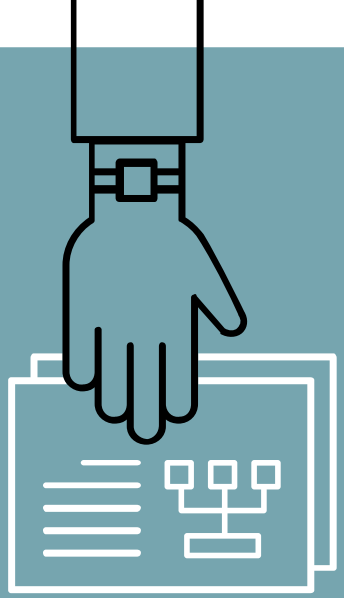
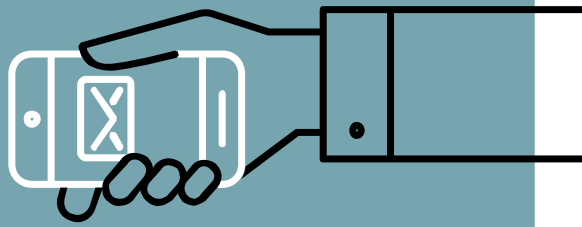
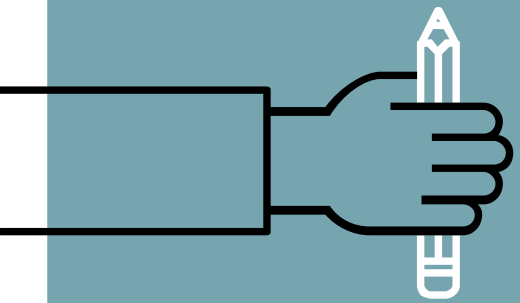
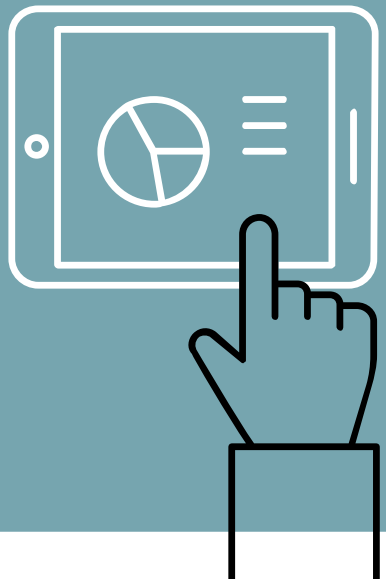




BST 270
Reproducible Data Science

Winter 2022
Session 4



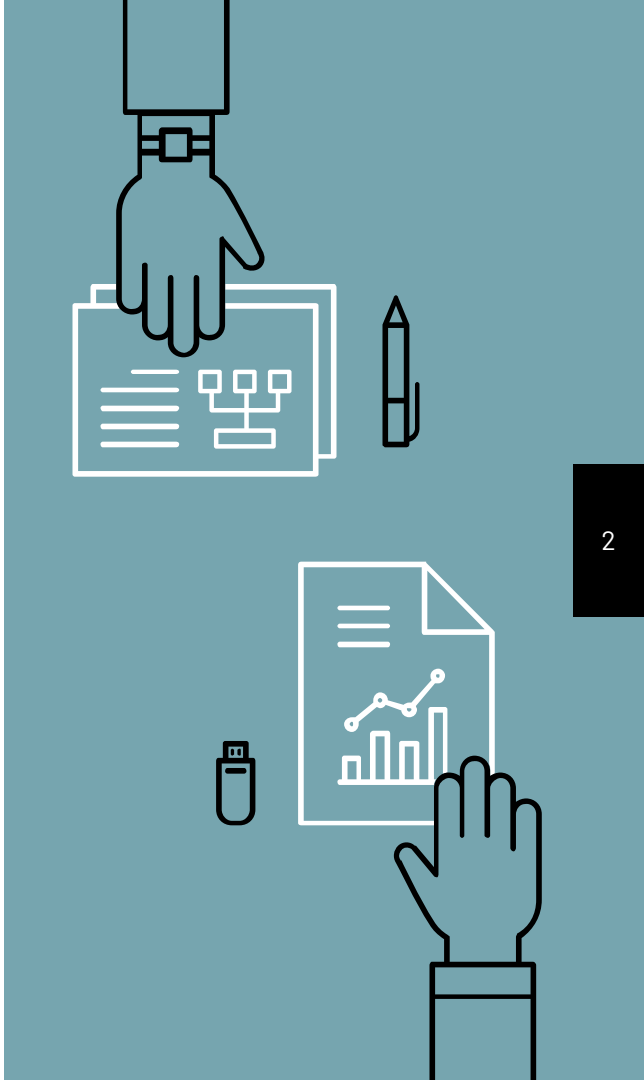


Module 4 Comments

Well documented data is really crucial for reproducible data science (not only the labels but also the names of the variables themselves), especially after going through the data wrangling part of the in-class project.

For the final project of one of my courses last fall, I used a COVID-19 rumor dataset stored on Github which consists of both the code for analysis and the raw twitter rumor text data itself. This serves as an example of GitHub being used for reproducibly creating and sharing an analysis code AND storing data as well, particularly the case for small text or text-like data sets that might frequently change.

Something I often forget to do when running code in RMarkdown is to make sure the code still works after clearing out the environment and/or closing out and reopening — sometimes I think I've successfully wrangled a dataset but it turns out I was really just still relying on cached objects from other attempts.



Module 4 Comments

One tip I learned to make it easier to keep up with documentation was to minimize the repetitive, annoying parts of it. For example, having a keyboard shortcut that adds a program header to code that can be filled out with the relevant information or 'utility' programs that mostly automate otherwise time-consuming processes such as creating codebooks and data dictionaries for most data sets.

I love the point about avoiding sending documents as attachments. I'm sure we've all experienced receiving feedback from multiple people on separate copies of the same document and then having to combine all of the comments into one version before responding (a nightmare!) I'm definitely going to think about using Google Docs, or something similar, for collaborative manuscripts moving forwards.

I never considered that including age instead of birthdate or days from start of study instead of date of procedure was a way of de-identifying data. I assumed it was just the raw information that had been collected.

Module 4 Comments

It was interesting to learn from Curtis that software development team spend 40% of their effort on specification and design before any code is written, another 20-30% on testing, 10%-20% for production and support, and very little time/effort on the actual coding part.

I didn't know that journals publish data sets (as well as some of the other outlets mentioned) that aren't accompanied by an article per say.

One side point in the videos that I found interesting was -- For genetic data, a research participant agreeing to share their data is also implicitly sharing their relatives' data in some sense as well. Makes you rethink the implications of data sharing in this context.

Module 4 Discussion

Choose which tier (1-5) the following example data sets would fall into according to the Harvard University Information Security Policy.

A. Identifiable medical records that include age, sex and mental health details of all patients at Mass General Hospital.

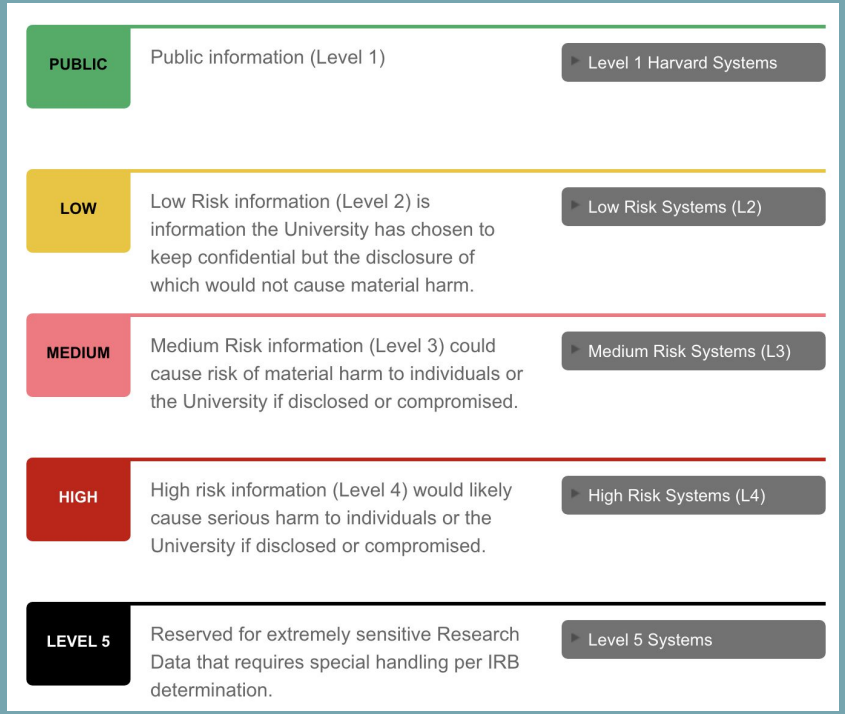
B. A data set including the date, time, duration, location and category of tornadoes in Oklahoma in 2011.

