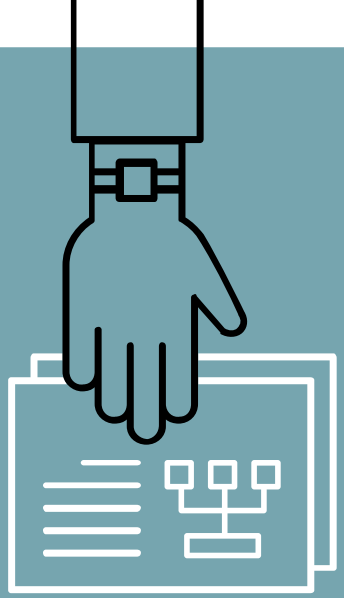
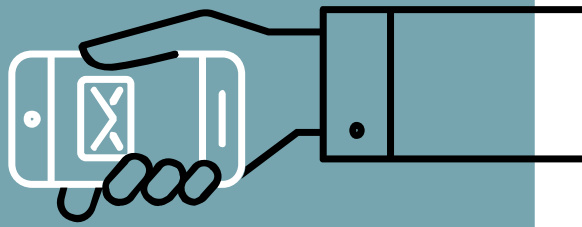
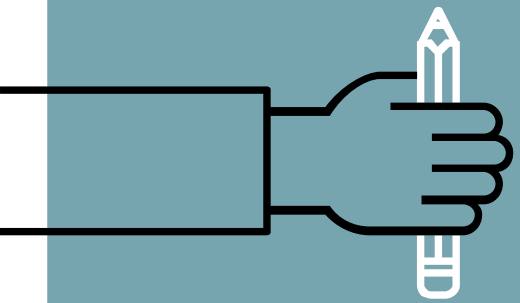
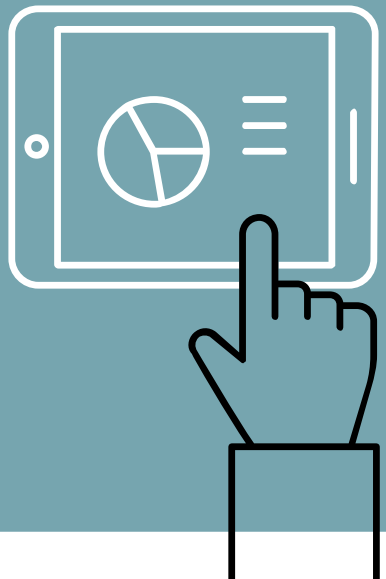




BST 270
Reproducible Data Science

Winter 2021
Session 4





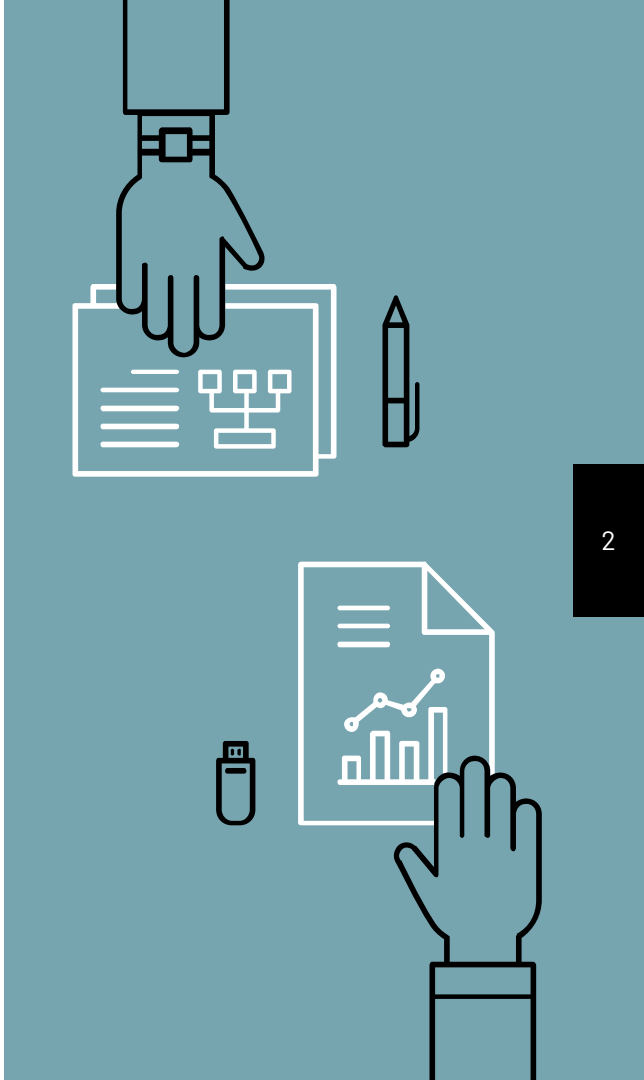
Module 4 Discussion

Positive and Negative Controls

Positive controls: Experimental inputs for which a known result or output is expected. These can be pre-specified and automatically checked throughout the analysis process. Also referred to as unit tests.

Negative controls: Null inputs for which no or an uninteresting result is expected and can be checked.

Controls should always be included, and checked automatically if possible, at each stage of the study intended to be reproducible.



Module 4 Discussion

Identify the following as either a positive or negative control.

1. You have written a function that transforms birth dates (MM/DD/YYYY) into age (Years). You input 10 different birth dates and find that your function outputs the correct ages in years.

2. You derive a new test statistic and write a function to calculate its value. You calculate the test statistic for a data set that you know is not significant under the null, and find that it is indeed not significant.

3. You create a test to detect cancer cells. You run the test on a batch of healthy, non-cancerous cells and find the test result is negative.

Module 4 Discussion

Identify the following as either a positive or negative control.

1. You have written a function that transforms birth dates (MM/DD/YYYY) into age (Years). You input 10 different birth dates and find that your function outputs the correct ages in years. Positive control

2. You derive a new test statistic and write a function to calculate its value. You calculate the test statistic for a data set that you know is not significant under the null, and find that it is indeed not significant.

3. You create a test to detect cancer cells. You run the test on a batch of healthy, non-cancerous cells and find the test result is negative.

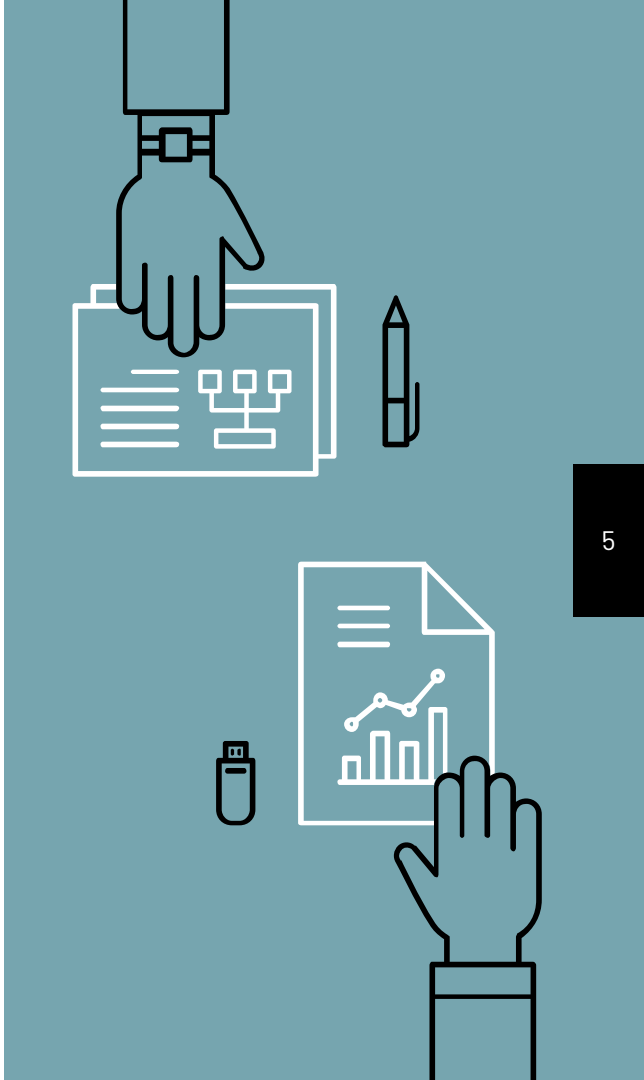
Module 4 Discussion

Identify the following as either a positive or negative control.

