Versioning of data and code using Git

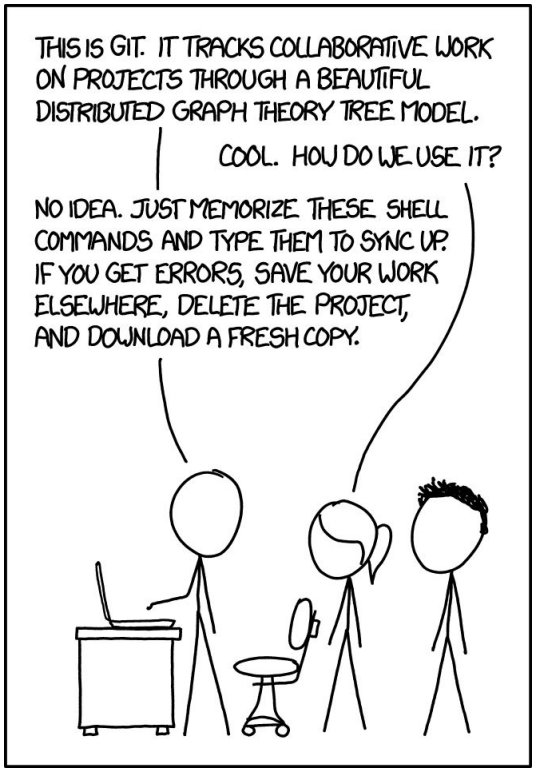
Introduction to Data Management Practices course

NBIS DM Team

data-management@scilifelab.se

https://nbisweden.github.io/module-versioning-dm-practices/

Learning Objectives



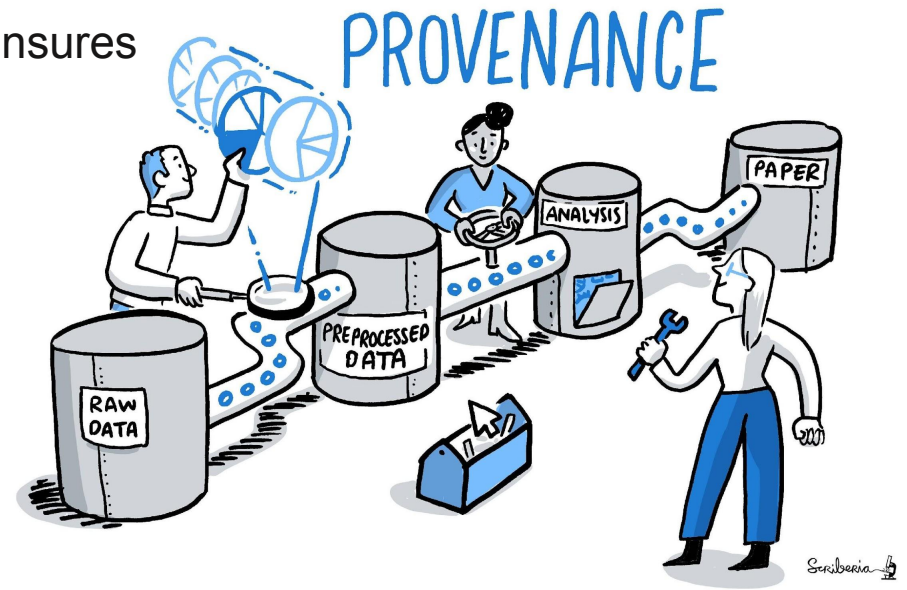
- Explain the benefits of version control
- Explain the differences between git and GitHub
- Create your first GitHub project
- Understand how one can collaborate with others using GitHub

Why use versioning for research projects?

Track project history and increase reproducibility

Keeping a record of the changes made ensures that the research findings can be:

- validated
- collaborated upon
- documented accurately



Why use versioning for research projects?

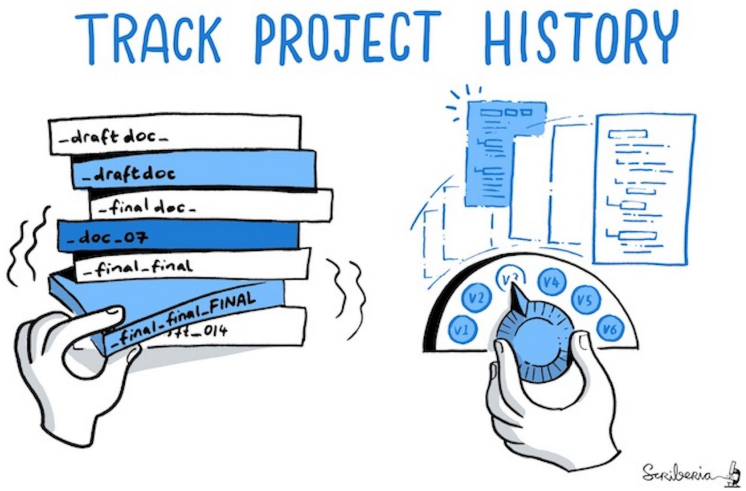


Image from The Turing Way community by Scriberia - CC-BY 4.0
DOI: 10.5281/zenodo.3332807

