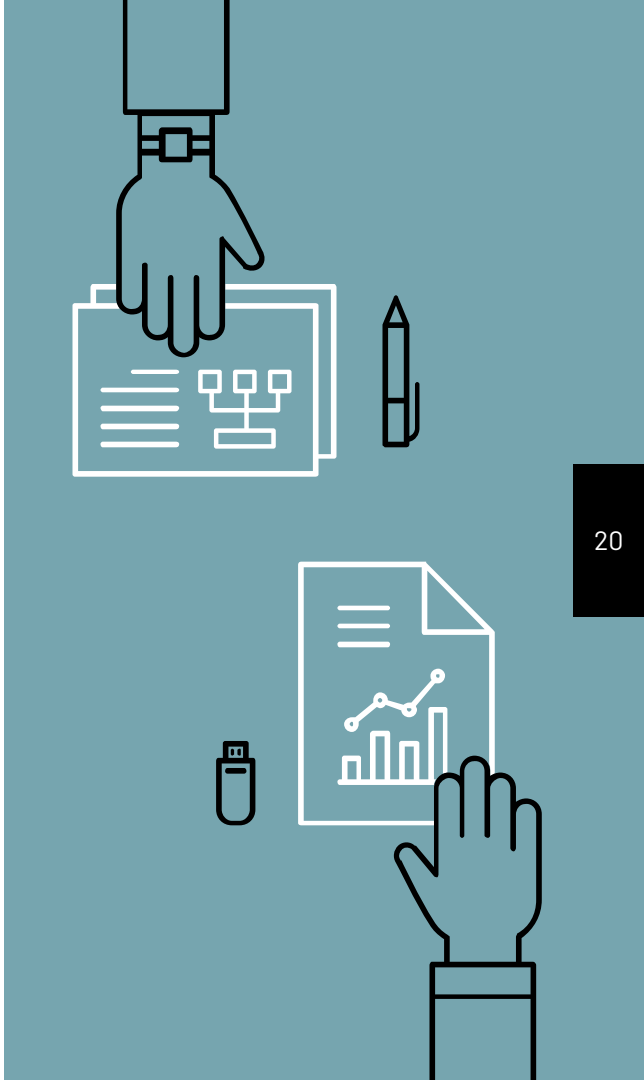
- Yes. This paper is a great resource. I think reproducibility is more important to some pharmaceutical companies than it is to academic researchers.

If you collected the data yourself, how do you obtain a DOI number for it to use for citations? Does the repository that you are publishing your data to generate a DOI for you?

- If you use Harvard Dataverse, you can open your data to the general public, or restrict access and define customizable terms of use. When you publish your data, you automatically get a standard data citation with a Digital Object Identifier (DOI), and your metadata is open and findable via search engines, even when the data are restricted. There are other ways to obtain a DOI and I found this resource.

It sounds good to maintain and manage data in a normative database. But would the accidental crash of it cause a big disaster and is it possible?

- Databases crash all the time. Usually it isn't a huge deal and can be dealt with quickly. Database managers are usually always available if this happens. Here's a nice list of why databases can crash.



