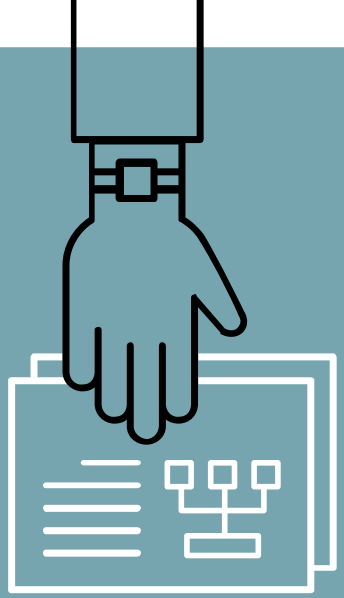
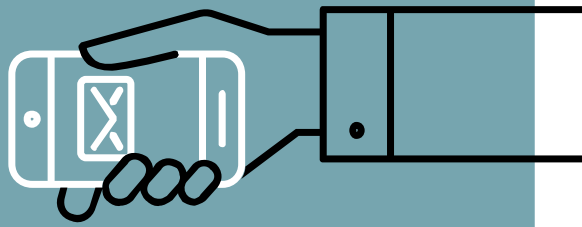
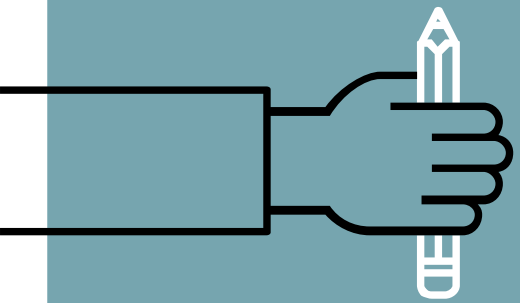
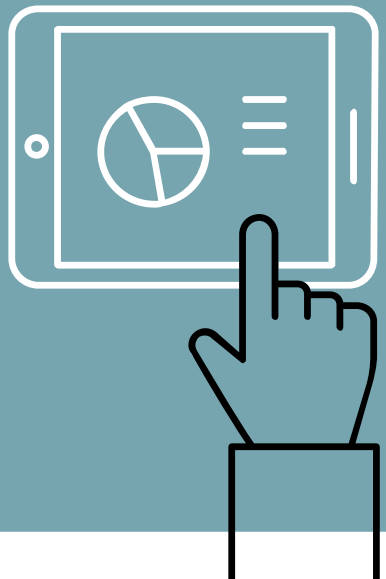




BST 270
Reproducible Data Science

Winter 2022
Session 7





Module 6 Part 2 Comments

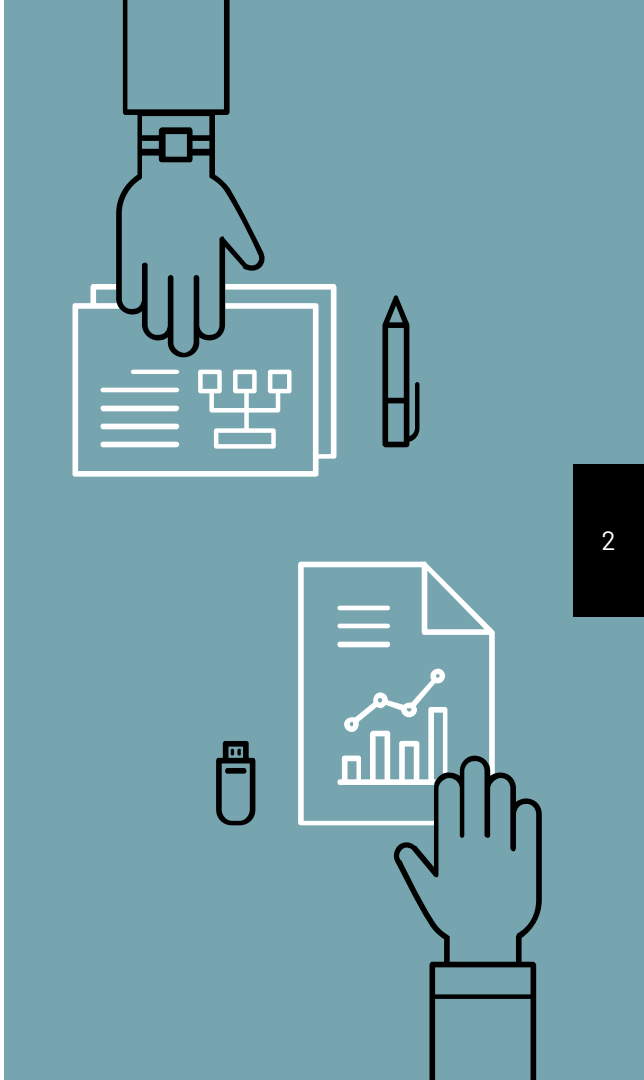
I like how user-friendly the Dataverse looks. Given the enormous amount of data on the Dataverse, I would not have expected the organization to be this neat.

Dataverse is a powerful tool for reproducible data science. It also reminds me that it might be a great resource when we work on some course projects when we need to find data sets on our own.

Using the cache=TRUE flag in R Markdown is a lifesaver, and has saved me a considerable amount of time in the past.