Benefits:

- Store history of changes, **who** changed **what** and **when**
- Go back to a previous version
- Find out when a problem was introduced
- Formalise ways of collaborating
- Useful for many types of files
- Access to same files from different computers - lightweight backup

What is version control?

- Automatically creates a **robust and rigorous log** of changes to a file, without renaming files (v1, v2, *final_copy*)
- **Minimal storage requirements** - securely stores only the necessary information to recreate previous versions
- **Prevents user errors** and performs conflict checks
- **Coordinate collaborative changes**

GitHub Desktop

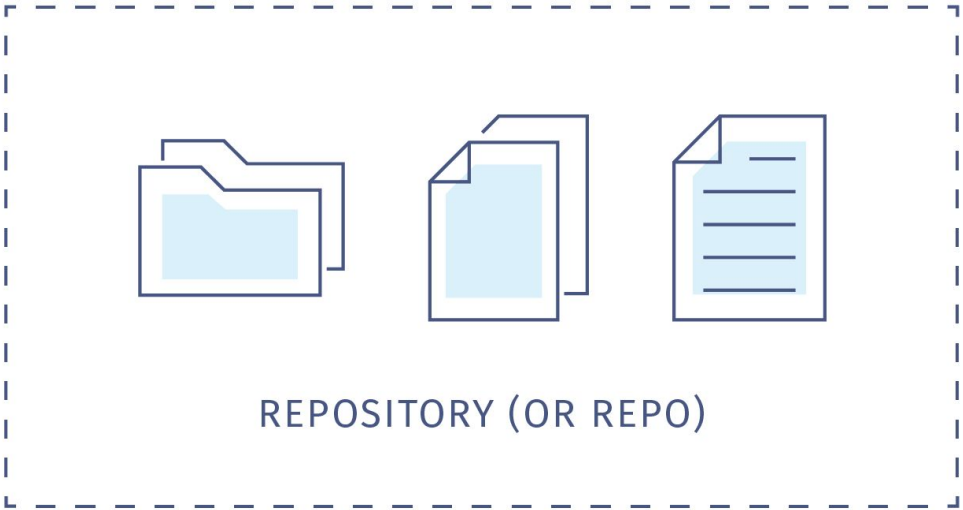
Git is one of the most widely used version control systems in the world. It is a free, open source tool you can install locally on your computer

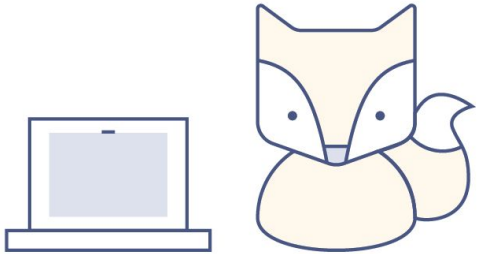
GitHub on the other hand is a **popular website for hosting** and sharing Git repositories remotely. It offers a web interface!

GitHub Desktop is an app on your computer and replaces the use of a terminal window to use git and GitHub

Slide adapted from: Malvika Sharan. (2020). GitHub-Developing-Collaborative-Document. See DOI: 10.5281/zenodo.3835657

