Tools and tricks for reproducible research

Session 11 of the

Basic Biostatistics &
Bioinformatics workshop series

A collaboration between

SLU's Center for Statistic and

SLU's Bioinformatics
Infrastructure

Amrei Binzer-Panchal, SLUBI

SLUBI - SLUs bioinformatic infrastructure

Created to support SLU's teachers, staff and students in the use of bioinformatics

Consultation, training, data analyses…

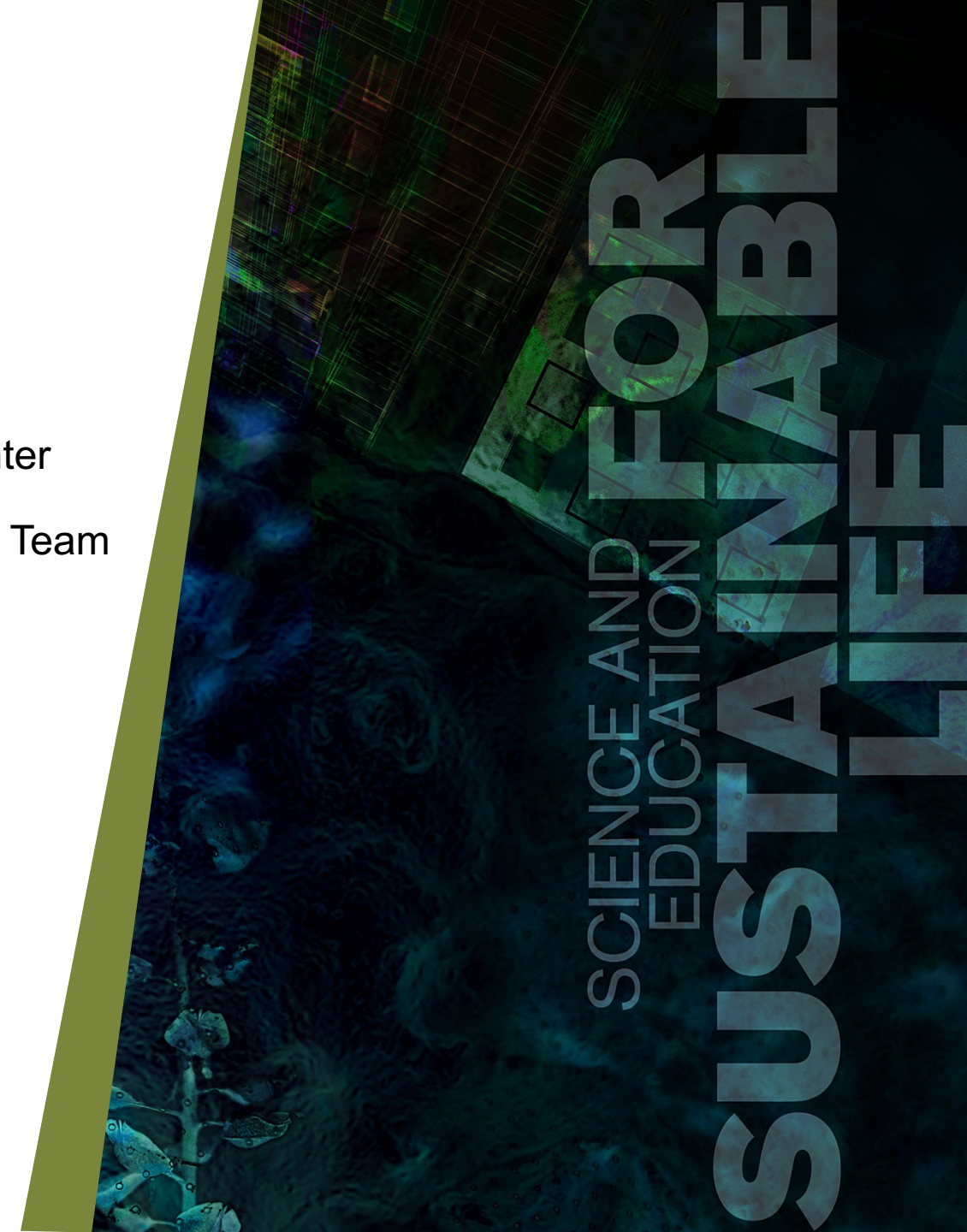
www.slubi.se

slubi@slu.se

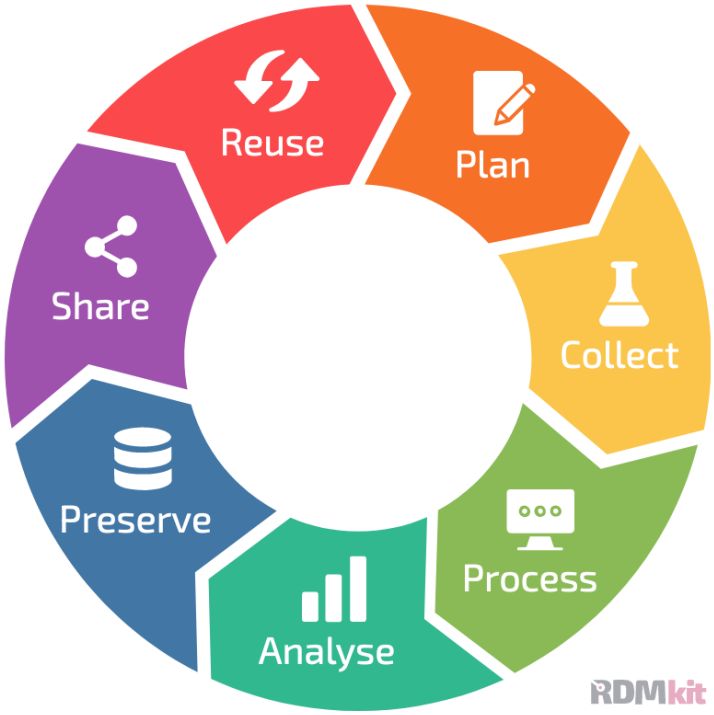
SLU BI BIOINFORMATICS INFRASTRUCTURE

Resources for you

- SLU's Data Management Center
- SciLifeLab Data Management Team & Data Centre



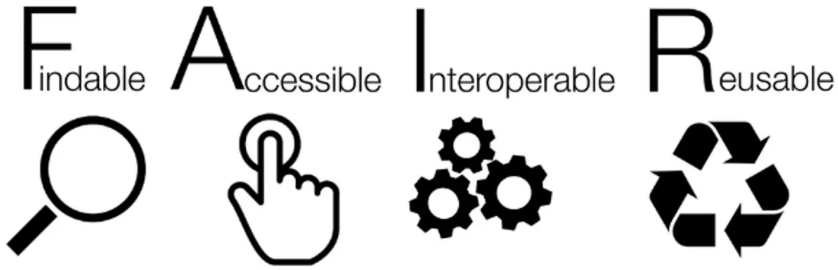
In a typical project, the data is going through a lot…



https://rdmkit.elixir-europe.org/

FAIR principles

- Promote **efficient data discovery and reuse** by providing guidelines to make digital resources



- Address aspects enabling software and infrastructure to automatically find and use research data

