# reproducibility lecture

wbg231

January 2023

## 1 Introduction

- explains her role
- how to convert from one file format to another
- can make an appointment about many topics
- reproducibility has been a consistent concern for a long time
- reproducing research is hard
- the majority of studies across all dispense were found to be irreducible
- these can have effects that are important
- reproducibility is when independent people use the same code and data to verify results
- code and data reproducibility are part of this, we want to make the entire pipeline reproducible
- replication, independent people use the same data and dirent analysis to validate results
- reviewable research is just the article
- open and reproducible research means we have the all the data and the computational frame work
- we need to be able to test the reproducibility of results
- even software results can vary depending on the hardware or operating system that is being used
- reproducibility allows one to use current work or code more easily down the road

## methods

- need to have set practices for data and code management
- at the start of project set the following
    1. storage adn backup
    2. file formats
    3. project structure
    4. version control
    5. documentation
    6. and group roles
- following these tools will make sure the work is understandable by machine but humans as well
- should have 3 copies of all research material one in a different physical location than you
- original copy, cloud service and external hard drive
- the life time of a external hard drive is like 3 to 5 years
- nyu google drive good for moderate use data
- box is good for big data with security concerns
- nyu hpc can also let you back up and store data
- if working with highly sensitive data ask questions

## project structure and documentation

- it is easy to louse your research data
- do not put documentation in file path put it at a readme file
- make sure to explain the version of your pyhton
- also comment how the data is collected and interpreted might help
- if using jupyter can write what you are doing while you are writing it
- file naming is important keep choices logical
- add prefixes to you data with the data created YYYY-MM-DD so you can sort it easily
- don't use spaces in your file names

- keep file names short, a file name that is
- keep a standard way of organizing projects
- try keeping each project in its own directory
- put documentation in docs folder
- put data in a data folder
- code in a src folder
- results in a results folder

### file types

- use a common file type that is file agnostic

### version control

- try to use version control if you are writing code
- github enables collaboration can undo and re-do version control
- git is a free open source tool that can be used to store gitrepos
- git works best with plane text formats
- use git to keep track over versions
- once we have established habits for code and code environments

### computational environments

- there are tools that will help with computational reproducibility
- containers are pretty common and liked
- web based ides are pretty popular as well
- when processing and analyzing tools dockers can be good
- web based are good for longer term reproducibility

### using containers

- a way to produce virtual operating system outside of the operating system
- docker is the most popular
- learning how to make containers work can be tough

### web based ides

- web based ides are a nice
- just ide's on the browser, allow users to export or share work in the web
- whole tale is popular for reproducibility
- web ides require that you work in there platform first. that can be tough if you need to different computing environments

### web based replay system

- take links to code hosted elsewhere and makes a runnable ide for git
- this is pretty flexible
- **binder** is pretty popular for this
- packaging tools export all of the computational environments you used
- they pack everything you used in one place
- reprozip traces all steps and dependencies and can make a protbale bundle for everything that needs to be re ran later
- try things out and see what works well.