C. Identifiable financial records including credit card numbers and social security numbers of people living in Los Angeles, CA.

D. The de-identified employment history of all adults 18-70 in Massachusetts.

E. A data set detailing microorganisms used for not yet published research.

Harvard University Information Security Policy



Module 4 Discussion

Choose which tier (1-5) the following example data sets would fall into according to the Harvard University Information Security Policy.

A. Identifiable medical records that include age, sex and mental health details of all patients at Mass General Hospital. Level 5

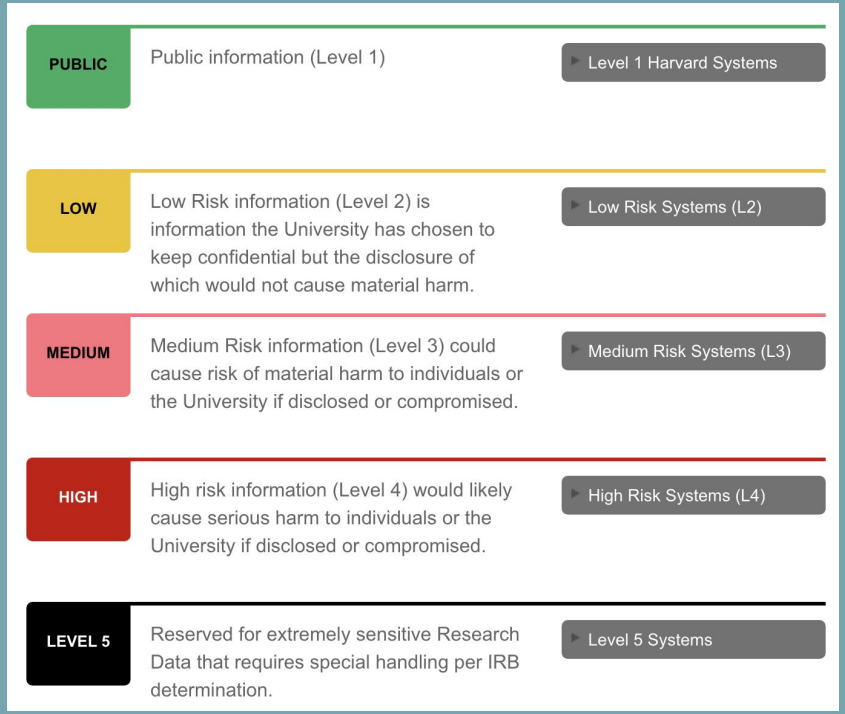
B. A data set including the date, time, duration, location and category of tornadoes in Oklahoma in 2011.

C. Identifiable financial records including credit card numbers and social security numbers of people living in Los Angeles, CA.

D. The de-identified employment history of all adults 18-70 in Massachusetts.

E. A data set detailing microorganisms used for not yet published research.

Harvard University Information Security Policy



Module 4 Discussion

Choose which tier (1-5) the following example data sets would fall into according to the Harvard University Information Security Policy.

A. Identifiable medical records that include age, sex and mental health details of all patients at Mass General Hospital. Level 5

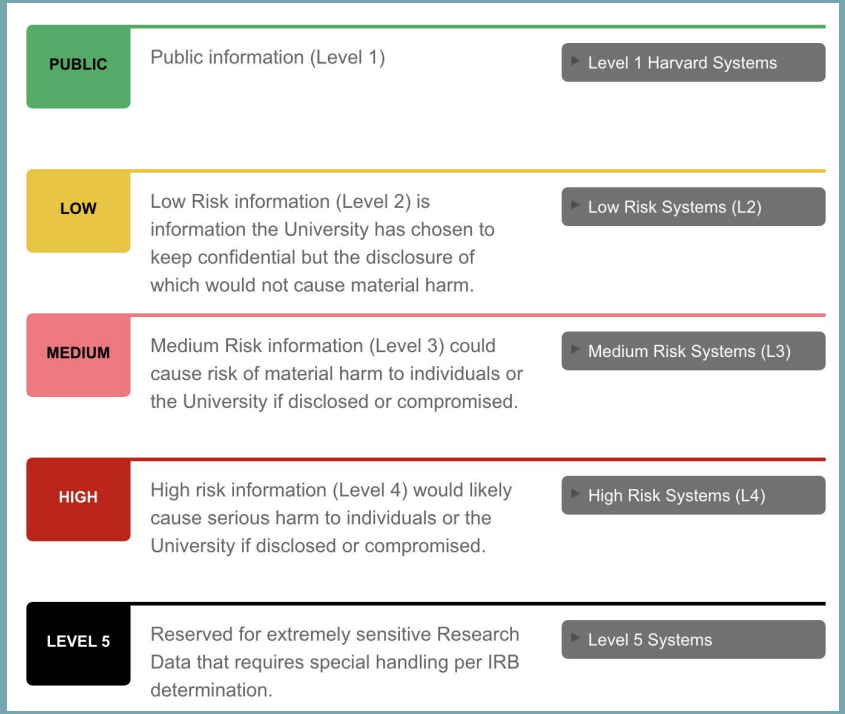
B. A data set including the date, time, duration, location and category of tornadoes in Oklahoma in 2011. Level 1

C. Identifiable financial records including credit card numbers and social security numbers of people living in Los Angeles, CA.

D. The de-identified employment history of all adults 18-70 in Massachusetts.

E. A data set detailing microorganisms used for not yet published research.

Harvard University Information Security Policy



Module 4 Discussion

Choose which tier (1-5) the following example data sets would fall into according to the Harvard University Information Security Policy.

A. Identifiable medical records that include age, sex and mental health details of all patients at Mass General Hospital. Level 5

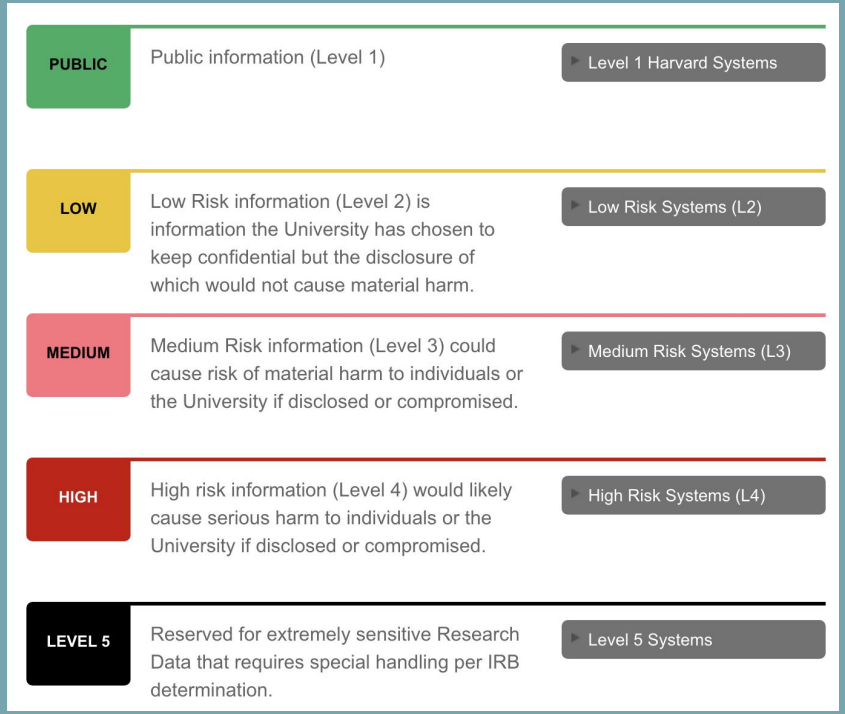
B. A data set including the date, time, duration, location and category of tornadoes in Oklahoma in 2011. Level 1

C. Identifiable financial records including credit card numbers and social security numbers of people living in Los Angeles, CA. Level 4

D. The de-identified employment history of all adults 18-70 in Massachusetts.

E. A data set detailing microorganisms used for not yet published research.

Harvard University Information Security Policy



Module 4 Discussion

Choose which tier (1-5) the following example data sets would fall into according to the Harvard University Information Security Policy.

