Module 4.3: Learn - Professional Development

Overview

In this installment of professional development, we want to introduce you to best practices to help you document the code and tools you use to do analysis such that you and others can successfully apply these techniques to other questions. Time erases your memory so it is best to be meticulous note takers as we go so we can have something to refer to later on. If we can develop habits that make it easy to figure out details about which tools and data files were used, that makes it easy to pass our knowledge on to others for collaboration and growth.

How to document methods for reproducibility

There are many contexts for when you would need to present information about how you did your analysis. You may be discussing how you did something with your boss or your labmate. You might be presenting your work as a poster for a conference or in a manuscript in a final project or publication. Whatever the reason, it is vital as a computational researcher to be able to document your methods in fine detail so that you can get feedback, spot errors, and contextualize your findings.

Documenting your code for reproducibility

Throughout this course, you have seen examples of coding practices that improve reproducibility:

- Detailed comments and descriptions in the code itself
- Organizing a project such that code and output are together
- High level descriptions of the objective of the analysis and how to run the code
- Descriptive variable names that help keep track of what is being done in the code

When presenting computational methods to others, you want to provide all the information you have needed or w this analysis from scratch. Programming environment tools like RStudio are designed to make it easier for you tc but they hide certain aspects of your code such that you might not realize are key to being able to reproduce you

When documenting your computational methods, be sure to include:

- Programming language you are using with the version you used
  - In RStudio, if you go to the Tools menu and select Global Options, you can see what R version you are us
  - If you are using programming languages in a Linux style environment, you can typically type in the comma followed by
    `--version` to see the version that you have been using (for example, entering R `--version` on the cor
- Non standard packages you are using with the version
  - We have been using the `sessionInfo` function in R to list all the packages we have loaded in our enviro analysis report was printed
  - When documenting your methods, pick out the ones that are important to the analysis and be sure to inclu package in case someone has an incompatibility issue when trying to run later somewhere else
- Source and format of input files
  - What software was used with version numbers and any important parameters used
  - Workflow manager like Snakemake if it was relevant to the analysis or results
  - Links to identifiers to a public data repository
  - References to papers where the data was in the supplementary materials
  - Projects that generate and store data for public use such as the The Cancer Genome Atlas (TCGA)
  - Format of input files
  - How they were generated:
  - Where input files were downloaded from, such as:
- Links to any data you stored in a public data repository
- Links to any applications you host on the internet to help others repeat your methods (web apps)

- General specifications of the computer system you used if the resources on that computer were necessary or
  - In this course, we are working on Agave supercomputer at ASU and we are running RStudio from there w for a particular version of R such, such that we are all working in the same programming environment.

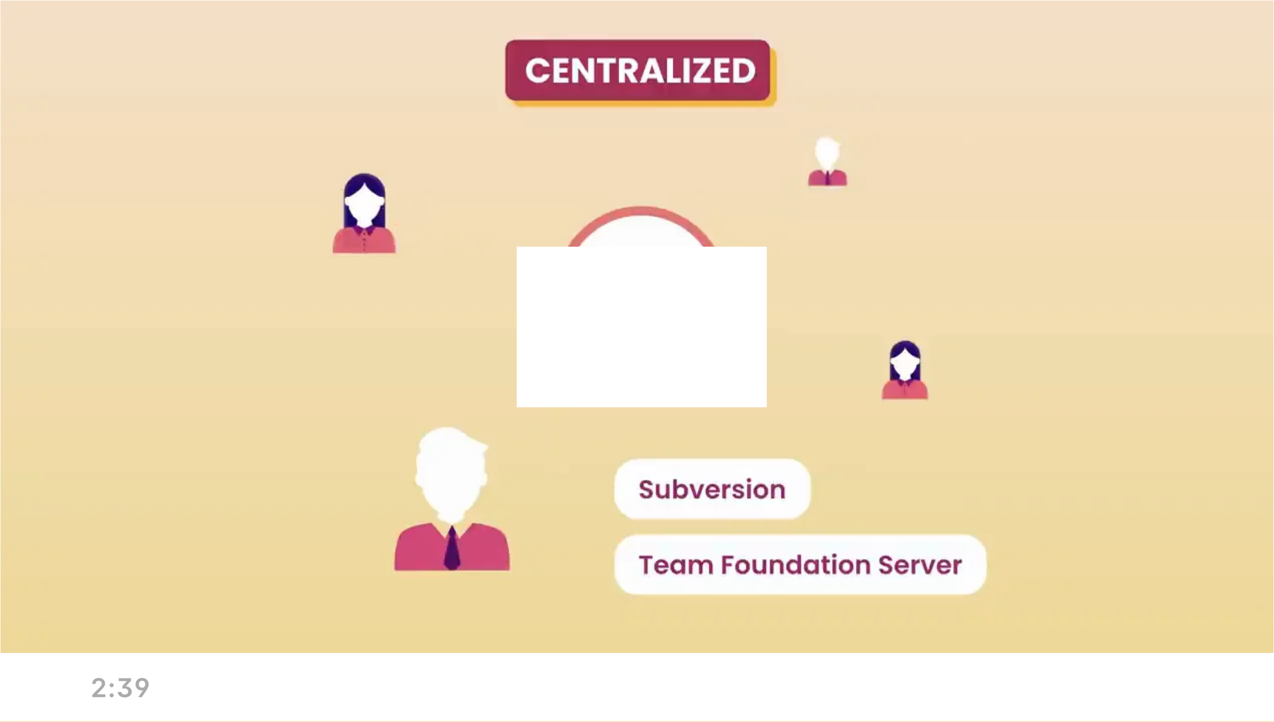
- In situations where we were working with people who do not have access to Agave, we can create a com environment using tools like Docker.

Often the best way to make sure that you have documented your methods well is to run them by someone else. repeat your methods will help you to find the parts that are confusing or have a dependency that you didn't consi

Platforms for sharing code with version control

There are many fantastic tools we can use to share code and coding environments. These tools are meant to he development timeline of the analytical method to help multiple users to view without breaking dependencies dow these tools offer some way to host what is called a "Readme" file, which is meant to contain descriptions of what being developed and what the purpose of those methods is.

One extremely popular tool to share code is Git with its public repository called GitHub. This short video summar



Video. What is Git? Git explained in 2 minutes.
This video will teach you about the Git version control system.

View Transcript. (https://canvas.asu.edu/courses/122165/files/54792235?wrap=1)
(https://canvas.asu.edu/courses/122165/files/54792235/download?download_frd=1)

We have a GitHub repository for our lab in which we have created a directory that contains the code to generate class. We create GitHub repositories for pipelines and tools we develop for working with different members of ou and when publishing techniques in manuscripts.

Module 4.3 Additional Resources

- **Workshop on writing reproducible code (https://utrechtuniversity.github.io/workshop-computational-repr**
- **100 second summary of Git commands (https://www.youtube.com/watch?v=hwP7WQkmECE)**
- **Good enough practices in scientific computing (https://journals.plos.org/ploscompbiol/article?id=10.137**