Workflows, reproducibility and best practice

Magali Hennion & Cédric Midoux

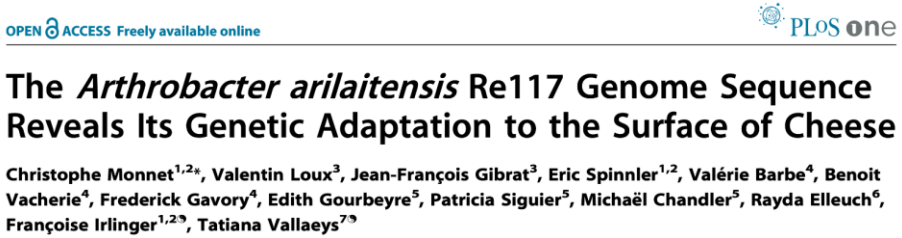
RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

INRAE

Reproducibility crisis

Everyone has had this experience

An interesting article ...



OPEN ACCESS Freely available online PLoS one

The *Arthrobacter arilaitensis* Re117 Genome Sequence Reveals Its Genetic Adaptation to the Surface of Cheese

Christophe Monnet[1,2*], Valentin Loux[3], Jean-François Gibrat[3], Eric Spinnler[1,2], Valérie Barbe[4], Benoit Vacherie[4], Frederick Gavory[4], Edith Gourbeyre[5], Patricia Siguier[5], Michaël Chandler[5], Rayda Elleuch[6], Françoise Irlinger[1,2☯], Tatiana Vallaeys[7☯]

... but a deceptive M&M

collaboration with the user community. Genome comparisons were performed using Origami, an in-house tool developed for microbial genome comparison. Orthologs were defined as reciprocal best hits with an e-value lower than 10^{-3}. Transposases were excluded from the analysis. Core genes were defined as orthologs shared between the four *Arthrobacter* strains. Synteny was studied using an in-house developed tool, Align, using dynamic programming to search conserved gene trains allowing gaps and "mismatches" (homology relation instead of orthology). Circular representation of the genome was produced using the Circos software [27].

Monnet (2010)

The data deluge

Paradigm of science

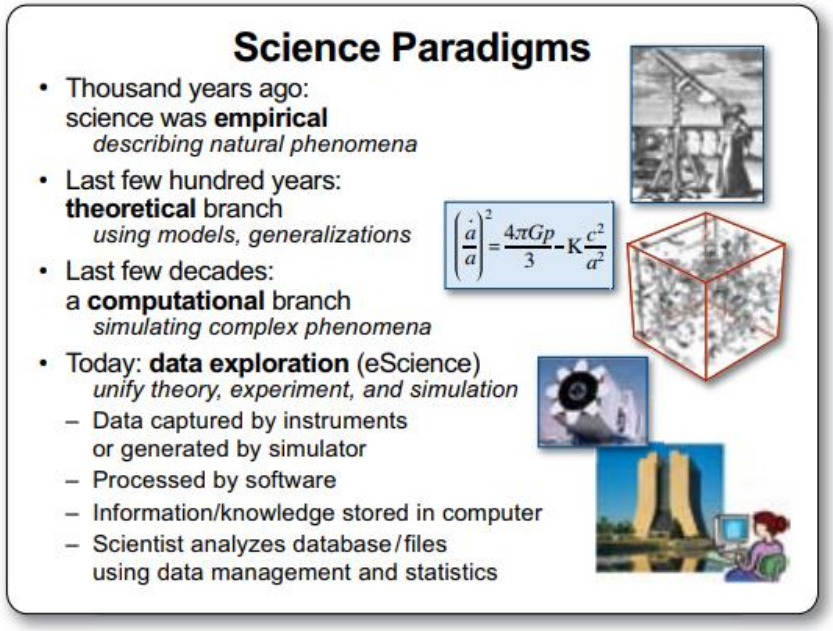
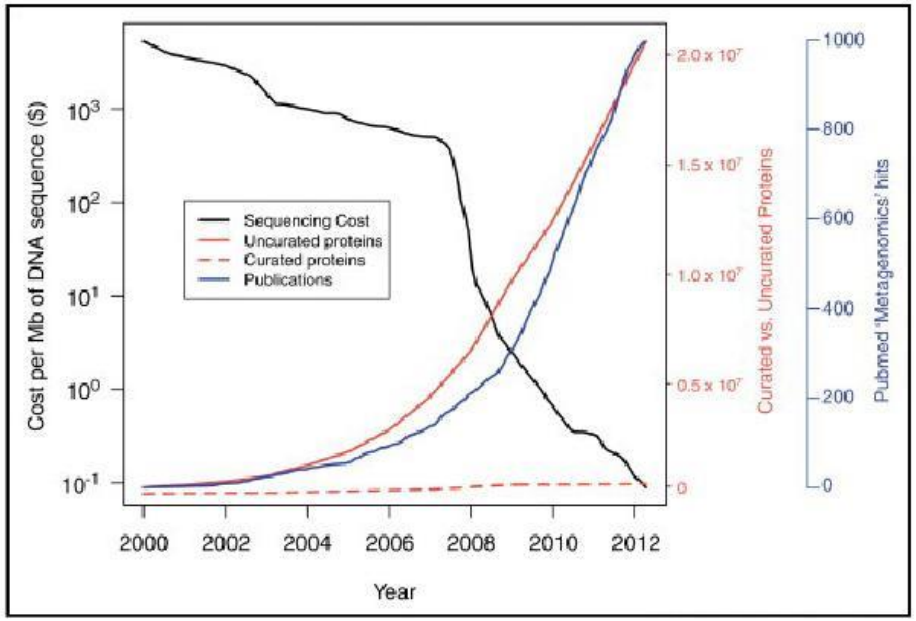


