Reproducible analysis and interactive reports using containers and R/Shiny in bioinformatics pipelines

PEC1

Luis Morís Fernández

# Title

Reproducible analysis and interactive reports using containers and R/Shiny in bioinformatics pipelines

# Keywords

reproducibility, interactive, pipeline, bioinformatics

# Context and justification of this work

Reproducibility is one of the cornerstones of the scientific method, nonetheless is one of the most commonly forgotten, sometimes shadowed by the novelty or alleged impact of the results. With the advent of informatics more powerful and flexible analyses are available to researchers in all fields. Unfortunately, this power and flexibility comes usually at the cost of complexity and length. Bioinformatics pipelines are paramount examples of the benefit of these powerful and flexible analyses and also of the cost of complexity. This issue is not solved exclusively by properly storing the scripts of analysis and the data, additional considerations must be taken, such as the pathing, the version of the language used, the version of the libraries, the operating system and its version, etc. All these problem hinders reproducibility and utmost care must be taken when preparing and storing these pipelines, so they can be run not only at the moment of the initial analysis but at any moment in the future when another researcher may need to verify the analyses, extend them or run them on new data.

In this work I will explore two possible ways of improving reproducibility when developing bioinformatic pipelines. The first one is making the analysis steps traceable from the start to end. This regards the actual scripts that will be executed during the analysis. To address this issue I will take advantage of pipeline creation tools that allows creating long reproducible analysis pipelines. The second is creating a reproducible a light-weight stable environment were the previous pipeline will through the use of containers.

The other objective of this work has to do with the interactivity of the the reports obtained during the analysis process. It is rarely the case, particularly in clinical environments, that the same person has the ability to program and run analyses (e.g.: a bioinfomatics expert) and also to make decisions based on those data (e.g.: a medical doctor). This usually creates a communication overhead between the data expert and the decision-maker until the exact data needed to make the decision is found. Here I will create an interactive report that will try to ameliorate this problem by providing the decision-maker with a small set of tools that would decrease this communication overhead.

# General description

The intention of this work is to create a proof of concept that would address two problems commonly found when developing bioinformatics pipelines. The first one is how to improve its reproducibility through the use of containers. And the second one is to explore interactive reports as a mean to reduce the overhead between bioninformatics and medical personnel improving the decision making loop.

# Objectives

## General Objectives

Create a proof of concept that includes a traceable and reproducible pipeline. This pipeline must run inside a container and produce an interactive report as a result.

## Specific objectives

1. Describe the reproducibility problem in bioinformatics
2. Explore containers as a mean to improve the reproducibility of bioinformatics pipelines
3. Explore interactive reports as a mean to improve the decision making loop in clinical settings
4. Create a pipeline using containers that produce an interactive report as a result in a real setting

# Approach and method

For this work, firstly, I will describe the state of the art in reproducibility with particular emphasis in the bioinformatics field. Secondly, I will create a proof of concept of a containerized pipeline with and interactive report. In this second part, I will count with the support of the *Plataforma de Bioinformática de la Unidad de Estadística y Bioinformática (UEB)* for having access to data and creating a useful pipeline that could be used in a real clinical setting.

From a technical point of view, albeit other tools may be explored during this work, the language R will be used as the foundation for all the software produced in it, with a particular focus on the targets package for creating reproducible pipelines. The docker container technology will be used for containerizing the pipeline. And finally the R/Shiny framework will be used to create an interactive report.