A. Identifiable medical records that include age, sex and mental health details of all patients at Mass General Hospital. Level 5

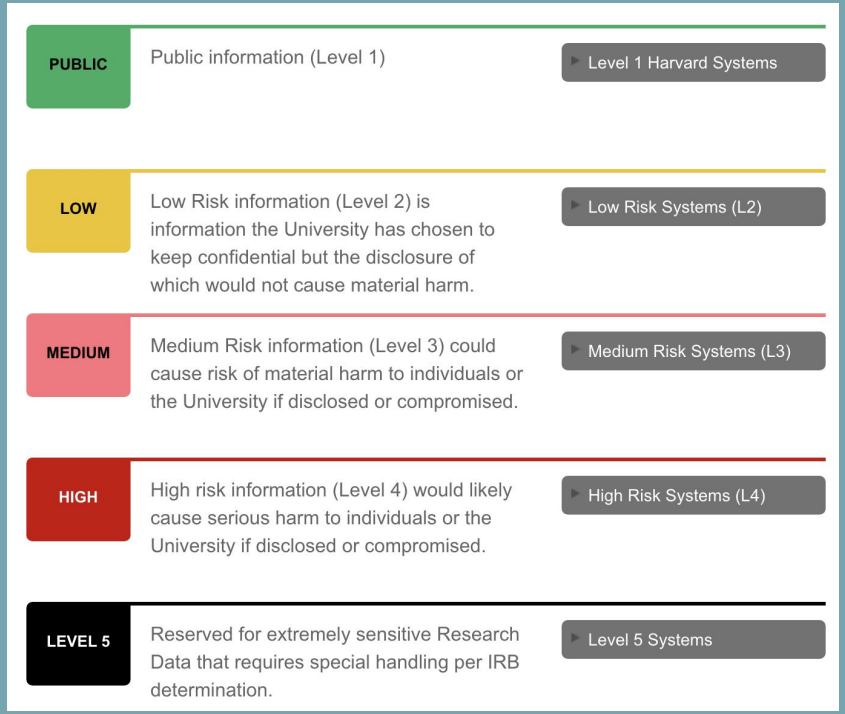
B. A data set including the date, time, duration, location and category of tornadoes in Oklahoma in 2011. Level 1

C. Identifiable financial records including credit card numbers and social security numbers of people living in Los Angeles, CA. Level 4

D. The de-identified employment history of all adults 18-70 in Massachusetts. Level 3

E. A data set detailing microorganisms used for not yet published research.

Harvard University Information Security Policy



Module 4 Discussion

Choose which tier (1-5) the following example data sets would fall into according to the Harvard University Information Security Policy.

A. Identifiable medical records that include age, sex and mental health details of all patients at Mass General Hospital. Level 5

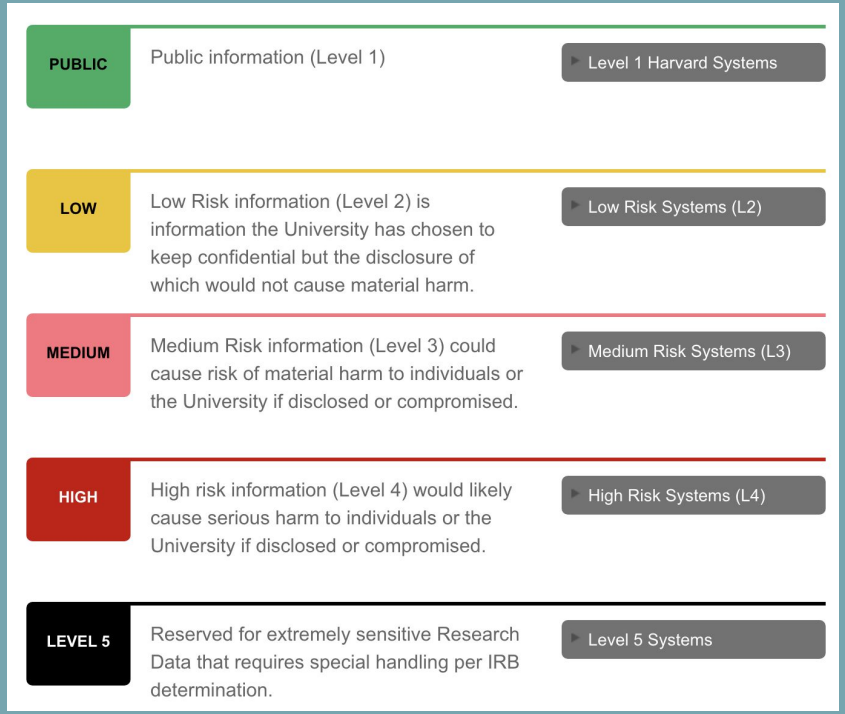
B. A data set including the date, time, duration, location and category of tornadoes in Oklahoma in 2011. Level 1

C. Identifiable financial records including credit card numbers and social security numbers of people living in Los Angeles, CA. Level 4

D. The de-identified employment history of all adults 18-70 in Massachusetts. Level 3

E. A data set detailing microorganisms used for not yet published research. Level 2

Harvard University Information Security Policy



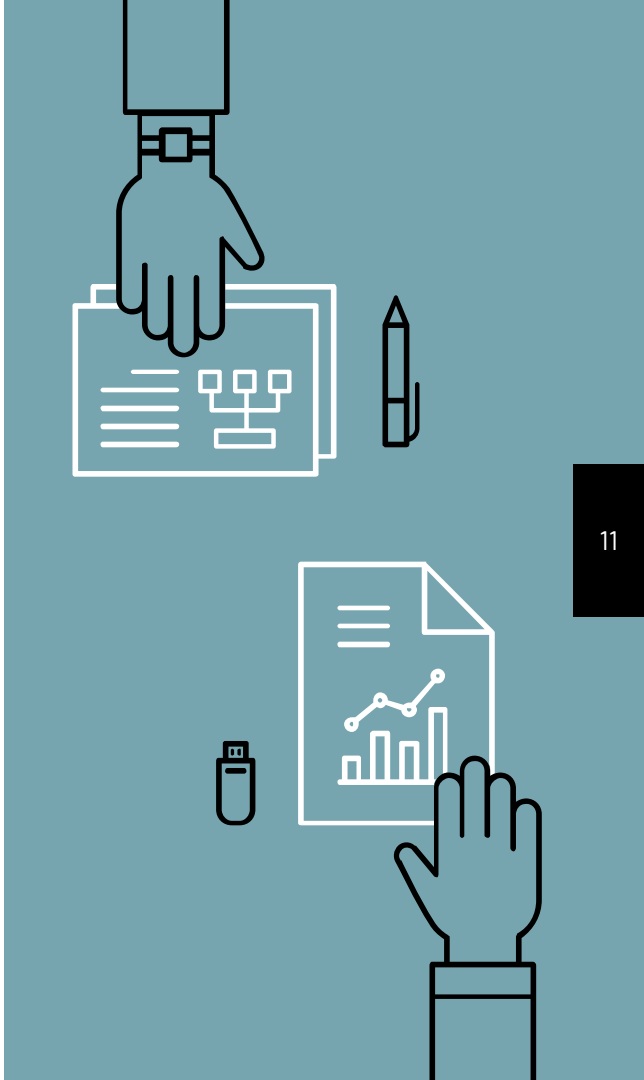
Module 4 Discussion

During the data privacy and security module, Dr. Huttenhower talks about the anonymization process for data security. With this anonymization process in mind, you mentioned during lecture that the authors of the MIMIC dataset will slightly alter some of the variable measurements to ensure data privacy. How do we know that this anonymization process does not change the results of any statistical analyses?
Can security measures, such as "Anonymization", change the data in a way that impacts the analyses? Do these procedures pose risks to the integrity of the data?

- The researchers who maintain the database do several types of checks to make sure results of analyses do not change. This isn't perfect, but a ton of work goes into this.

Have you ever dealt with any complications around private data?

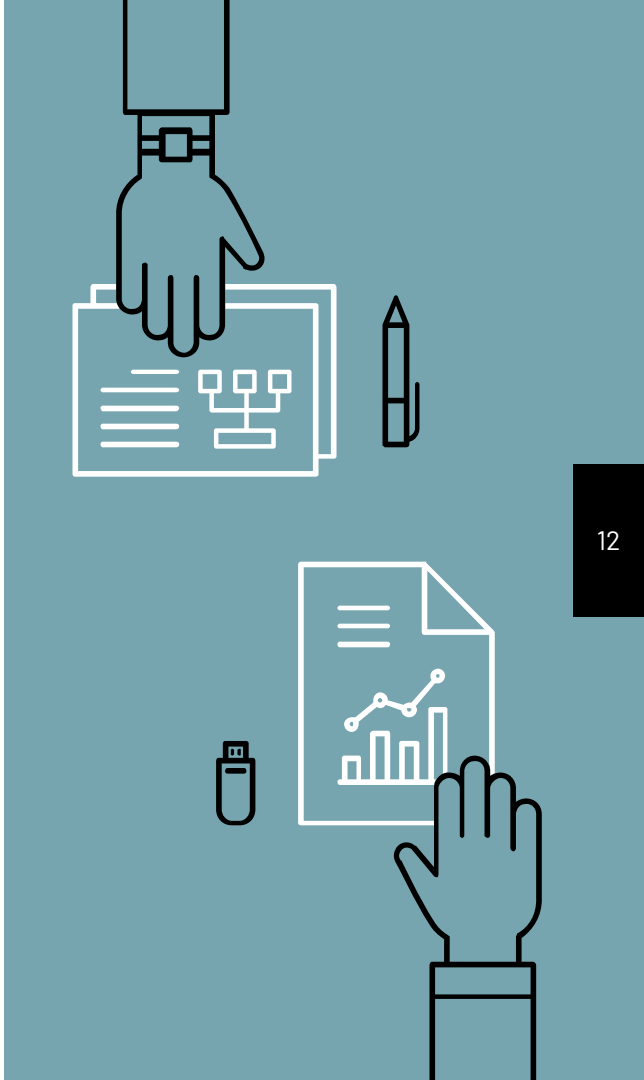
- Yes. As a data science consultant for a healthcare startup it took me 10 months to get access to medical claims data. I completed all of the required training, but my title of "consultant" wasn't deemed secure enough and my title had to be changed to part-time Data Scientist.