I love the cache option in Rmarkdown, which saves me a lot of time when some particular chunks of code are slow to run. However, it may leave behind very large cache files.

Notebooks seem like a really great way to promote reproducible data science since you can just run the whole notebook (or rmd file) to reproduce results.



Module 6 Part 2 Comments

I was amused when I found out what YAML stands for, and slightly peeved that they chose the Y in YAML when there were so many options.

I also like to occasionally use a system called Org for literate programming: https://orgmode.org/. It integrates well with any programming language (including R, Python, etc.) has good LaTex support, and you can export to PDF, HTML, etc.

Sonia's presentation is very clear and thorough, great tutorial on Harvard Dataverse. One great advantage of Harvard Dataverse compared to Github is that in Dataverse metadata, dataset, reports, and other supplementary materials are clearly labeled at the time when the repository is created so that those who wish to access the dataset will not feel overwhelmed, while in Github all the files are juxtaposed.

Module 6 Part 2 Comments

It's interesting to think about researchers decades down the line looking back at COVID data, especially in the beginning, and whether or not the data is of enough quality or has enough information about it to be used without asking questions from the data contributors who may be deceased.

Something I found interesting when working in Institutional Research was that the hosting university for the shared consortium data is always a private university, because the idea is that any data hosted by a public university would be inherently government-owned.

I have never thought about how data was managed in the past so it was surprising for me to hear Dr. Barbosa talk about her experience with data management about 20 years ago. And it is quite stunning to see how that has evolved to the digital eras that we have here right now.

Module 6 Part 2 Comments

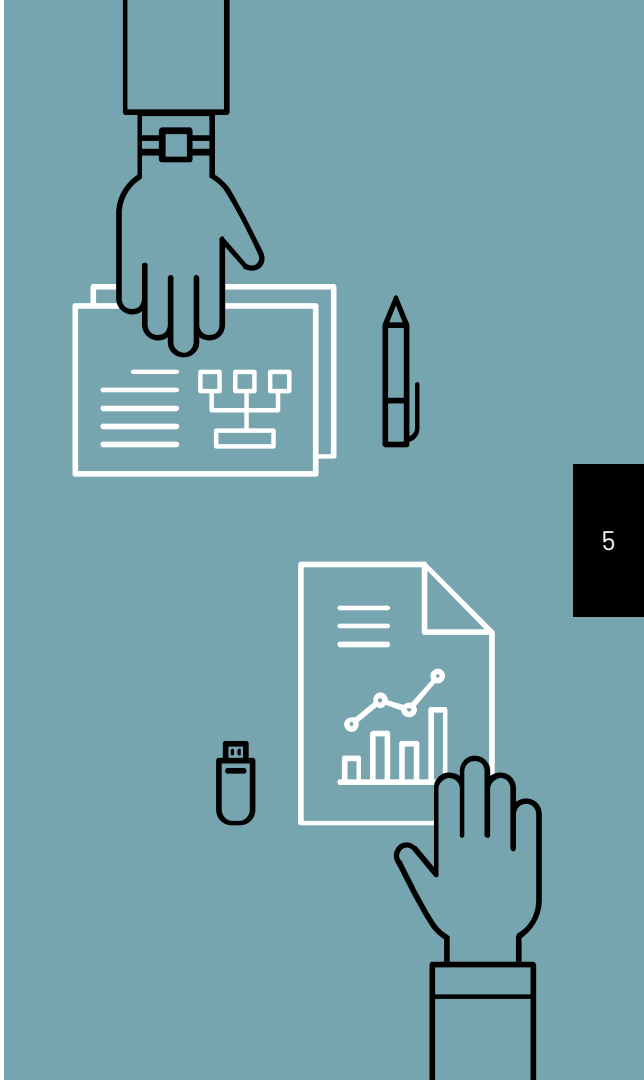
I thought the note that once the data is published on the dataverse it cannot be unpublished was important. This follows publishing guidelines and we should get into the habit of double checking things and making sure code and data are clean and reproducible before publishing.

I didn't know you can put table of contents and bibliographies in rmd files. That looks real neat.

Being able to search Harvard's DataVerse installation in R based on a very vague topic is super cool.

I am very excited about Harvard dataverse. I wasn't aware of this repository and it could be useful as I progress through the program.

These videos really make me feel good about how paranoid I used to be at my previous job as a data manager documenting everything and adding detailed comments to my code, even when it would take me longer to finish my tasks.