Repository

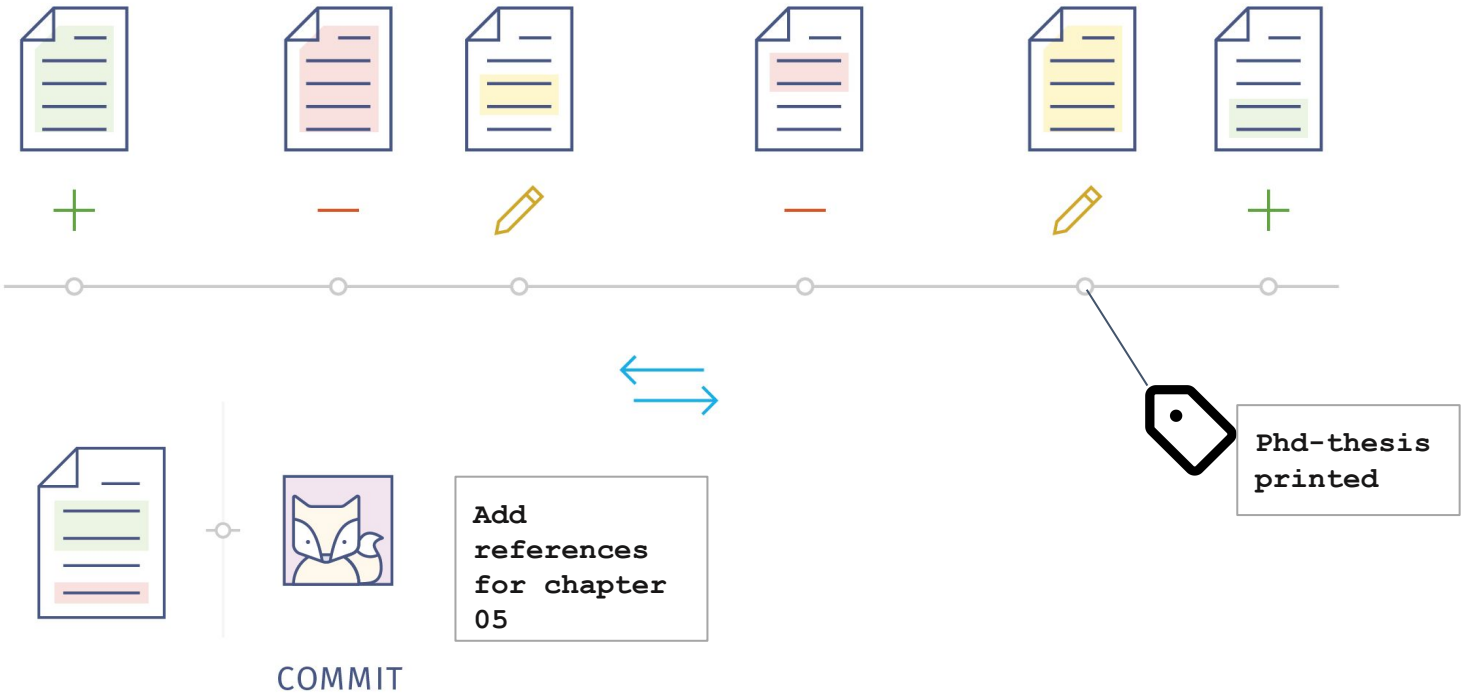




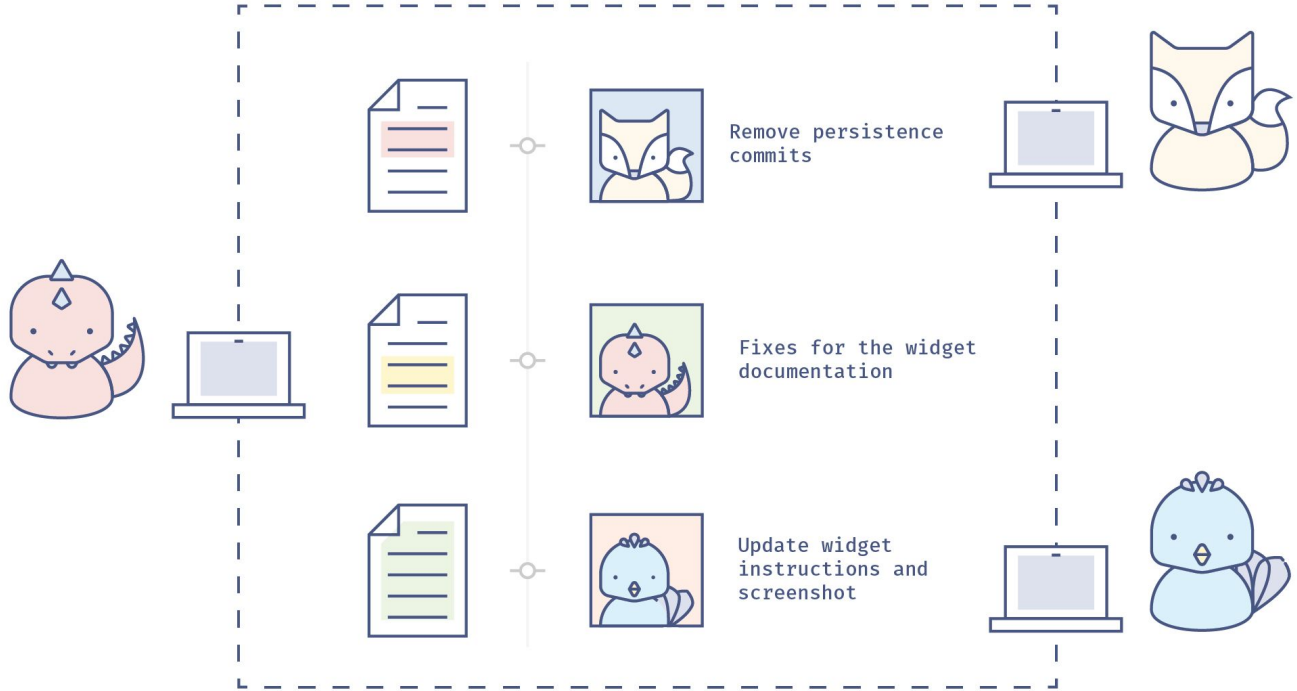
A repository can be **private** or **public**.

