Using environments in science

Ulises Hernández - Roots for Resilience program

Acknowledgements

Roots for Resilience (R4R) and FOSS

R4R cohort

Director

Tina L. Johnson

Instructors

Dr. Jeffrey Gillan

MSc Michele Cosi

I encourage you to apply. My proposal and info is available at the Graduate students repo. You can also contact me for more information.

Open Science

"Open Science is transparent and accessible knowledge that is shared and developed through collaborative networks"

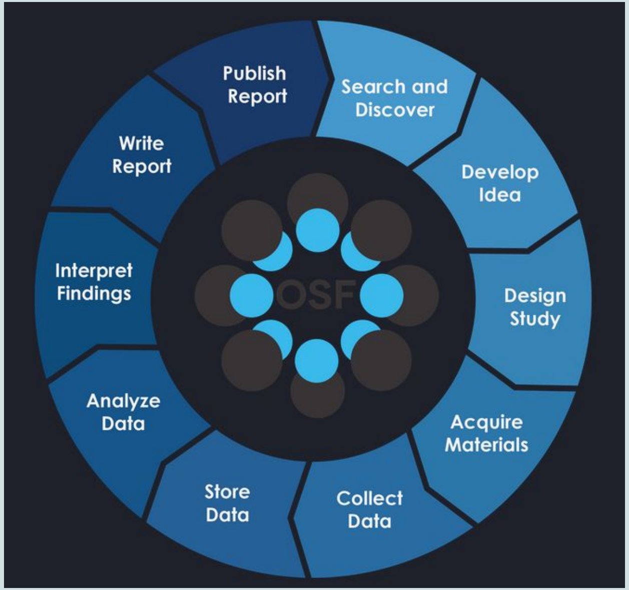
-Vicente-Saez & Martinez-Fuentes 2018

Nelson Memo states that research funded through tax-payers must be open access by 2026.

- For example by publishing in Open Access Journals or distributing the Author's Accepted Manuscript

University of Arizona Libraries offers information on Open Access publishing.

The Research Life Cycle



Open Science Framework

Open Data

FAIR Guideline Principles (Wilkinson et al. 2019). Data should be:

- Findable: Easy to find by the public. Data is indexed or registered in a searchable source.
- Accessible: Use of metadata and clear language for how to access and use the data.
- Interoperable: Use of standards to encode and exchange the data.
- Reusable: Able to be used again to maximize the research outcome.

Exceptions for sensible data such as data on human health, groups that do not want their data to be public, and endangered or cultural keystone species.

If applicable Data Storage should follow the FAIR principles.

ReDATA project by the University embodies the FAIR principles for data collections.

Open Methodology

Jonathan has been struggling to build the code to analyse their data.

They found a blog stating that someone worked on the same code. Hurray! A big smiles draws on their face.

They go to the repository for the code and founds out there is no information about how to use the tool.

Optimistically they try to build the code by themselves.

After hours of hard work they failed :(. The big smile leaves their face.

Have you been in Jonathan’s position?

Open Methodology

"An open methodology is simply one which has been described in sufficient detail to allow other researchers to repeat the work and apply it elsewhere." - Watson (2015)

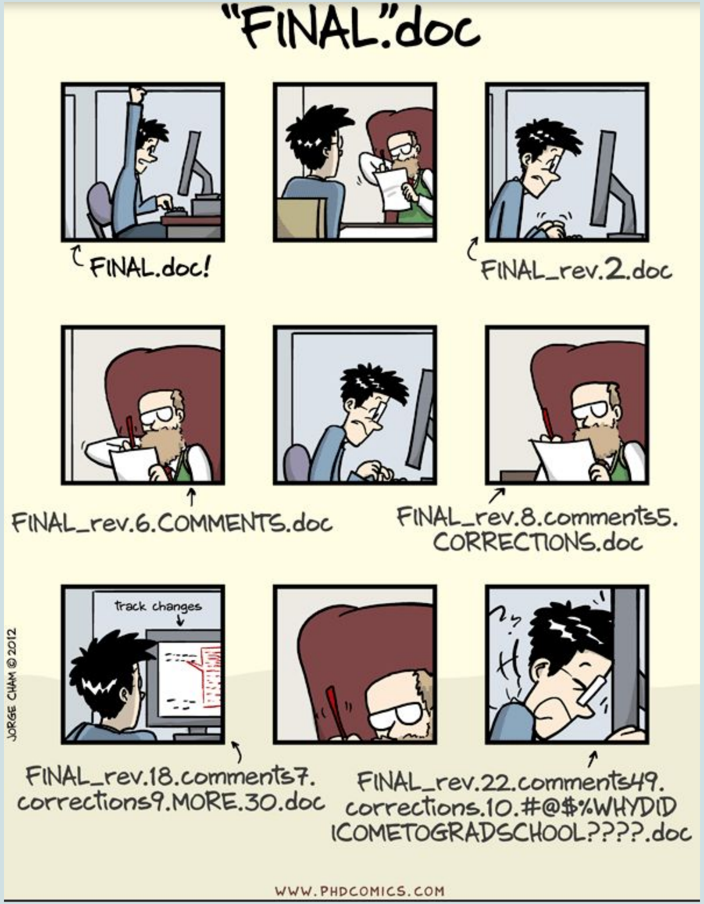
- Describe how to use or build your computer code. Should focus on clarity, completeness, accuracy and relevance.