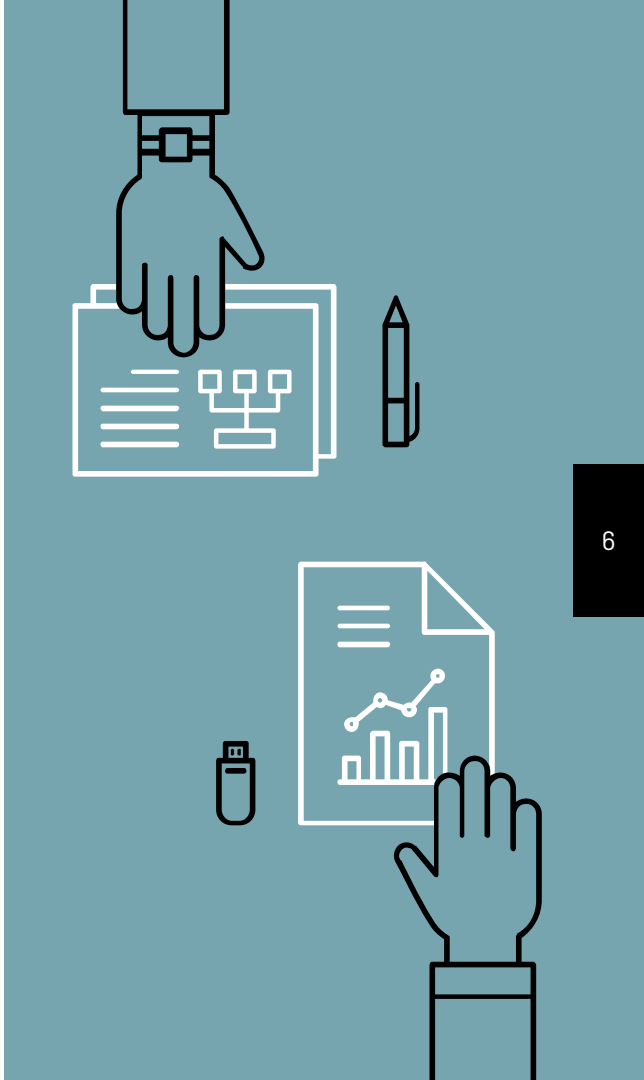


Module 6 Part 2 Comments

One aspect of R Studio that I can't live without is the ability to run individual lines of code in a script or .RMD file. My experiences with Eclipse and Pycharm, though mostly positive, still left me frustrated when I couldn't diagnose a problem by looking one line at a time. These IDEs make me run the entire script at once, which requires lots of commenting. I know that Jupyter notebook lets you run individual code, but it's still not the same as with RMD files.

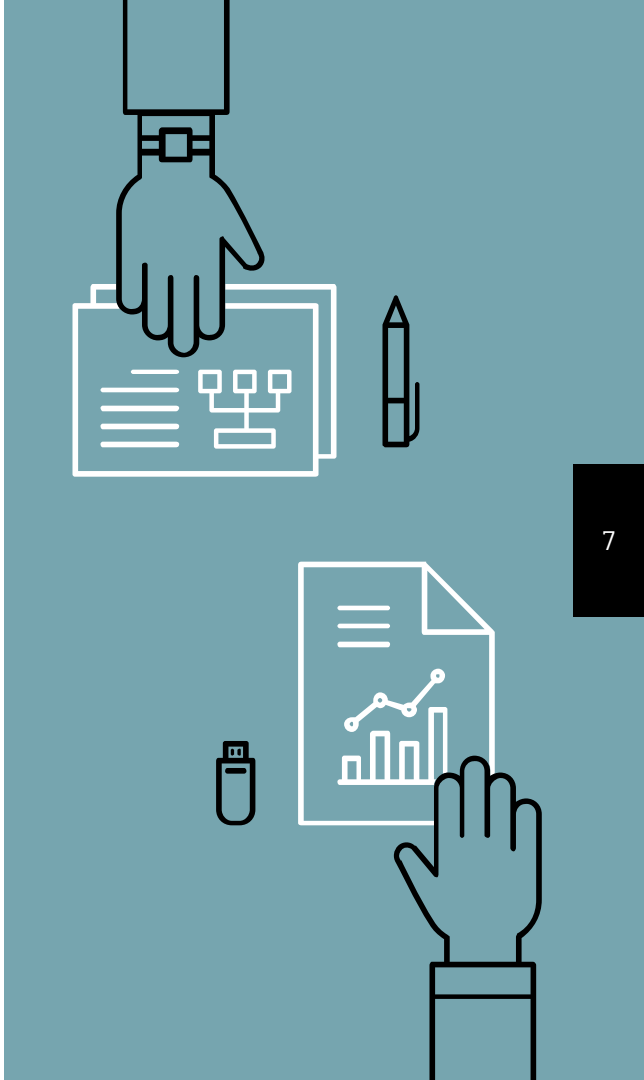
Will entire degree programs be devoted to data engineering? There seems to be many tools out there that people could spend a whole semester learning about. The future is now!

Watching the video with Sonia Barbosa, took me back to when I used to collect data using paper questionnaires, and then enter all the data into a database. This was so tedious especially when we had to look at codebooks at the same time to make sure the data was entered correctly. This is to say that I really appreciate the efforts and the use of software to help making the data management process flow more smoothly.



Module 6 Part 2 Comments

I find the introduction of Rmarkdown by Dr. Gandrud very useful. Though I use Rmarkdown almost every day, it is my first time to learn some of the tools and specifications that were introduced. It is also my first time to learn how Python code can be executed in R environment.