Wilkinson et al. (2016)

https://www.tpximpact.com/knowledge-hub/insights/fair-data-guide/

FAIR data life cycle

- FAIR principles rely on **good data management practices** in all phases of research

– Research documentation
– Data organisation
– Information security
– Ethics and legislation

https://rdmkit.elixir-europe.org/

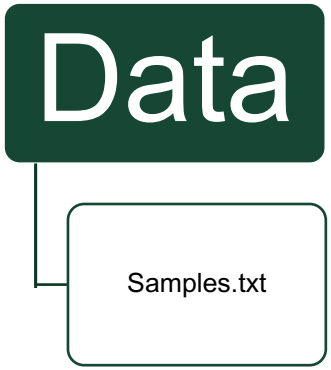
good data management practices: who benefits?

Colleagues	people I collaborate with need to understand what I do
Scientific community	scientists can find and re-use my data
Society	has the right to know what happens with the data
Myself	Future-me will not always remember what present-me did

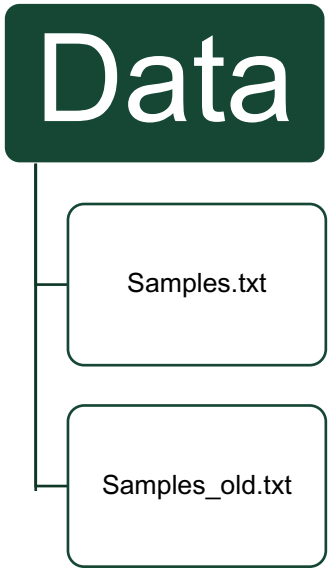
“Your primary collaborator is yourself six months from now, and your past self doesn’t answer e-mails”

Rachael Ainsworth

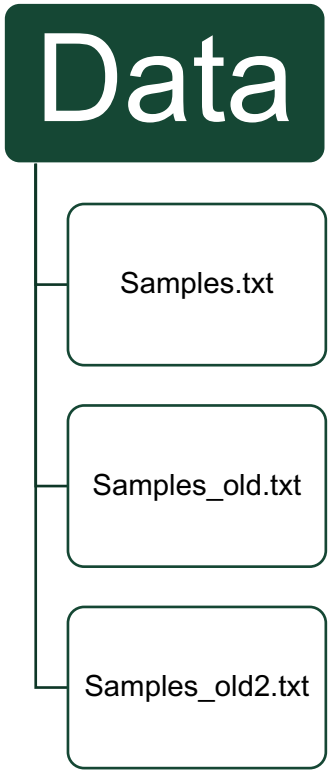
First step: organization



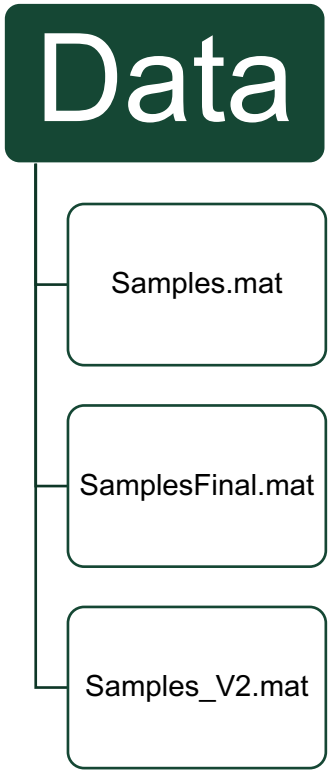
After a while



More time has passed



Which one is the most recent?