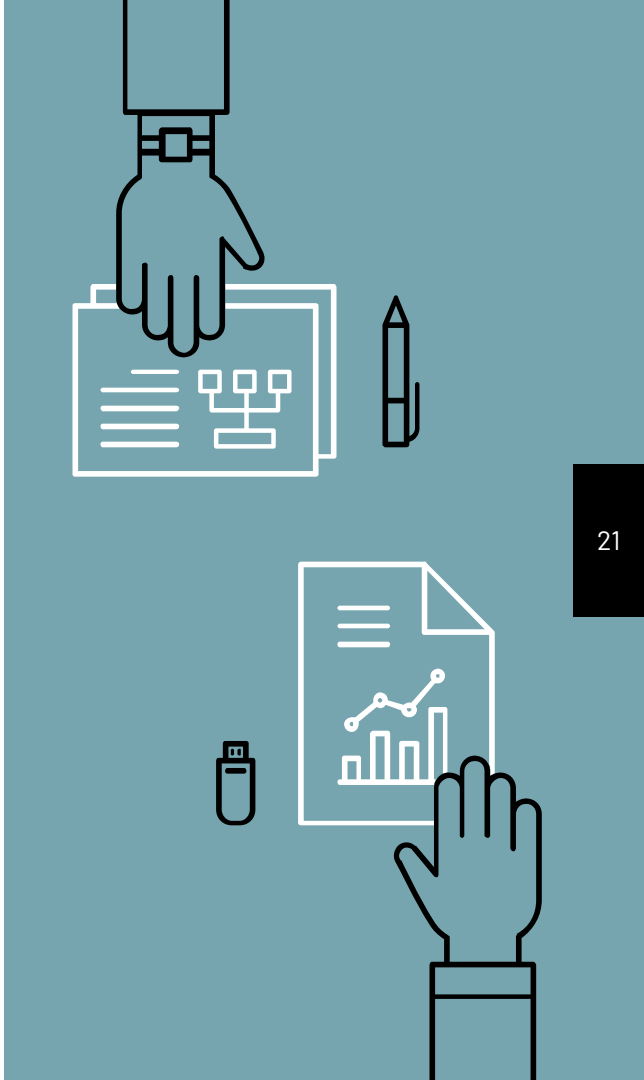
Module 6 Part 1 Discussion

Would you suggest taking data management courses or is that something you learn through experience?

- Only if you will be a data engineer/managing data a lot in the future. If you will occasionally pulling data from a database I would suggest learning SQL basics. The fundamentals are pretty easy to learn and a few commands will go far. You can watch free videos online.

How much time typically do you end up splitting between organizing all the data and files, understanding the field/context you're working in, and actually doing analyses?

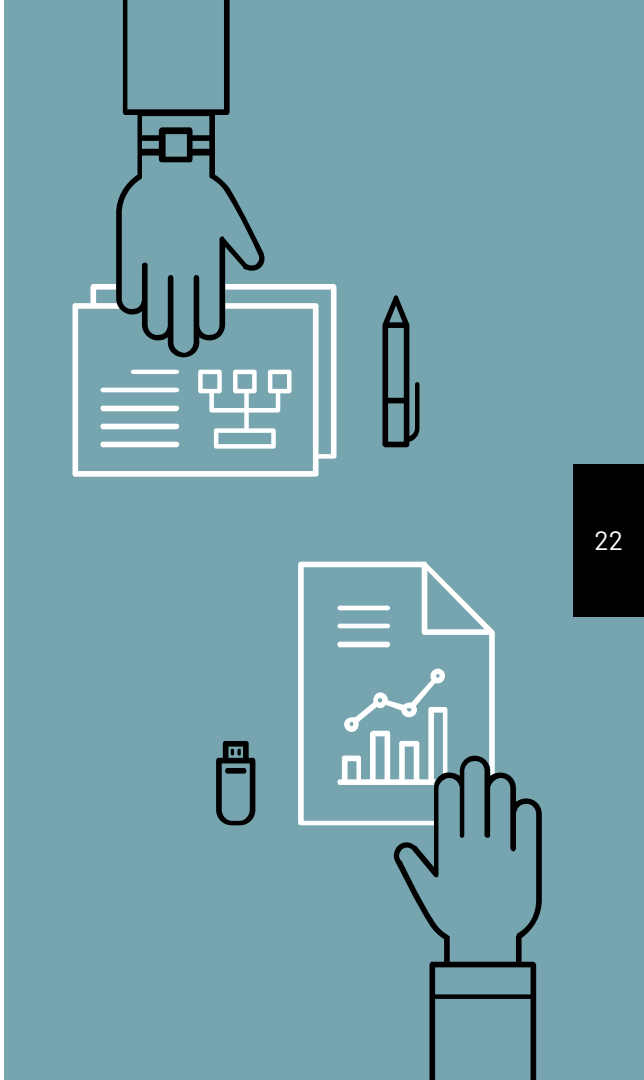
- When I start a project I spend a lot of time discussing the goals/purpose of the project with my team and doing a lit review and overview of what data is available. I keep notes during each meeting and update any project progress on a team Confluence page. When I start to code I organize my files as I go. This saves me a lot of time in the long run. I spend most of my time cleaning the data, exploring the data and figuring out what kinds of analyses are appropriate. I spend the least amount of time on actually running the analyses. All of the prep work before the analyses takes up ~60-70% of my time. Data analysis ~10% and writing it all up ~20-30%.