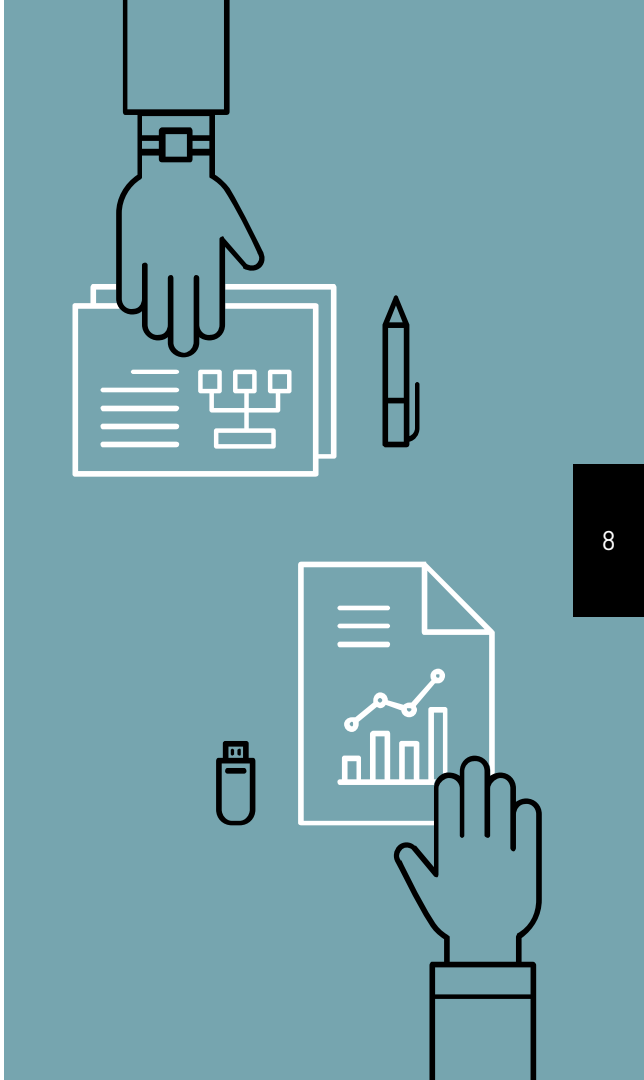
Module 6 Part 2 Discussion

Notebooks seem to require a GUI system to interact optimally while make files do not (can use terminal/shell). Make files seem game changing too, but tricky to manage. What are some more complex examples of make files and what are the benefits over managing in make vs several notebooks?

- I think Makefiles are best for building a pipeline and executing code. Notebooks are good for all data processing and analyses, but executing them one-by-one is time consuming. Makefiles can get complicated, but it does automate the process which can save a lot of time. It also deals with dependencies which can be a lifesaver.

Is there any community preference for exporting R Markdown files as PDF vs HTML? Is either of the two more reproducible or portable than the other? I generally default to exporting as a PDF, but I wonder if the dependence on LaTex makes it more difficult to compile across different systems.

- Exporting to a PDF will not work on someone's computer if they don't have LaTeX set up, so that does hinder reproducibility. If it outputs to HTML, anyone can knit it and/or open the HTML in a browser. I personally use HTML more often because of this, and because the formatting usually looks better.



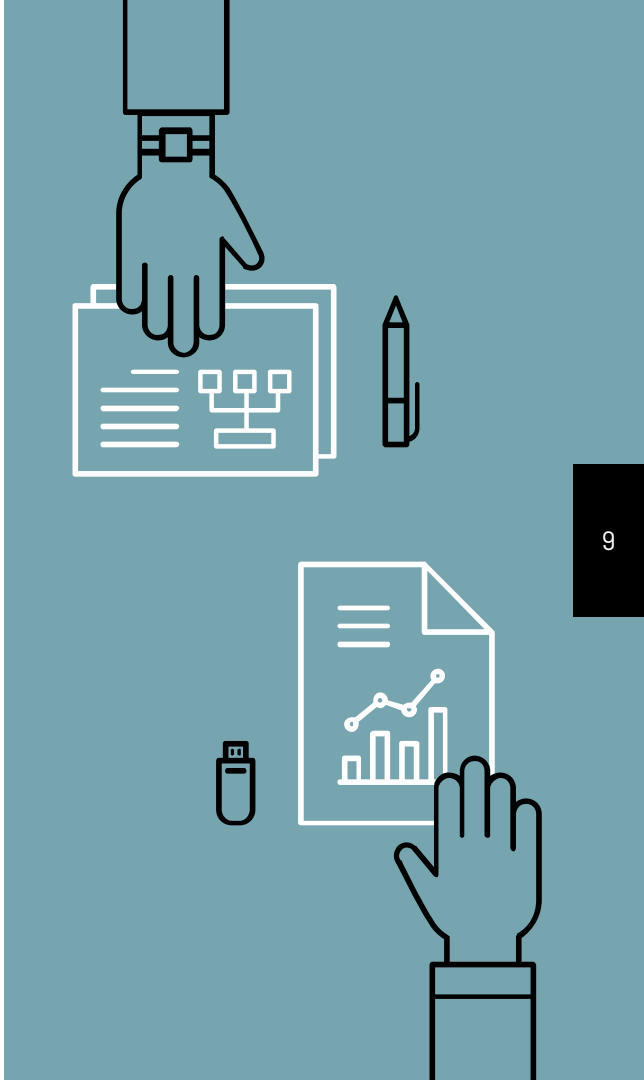
Module 6 Part 2 Discussion

Thoughts on Christopher Gandrud's suggestion to store all data and code in plain text files?