best practices - file organisation

- There is a folder for the raw data, which **does not get altered.**
- Code is **kept separate** from data.
- Use a **version control** system (at least for code) – e.g. git.
- There should be a **README** in every directory, describing the purpose of the directory and its contents.
- Use **file naming** schemes that makes it easy to find files and understand what they are (for humans and machines) and document them.
- Use **non-proprietary formats** – .csv rather than .xlsx

https://figshare.scilifelab.se/articles/presentation/Data_Management_in_the_life_sciences/21976727/1?file=39011624

Directory structure sample project

```
Project name
|
| - code/                    all code from input to final results
| - data/                    raw and meta data, NEVER EDIT
        \ - README.md                data details summary
| - docs/                    project documentation
| - results/                 output from workflows and analyses
        \ - README.md                results details summary
```

Noble WS (2009) A Quick Guide to Organizing Computational Biology

Reproducible research

The term *reproducible research* refers to the idea that scientific results should be documented in such a way that their deduction is fully transparent.

This requires a detailed description of the methods used to obtain the data.

https://en.wikipedia.org/wiki/Reproducibility

Reproducible research

- **Codify** everything
- Shell scripts, r scripts, python scripts…

- **Use notebooks** to integrate scripts and documentation

SLU

quarto®

- Creates reproducible documents that can be regenerated when underlying assumptions or data change.
- Supports executable code blocks within markdown, use python, R, Julia, observable, bash…
- Can be edited in any text editor, such as VS Code, RStudio, Jupyter Lab…
- Uses a single source document to target multiple formats, such as articles, reports, presentations, websites, and books in HTML, PDF, MS Word, ePub…
- Supports interactive data exploration in documents using Jupyter Widgets, htmlwidgets for R, Observable JS, and Shiny.

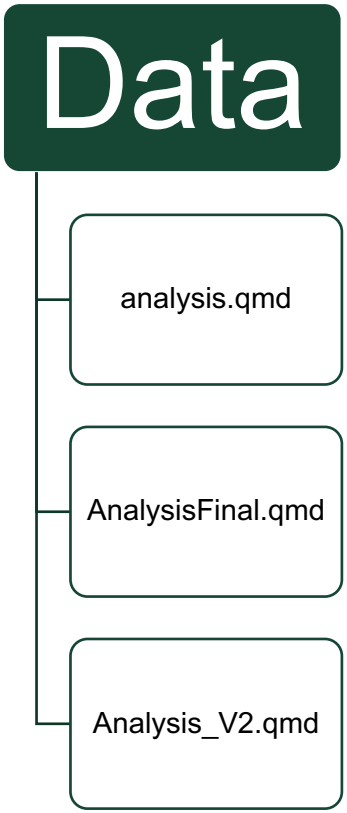
https://quarto.org/

Now I have reproducible documents.

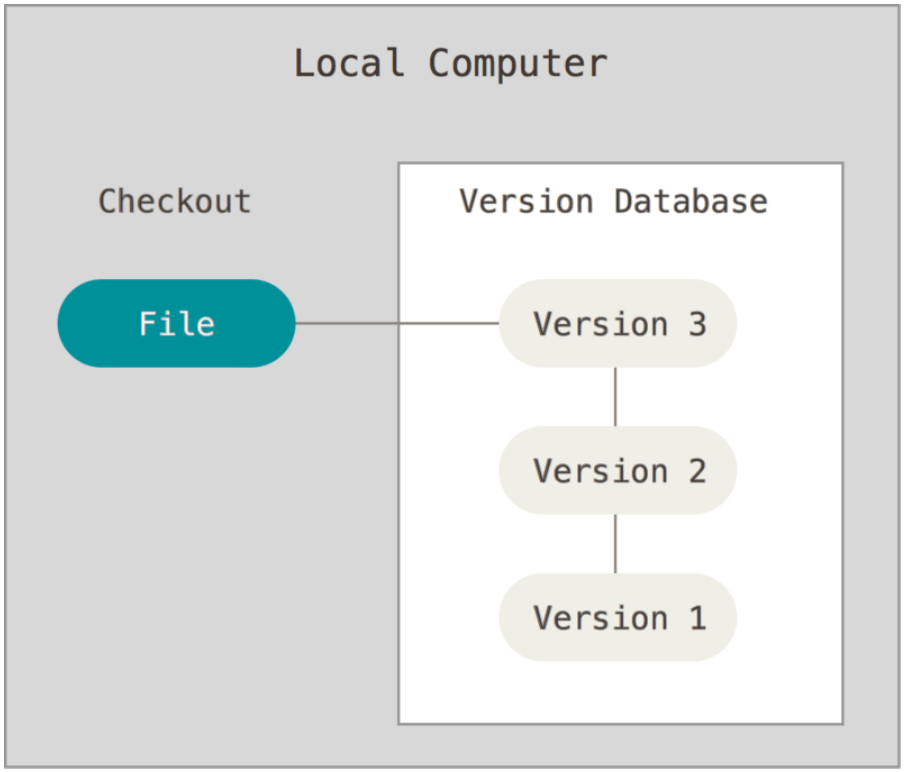
What if I modify my analyses?