FIGURE 1

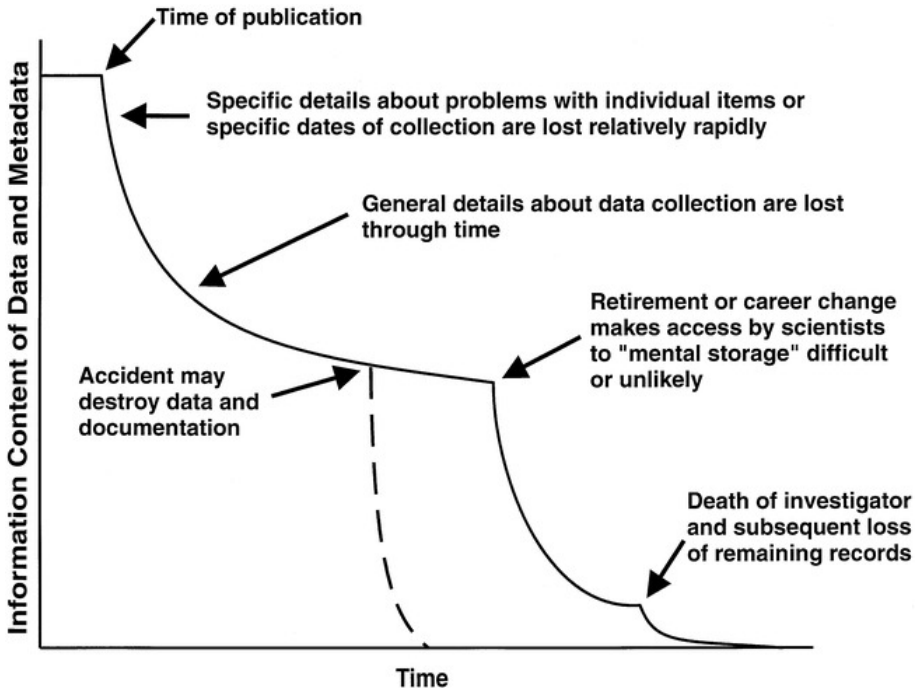
More and more data



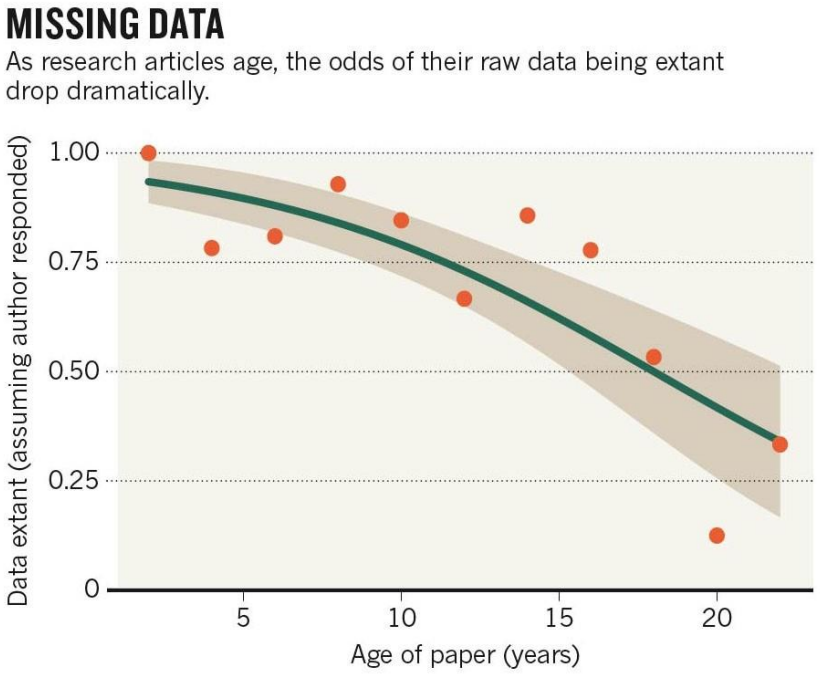
Hey (2009)

Murphy (2014)