1. You have written a function that transforms birth dates (MM/DD/YYYY) into age (Years). You input 10 different birth dates and find that your function outputs the correct ages in years. Positive control

2. You derive a new test statistic and write a function to calculate its value. You calculate the test statistic for a data set that you know is not significant under the null, and find that it is indeed not significant. Negative control

3. You create a test to detect cancer cells. You run the test on a batch of healthy, non-cancerous cells and find the test result is negative.



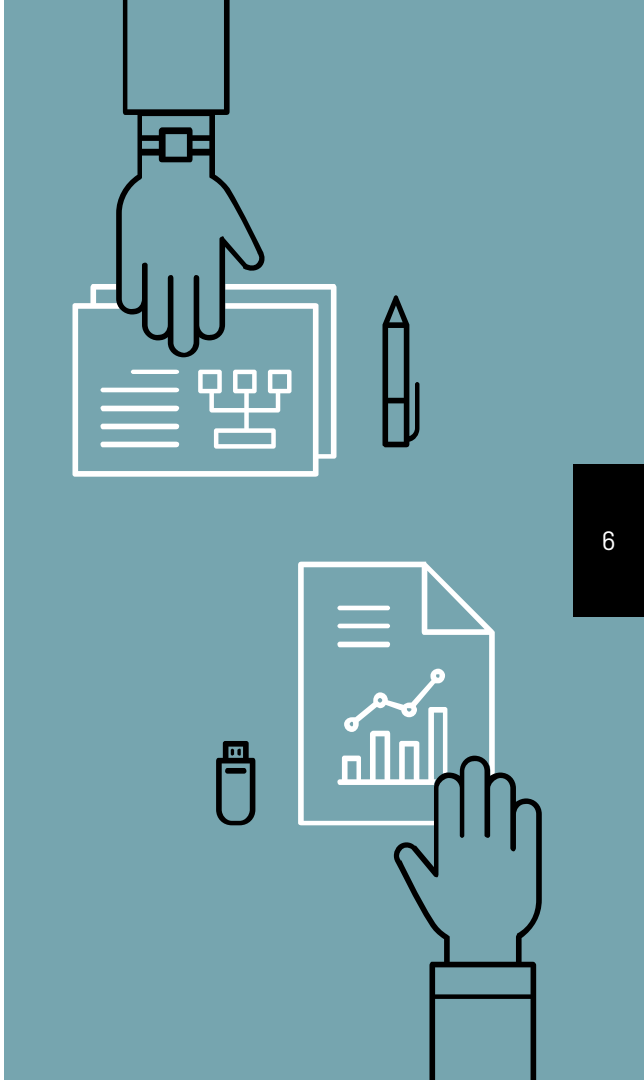
Module 4 Discussion

Identify the following as either a positive or negative control.

1. You have written a function that transforms birth dates (MM/DD/YYYY) into age (Years). You input 10 different birth dates and find that your function outputs the correct ages in years. Positive control

2. You derive a new test statistic and write a function to calculate its value. You calculate the test statistic for a data set that you know is not significant under the null, and find that it is indeed not significant. Negative control

3. You create a test to detect cancer cells. You run the test on a batch of healthy, non-cancerous cells and find the test result is negative. Negative control



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.

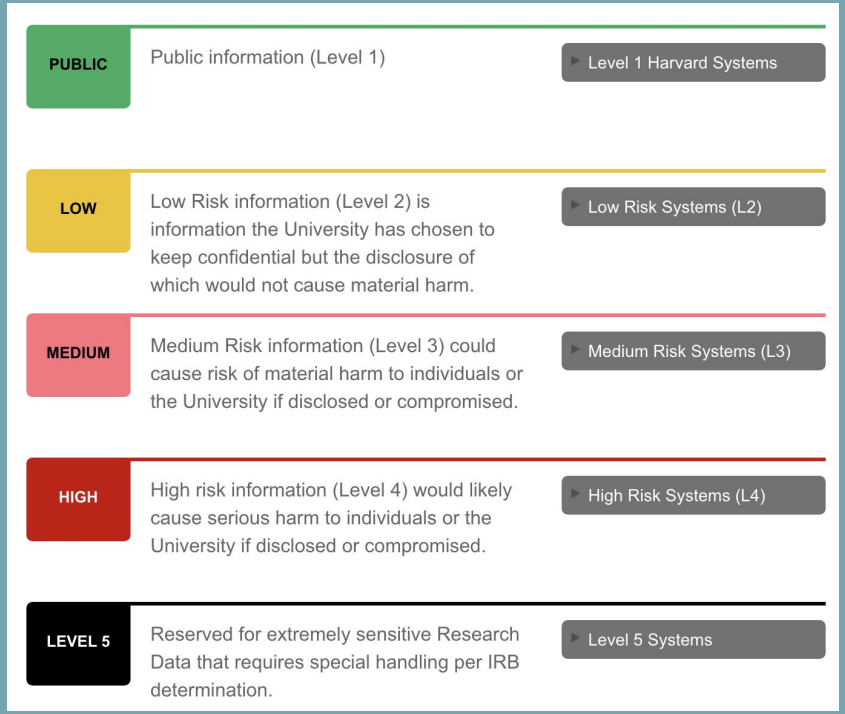
B. A data set including the date, time, duration, location and category of tornadoes in Oklahoma in 2011.

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

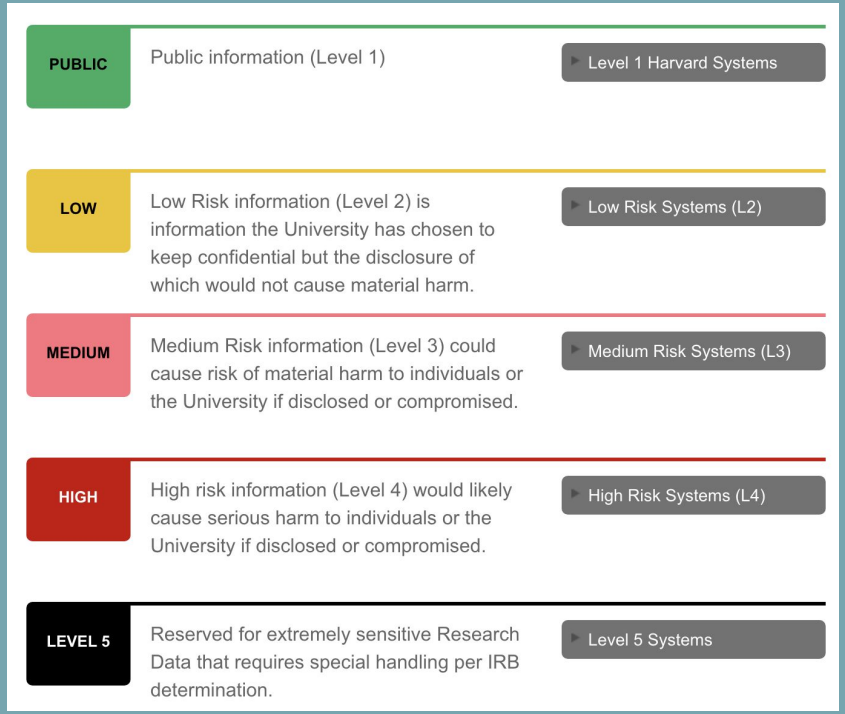
B. A data set including the date, time, duration, location and category of tornadoes in Oklahoma in 2011.