Version control

- the practice of tracking and managing changes to files

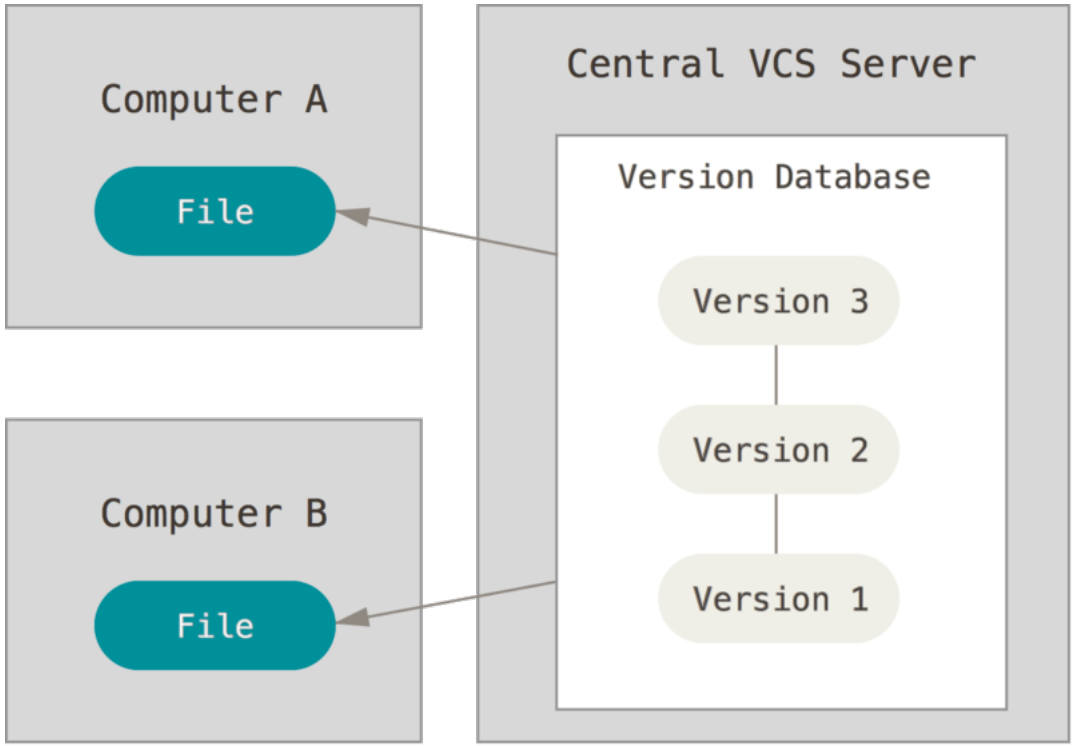


Version control - local



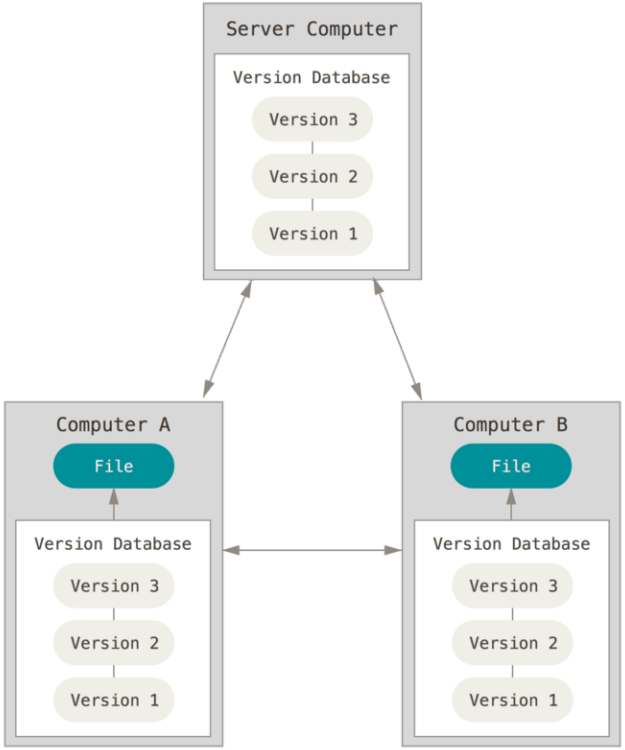
https://git-scm.com/book/en/v2

Version control - centralized



https://git-scm.com/book/en/v2

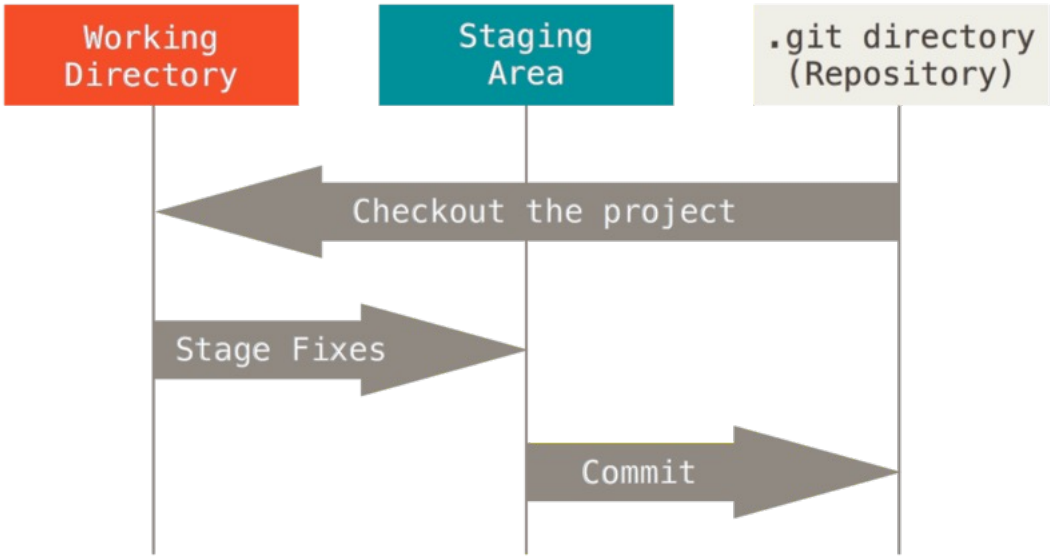
Version control - distributed



https://git-scm.com/book/en/v2

git

- version control system that handles all from small to very large projects
- Simple to use

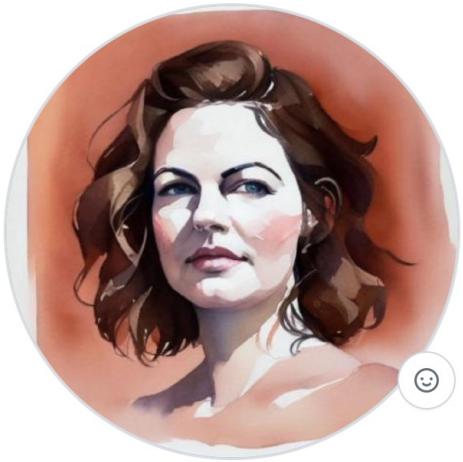


https://git-scm.com/book/en/v2

Github

- code hosting platform for version control and collaboration
- built on git
- For each project you create a repository, which can be public, or private

https://docs.github.com/en/get-started



Amrei

amrei-bp

Bioinformatician for **@SLUBioinformaticsInfrastructure**.

Edit profile

1 follower · **0** following

SLU, the Swedish University of Agricultural Sciences

Uppsala

www.slubi.se

Organizations

amrei-bp / README.md

Hi there 👋 I am Amrei Binzer-Panchal

A bioinformatician working for SLU's bioinformatics infrastructure.

I support research groups at SLU by performing bioinformatic analyses and developing workflows on request. I have started my bioinformatics career with microarrays, but have since worked on sequence based data, performing analyses such as expression analyses, genotyping, single cell analyses, GWAS, metagenomics analyses, and more.

Pinned

Customize your pins

SLUBioinformaticsInfrastructure/Slubi_distill_homepage Public

CSS

nf-metavir Public

Forked from jhayer/nf-metavir

Nexflow pipeline for viral metagenomics analyses

Nextflow

SLUBioinformaticsInfrastructure/Nextflow_training_qc_pipeline Public

This is a short pipeline for students to improve their understanding on using Nextflow as a workflow manager.

Nextflow

SLUBioinformaticsInfrastructure/SBD_booklet Public

This is the code for the booklet for SLU's Day of Bioinformatics in September 2023.

TeX

SLUBioinformaticsInfrastructure/Three_Bees_Workshop_Series Public

This is the repository for the collaborative workshop series between SLUBI and SLU's Center for Statistics.

rnaseq-exercises Public

A repository to practice RNASeq exercises using Gitpod

Shell

Session 12 of the 3B's

Basic Biostatistics & Bioinformatics

Session 12: git

Swedish University for Agricultural ciences

Now I can share my reproducible code.

What about the tools?

Sharing of bioinformatics tools

- Different environments
- Different OS
- Different packaging
- Conflict between tools or versions
- Impact on usability and reproducibility