Module 6 Part 1 Discussion

Merce Crosas's discussion of the challenges of big data got me thinking about the use of cloud computing in data science/computational research. Is it common for researchers in biostatistics to use cloud computing environments (for example, AWS) when developing new statistical methods? Are there any resources at Harvard for learning more about this?

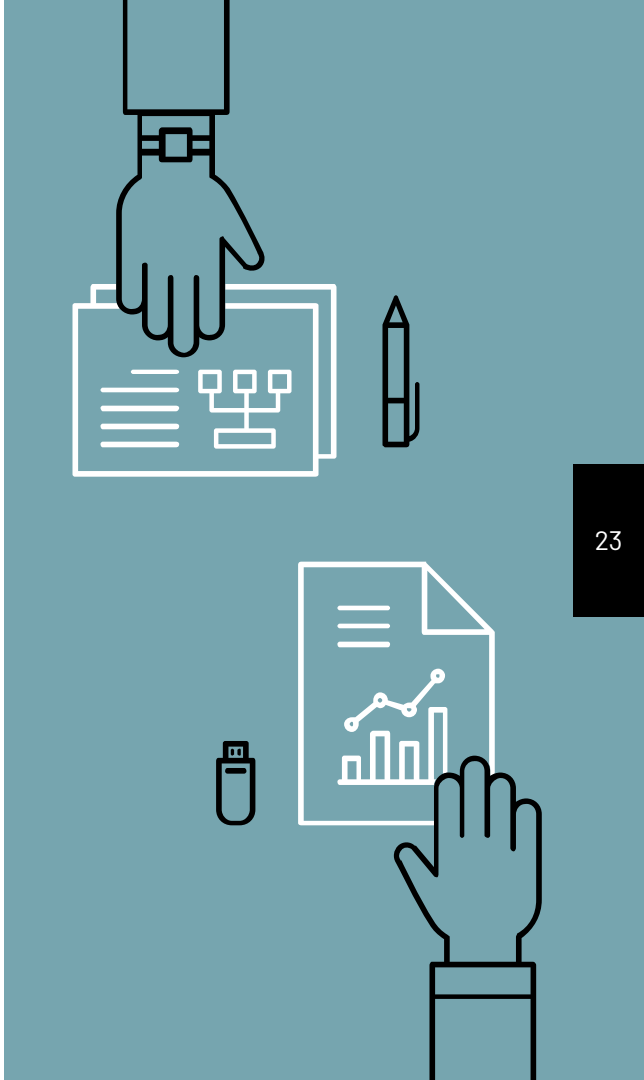
- I'm not sure I would say *common* - yet. I think more and more researchers are moving to cloud computing as time goes on. Harvard now has AWS cloud resources that can be used by Harvard researchers. I think Harvard is also dismantling one or more of their clusters but I could be wrong - that's something I heard they were considering a couple of years ago.



Module 6 Part 1 Discussion

I know many research groups (and companies) have collections of code that people have made over the years to make figure/table formats, do basic statistics, data cleaning, etc. Is it common practice to make these public? (For my last job, we would not have been able to publish our analysis code without also publishing our group codebase). Is there any reason not to publish this code?

- I don't think this is very common since most (to my knowledge / in my experience) researchers format everything later on or don't have code they reuse that often, but some do. There are packages in R that create really nice LaTeX tables, for example, and that code could be published with the rest of the code for the paper.



Module 6 Part 1 Discussion

The DataTags project details 6 DataTags levels. Can you provide some examples of these 6 levels (especially how to differentiate between blue and green)? And how do these 6 levels correspond to the 5 levels by Harvard University Information Security Policy?