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

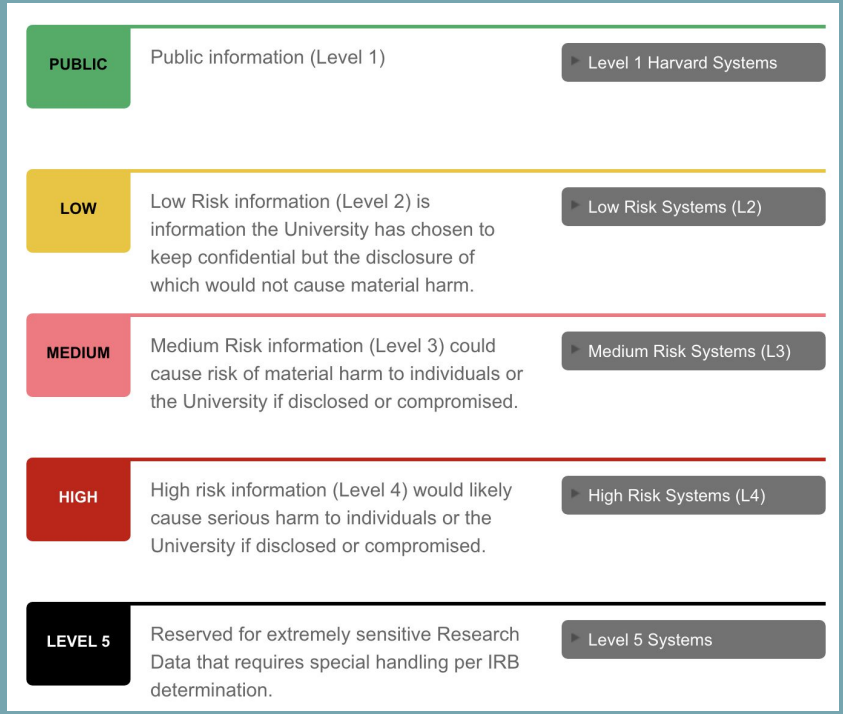
B. A data set including the date, time, duration, location and category of tornadoes in Oklahoma in 2011. Level 1

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

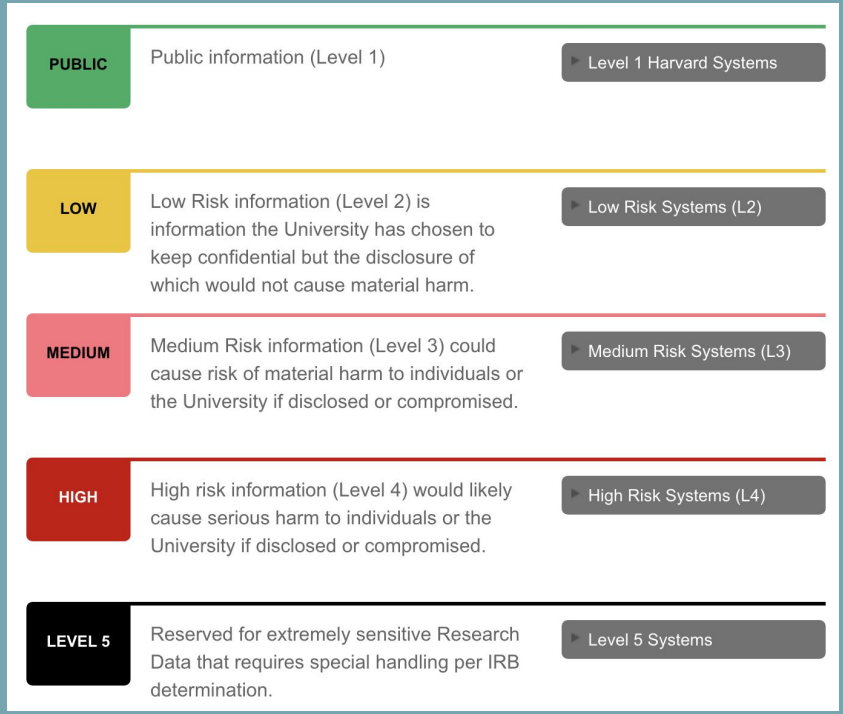
B. A data set including the date, time, duration, location and category of tornadoes in Oklahoma in 2011. Level 1

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

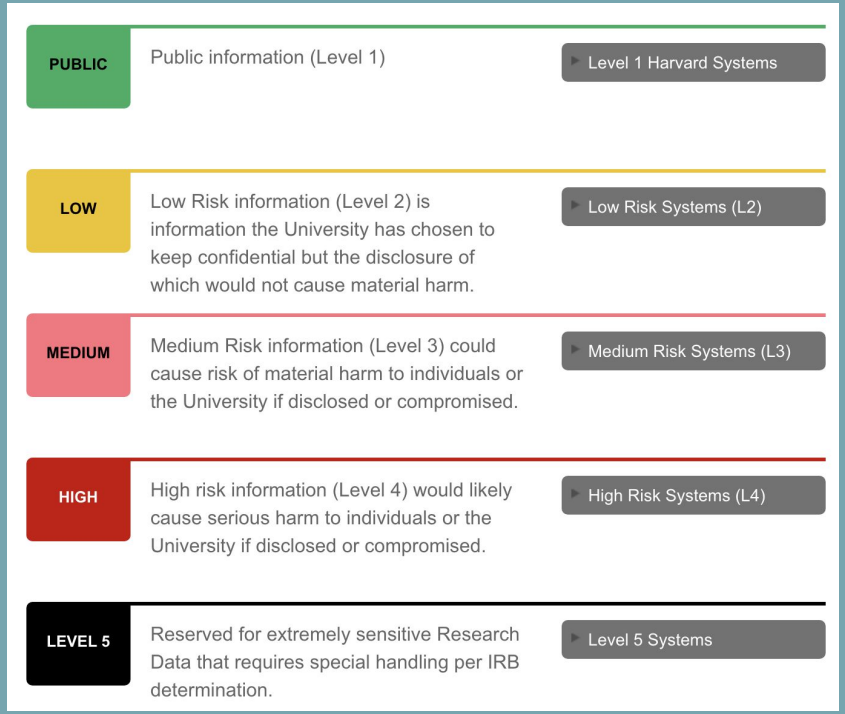
B. A data set including the date, time, duration, location and category of tornadoes in Oklahoma in 2011. Level 1

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

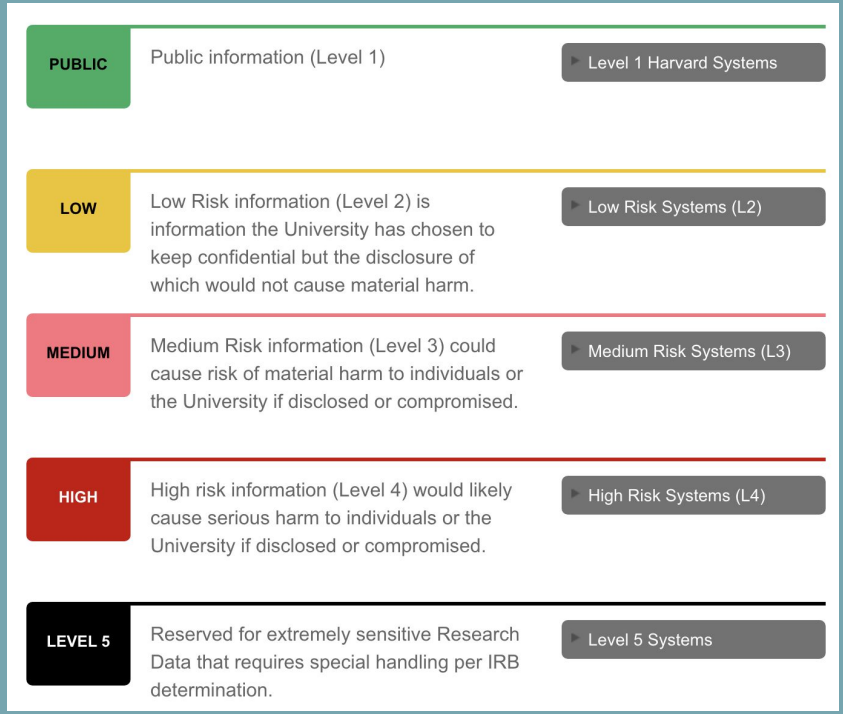
B. A data set including the date, time, duration, location and category of tornadoes in Oklahoma in 2011. Level 1

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



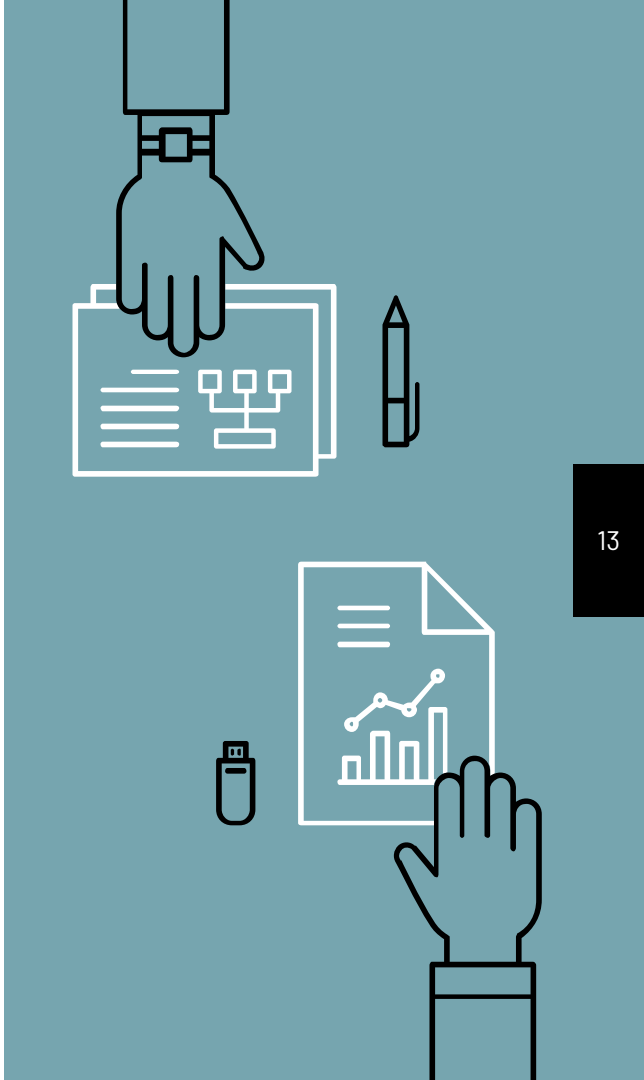
Module 4 Discussion

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?