The ravages of time …

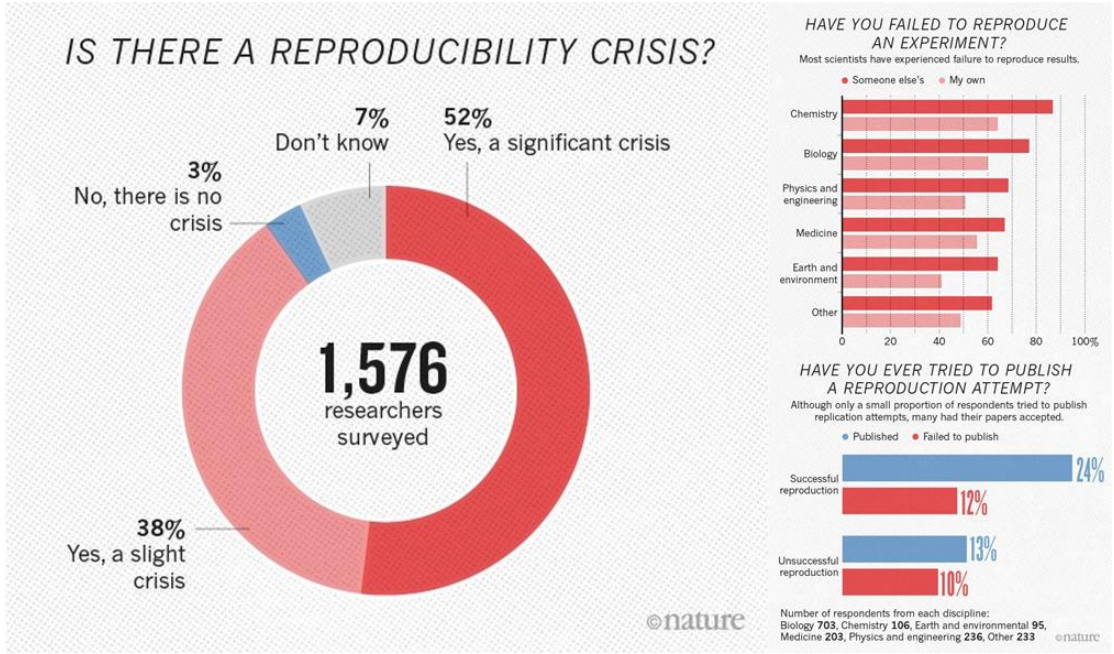


Michener (1997)

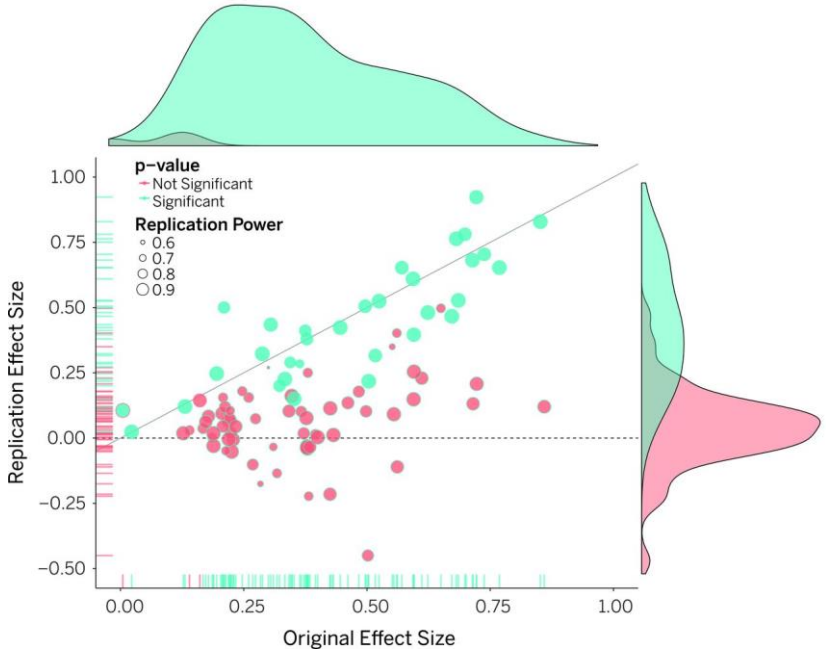


Gibney (2013)

Reproducibility ?



Estimating the reproducibility of psychological science

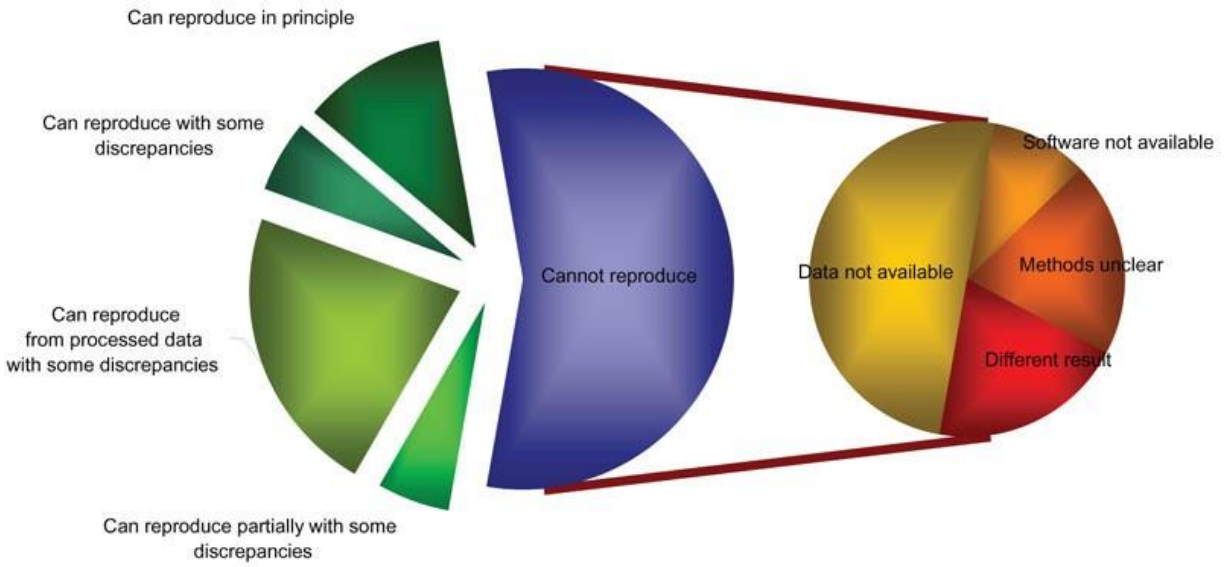


Baker (2016)