- Check out Merce Crosas's slides

DataTags Levels

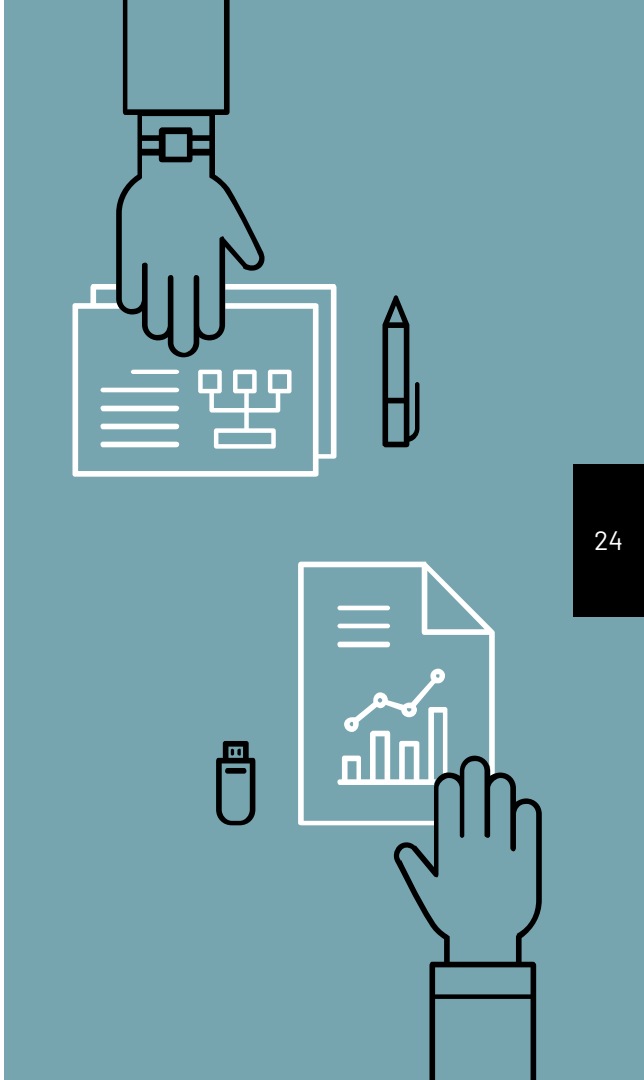
Tag Type	*Description*	*Security Features*	*Access Credentials*
Blue	Public	Clear storage, Clear transmit	Open
Green	Controlled public	Clear storage, Clear transmit	Email- or OAuth Verified Registration
Yellow	Accountable	Clear storage, Encrypted transmit	Password, Registered, Approval, Click-through DUA
Orange	More accountable	Encrypted storage, Encrypted transmit	Password, Registered, Approval, Signed DUA
Red	Fully accountable	Encrypted storage, Encrypted transmit	Two-factor authentication, Approval, Signed DUA
Crimson	Maximally restricted	Multi-encrypted storage, Encrypted transmit	Two-factor authentication, Approval, Signed DUA

DataTags and their respective policies

Sweeney L, Crosas M, Bar-Sinai M. Sharing Sensitive Data with Confidence: The Datatags System. Technology Science. 2015.

DataTags vs Harvard Security Levels

Blue	**Level 1:** No sensitive data; open data
Green	**Level 1:** Low risk de-identified data
Yellow	**Level 2:** Confidential information by University standards; no material harm
Orange	**Level 3:** Confidential information that could cause material harm (non-level 4 FERPA)
Red	**Level 4:** High-risk confidential information (SSN)
Crimson	**Level 5*** (Level 4.5, on the network) Information that would cause severe harm