- 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.



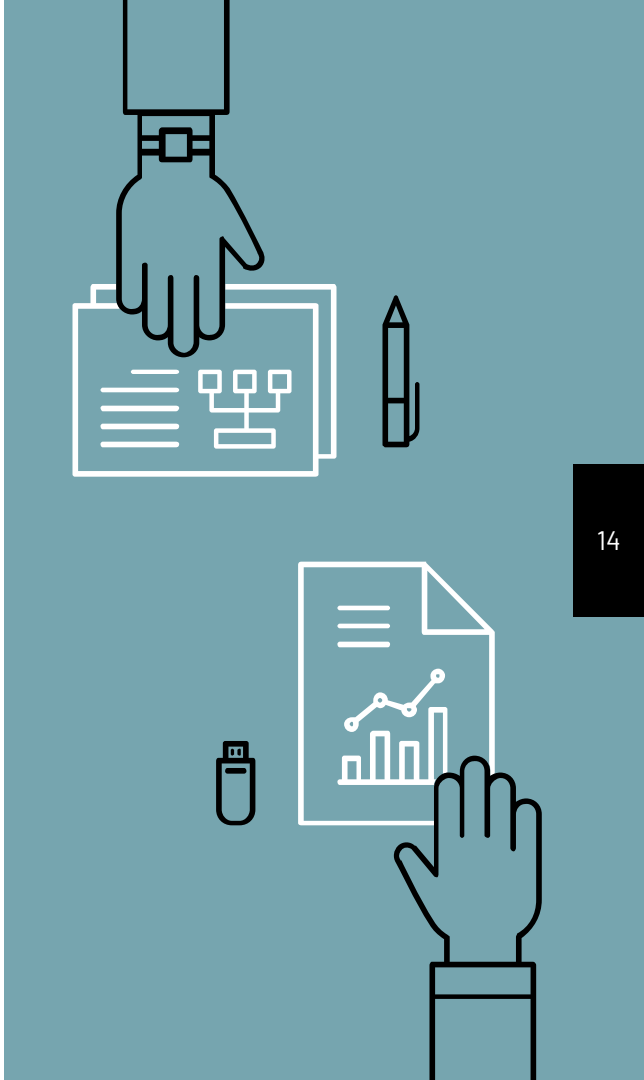
Module 4 Discussion

I've worked with biobank data that we had to both pay and register to genotype and use. How do we make that reproducible? We can't exactly just make that data available on dbGAP.

- This is tough. Being as detailed as possible either in the Methods section of a paper or in the Supplementary Material is a good start. I think this is a situation where the authors are not able to ensure complete reproducibility and anyone wanting to reproduce your results will have to talk to you and your team and then attempt it themselves, and you wouldn't be at fault if the results differ a bit.

Are papers these days generally more reproducible than in the past? In what aspects are they still lagging?

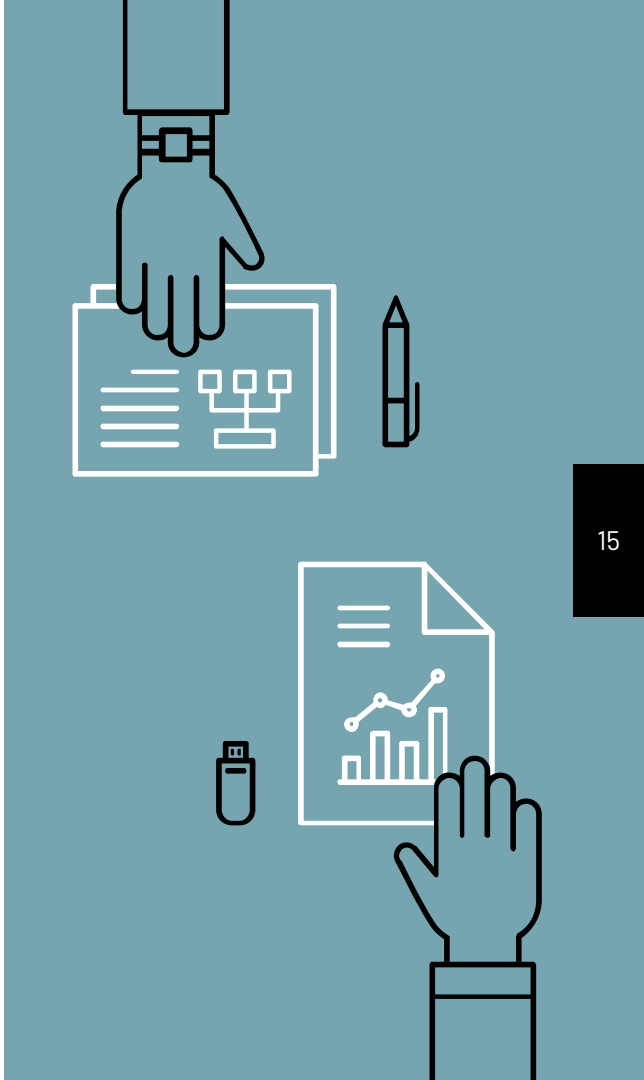
- Yes. While progress is slow, many journals require you to publish your data (when you can) and code in a publicly available repository. Some journals have also increased the page limit of the manuscript and Supplementary Material section in an effort to encourage researchers to be as detailed as possible.
- Some journals don't have very rigorous transparency requirements yet, and we don't have a central database for papers and data/code.



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?

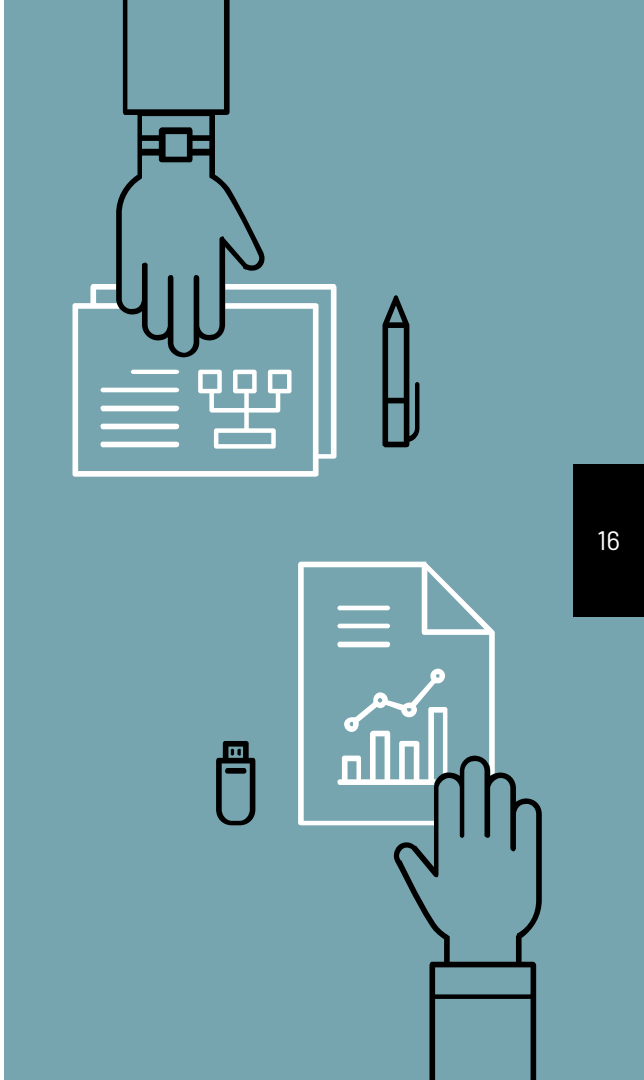
- I don't know of any requirements in place yet, but some journals do ask you to keep technical jargon to a minimum (if you can) and manuscripts that are easier to read tend to be published by better journals.
- There is a push for more "storytelling" in manuscripts that make the paper easier/more enjoyable 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?