OPEN SCIENCE COLLABORATION (2015)

Also with bioinfo

- *Given the complexity of microarray-based gene expression studies, guidelines encourage transparent design and public data availability. […]*
- *Here we evaluated the replication of data analyses in 18 articles on microarray-based gene expression profiling published in Nature Genetics in 2005–2006.*
- *One table or figure from each article was independently evaluated by two teams of analysts.*
- *We reproduced two analyses in principle and six partially or with some discrepancies; ten could not be reproduced. […]*



Ioannidis (2009)

Cost of not having FAIR research data

Following this approach, we found that the annual cost of not having FAIR research data costs the European economy at least €10.2bn every year.

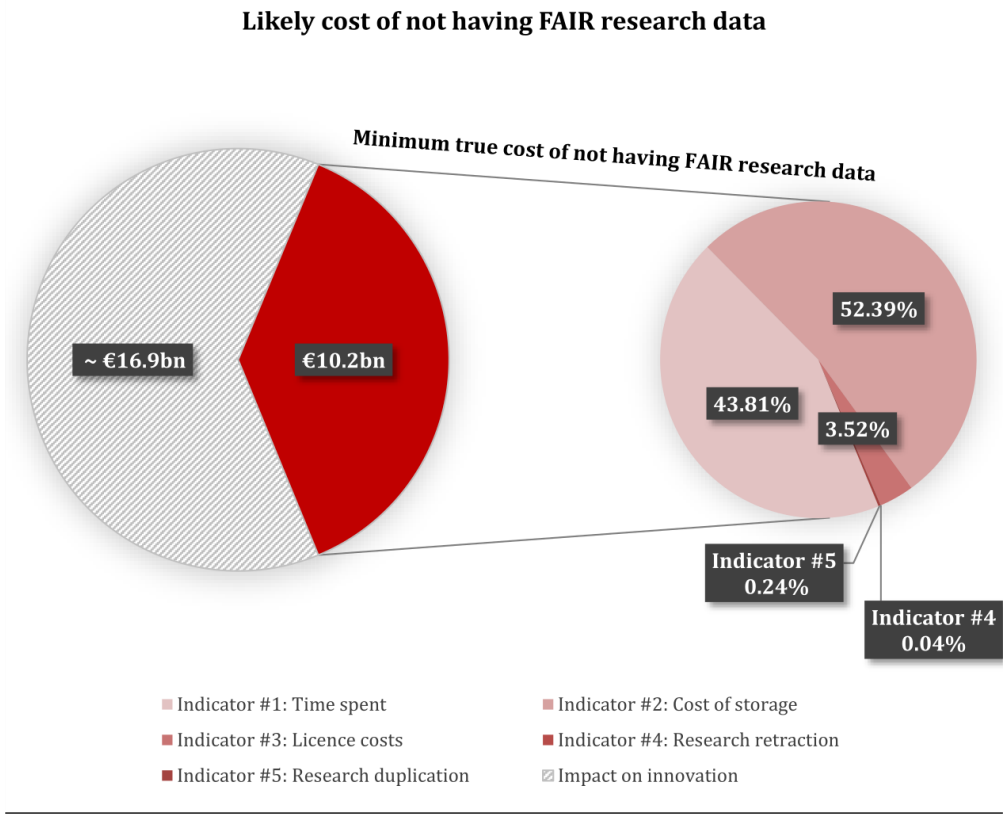


Figure 5: Cost breakdown

PwC UE Services (2018)

KC Green (2013)

UNESCO Recommandations

According to the UNESCO Recommendation, open science is a set of principles and practices that aim to make scientific research from all fields accessible to everyone for the benefit of scientists and society as a whole. The Recommendation aims to ensure not only that scientific knowledge is accessible but also that the production of that knowledge itself is inclusive, equitable and sustainable.

By promoting science that is more accessible, inclusive and transparent, open science furthers the right of everyone to share in scientific advancement and its benefits, as stated in Article 27.1 of the Universal Declaration of Human Rights.

UNESCO (2021)

Loi Pour une République Numérique 2016

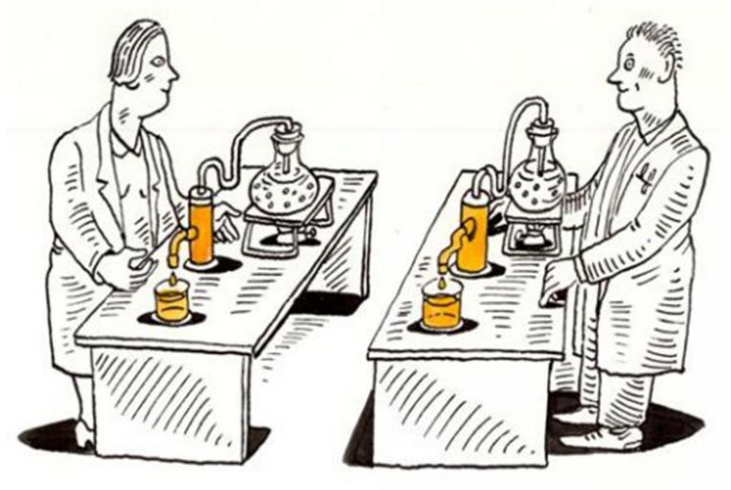
- *Le premier volet concerne la circulation des données et du savoir. Il comprend des mesures sur l'ouverture des données publiques, la création d'un service public de la donnée. Il introduit la notion de données d'intérêt général, pour optimiser l'utilisation des données aux fins d'intérêt général. Une partie est également dédiée au développement de l'économie du savoir, avec la possibilité pour les chercheurs de publier librement leurs articles scientifiques dans un délai de six à douze mois. Le Sénat a voté en faveur de la facilitation de l'ouverture et de la réutilisation des données des administrations ainsi que des décisions des juridictions administratives et judiciaires. La diffusion de ces données sera circonscrite aux données dont la publication présente un intérêt économique, social, sanitaire ou environnemental.*