Use of README files to describe the project.

Version control: Git

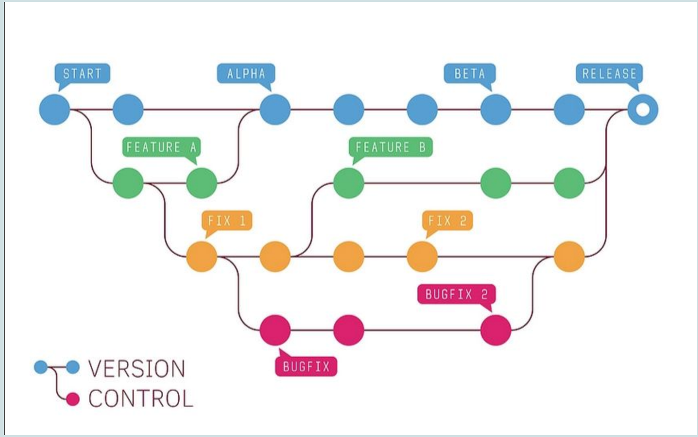
Git offers a way to track the changes among different versions of a project through tree structures.

Allows to recover previous versions of the project and see how the project have evolved through time.



Version control: Git

Git offers a way to track the changes among different versions of a project through tree structures.



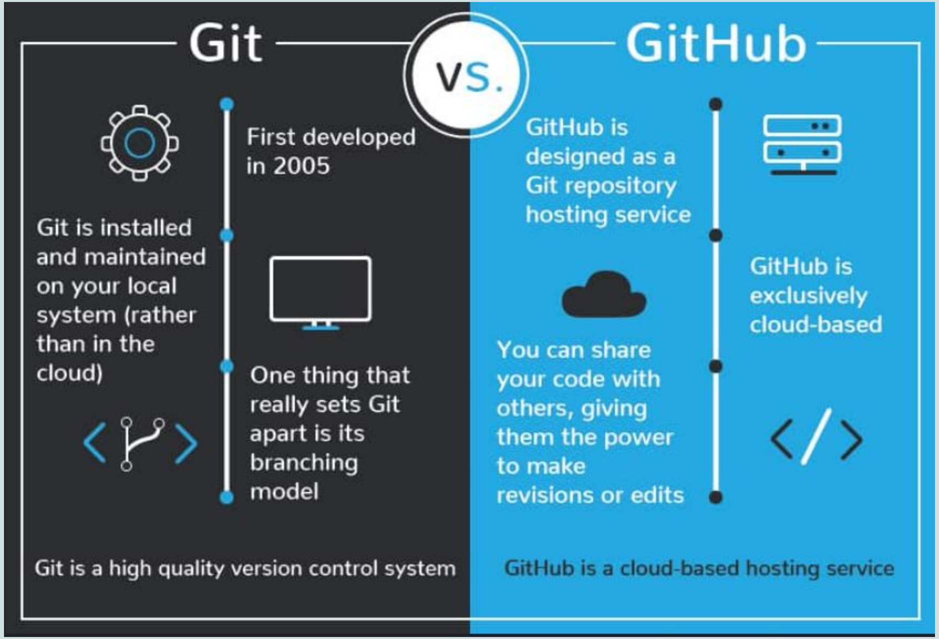
Allows to recover previous versions of the project and see how the project have evolved through time.

Allows collaborative science. People can work in parallel in the same code using local copies and then merge changes to the main project.

Where is the main project stored?

Git vs Github

GitHub is and option to store your project on the cloud. Git tools are also available through the GitHub website.



Git

VS.