Make sure public repositories don't contain any secrets or 'sensitive' information.

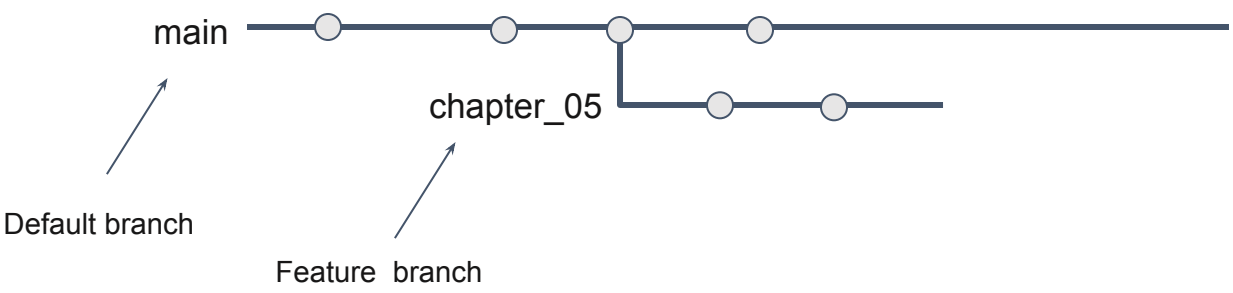
Real-time tracking of changes - commits and tags



Collaborating



Collaborating using branches



merge

Pull Request (PR)

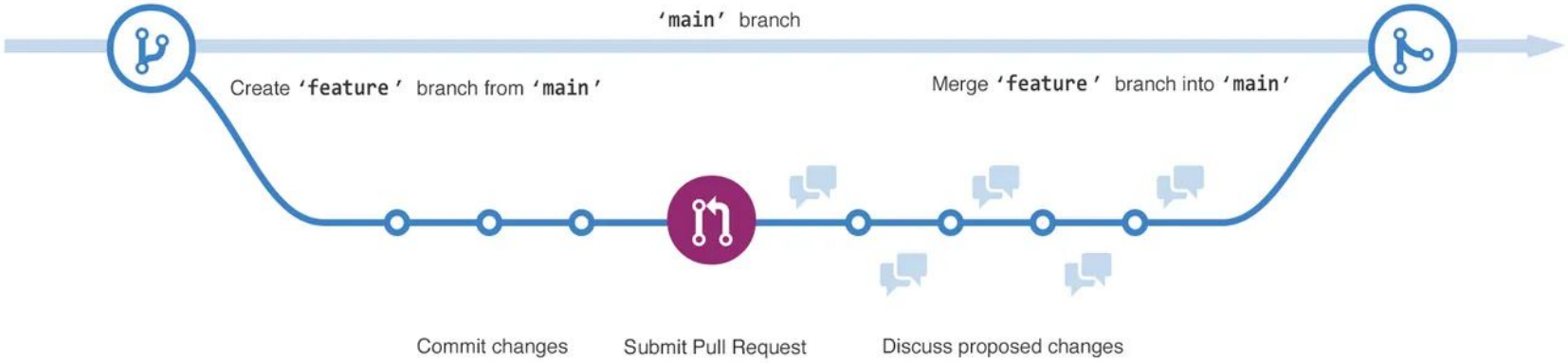
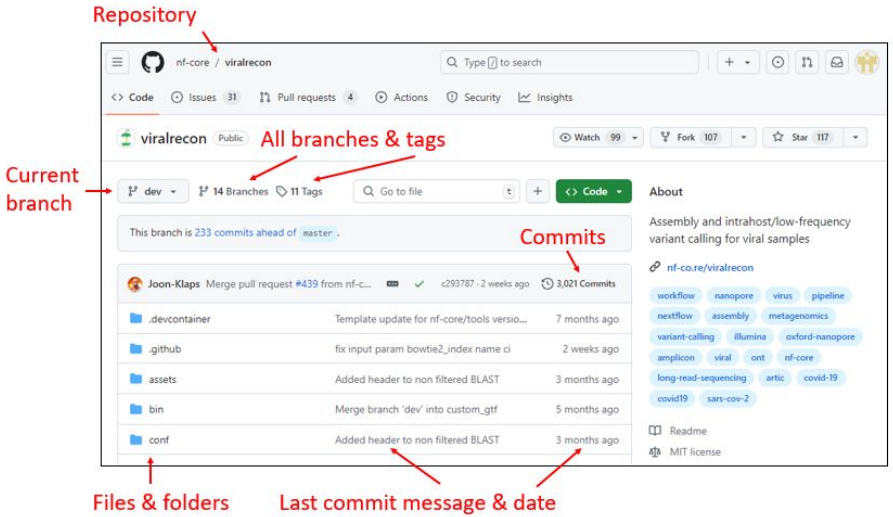