Module 4 Discussion

Is there any discussion to be had regarding this call for lengthier and more detailed directions/write ups for reproducibility and problems with journal accessibility? I know a fairly big issue in academia is that some journals can use complicated language that makes them hard to follow, which can cause a barrier for people to enter academia. Obviously increased requirements for including information on reproducibility and more rigorously describing your methods is a good thing, but are there any requirements, or plans to create requirements that would ensure that while these papers may need to be lengthier and more descriptive, they can still be accessible to a general audience?

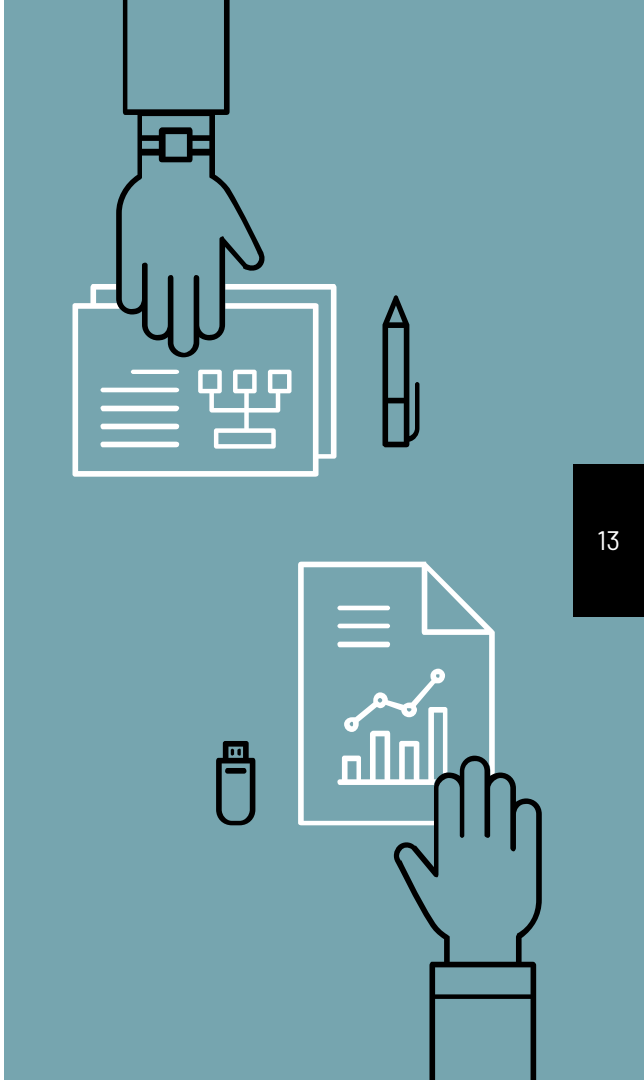
- I don't know of any requirements in place yet, but some journals do ask you to keep technical jargon to a minimum (if you can) and manuscripts that are easier to read tend to be published by better journals.
- There is a push for more "storytelling" in manuscripts that make the paper easier/more enjoyable to read, while still presenting pertinent results.
- I suggest reading Writing Science if you want to improve your own technical writing. It was life changing for me.



Module 4 Discussion

The "Tools and Standards" submodule mentions the importance of a Read Me file. Do you have any suggestions on how to create an effective / well-organized Read Me file?

- I treat README files as a major source of documentation for a project
- I usually turn the document I use to keep track of which files are where and in what order to execute code into a README file.
- Think of a person who knows nothing about your project and clicks on your project repository. They have no idea how to navigate the folders/files and don't know what the project is about. In your README file include a summary of the project, including how to contact you or the corresponding author/colleague, instructions on how to navigate the folders/files in order to reproduce your work, and any other details that will help them understand your work (ex: which programming language you used and which version).
- I think this and this are good resources.



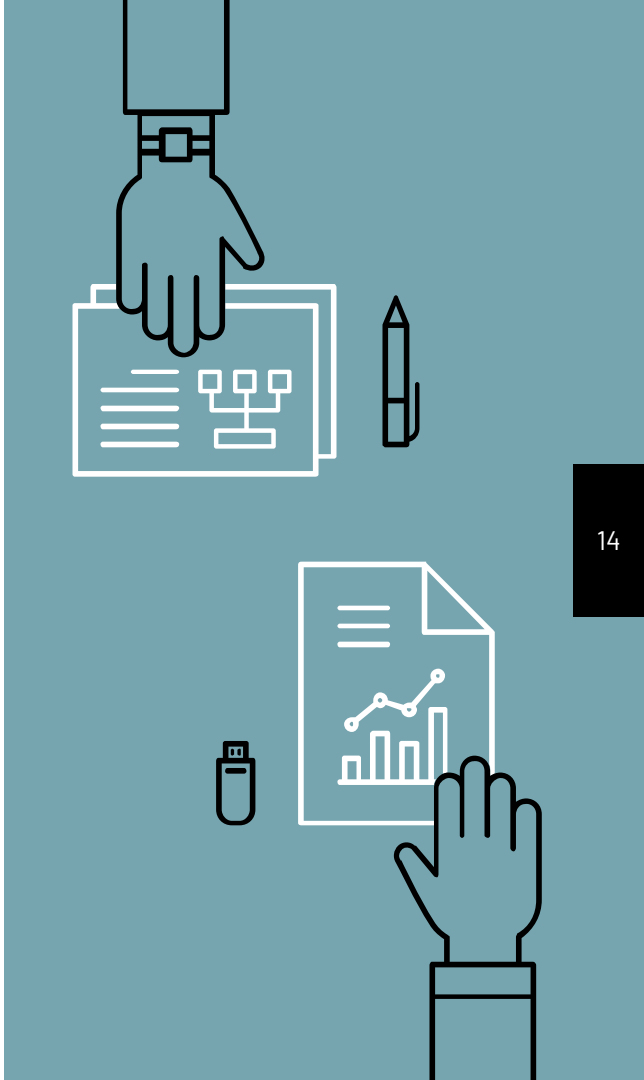
Module 4 Discussion

What are some examples of revision control repositories?

- Repositories: GitHub, Bitbucket
- Version control systems: git, Subversion (SVN), Perforce, CVS, RCS
- Check out this resource

Can you elaborate on the executable data pipeline?

- Think of this as a workflow for running (executing) your code.
- Ideally, you would have a driver script (a file with code) that when you or someone else runs the code in this file, all of your data cleaning, analyses and results (including plots) are created using the files needed for the project.



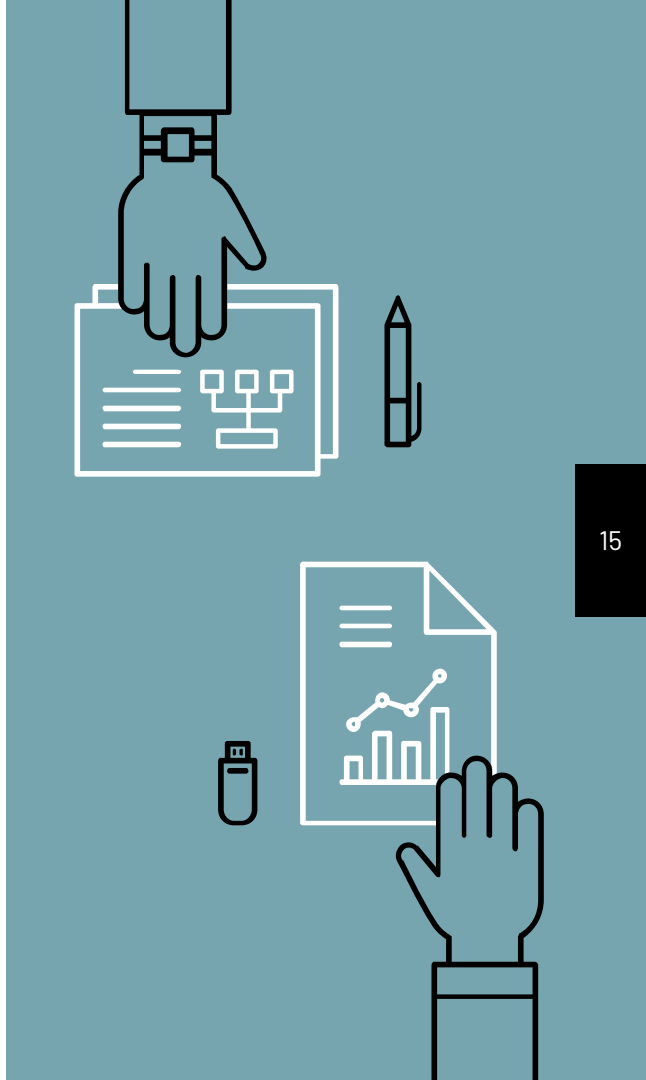
Module 4 Discussion

In what case is anonymized data still in need of data security mechanisms? How does this kind of data differ from randomly generated data?