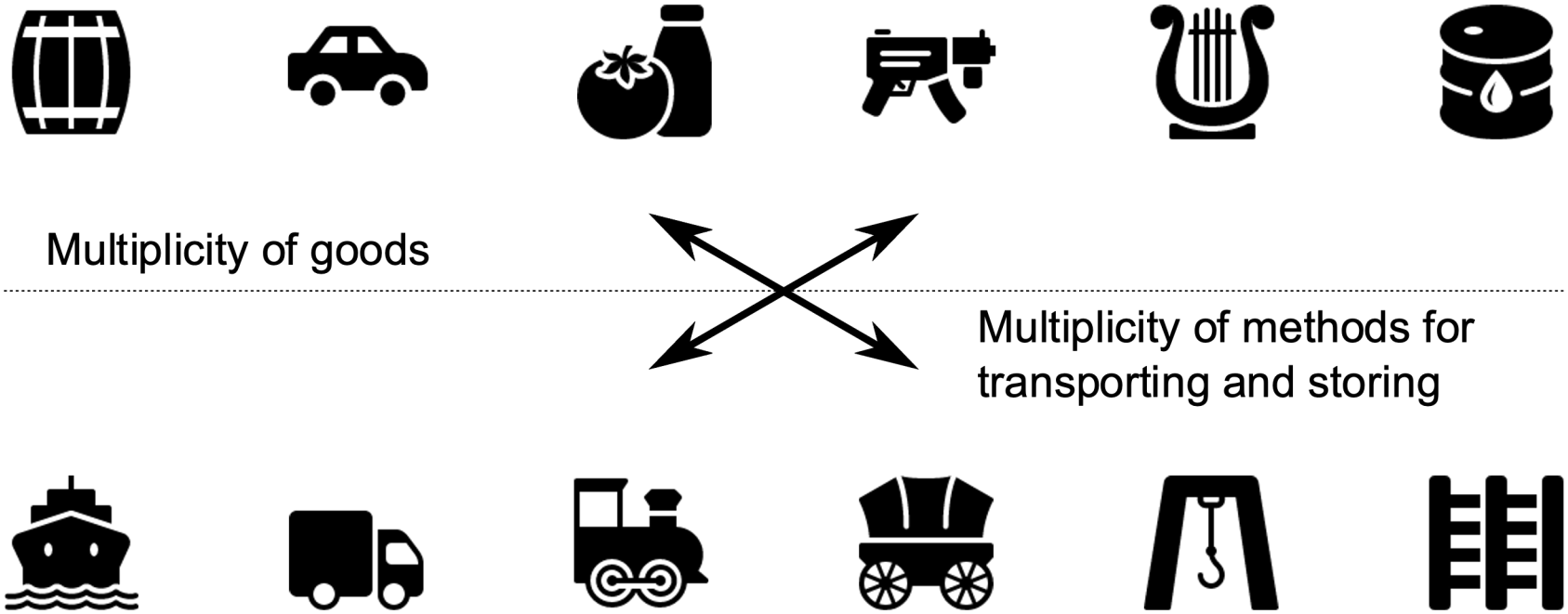
Package/environment management system

- helps you find and install packages
- use the same version of a tool as your collaborator
- !! Might still give different results, depending on the OS