GitHub

First developed in 2005

Git is installed and maintained on your local system (rather than in the cloud)

One thing that really sets Git apart is its branching model

GitHub is designed as a Git repository hosting service

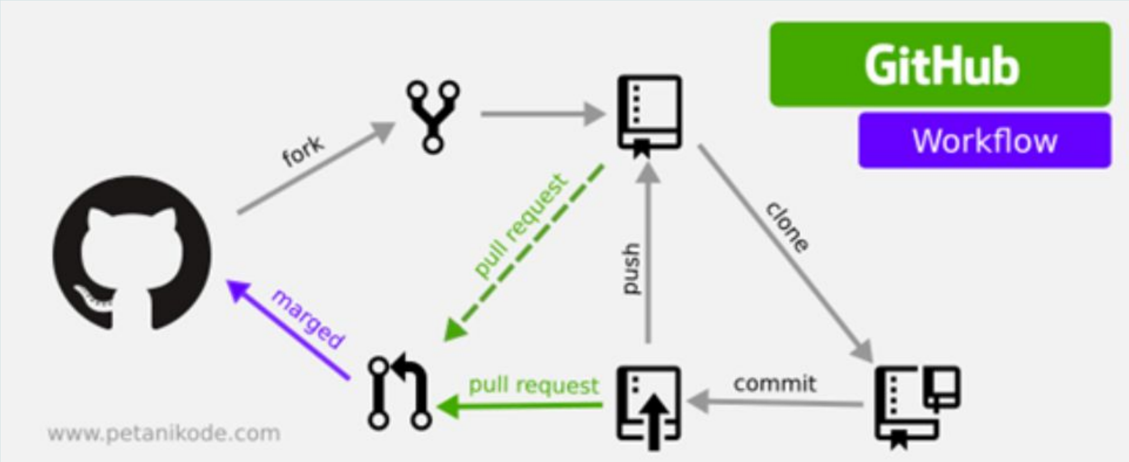
GitHub is exclusively cloud-based

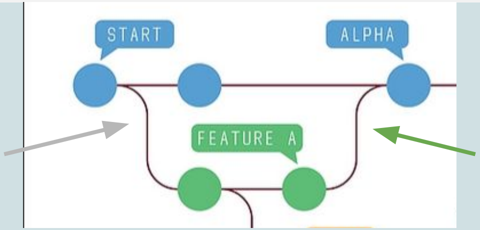
You can share your code with others, giving them the power to make revisions or edits

Git is a high quality version control system

GitHub is a cloud-based hosting service

Github workflow





Fork

Pull request

Example: Go to github.com/ulisesjiu, then to repositories and click on HighUd_mutationaccumulation project.

Reproducible Science: Computing environments

Combination of hardware, software and network resources used in a project.

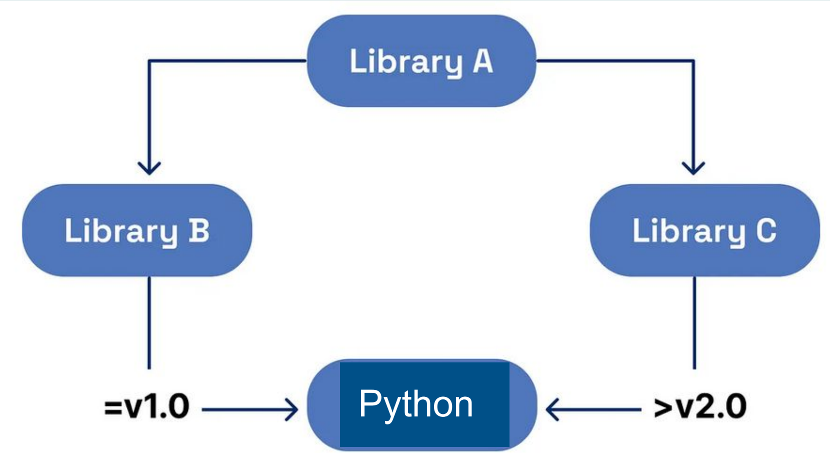
- CPUs, GPUs, RAM
- Operating system and version: Linux, MacOS, Windows?
- Software versions: R, Python, etc
- Package versions

Everytime we build a script we tested it in our own computer environment. May not work on someone's else computer or even on your own computer in the future.

Software Dependency Hell: Inability to reproduce a project due to lack of shared environment.