- If it's health/biomedical data, always. This is due to the risk of re- identification.
- Anonymized data is still the original, collected data. Just the identifiers have been removed for privacy concerns. Randomly generated data is completely fabricated and not collected from a human.

The speaker keeps mentioning that we should document who is running what. Is this common to do? Where do people document who runs each analysis? Is this really an area where issues arise?

- Yes, people do this and yes, this is an area where issues arise, especially if you are part of a bigger team.
- In industry, there are different software that teams use to assign specific project tasks to each team member and a deadline for completing that task. These are typically called "tickets" and it's a great way to track who did what and how long it took.
- In academia, this can still be the case but it isn't as common (in my experience). The default is for everyone to keep track of what they do and then notify the team during a meeting or email chain.

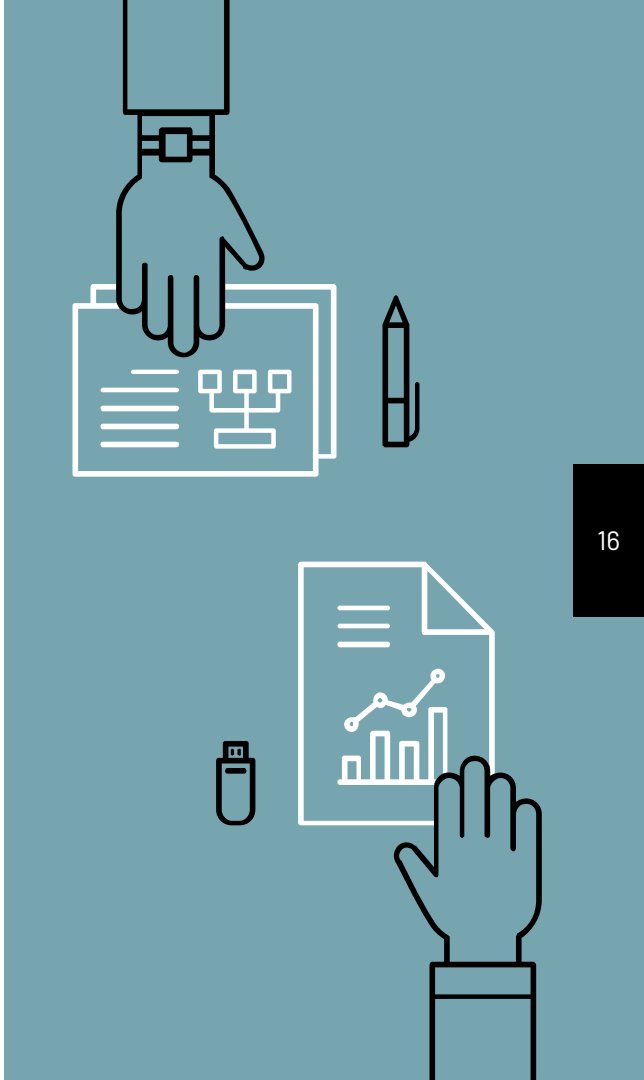


Module 4 Discussion

Do you set up a revision control repository? If so, what tools have you had success using?

- I do! For academic projects I set up a repository (private while I'm still working on the code) on GitHub. It's been the easiest platform for me to learn how to use and the community is very helpful.

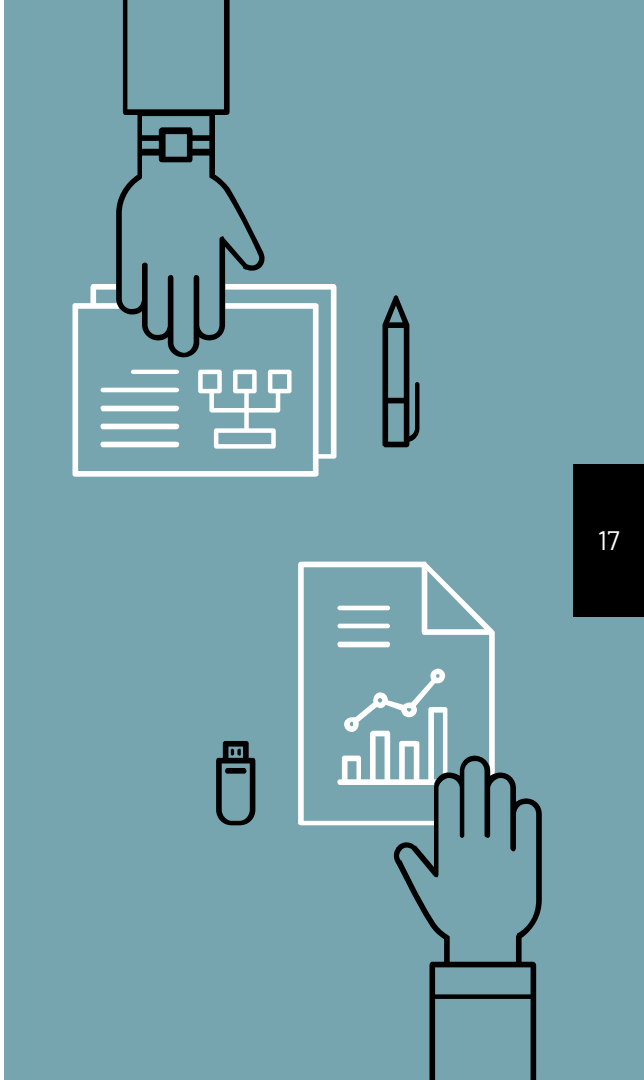
- For consulting work I use GitLab because my this is what my team uses. It's just like GitHub but has more features that are helpful for larger teams or multiple teams at one company.



Module 4 Discussion

Could you please share some your experience of keeping privacy and security during your research?

- All of the sensitive data I've worked with has been kept secure on Google Cloud Platform or on Harvard's computing cluster. I haven't had to maintain anything myself yet. I did recently submit an IRB application and received an exemption to keep de-identified survey data on my Harvard iMac that is maintained by Harvard IT. I had to be very detailed about how I would keep it secure (I would only store it on the Harvard iMac, in a password-protected folder, and I wouldn't share it with anyone).



Module 4 Discussion

How did the publishing house or the journal not notice so many flaws in the paper? is that not one of the main reasons for going through the whole process of being published by a reputable journal?

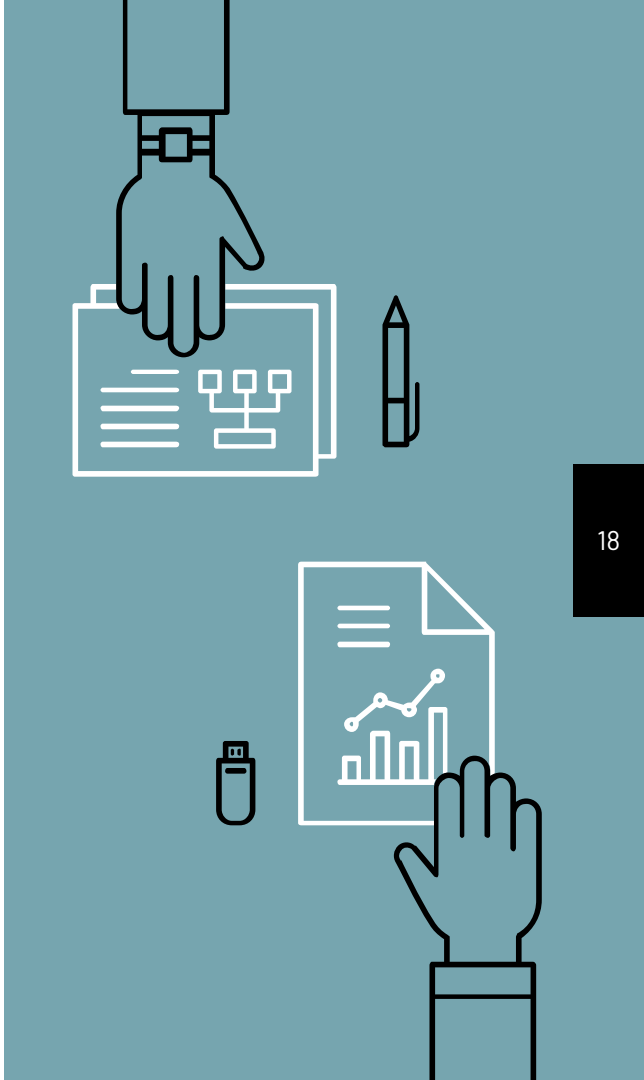
- Journals leave it up to 2-4 reviewers and an editor to review/accept papers. I'm guessing they didn't notice because they didn't try to reproduce the results. You would think so, but at that point in time reproducibility wasn't as much of a concern for journals, unfortunately.