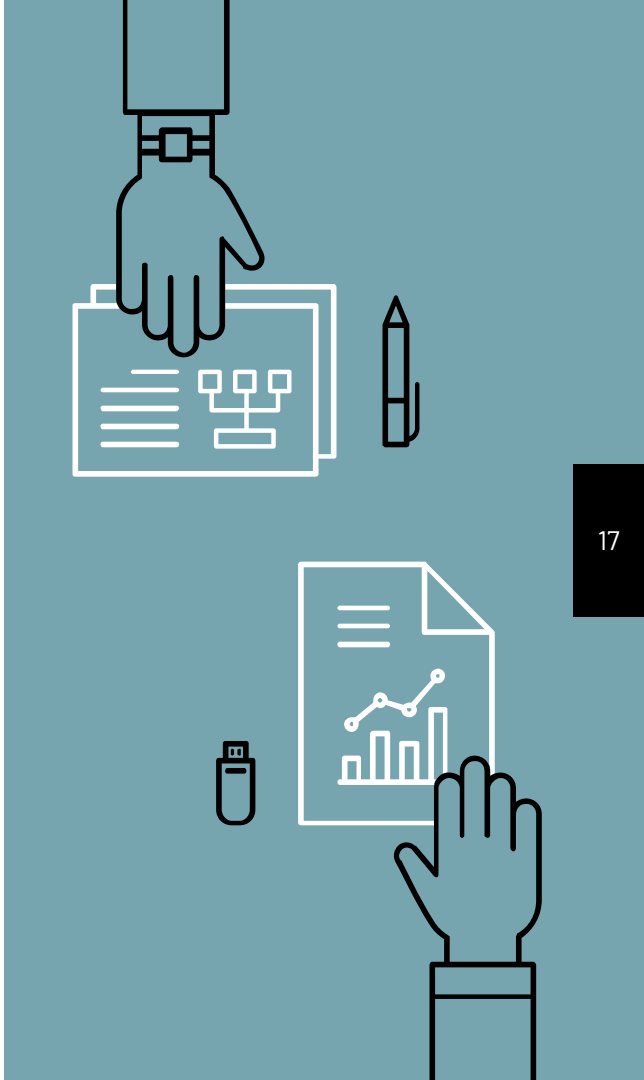
- I treat README files as a major source of documentation for a project
- I usually turn the document I use to keep track of which files are where and in what order to execute code into a README file.
- Think of a person who knows nothing about your project and clicks on your project repository. They 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

It is my first time hearing a real scientific scandal about how a chemotherapy drug responds on patients (Potti 2006). After finishing this module, I had a question that since the journal article was always peer reviewed by other scientists, if there was any significant issues with those paper by Potti, why were those essays still published? Why were there no authority from the journals to deal with those issues before or right after the article was published by Potti?

- The reviewers and editor of the journal did not check or try to reproduce the results during the review process so they didn't know the results were wrong.
- I'm not sure why the journal waited so long to listen to the scientific community to retract the Potti paper; it may have been because retracted papers stain the reputation of a journal.
- These checks should have happened before the paper was published and any responses to the paper after publication should have been taken seriously, but for some reason these things didn't happen.



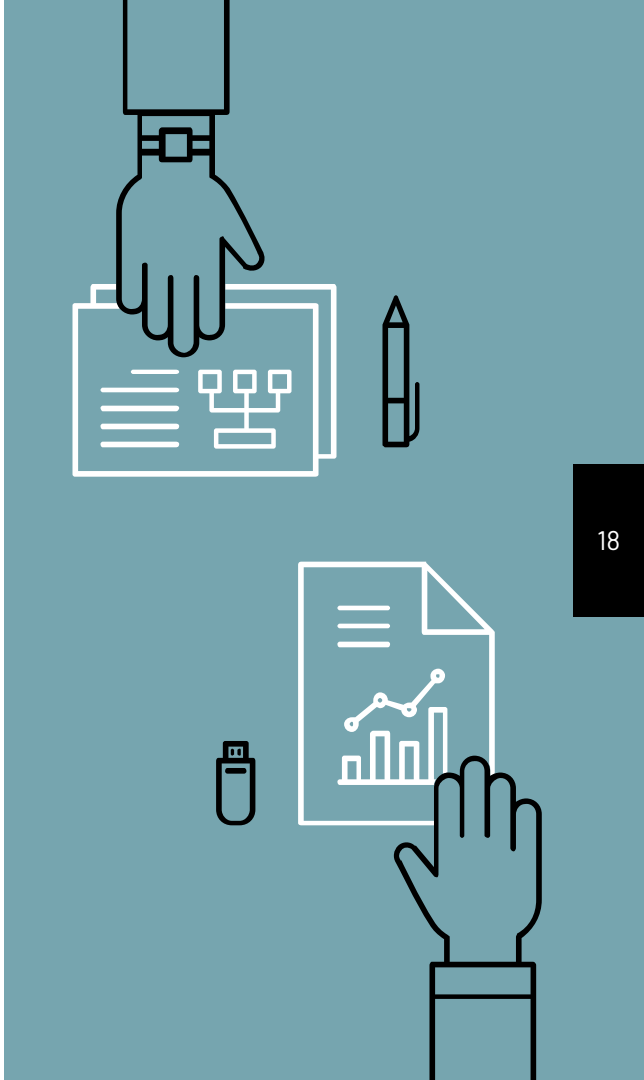
Module 4 Discussion

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?

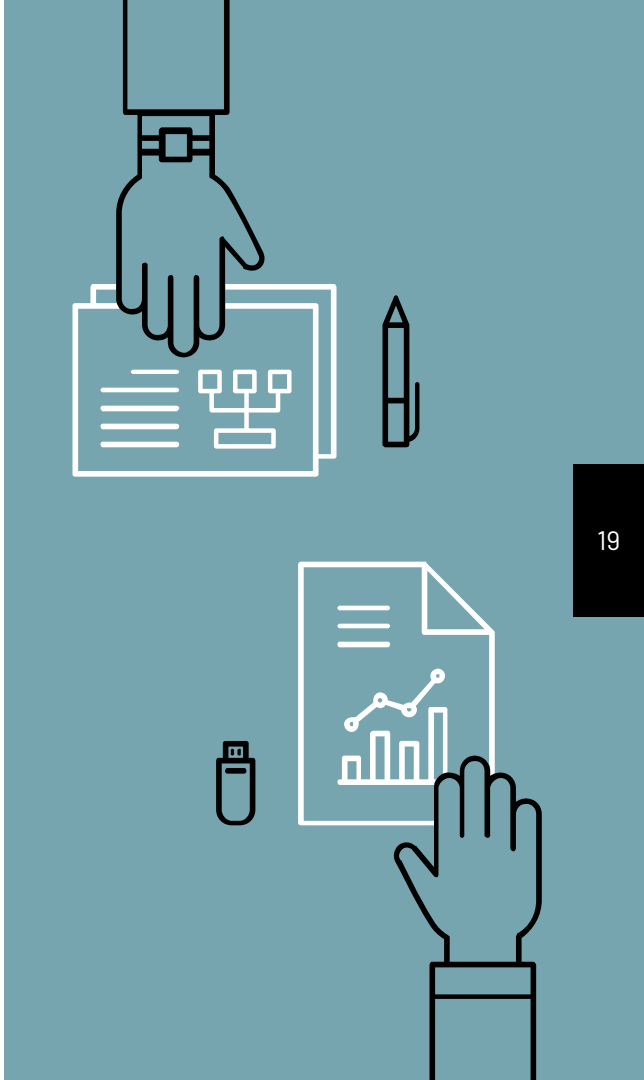
- Think of this as a workflow for running (executing) your code.
- Ideally, you would have a driver script (a file with code) that when you or someone else runs the code in this file, all of your data cleaning, analyses and results (including plots) are created using the files needed for the project.



Module 4 Discussion

I'm not sure when these video lectures were recorded but has there been any progress made towards this idea of a joint database that links publicly available data and scientific journals with each other? Is this a feasible goal for the data science/scientific community?

- These videos were recorded late 2016, early 2017
- Yes, there has, but I don't think there is one journal -> data/code repository link yet. Some journals require you to publish/host your data/code in a publicly available repository, but this can be on GitHub or other platforms and not a journal-specific platform.
- I believe this is a feasible goal for the data science/scientific community, but I think it will take decades and I'm not sure there will ever be just one database. I think what will happen is bigger journals will host their own databases and link papers and data/code and smaller journals will have the authors use GitHub and other platforms.