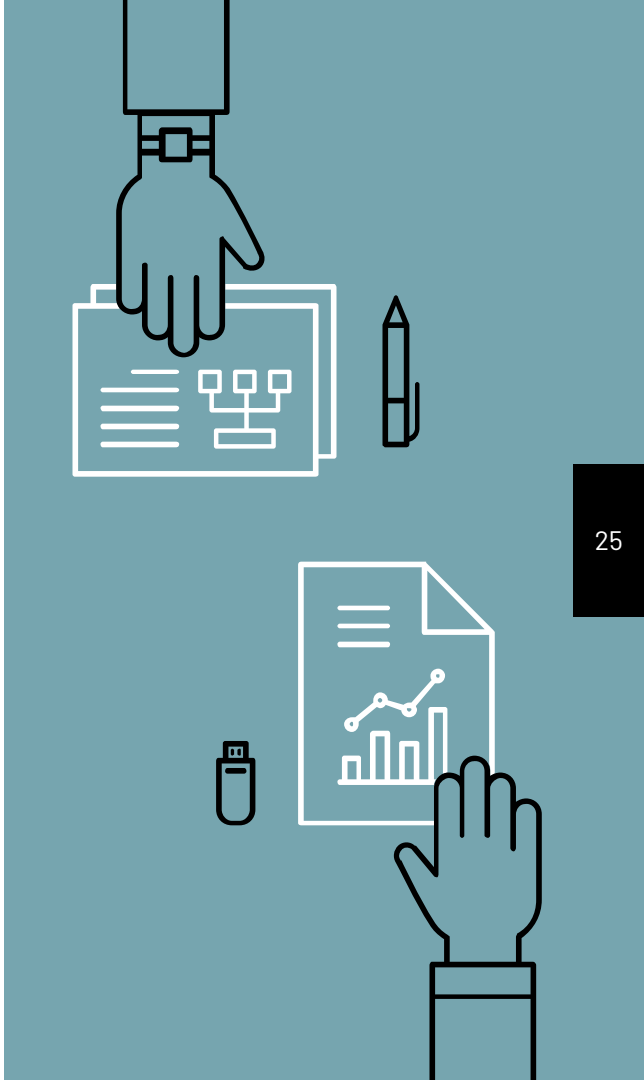
Module 6 Part 1 Discussion

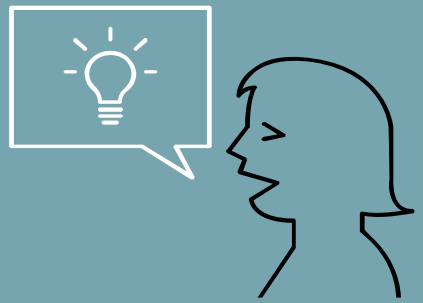
It is not uncommon for the analysis of a paper to be carried out with commercial software like Matlab or Stata. In these scenarios, even if the authors make their code available, there are likely to be researchers wanting to reproduce the paper's results that are unable to do so because they don't have the necessary software. Do any journals have policies that address this? e.g. by requiring a descriptive methods document one could follow.

- I don't think journals think about this and assume everyone interested in reproducing the analyses or using the code in some way has access to software like Matlab or Stata - or that they can convert the code using R or Python, which are freely available.

In scenarios where a team of authors is writing a paper while conducting their analyses in R, what are the relative merits of producing the paper directly in R Markdown vs. maintaining a tex file and conducting the analysis in a separate R file? Are there scenarios/types of projects where one approach makes more sense than the other?

- I think a couple of things come into play:
  - The length of the file. If you have a very long file for analyses, it may not be worth it to try to also use that document to produce the paper.
  - The experience of the team. I'm not sure how many researchers use R Markdown to produce a paper; I think the majority would feel more comfortable using a tex file or overleaf and copy and pasting from R.