Not sure how important it is, but could we go over more how supplemental provenance annotation systems like JSON, Taverna, etc work/are used for data provenance?

- They help keep track of who has handled the data and when, and how different files relate to each other. Check out the screenshot on slide 27.

What is an alternative method to share large non-textual data files except repositories such as Github and Bitbucket?

- Cloud services like Google Cloud Platform and AWS, and Dropbox.



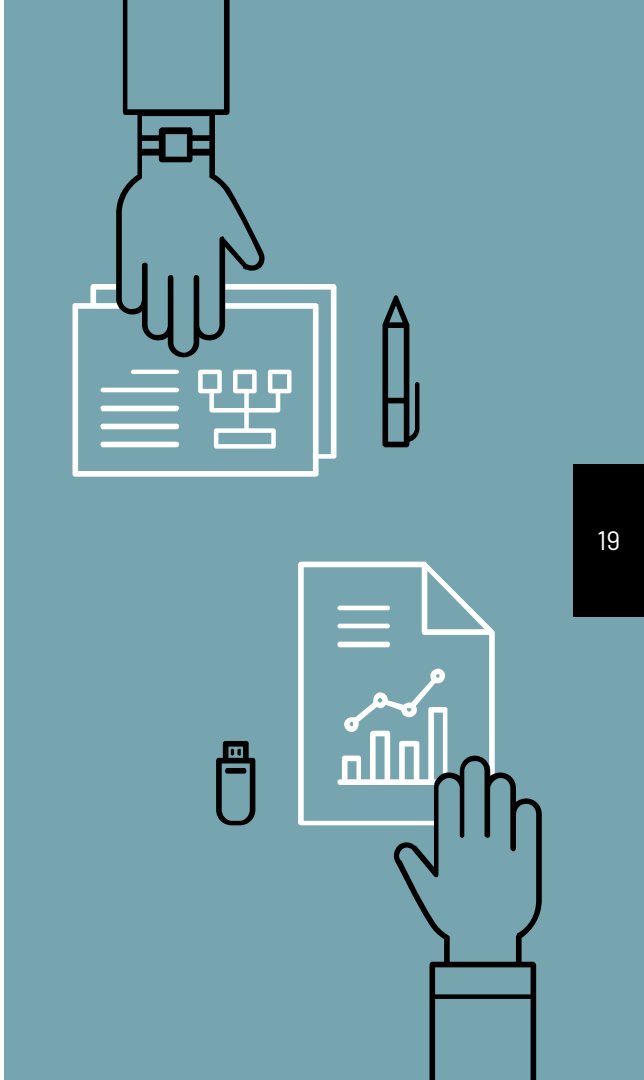
Module 4 Discussion

How can we properly manage and store the figures in our research articles by online repositories? Should we treat them as the usual internet images with copyrights or otherwise?

- I think once they are published the journal handles all of the copyright details and issues. You could store them in a repository with a link to the published paper.

Is this ok to separate a project into different parts (data, code, documentation, ...) and store them at different online repositories, or they should be better kept in one place?

- I personally think having them all in one place is easier/better for reproducibility and project management. However, sometimes (most times if working with collaborators) this isn't possible. The data may have to be stored in a specific, safe, place or a specific repository, and the code and documentation may have to be in different places so that others can also access them if needed.



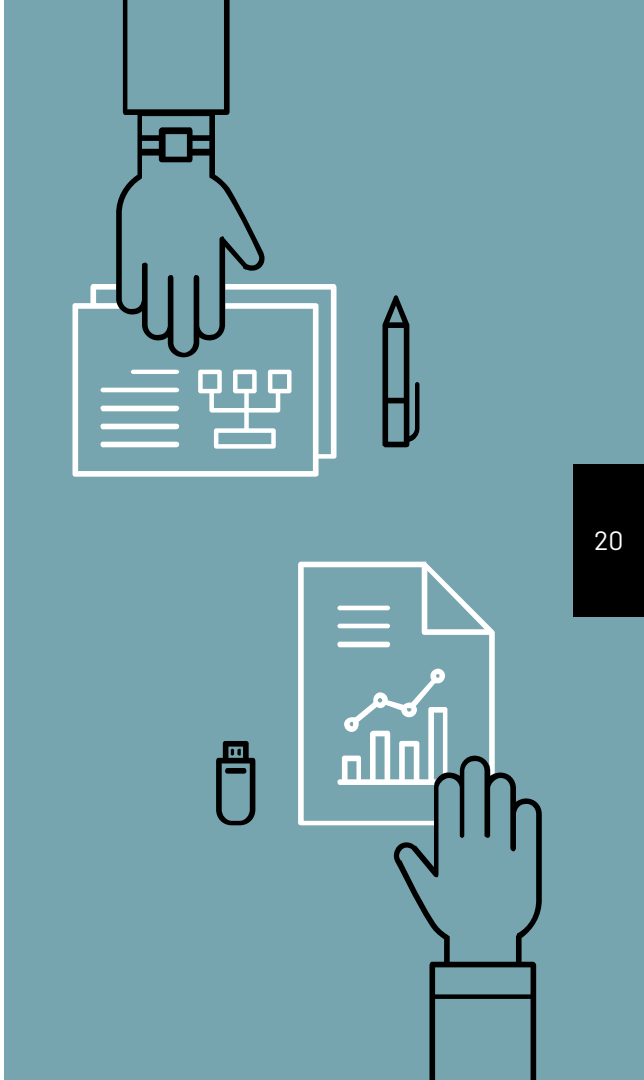
Module 4 Discussion

Unit testing in big data/streaming data seems like requires more rigor since you cannot use all data during positive and negative result unit checks, simulations, and sensitivity analyses. If we are just taking a subset in analyses, wouldn't we want to perform those checks in other subsets too, or is one enough?

- More is always better, but you may be constrained by computational resources. For example, I would use a subset (or subsets) of the data to write the code and test the code on my personal computer. Then I would send my code to the software team to implement using more cloud computing resources. If something didn't work they would let me know and I would look at the error and my code to fix it.

Can you walk us through your own data science pipeline?

- Too much to type so I'll talk you through it :)



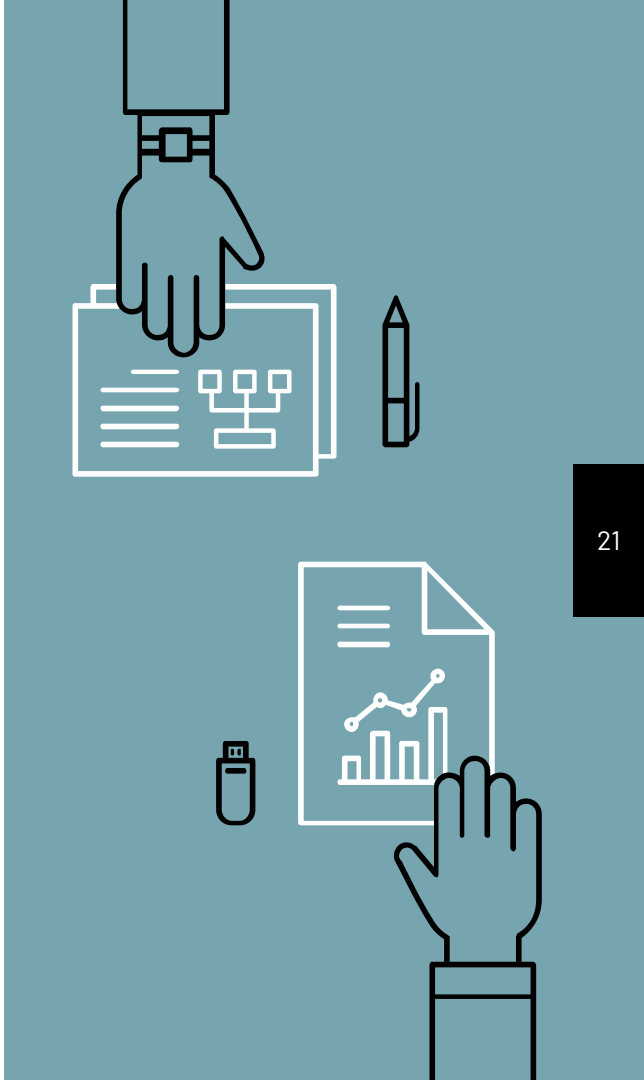
Module 4 Discussion

In one of the videos, they seemed to suggest that storing data on GitHub is not usually best practice. While this is certainly the case when dealing w/ very large data sets in fields like Genetics, I would argue that for smaller projects, it is probably the easiest place for researchers to store and access data, especially if the data is not sensitive. What are your thoughts on using GitHub for data storage?

