Ready for BioData.pt Management?

Intensive Course

Data Processing & Analysis

Daniel Faria, Jorge Oliveira, Gil Poiares-Oliveira

BioData.pt | elixir PORTUGAL

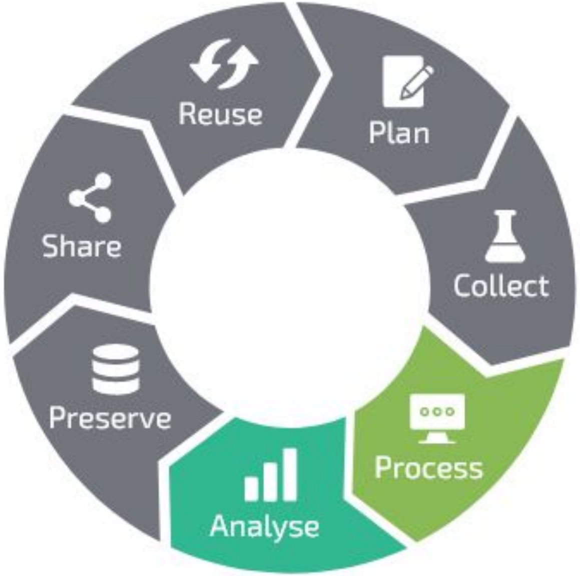
I – Challenges

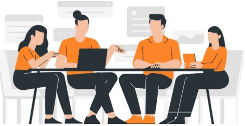
Learning Outcomes:

- Tackle the RDM challenges that arise in data processing and analysis

Data Processing & Analysis

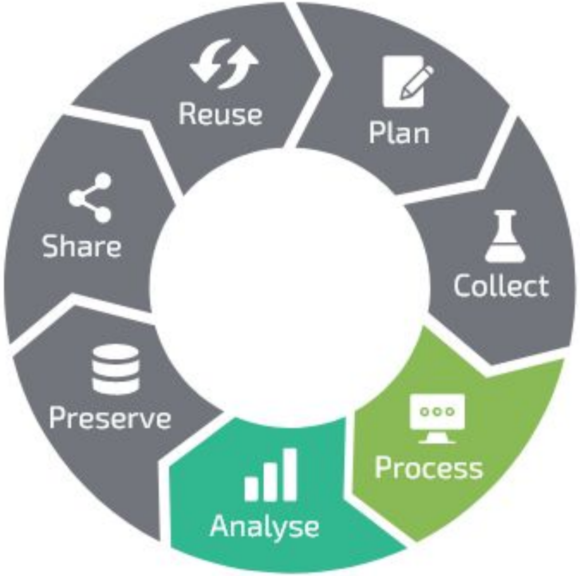
- **Data Processing:** transform the data in preparation for analysis
- **Data Analysis:** extract knowledge from the data
- In practice the distinction between the two is immaterial
  - They can be executed in a single workflow
  - Most of the challenges are shared

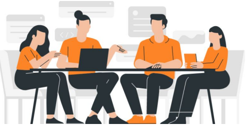




Data Processing & Analysis

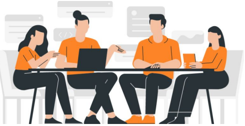
- **Challenges:**
  - Data anonymisation/pseudonymisation
  - Data cleaning / quality control
  - Data integration
  - Documentation & workflow management
  - Data organization
  - Computing & storage





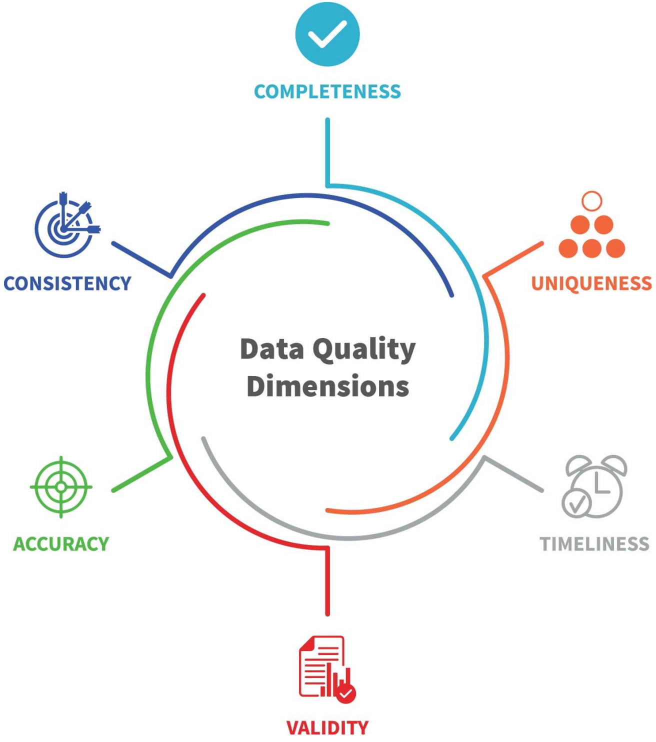
Data Anonymisation / Pseudonymisation

- Should be carried out immediately after data collection to minimise access to personally identifying data
- Under the GDPR, data is anonymised only if not even the data controller can re-identify the data, otherwise it is only pseudonymised
- In the world of big data, re-identification may be possible given sufficient “non-identifying” details

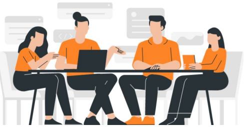


Data Cleaning / Quality Control

- Data quality issues that could not be fixed during data collection should be addressed in the early stages of data processing
  - Fix missing or erroneous values
  - Remove outliers
  - Remove low-quality data (e.g. in nucleotide sequencing results)



Ready for BioData.pt Management?