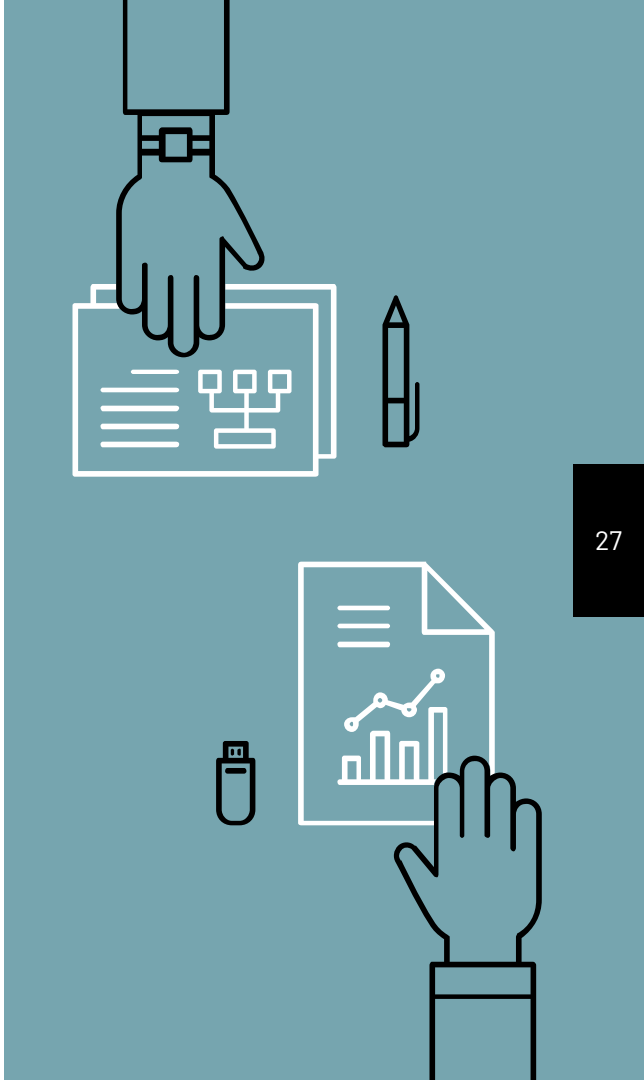
In-Class Project

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 1-3
- Critique reproducibility

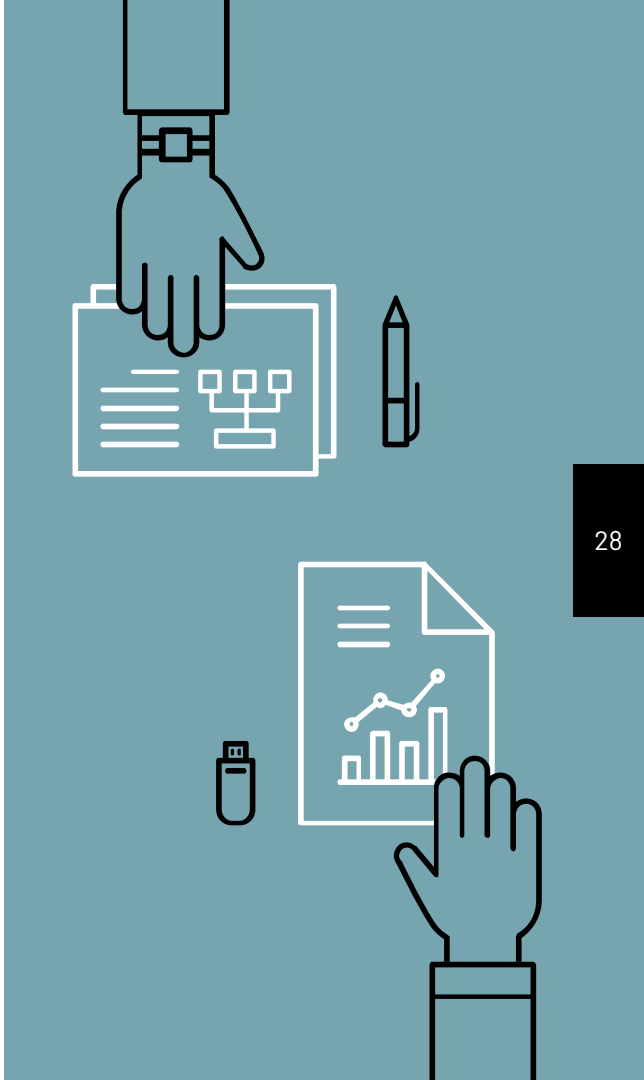


We will attempt to reproduce the results from and critique the reproducibility of:

[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 1-3
- Critique reproducibility



Homework

- Watch Module 6 part 2 videos
  - 6.3.11 - 6.5.1

- Submit Module 6 part 2 discussion points