- I think GitHub is fine for smaller, less sensitive data. I agree, it is one of the easiest platforms for researchers to access data.

Suppose I am working on 2 projects on a data set. I publish the first paper w/ the accompanying dataset, and then someone beats me to the 2nd paper w/ that data set. Is there some protection because it seems like this incentivizes people not to share data, even thought that should be incentivized?

- This is one of the reasons researchers have been hesitant to share data. And there isn't really any protection against it. It does usually take at least a month for a paper to be published (and usually longer) so hopefully that would help with getting at least a preprint of the second paper out before someone else is able to.



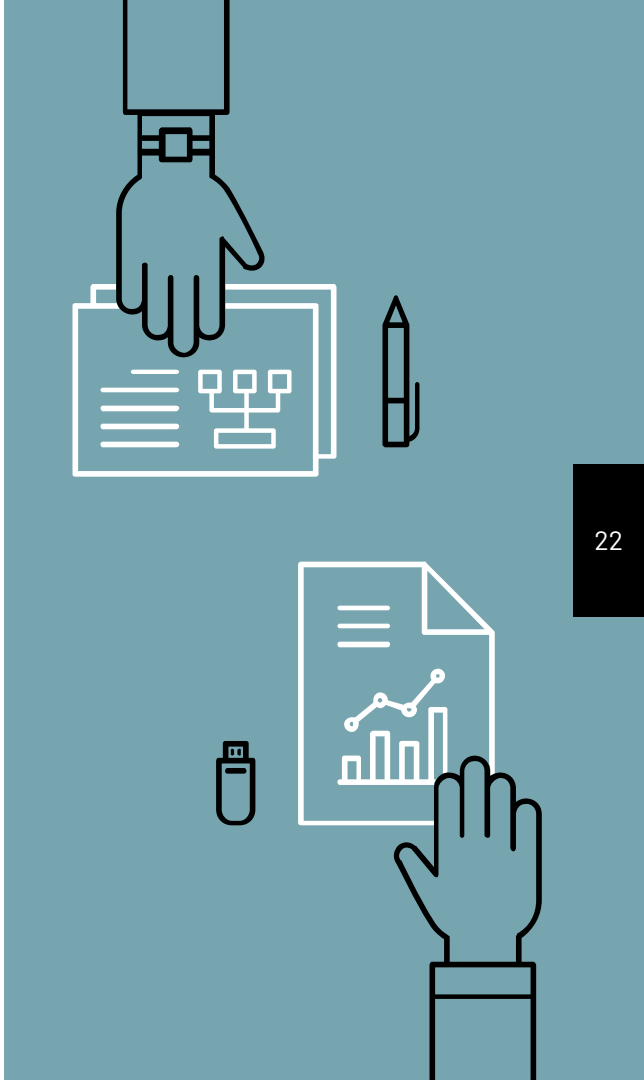
Module 4 Discussion

How much of a future data provenance plan is required in the IRB submission process?

- Depends on the size of the study/sample but I'll show you my IRB submission. There is a whole section on data provenance and that includes what happens to the data when the study has ended.

If a paper is published that contains sensitive information that can be traced back to participants, does it need to be retracted or just modified? Is this still true for papers that were published a long time ago?

- Great question. I'm not totally sure. My best guess is that the paper would have to be at the very least edited/updated. If the data was also publicly available, it would have to be removed from whatever repository it lives in and either de-identified or not made available except to certain individuals.



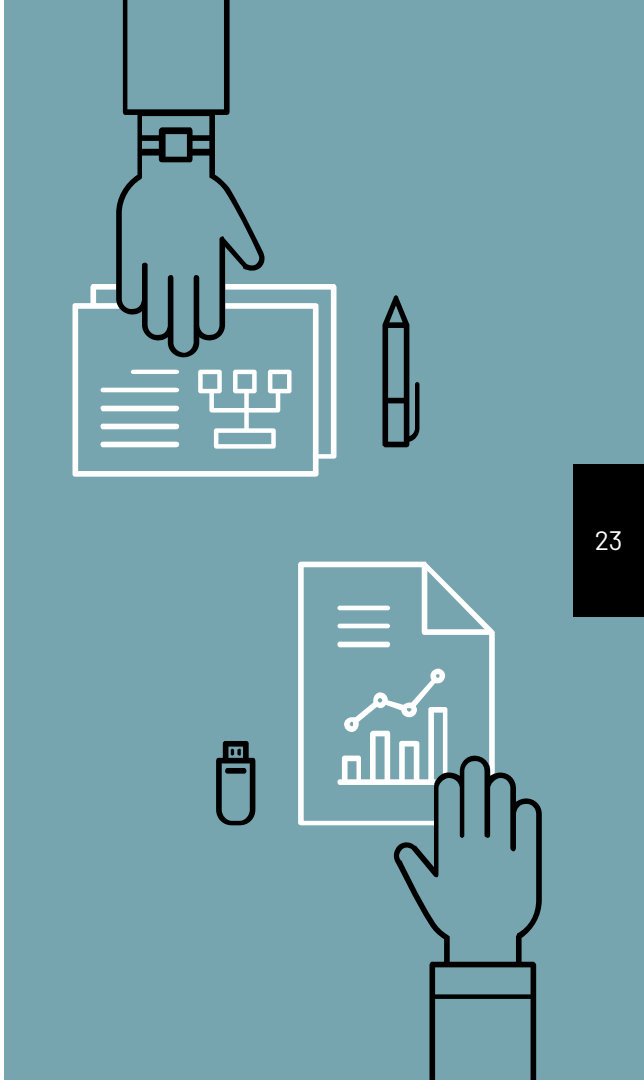
Module 4 Discussion

How do you prepare for packages/software being obsolete or not supported? For example, if your analysis is run in R, sometimes (thought not often) packages update and then do not function as expected. Are there any best practices for this?

- Documenting the version of R and the versions of the packages helps. If something is updated and your code breaks, you can usually revert to a former version and get it working again. If this happens, I try to update any code that I need to. But this is exactly why documenting versions is important.

I found the de-anonymization portion pretty interesting, are there any other high profile cases of being able to de-anonymization sensitive datasets?

- Yes. Check out Wikipedia for more information.



Module 4 Discussion

Module 4.5 showed the 18 identifiers HIPPA required to be removed from PHI. I have recently been doing research on COVID-19 where we have every patient's infection date. This is a crucial part of our data analyses as we are doing a survival analysis. In this case, does that mean we cannot share our data (with all other 17 identifiers removed) when we publish our work? But these data can still be made available as controlled access data, is that right?

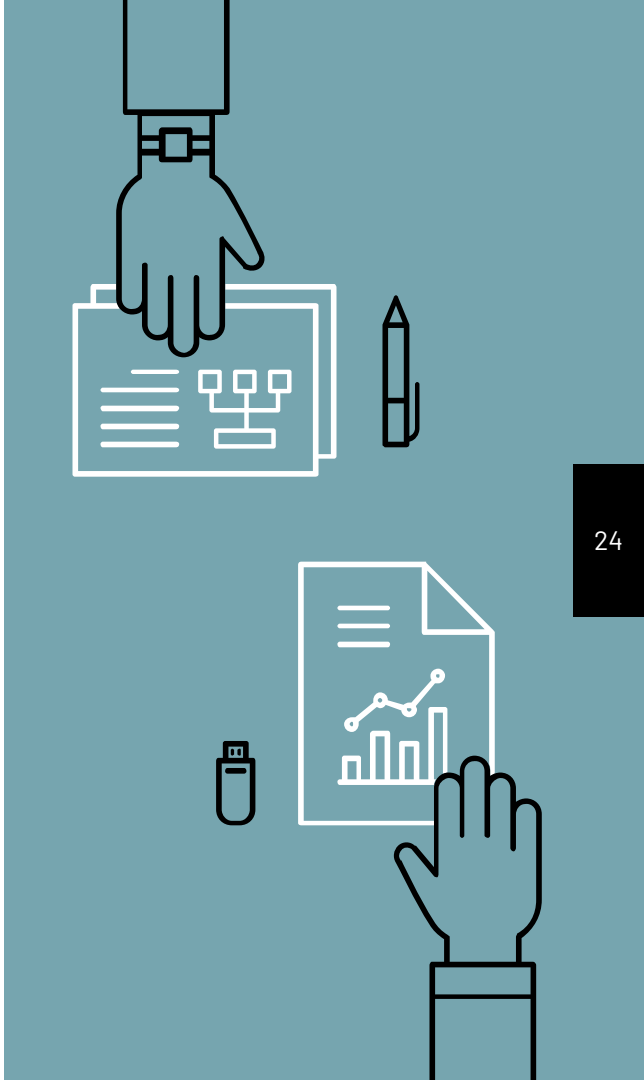
I don't think the infection date counts as PHI or identifiable information. Check out this link for more info.