« Aussi ouvert que possible, aussi fermé que nécessaire »

[LOI n° 2016-1321]

Plans Nationaux pour la Science Ouverte

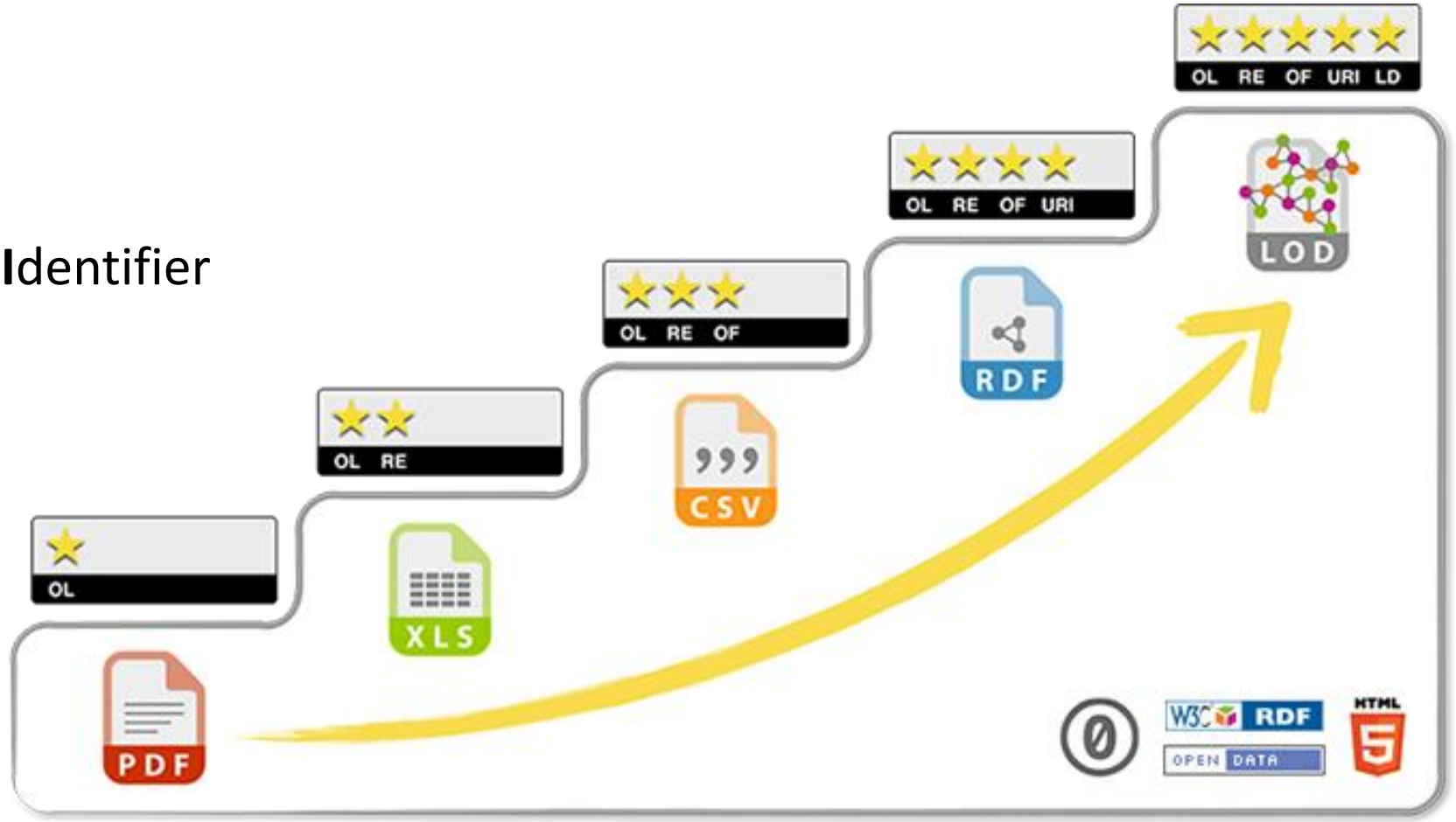
[Ouvrir la Science (2021)]



Basic concepts of reproducibility

Open Data 5★

★ **O**pen **L**icense
★ machine **RE**adable
★ **O**pen **F**ormat
★ **U**niform **R**esource **I**dentifier
★ **L**inked **D**ata