Reproducible Science: Computing environments

Collaborator 1



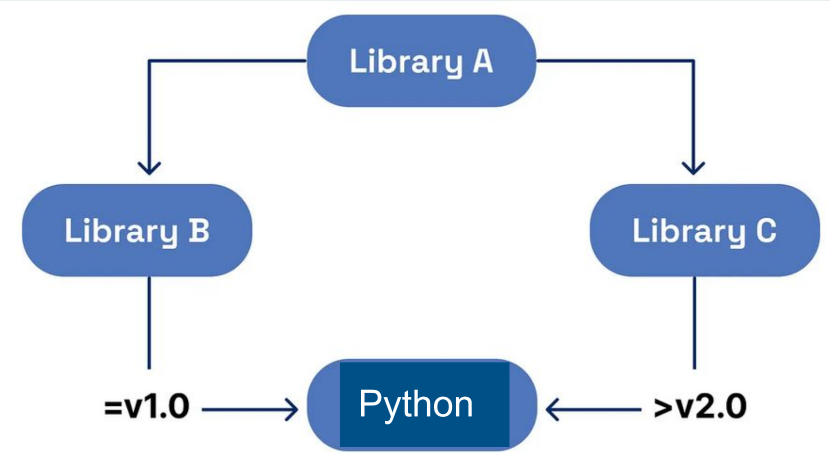
Collaborator 2

Software Dependency Hell: Inability to reproduce a project due to lack of shared environment.

Reproducible Science: Computing environments

Collaborator 1

Collaborator 2



Software Dependency Hell: Inability to reproduce a project due to lack of shared environment.

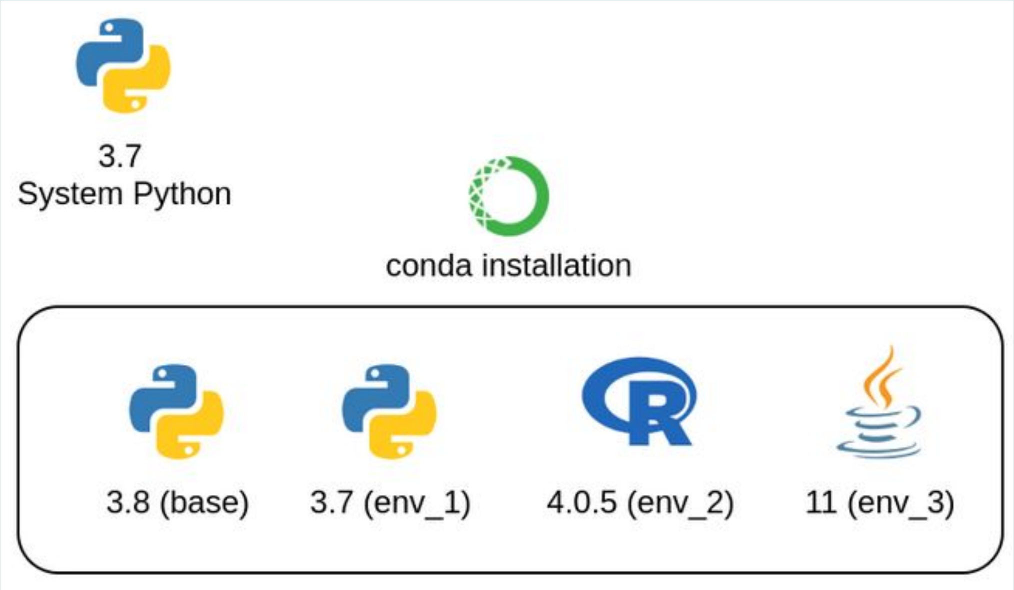
A solution: Environment Managers

Environment managers: Conda

Conda offers a solution to manage environments on Python based projects.

For R based projects renv is a better option. Its use is similar to conda.

Your computer's default environment



Your project's environment

Conda: working with environments

Creating an environment in conda is as simple as typing in the terminal:

- conda init
- conda create --name myenv
- conda activate myenv

After this you can install all your libraries using conda as:

- conda install python=3.9

Conda will manage your environment and make sure it stays independent from the rest of software installed in your computer.

Once you are ready to publish your work you can share your environment in a yml file:

- conda env export --no-builds > myenv.yml

Conda: reproducing environments

Reproduce someone’s else environment in conda is as simple as typing in the terminal:

- conda init
- conda env create --file myenv.yml
- conda activate myenv

Let’s work with an example using GitHub codespaces.

Open your forked HighUd_mutationaccumulation project.

Conclusions

Git plus GitHub allow us to use version control allowing a great option to organize our projects. They also serve as a platform for Open Science.

Environment managers allow a powerful tool to keep control of the dependencies used in a project. They also serve as a platform for Reproducible Science.

These platforms are not complete without the FAIR principles and Open Methodology.

Scientists should focus into practicing Open and Reproducible Science when possible.