Best Practices for Reproducibility in Computational Research

Concept

For any research program, an independent researcher should be able to **replicate the experiment**, under the same conditions, and achieve the same **results**.

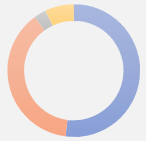
Reproducibility in numbers

70% - researches that failed to reproduce another scientist's experiments;
50% - researches that failed to reproduce their own experiments.

What can be done?

- More robust experimental design
- Better statistics
- Better mentorship

Is there a Reproducibility Crisis?



- Yes, a significant crisis
- Yes, a slight crisis
- No, there is no crisis
- Don't know

Source: Nature, 2016, "1,500 scientists lift the lid on reproducibility"

Source: davidebonazzi.com

The Ten Simple Rules

- Rule 1: For Every Result, Keep Track of How It Was Produced
- Rule 2: Avoid Manual Data Manipulation Steps
- Rule 3: Archive the Exact Versions of All External Programs Used
- Rule 4: Version Control All Custom Scripts
- Rule 5: Record All Intermediate Results, When Possible in Standardized Formats
- Rule 6: For Analyses That Include Randomness, Note Underlying Random Seeds
- Rule 7: Always Store Raw Data behind Plots
- Rule 8: Generate Hierarchical Analysis Output, Allowing Layers of Increasing Detail to Be Inspected
- Rule 9: Connect Textual Statements to Underlying Results
- Rule 10: Provide Public Access to Scripts, Runs, and Results

The five key elements of reproducibility

DATA
IEEE Dataport
Figshare
Dryad

- Should be findable, accessible, and re-usable
- Consider using cloud storage services to make dataset always available

CODE
Git
Mercurial
SourceForge

- Keep the entire code available
- Adopt versioning control
- Document the code

DOCUMENTATION
Jupyter
NextJournal
CodaLab

- Describes all the paper sections
- Write it in an iterative program
- Includes the code to generates the results

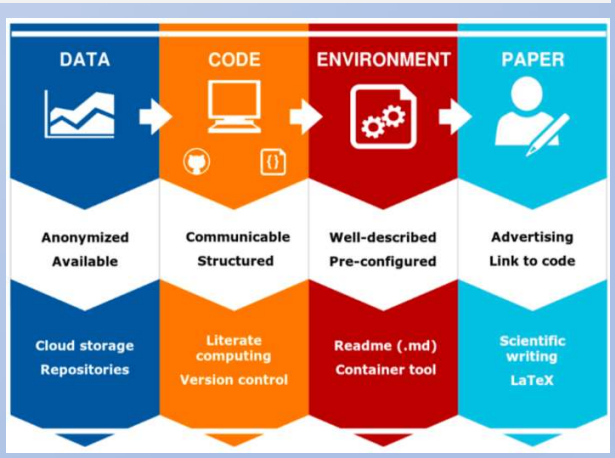
WORKFLOW
Sacred
Reana
MLFlow

- Describes how the modules are integrated
- Shows a great overview of input and outputs
- Demonstrate the entire process pipeline

ENVIRONMENT
Docker
VirtualMachines
Anaconda

- Share or detailed the entire process to reproduce the research steps
- Includes all the Licenses
- Be aware of all the packages versions

Icons source: cleanpng.com



William Herrera, 2019. Best practices for reproducible research.

Tips:
- Think in reproducibility before start
- Keep the directory organized
- Try your own application in a new environment
- Ask someone to test it too
- Keep it safe, avoid privileged commands

Example:
github.com/marcofrk/ia369_final_project

IA369 – 2S/2020 – MARCO A FRANCHI – 092207