- In my experience, plain text or csv files are preferred because they can be used with (almost) any language/software easily. For example, .R data is very specific to R and .dta is specific to Stata, and if you don't have those software set up on your computer you won't be able to read/open them unless you have another type of software that can convert them into a readable format. A lot of README files are plain text files so anyone can open them and see any information needed about a particular study/analysis/data.

Sonia mentioned in the video that CC0 means when someone gets access to the data he/she can put it in their own website and make it available while the owner of the data is no longer able to track the subsequent downloads directly on Dataverse. However, I wonder how Dataverse can allow owners to track such 'secondhand' downloads of their data when the data is not CC0?

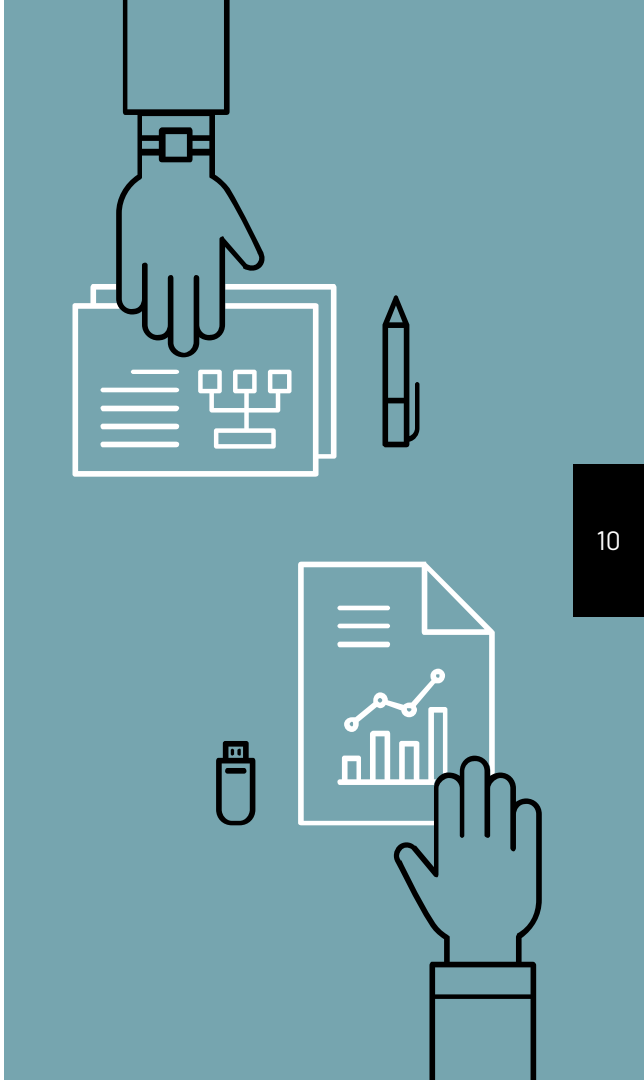
- I believe Dataverse integrates Make Data Count. Read more about it here.



Module 6 Part 2 Discussion

I'm getting a bit lost on the pros/cons for the different platforms for archiving data. Compared to GitHub, is it fair to say that the main advantages of Harvard Dataverse over GitHub are the ability to store confidential/secure data, some additional features for particular types of data, and ability to store bigger data?

- From the comment on slide 3: One great advantage of Harvard Dataverse compared to Github is that in Dataverse metadata, dataset, reports, and other supplementary materials are clearly labeled at the time when the repository is created so that those who wish to access the dataset will not feel overwhelmed, while in Github all the files are juxtaposed.



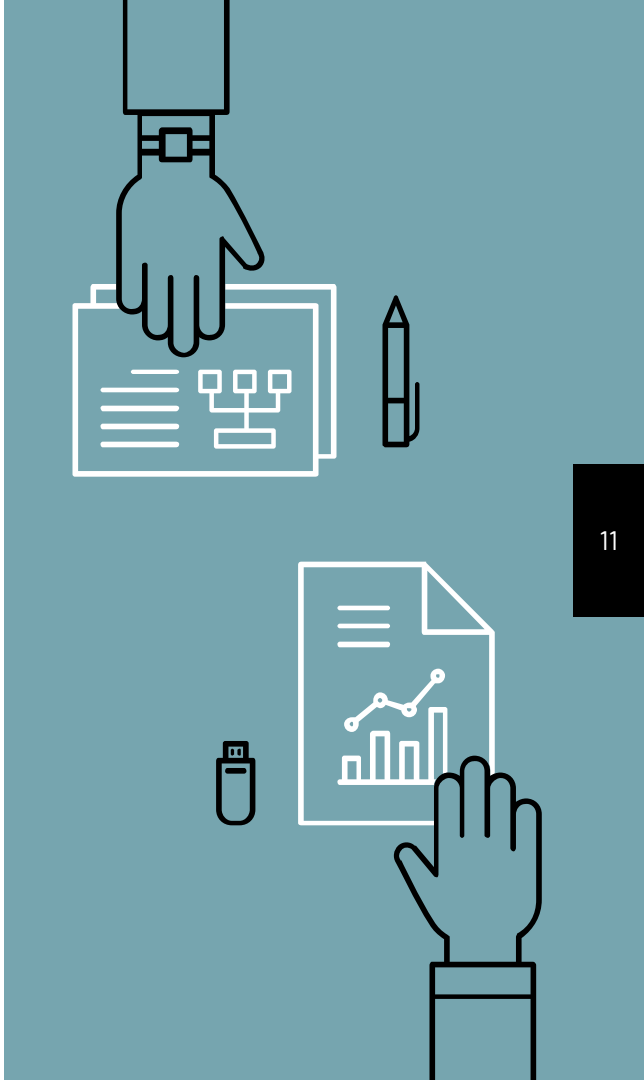
Module 6 Part 2 Discussion

For those sharing data on a platform like Harvard Dataverse, why would they reject someone's requests for accessing their data? The videos mention that authors may ask for further clarification before deciding whether they will grant access to someone.