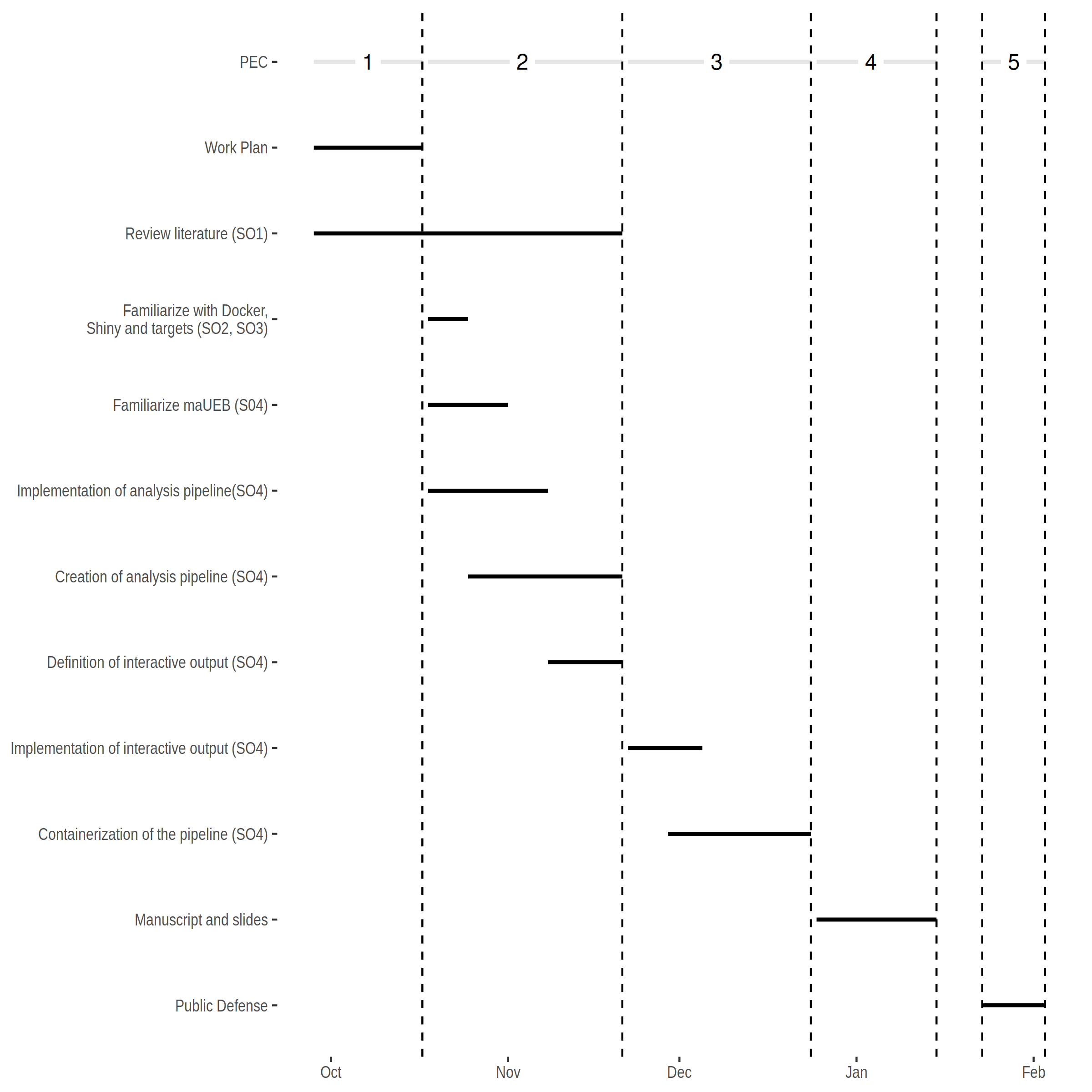
# Planning

## Tasks

Each task will be related to a specific objective using the acronym *SO*

1. Review literature for:
   1. **(SO1)** General problems of reproducibility in science and in bioinformatics
   2. **(SO1)** Current approaches to solve this problem
2. Software and technical knowledge and familiarization with:
   1. **(SO2, SO3)** Docker, R/Shiny and targets R package
   2. **(SO4)** maUEB package
3. **(SO2, SO4)** Definition and implementation of the analysis pipeline
4. **(SO3, SO4)** Definition and implementation of the interactive report
5. **(SO2, SO4)** Containerization of the pipeline

## Schedule



## Milestones

1. Workplan **(17/07/2022)**
2. Literature review and state of the art
3. Creating a pipeline using targets
4. Defining the interactive output
5. PEC2 **(21/11/2022)**
6. Implementing interactive output
7. Containerize the pipeline
8. PEC3 **(24/12/2022)**
9. Manuscript and slides **(27/12/2022 - 15/01/2023)**
10. Public defense **(23/01/2023 - 03/02/2023)**

## Risk analysis

Here I present possible risks and proposed solutions.

1. Oversized scope. This work tries to address two well-diferentiated problems, the reproducibility and the report interactivity. The interactivity scope can be reduced or removed from the work if the amount of time is not enough to cover both. The solution proposed is to reevaluate the scope when PEC2 is finalized and the scope is better defined.
2. Package availability. The initial intention is to use a recently developed package maUEB. If the package is not available I propose to use other alternatives covered during the master classes.
3. Data availability. If data is unavailable, for example confidentiality for confidentiality reasons, I propose to use open data sources.
4. Pipeline definition. If the the pipeline is not possible to define in a real clinical setting, I propose to follow steps similar to those used in the pipelines during the master classes.

# Expected results

1. Work plan

* A detailed workplan for the thesis work.

1. Thesis document

* A document containing an state of the art, methodology, results and conclusions obtained during the master’s thesis.

1. A working proof of concept

* A containerized pipeline, delivered as an R package.

1. Slides and virtual presentation

* Slides and a video presentation for the defence.

# Bibliography

IN PROGRESS