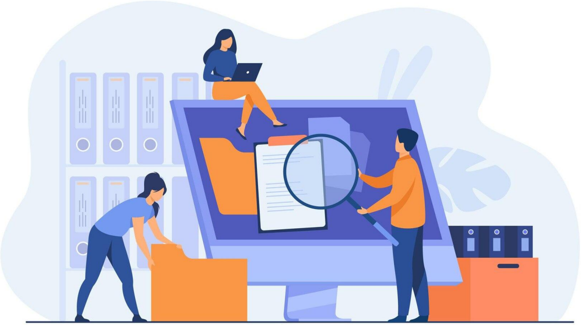


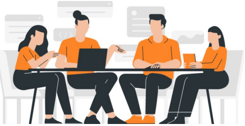
Data Integration

- Data integration is necessary whenever we need to combine multiple datasets
  - E.g. in a data reuse scenario, where we are combining data from multiple studies
  - E.g. when combining data from replicates or related samples in a single experiment
- You must ensure that the experimental conditions of the datasets being combined are reconcilable

Documentation & Workflow Management

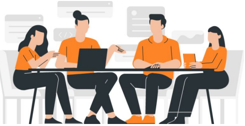
- Data documentation is (unsurprisingly) also one of the biggest challenges in data processing and analysis
  - At these stages the primary focus is documenting all transformations the data undergo, i.e. the data processing/analysis **workflow**





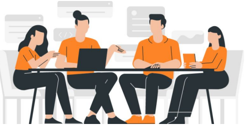
Documentation & Workflow Management

- You should document all:
  - Processing/analysis software you use
  - Operations executed in each software
  - Settings used in each software
  - Inputs and outputs of each software
- If you use in-house code/scripts for processing/analysis then you should document the code and publish it



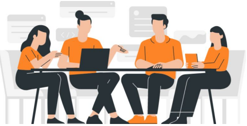
Documentation & Workflow Management

- The best practice for reproducibility is to create a “proper” workflow, document it and share it
- Options for workflow management include:
  - **Galaxy** (no programming skills)
  - **Jupyter Notebook** (low programming skills)
  - **RStudio** (medium programming skill)
  - **Bash** script (high programming skill)



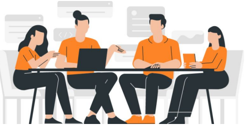
Documentation & Workflow Management

- The best options for sharing workflows are:
  - **GitHub** or **GitLab** both of which include version control and allow you to keep your workflow private until you wish to publish it
- You can also make a formal release of your workflow on Zenodo, though this is better suited for sharing non-computational protocols (namely for data collection)



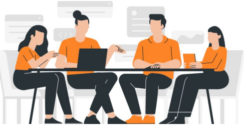
Data Organisation

- The challenges are the same we discussed for data collection, only amplified in scale due to data processing and analysis workflows typically multiplying the number of data files
  - This puts more stress on having an adequate folder structure and file naming conventions
    - Make sure there is a clear distinction between the names of the input and output files of each processing/analysis step (even if you put them on separate folders)



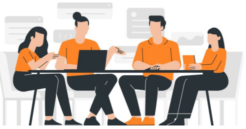
Computing & Storage

- Data processing and analysis are typically the only stages that require computing
  - Depending on the volume of data and type of analysis, you may be able to run them on your laptop or lab server, or you may need access to a HPC cluster or cloud computing service (due to high CPU, GPU or RAM requirements)
    - In Europe, academia can often access HPC services of research infrastructures free of charge, but pricing of cloud companies is not prohibitive



Computing & Storage

- One critical challenge is the connection between computing and storage
  - If you run the analysis in a HPC cluster or on the cloud, you need to move the data to and from the cluster/cloud and ensure there is enough storage there
- You need to factor in the additional storage requirements due to the files produced in data processing and analysis
  - Both for temporary storage linked to compute and for “permanent” storage



Thank You!

Questions?

II – Hands-On

Learning Outcomes:

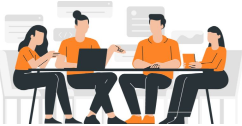
- Use Galaxy to create a data processing and analysis workflow

Galaxy

useGalaxy.*

- Galaxy is an open-source web platform for data processing and analysis, with workflow management functionalities
  - It is essentially a web interface for tools that normally run through the command line
- There are many public instances around the world (.org, .eu, etc) but you can also self-host it
- We have a training instance at https://dev.galaxy.biodata.pt/

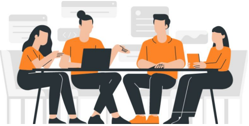
use it beyond this course)



Group Exercise

Set-up:

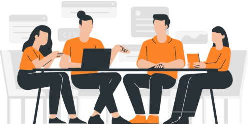
- Access our Galaxy instance https://dev.galaxy.biodata.pt/
- Register
- Explore the tool



Group Exercise

Part 1:

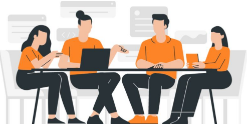
- Search for Quality Control workflow (Shared Data):
    - EGA Download Client (download fastq file)
        - Download file with ID “EGAF00004859455”
    - FastQC (reads quality control)
- Run the workflow and check the results
- Modify the workflow to improve it and share it



Group Exercise

Part 2:

- Create a complete workflow in Galaxy for WP2 Task 1 of the mock project, including the following data processing and analysis steps:
  - Quality control
  - Genome assembly
  - Variant calling



Thank You!

Questions?