- It may be because of the sensitivity of the data. There may be restrictions on who can access it.

The dataverse and other data sharing services seem tailored to one-off studies that use a set of data. Are there systems for sharing data that is continuously updated (e.g. NYT covid data)? There was a paper that was accepted into Nature Data promising to continuously update and report line list covid data and (surprise, surprise) as soon as the paper was accepted the group stopped updating the data. Is academia just poorly suited for continuous data sharing?

- I think in some cases it is. I think a big issue is turnover. Students and postdocs or even people in industry positions don't stay in those positions for a long time and it can be difficult to keep things up and running smoothly, especially if you have a larger team. Some teams are really good about it but most I've seen are not.



Module 6 Part 2 Discussion

I found the conversation on Docker really interesting, is that something that we would have access to when submitting a paper? It seems like the best way to make sure code works on everyone's different system.

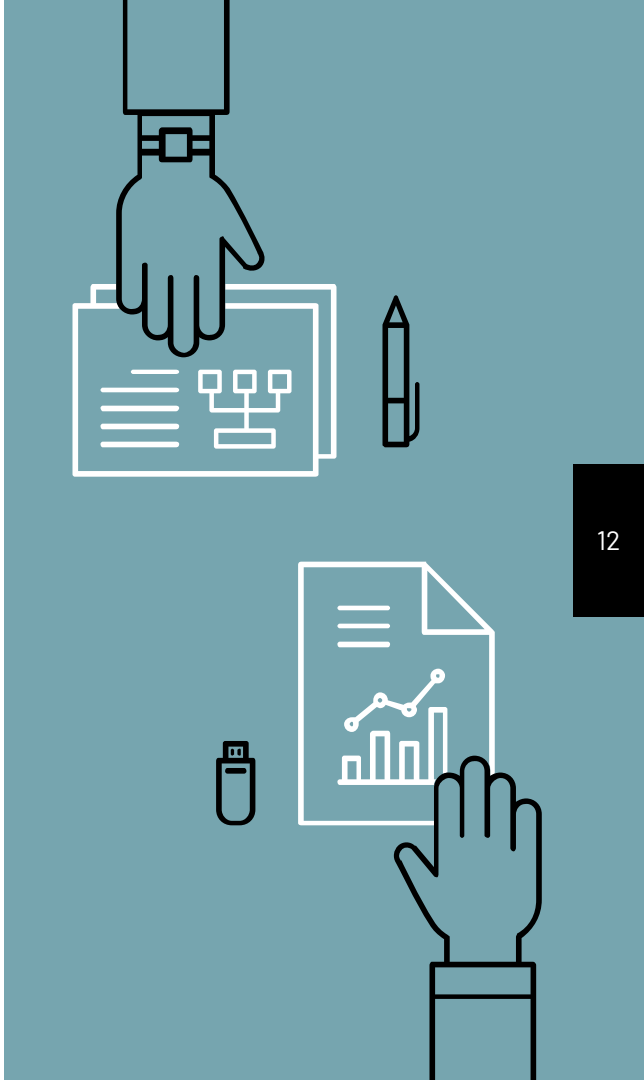
- It's something anyone can have access to at any time in the research stage.

Is there a usable equivalent to google docs but for code? I have heard of some ways to implement this when collaborating with code, but wondered if there were any others that are popular in practice.

- Here's a good resource. I personally haven't used any of them yet.

Does the Dataverse have any stats on how often the reviewers who are sent a private link to the raw data actually look at the raw data?

- No idea but I'm sure there has to be something that tracks that.