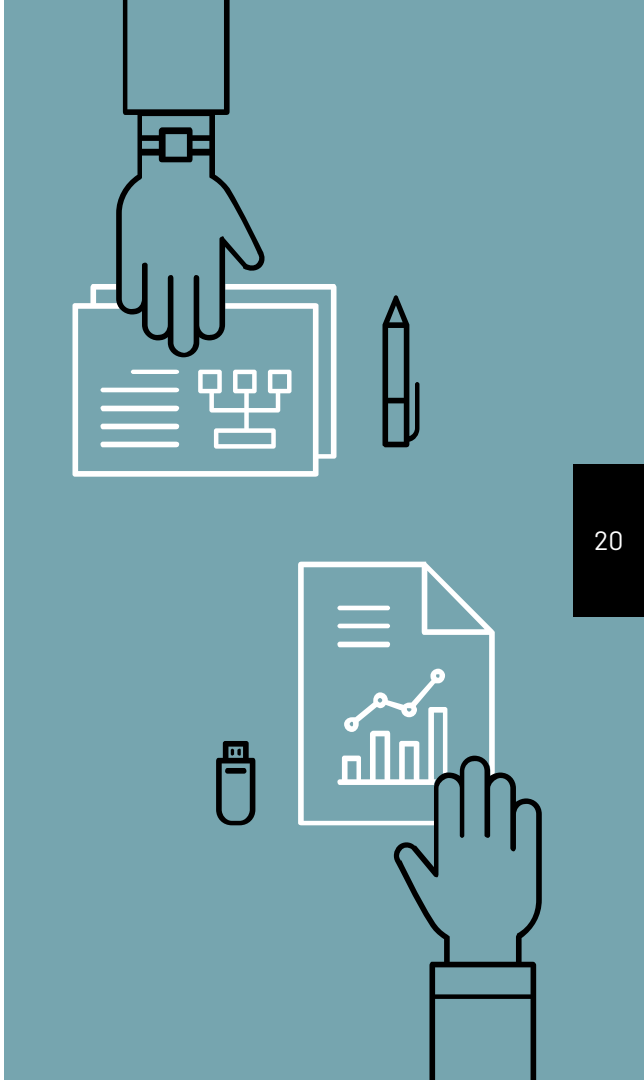
Module 4 Discussion

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.



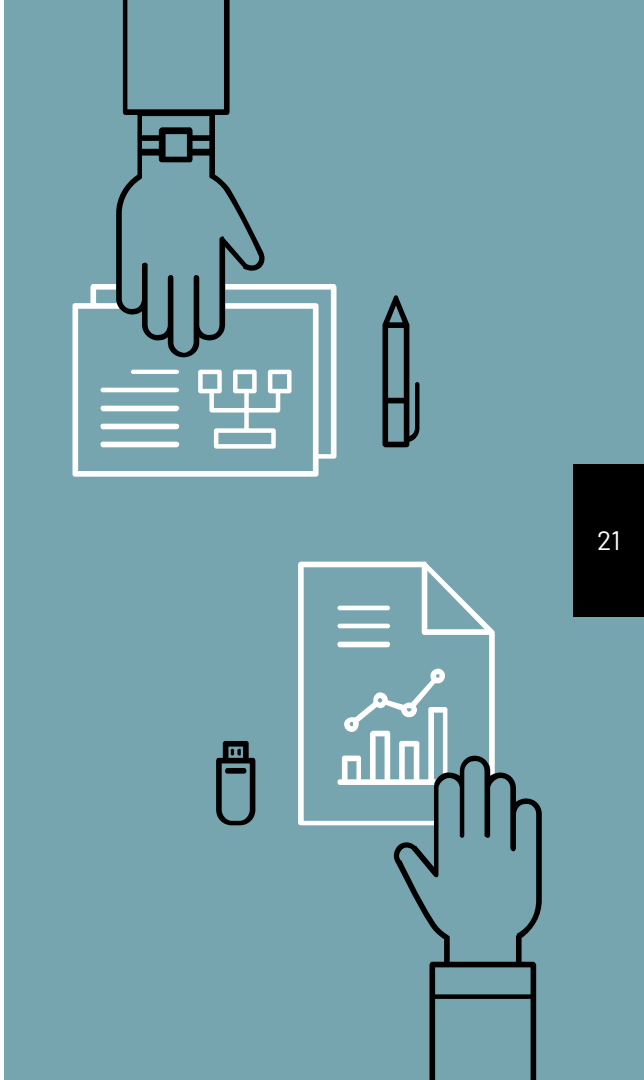
Module 4 Discussion

Does how secure data is affect how easy/hard it is to reproduce studies/identify studies that may not be reproducible?

- Yes. It makes the whole process longer or impossible, depending on who can be granted access to the data.

What happens if data is truly lost due to some unforeseen circumstances years after publication and the research is under investigation?

- Great question, and I'm not sure. My guess is the investigation wouldn't be able to move forward, but the researcher(s) would face legal consequences for losing the data (if they did something to cause the loss, intentional or not). If it's something like a fire the researchers had no part in, I don't think the investigation could move forward, unless errors are able to be proven some other way without the data.



Module 4 Discussion

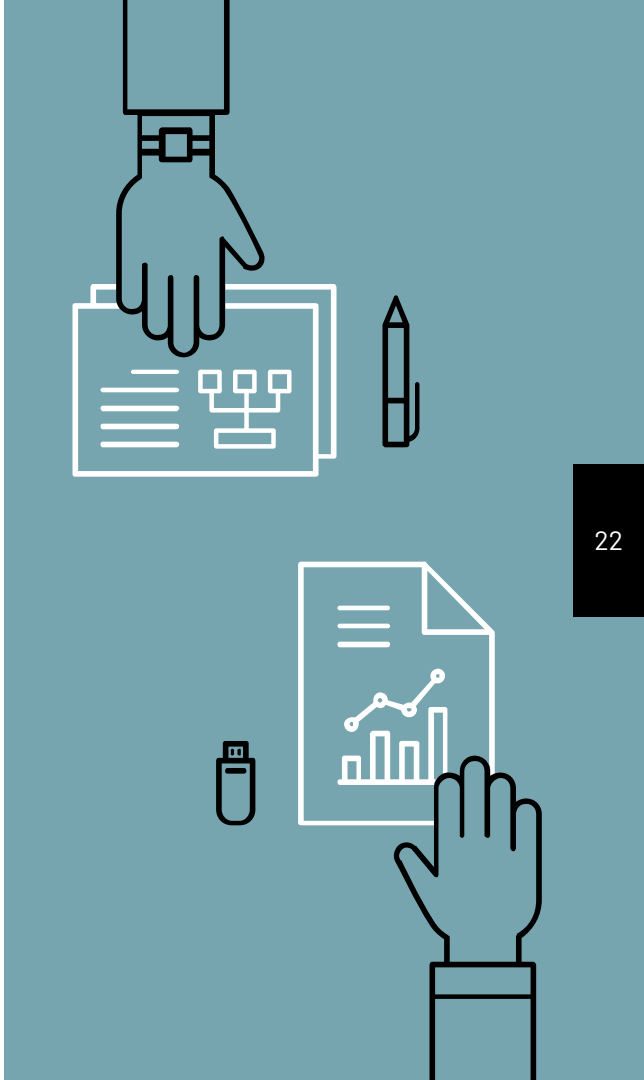
What are the current legal standards regarding consent of family members when it comes to identifiable genetic information?

- This is a fantastic question and I'm not sure. I know you don't need consent from family members to get genetic testing (like 23andMe, Ancestry, etc.) done. But, if someone in your family does get testing like this done, the DNA they submit can be used to find relatives who have committed a crime.

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.