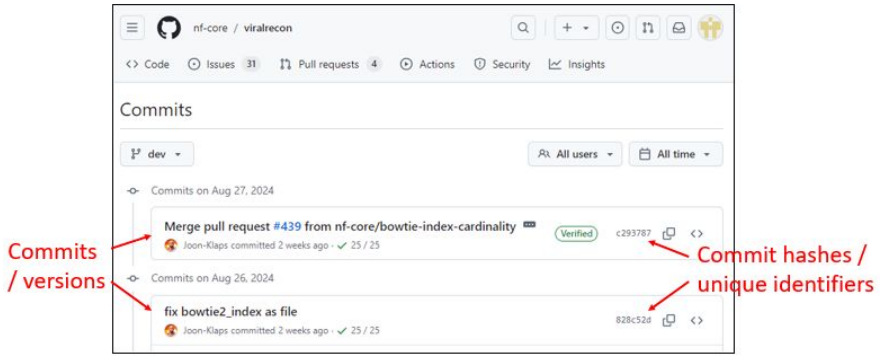
Image from GitHub Docs https://docs.github.com/en/get-started/start-your-journey/hello-world - CC0 1.0

Navigating the GitHub web interface

- Login to GitHub and explore the nf-core/viralrecon repository
- Change branch from master to dev

- Click on the commits of viralrecon-commits-dev





Glossary for Git & GitHub

repository/repo: a collection of documents related to your project, contains all data and history (commits, branches, tags)

cloning: copying the whole repository to your laptop - the first time

commit: a saved change/version of the project, gets a unique identifier (commit hash)

tag: a pointer to one commit, to be able to refer to it later. Like a sticky note that you attach to a particular commit (e.g. *phd-printed* or *paper-submitted*)

Glossary for Git & GitHub

branch: a copy of a repo that is contained within the original repo. Branches are used to work on different project features without altering the original or "main" repo.

merge: the act of incorporating new changes (commits) from one branch to another.

pull request (PR): a request to merge a commit or collection of commits to a repository

Quiz

Which of the following statements are correct concerning version control?

1. Version control serves as a standalone backup system.
2. Version control systems makes it easier to collaborate when working with the same file.
3. Using a version control system, I don't have to wonder which version of the manuscript was sent to publisher.
4. Git is exclusively designed for managing software development and cannot track changes in other types of files like documents or data.
5. GitHub Desktop is a version control system.
6. Cloning a repository in GitHub copies the repository to your local machine.

Quiz - solution

1. No, version control keeps track of different versions, but your work still needs to be backed up. Relying on only the remote repository at GitHub, does not comply with the 3-2-1 backup rule of having at least 3 copies, on 2 different types of digital media, and at least 1 separate location.
2. Yes, collaborations, and resolution of conflicts when editing the same file, is much easier with a version control system.
3. Well no, all versions are safely kept, but unless you put a (name) tag on a specific version, the system will not be able to help you on which version was sent to publisher.
4. No, Git started out as a software versioning control system, but can track a lot of other types of data as well.
5. No, GitHub Desktop is an app that provides a graphical user interface to the version control system Git, by communicating with the web service GitHub.
6. Yes, the cloned repository is stored on your local computer and connected to the remote repository on GitHub. Changes made locally can then be transferred by pushing them to the remote repository and in the other direction by pulling commits.