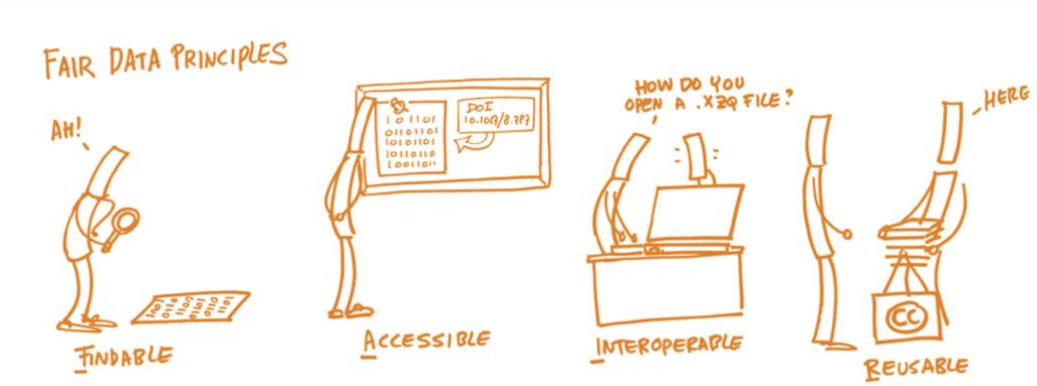
Berners-Lee

Qualinous (2024)

Scriberia

The Turing Way (2022)

FAIR IRL



Findable

- **Findable** (for humans and computers)
  - Persistent Identifier (DOI with zenodo or other)
  - Metadata describing the analysis and the tools (README)
  - A versioned script (git)
  - Available on a forge (GitHub, GitLab) or archive (Software Heritage)

Accessible

- **Accessible**
  - License and access rights
  - Standard communications protocol
  - Metadata accessible

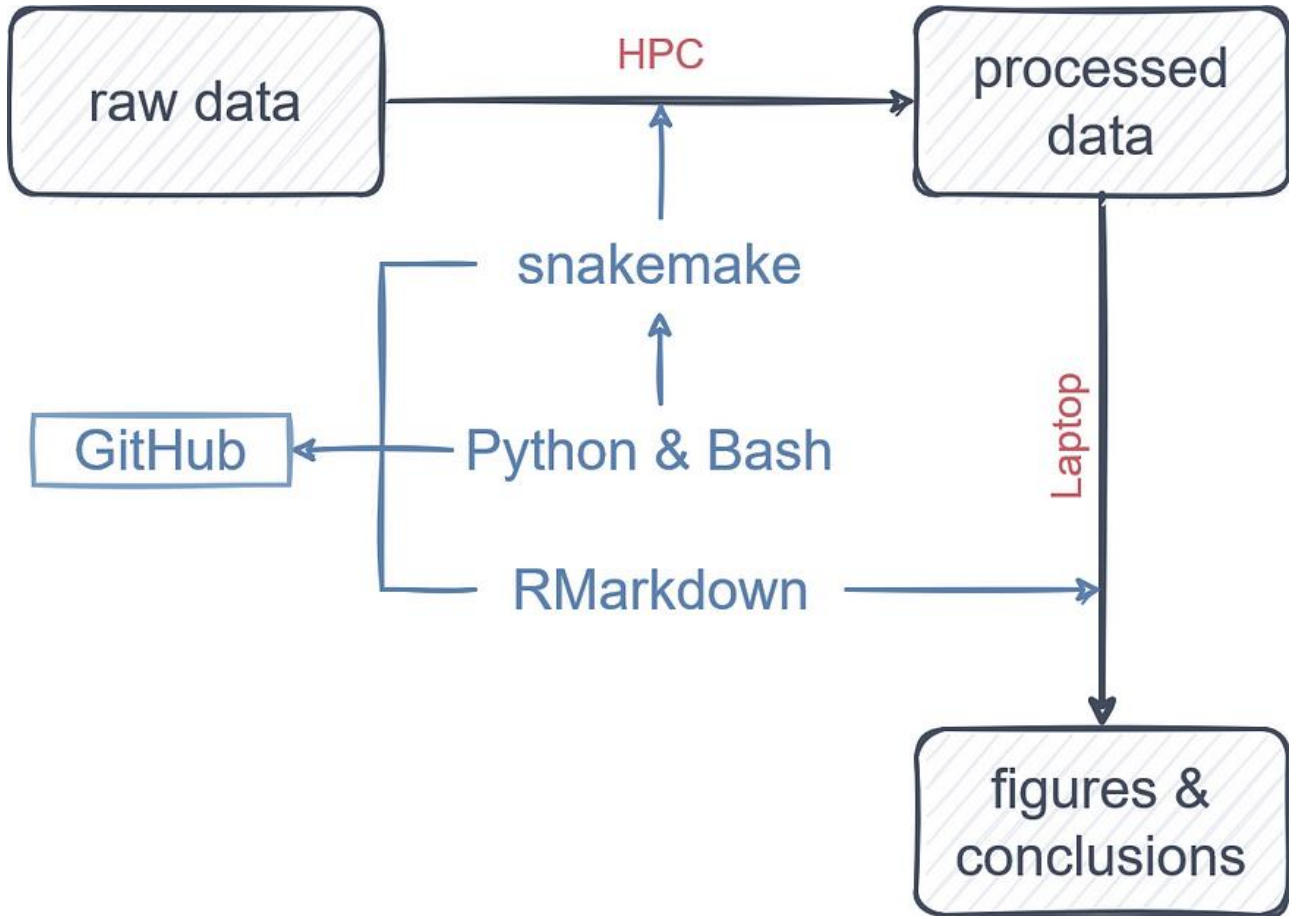
Interoperable

- **Interoperable**
  - Controlled vocabulary, onthology and linked metadata
  - Tools work together (snakemake or nextflow) in a controlled environment (conda or docker) locally or on a server (cloud or cluster)
  - Open and documented formats