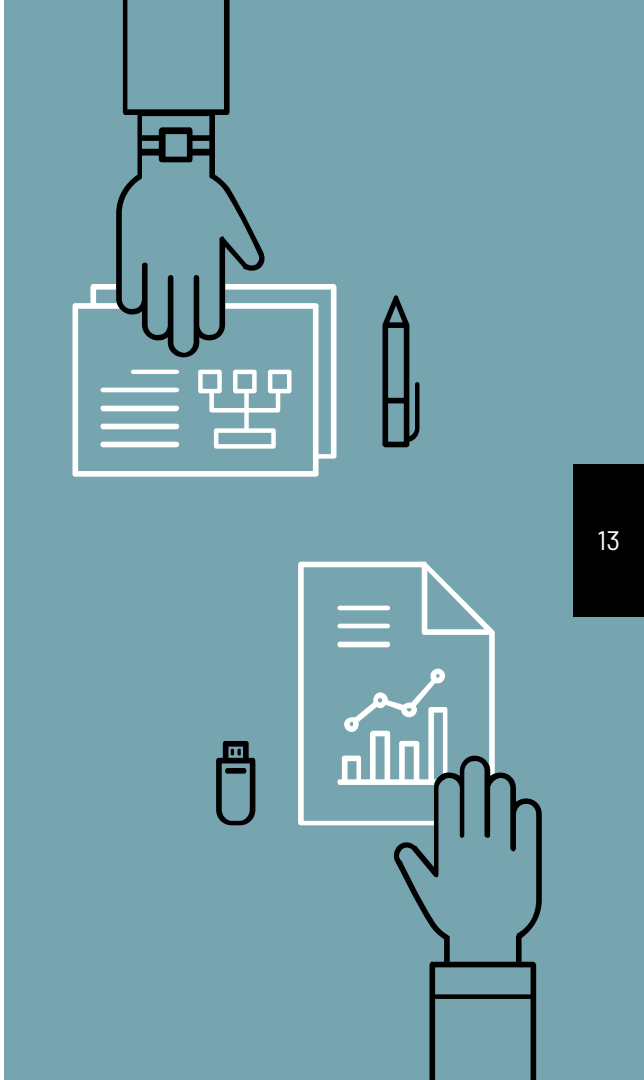
Module 6 Part 2 Discussion

From the video where Dr.Choirat showed how to connect dataverse to R, can we do version control here the same way we can do for github?

- Yes! Functions like create_dataset and dataset_versions allow you to access and save different versions of the dataset. Check out the documentation here.

I didn't know BibTex could be used with RMarkdown! Are there any accepted "best practices" for maintaining large bib files as part of a research team to keep things from getting disorganized? In practice, do you maintain bib files over time or do you start from scratch for every new project?

- So cool, right?! Here's a great resource for it. I don't think there are any accepted best practices yet; it all depends on the team. I personally created 1 BibTex file for my dissertation and used it when publishing the papers separately. It depends on the projects I'm working on - if I will be writing multiple papers with the same refs, I'll use 1 big file for multiple projects. But if the projects are completely different they have different BibTex files.



Module 6 Part 2 Discussion

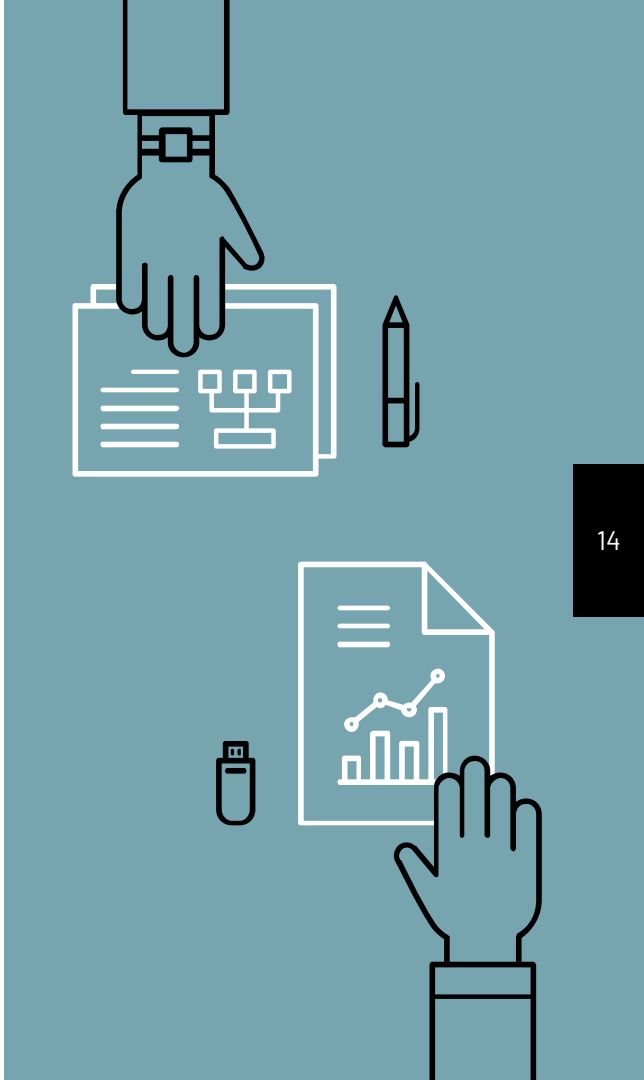
In RMarkdown, is there a way to loop through the metaparameters? For example, if you wanted to have the file automatically generate several output formats (PDF, HTML, Word doc, etc) with one calling of the file.

- In the header you can list which formats you want it to output and it will output all of them when you knit the file.

Other than the "preview" function, what is the difference between an R Notebook and an R Markdown? I have used R Markdown many times but don't seem to encounter any use cases for R Notebook.

Are they currently any major differences in R Notebook vs. R Markdown that would compel someone to use the former? R Markdown can do everything interactively now as well, which may not have been the case at the time of these videos?

- Great explanation here.