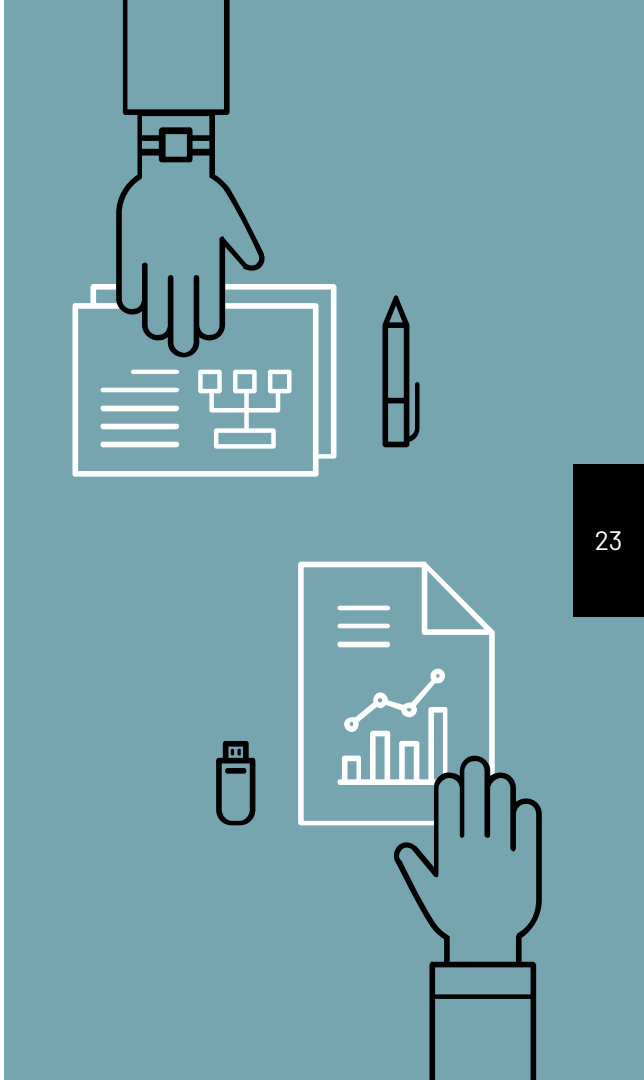
Module 4 Discussion

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).

Consider the situation that we are training models given the dataset. Our results may be based on some packages (many of which are black boxes), and there could be randomness or other issues we have not yet considered. What shall we do in this situation?

- Absolutely. To overcome some of this, you can set your random number generator (set.seed() in R), and document which version of software and package(s) you used.



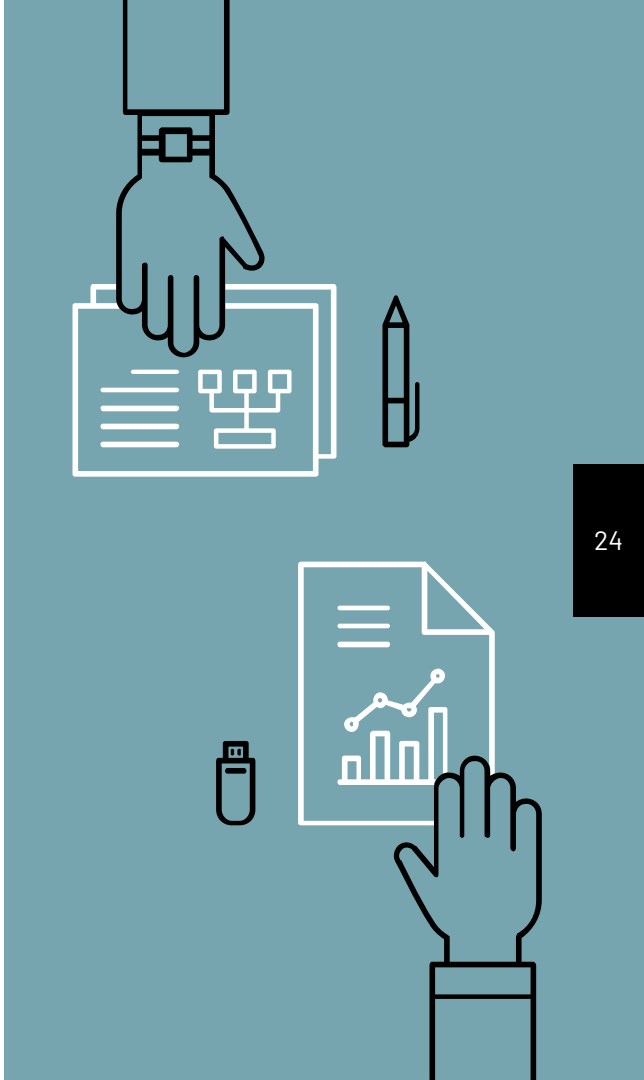
Module 4 Discussion

Though the videos offered some mechanisms by which researchers could ensure reproducible analyses (e.g., attaching code as a supplement), I would be curious to learn if these were often REQUIRED?

- Journals are starting to make them required. I know journals like Science, Nature, PLOS, etc. all require code to be published publicly/

Is pseudo-code an acceptable form of supplemental material, if you were not comfortable sharing your actual code?

- Sometimes. It depends on the journal. It's a nice alternative, though.



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?

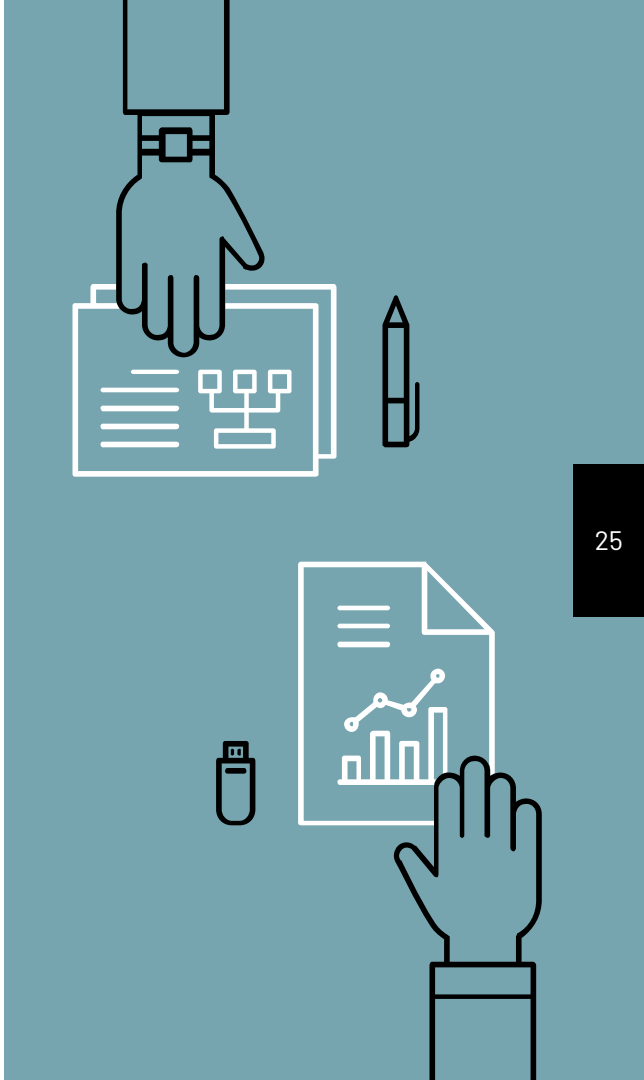
- Journals leave it up to 2-4 reviewers and an editor to review/accept papers. I'm guessing they didn't notice because they didn't try to reproduce the results. You would think so, but at that point in time reproducibility wasn't as much of a concern for journals, unfortunately.

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.

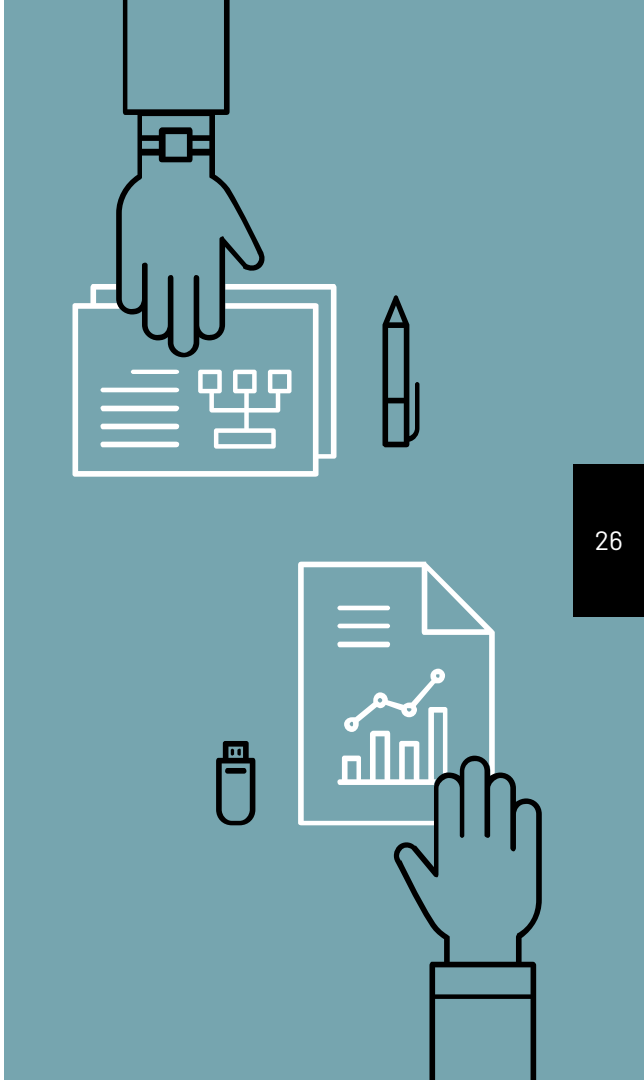


Module 4 Discussion

What are your personal favorite data provenance tools? (along the lines of Taverna, myGrid, etc.) Can we got over an example of formal scientific workflow environments that you have used?