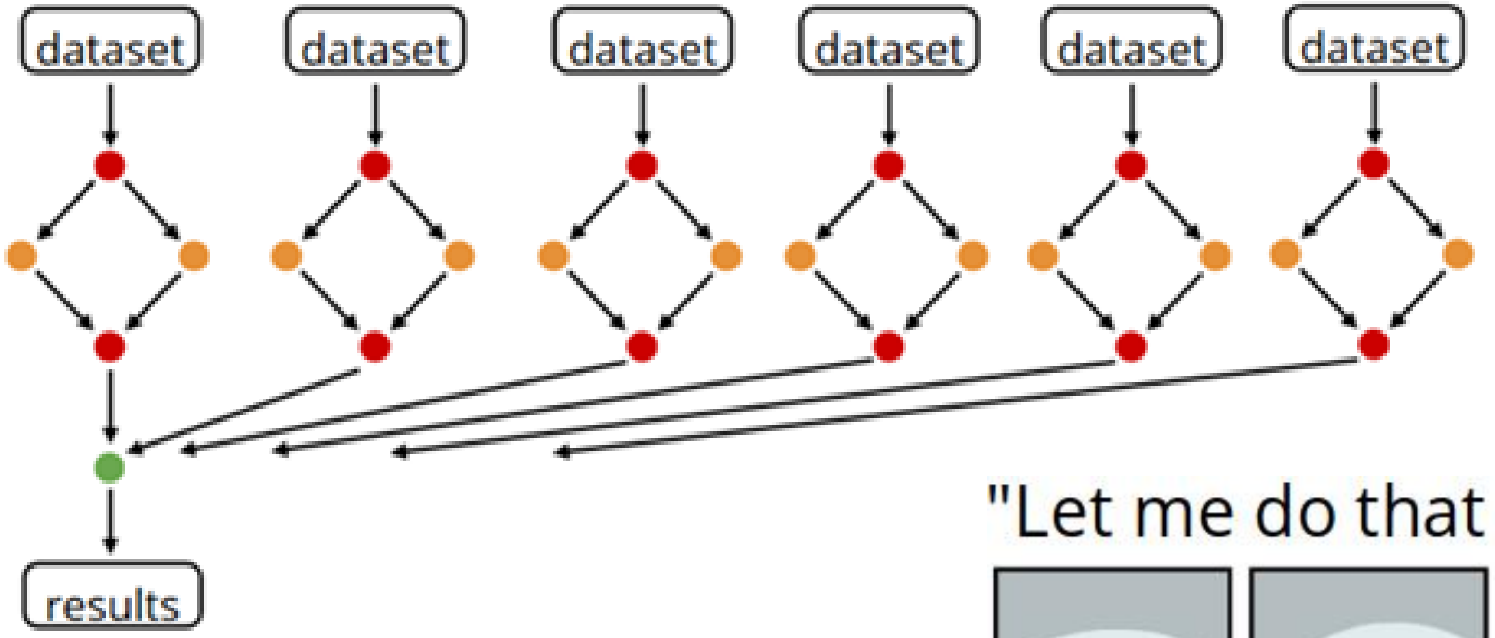
Reusable

- **Reusable**
  - Protocol can be replayed identically (Jupyter and Quarto) in a virtual environment
  - At any time : CI/CD (GitHub actions or GitLab workflow)

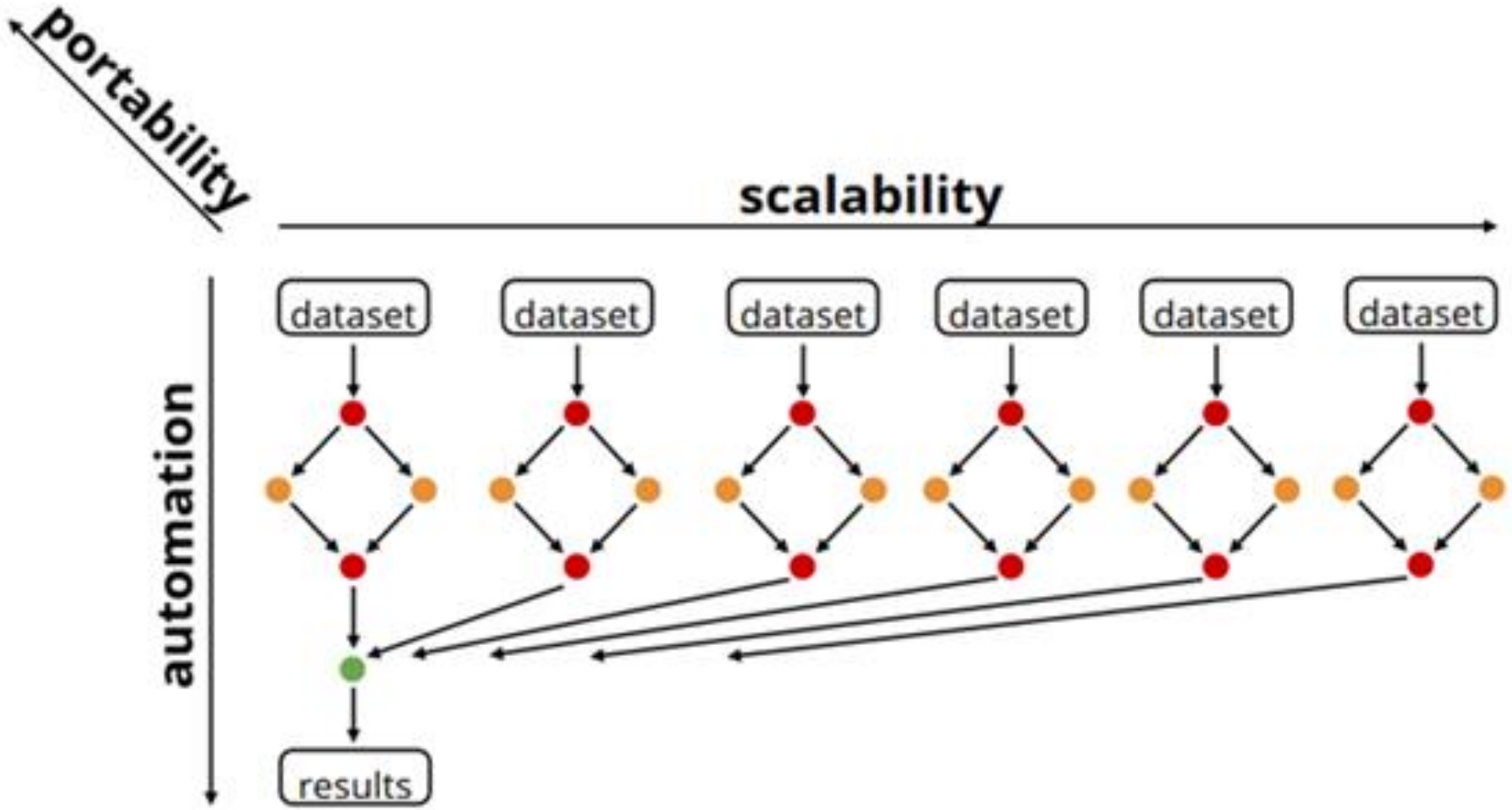
Workflow for Open and Reproducible Science