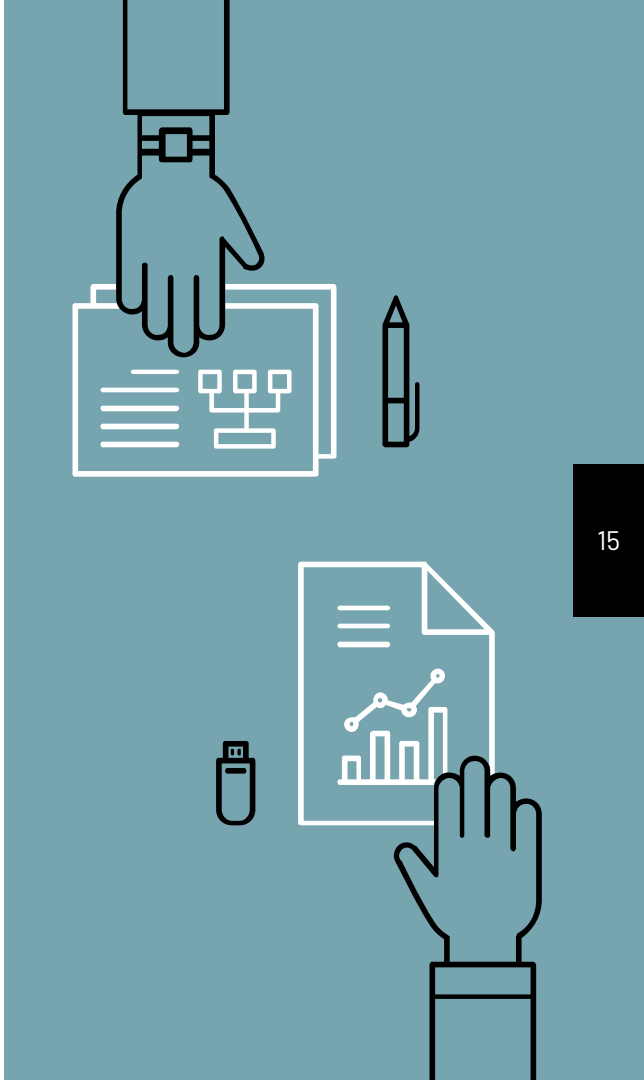
Module 6 Part 2 Discussion

Have all of the data sets in the dataverse been analyzed? More importantly, have they been cleaned?

- I have no idea if they have all been analyzed. I would lean towards yes but there's a ton of datasets on there. In terms of being cleaned, I have seen a few of them with the raw data files available so, not super clean.

What is the purpose of Makefiles vs. something like a bash/shell script.

- Make files have the added capability of dependency checking. This looks like a good resource. The short answer is that you can use a bash/shell script, but when there are complex dependencies, Makefiles do the job better.



Module 6 Part 2 Discussion

I haven't used Make before, but it sounds like such a powerful tool to automate the workflow! Are there any resources that you would recommend for people who are interested in learning Make?

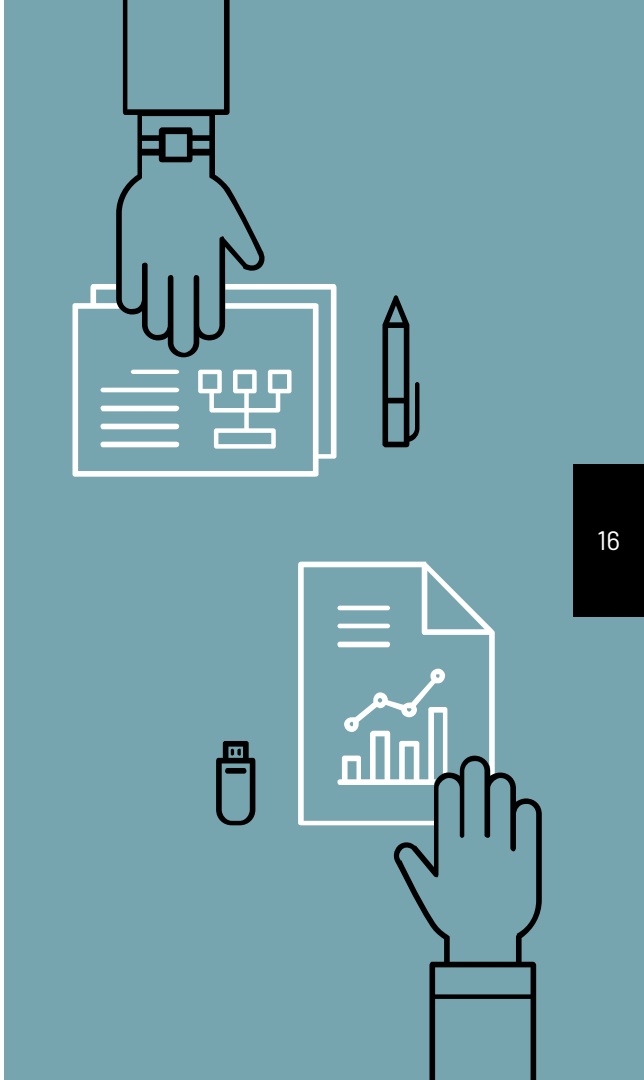
- Here's a good resource.

Can you make different seeds? Are these different or just 123?

- Yes. You can set any number as the seed.

What tools you use to manage your papers? Like in bib format?

- If I'm using LaTex, I create a .bib file in Overleaf. If I am forced to write the paper in Word, I use EndNote.



Homework

- Watch Module 6 part 3 and Module 7 videos
  - 6.5.2 - 6.6
  - 7.1

- Submit Module 6 part 3 discussion points

- Individual Project