- I haven't used any of the tools like Taverna personally, but I have used tools like Jira and Confluence when doing consulting work in industry. They allow teams to track who does what and when, and how long it takes. You can also set up a project page and add progress/updates to it using Confluence. It's worked well so far. I think Taverna, myGrid, etc. are more commonly used in genetics/genomics/omics research.

Given how complicated the trial sounds, is it possible that not enough details were provided in the methods section and not that the trial itself was fraudulent? was it at all possible that the original researchers themselves could have reproduced the research or provide the new investigators with more guidance to reproduce their results? that woful have been more beneficial to both group of researchers as well as to the general advancement of science.



Module 4 Discussion

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?

- I don't have any experience working with either, or any other software that makes a visual workflow of files (I haven't done an analysis that complex yet), but here are the links to check them out and a screen shot of a Taverna workflow example on the bottom right.
- To my knowledge, these tools are great for genetics/genomics projects with many files/inputs affecting the analyses and outputs.
- Taverna
- JSON

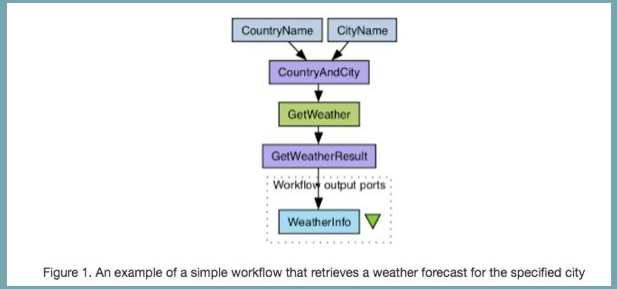
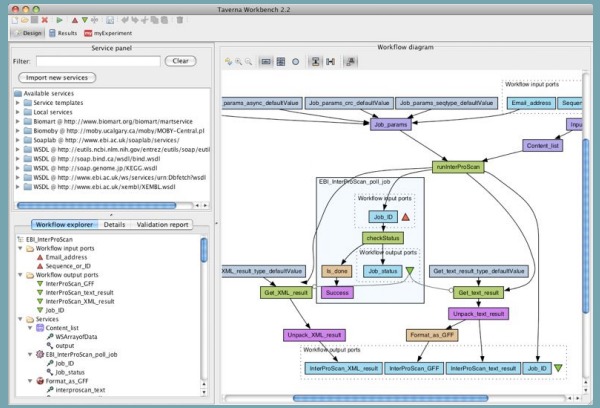
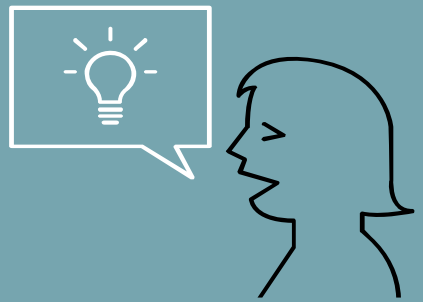


Figure 1. An example of a simple workflow that retrieves a weather forecast for the specified city





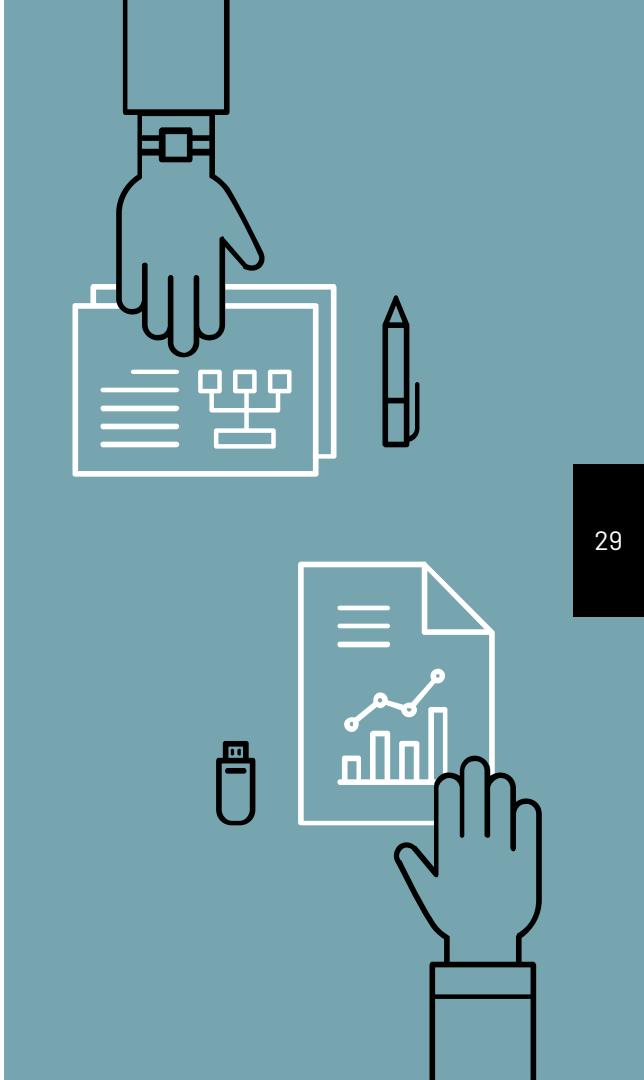
In-Class Project

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

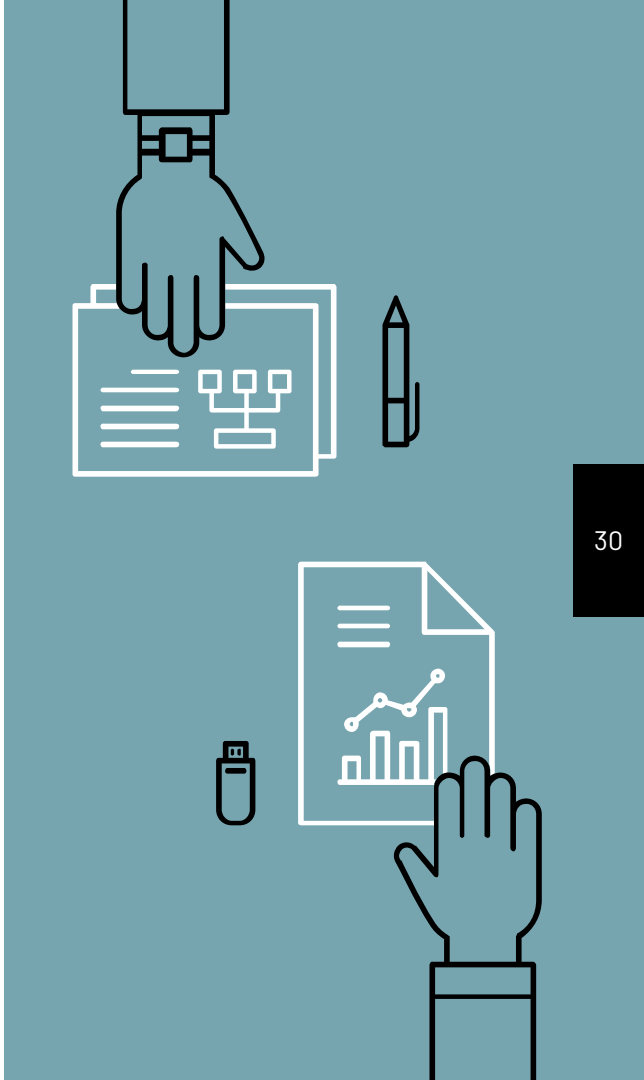


We will attempt to reproduce the results from and critique the reproducibility of:

[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