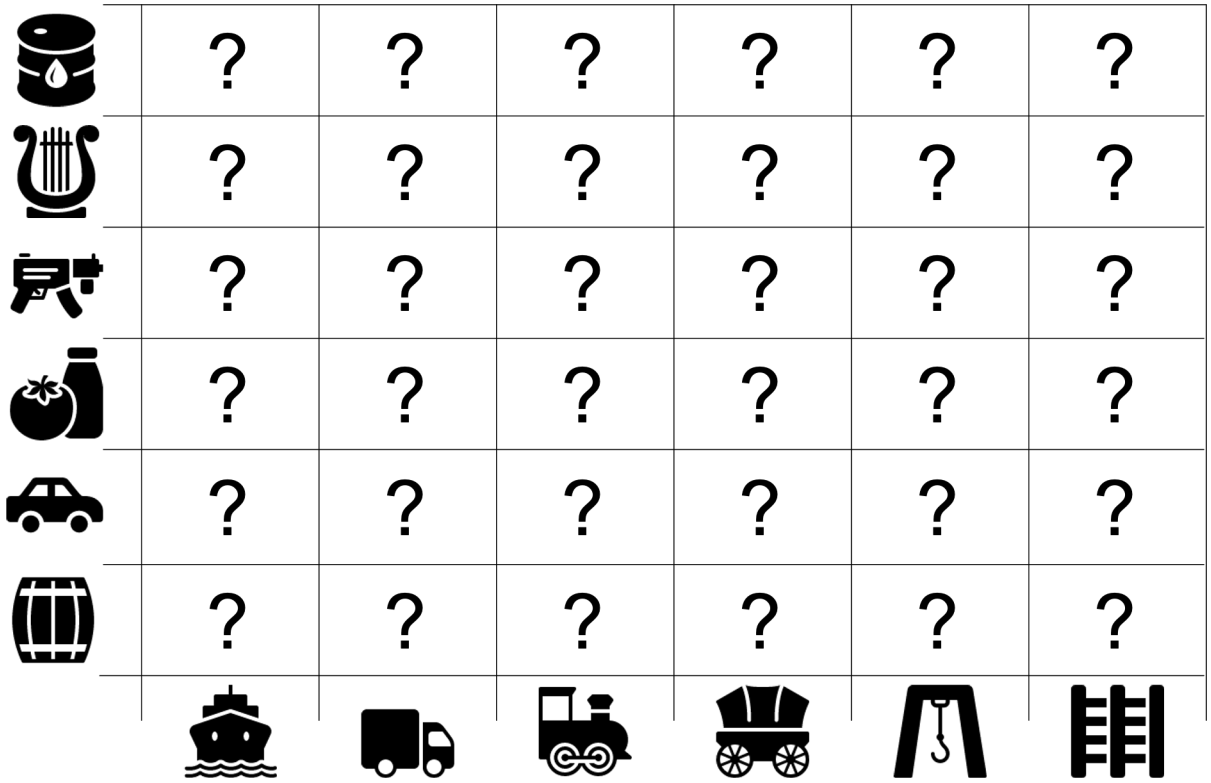
CONDA

BIOCONDA®

Deployment issues of bioinformatics tools

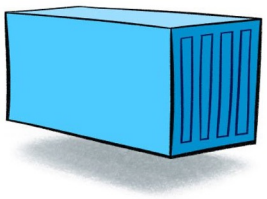


The matrix from hell



Intermodal shipping container

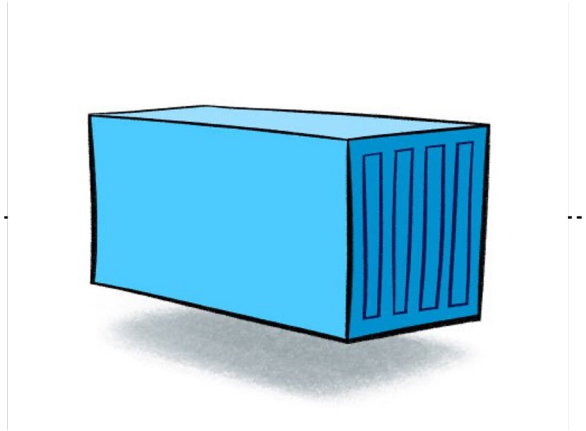
Multiplicity of goods



Multiplicity of methods for transporting and storing

Intermodal shipping container

- A standard container
- Virtually loading any goods
- Sealed until it reaches final delivery
- Can be
  - loaded and unloaded
  - stacked
  - transported efficiently over long distances
  - transferred from one mode of transport to another



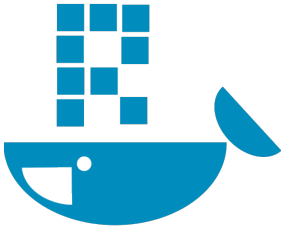
Containers in bioinformatics

Containers in bioinformatics

- separate applications from infrastructure
- contain everything needed to run the application
- do not change over time
- lightweight
- everyone gets the same container that works in the same way

Where can I get containers?

- Build your own
- Use pre-build ones



Now I can share my scripts and my tools.

Can someone make this even easier for me?

Bioinformatic workflows

- Chaining together a collection of tools / scripts.