I can't think of any circumstances under which Tier 5 (personally identifiable) data would be necessary for a research project??

- Any research that uses identifiable medical records.

Can we briefly discuss differences between literate programming and formal scientific workflow environment?

- Literate programming involves commenting your code and including text that explains what your code is doing. A formal scientific workflow environment is a tool that organizes every piece of your project - the code, data, analyses, etc.



Module 4 Discussion

What are some of the type-specific repositories for storing published data on protein crystal structures, families and domains?

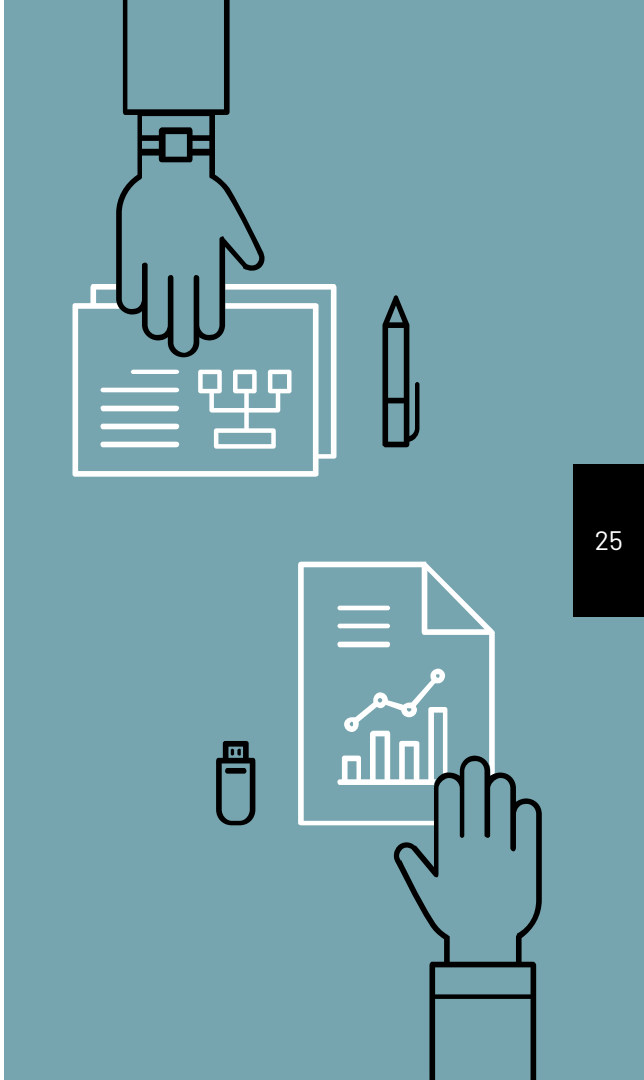
- This isn't my area of expertise but check out this paper - it may be relevant.

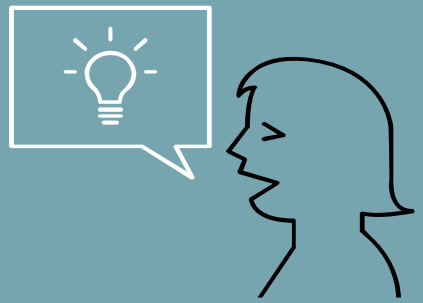
Do you have any example papers/githubs where there is elegant and simple code and data providence tracking for reproducible research?

- Yes. But I need to search for it because I forget the specific paper.

I've tried to keep track of analyses decisions outside of the analysis software (i.e. tracking my decisions for analysis in word while doing my work in R), but I seem to always fall off on this habit and have my practical analysis work go ahead of my tracking. Do you recommend always trying to integrate data providence with the analysis software (i.e. using R markdown and github so that your work can automatically be tracked)?

- I do! I also start off strong with my reproducibility efforts and then start to slack off as I get further into a project. What I make sure to do though is write down what I'm doing and why and I always save any progress I make before closing my computer. Nobody is going to be perfect - and you aren't expected to be! - but as long as you make an effort, you're good.





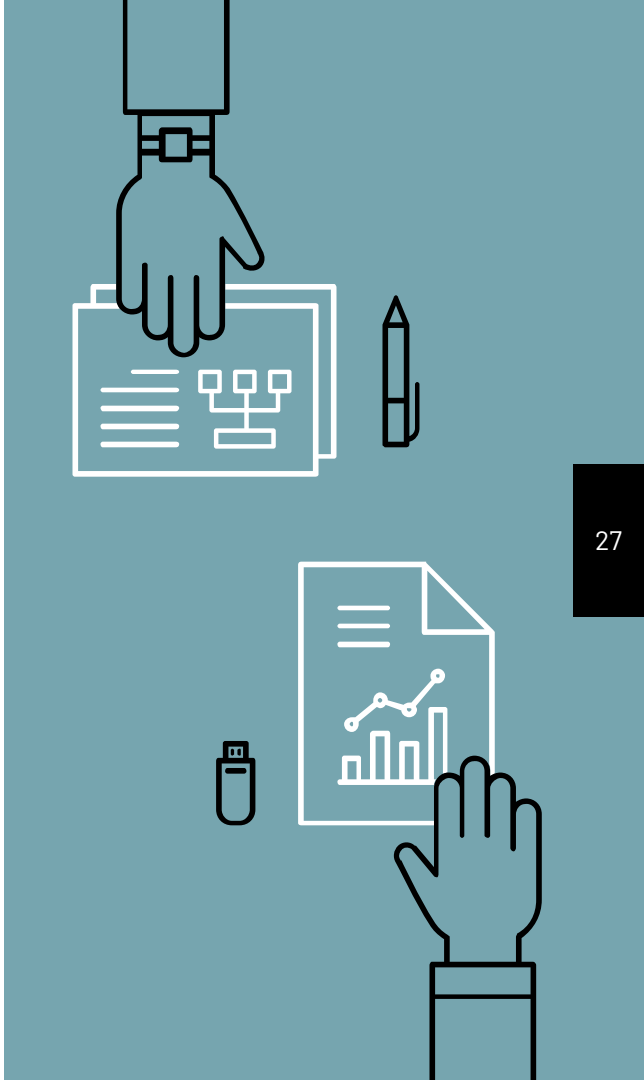
In-Class Project

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

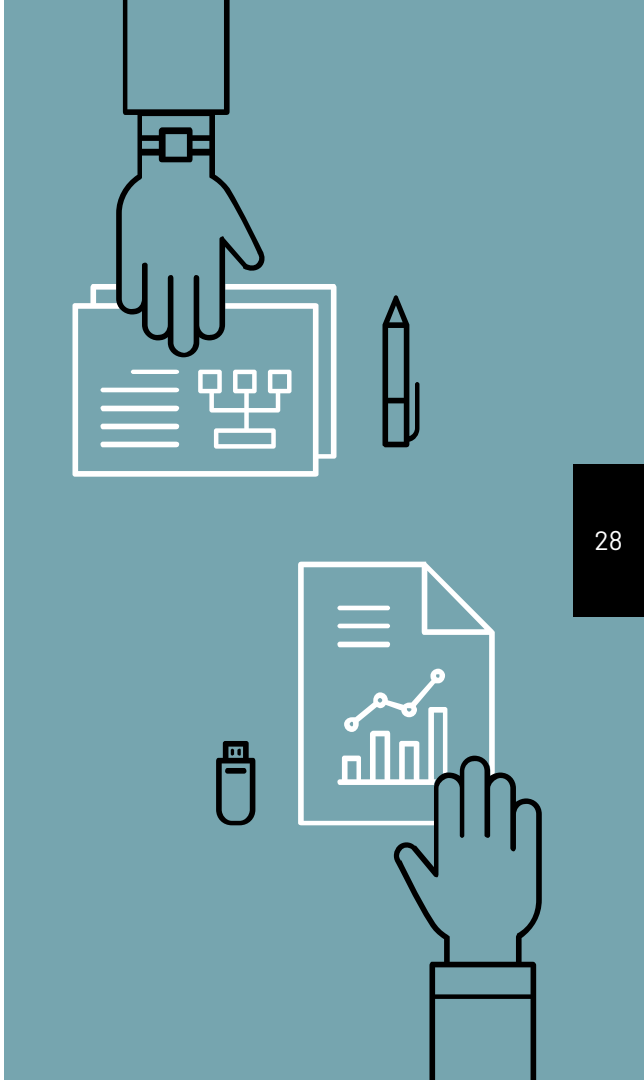


We will attempt to reproduce the results from and critique the reproducibility of:

[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



Homework

- Watch Module 5 videos
- Submit Module 5 discussion points