- Traditionally they are run and managed manually, slowing down the process.
- Time intensive to administer workflows and handle errors on a day-to-day basis.

Workflow manager

- optimize resource usage
- handle software installation and version control
- run on different compute platforms
- enable workflow portability and sharing

Wratten, Wilm & Göke, 2021

Nextflow is a workflow manager which

- enables the writing of scalable and reproducible scientific workflows.
- integrates various software package and environment management systems such as Docker, Apptainer, and Conda.
- allows for existing pipelines written in common scripting languages, such as R and Python, to be seamlessly coupled together.

Reproducibility

- Pipeline tracked and version controlled (Git)
- Native support for containers and conda environment
- The application handling and the configuration/deployment are separated
- The only parameter to change are the resources and environment

nf-core

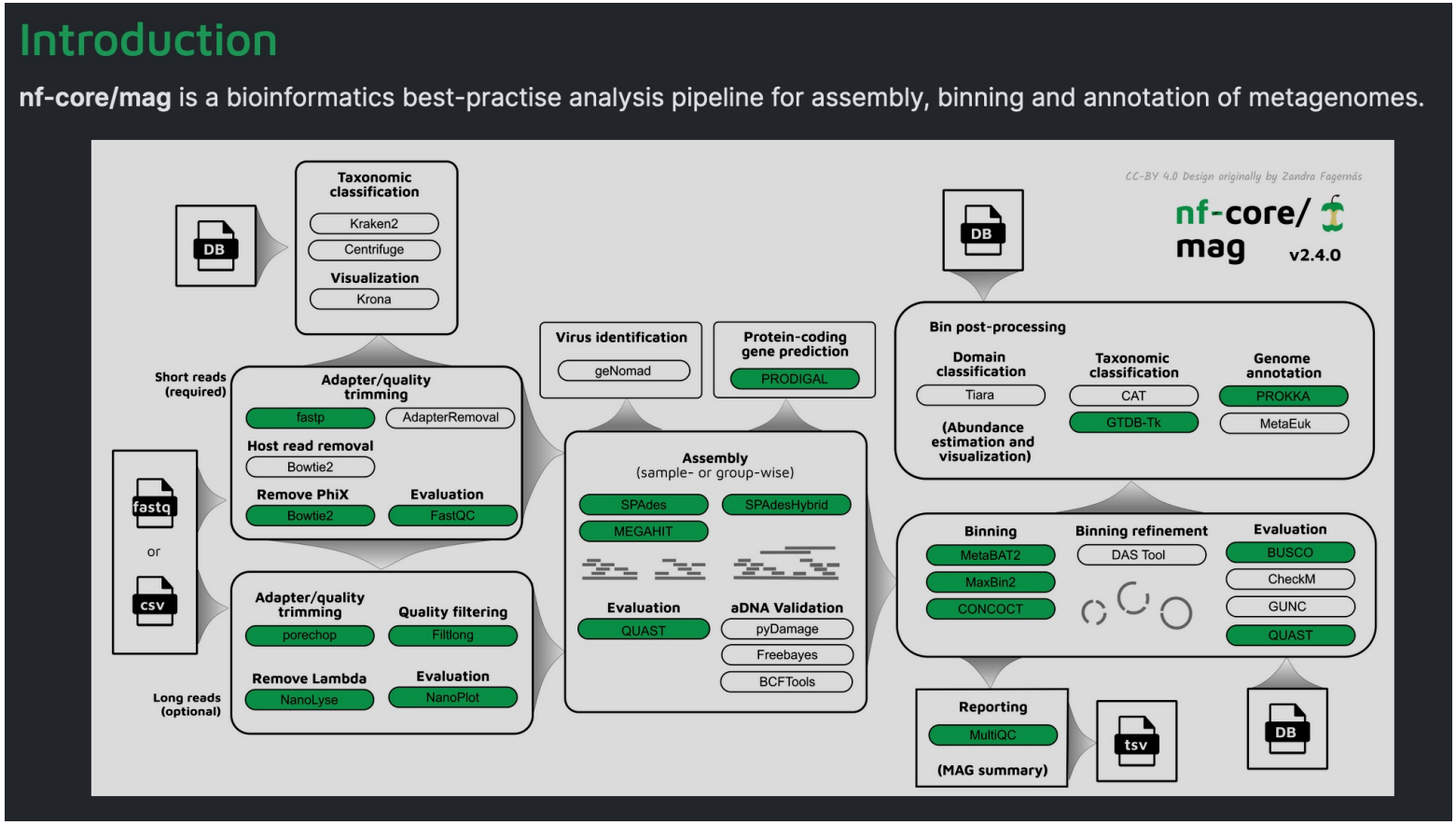
A community effort to collect a curated set of analysis pipelines built using Nextflow.

Wide variety of well documented pipelines.

Open source and version controlled on github.

They offer training in Nextflow.

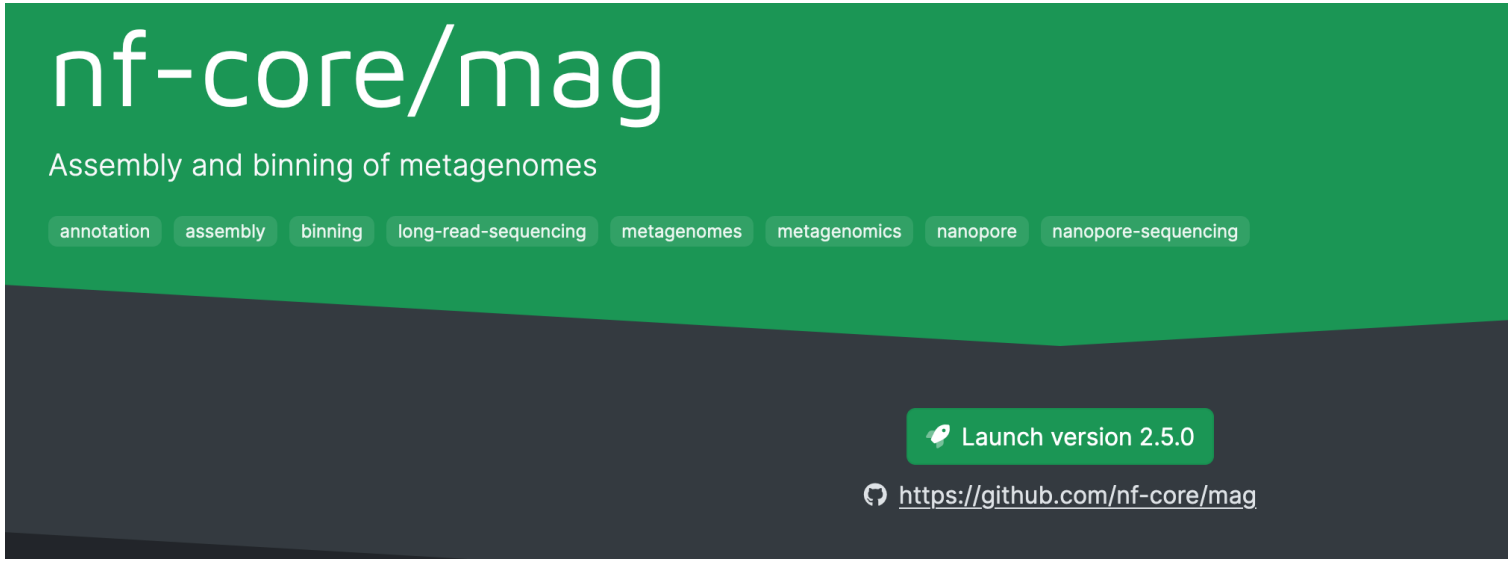
nf-core/mag



https://nf-co.re/mag/2.5.0

nf-core Launch!

Command line tool, or online user interface to configure workflow parameters for a pipeline run



https://nf-co.re/mag/2.5.0

Cloud development

- Gitpod
- Github codespaces
- Ephemeral containers

Take home

- Organize your projects
- Comment and codify
- Version control files with Git
- Containers are your best friends to win time on installation
- Use a workflow manager for a scalable and reproducible pipelines
- Help your future self remember what you did

Thank-you!

- Thanks to Renaud VanDamme & data-managment@scilifelab for slide ideas and content.

- Bonus tools for you:
- Google “github awesome-genome-visualization”


SLU
SCIENCE AND EDUCATION FOR SUSTAINABLE LIFE