Van Paemel (2021)

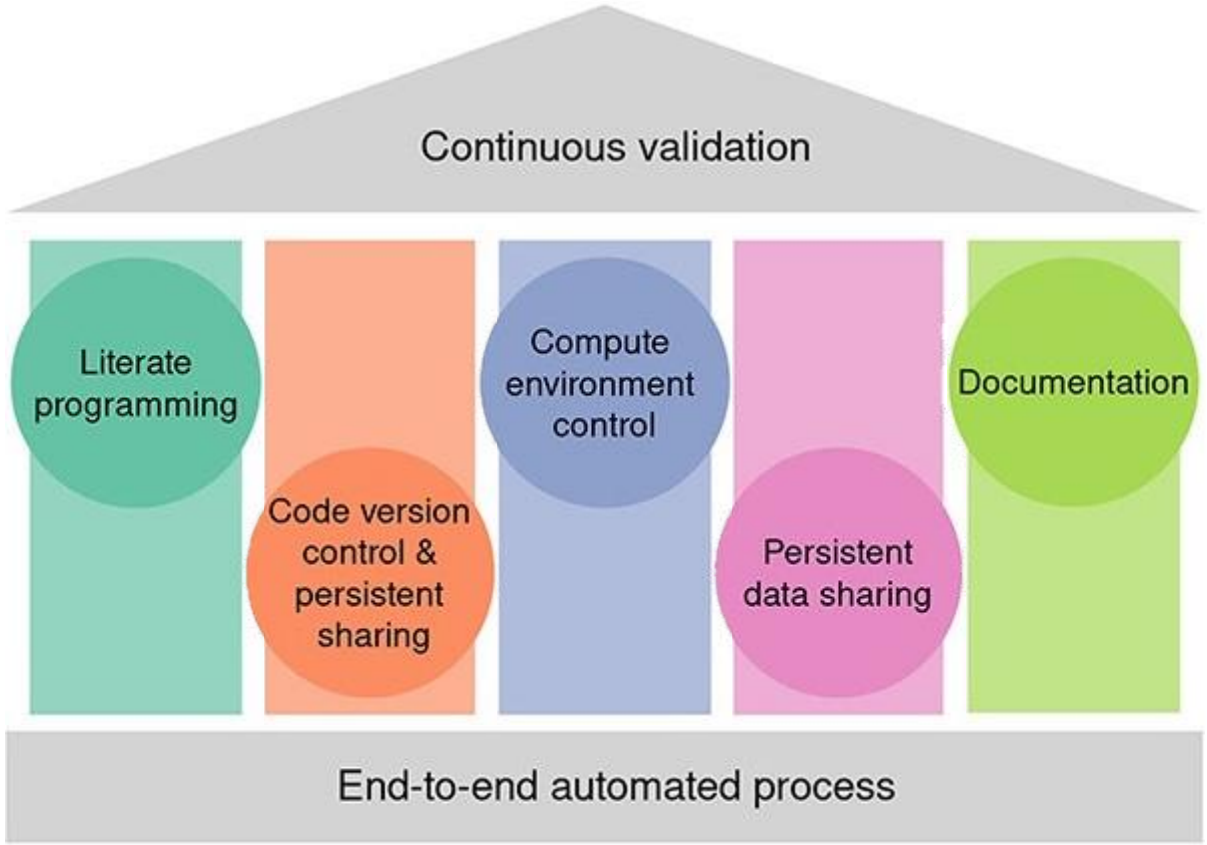


"Let me do that by hand..."



Workflow management:

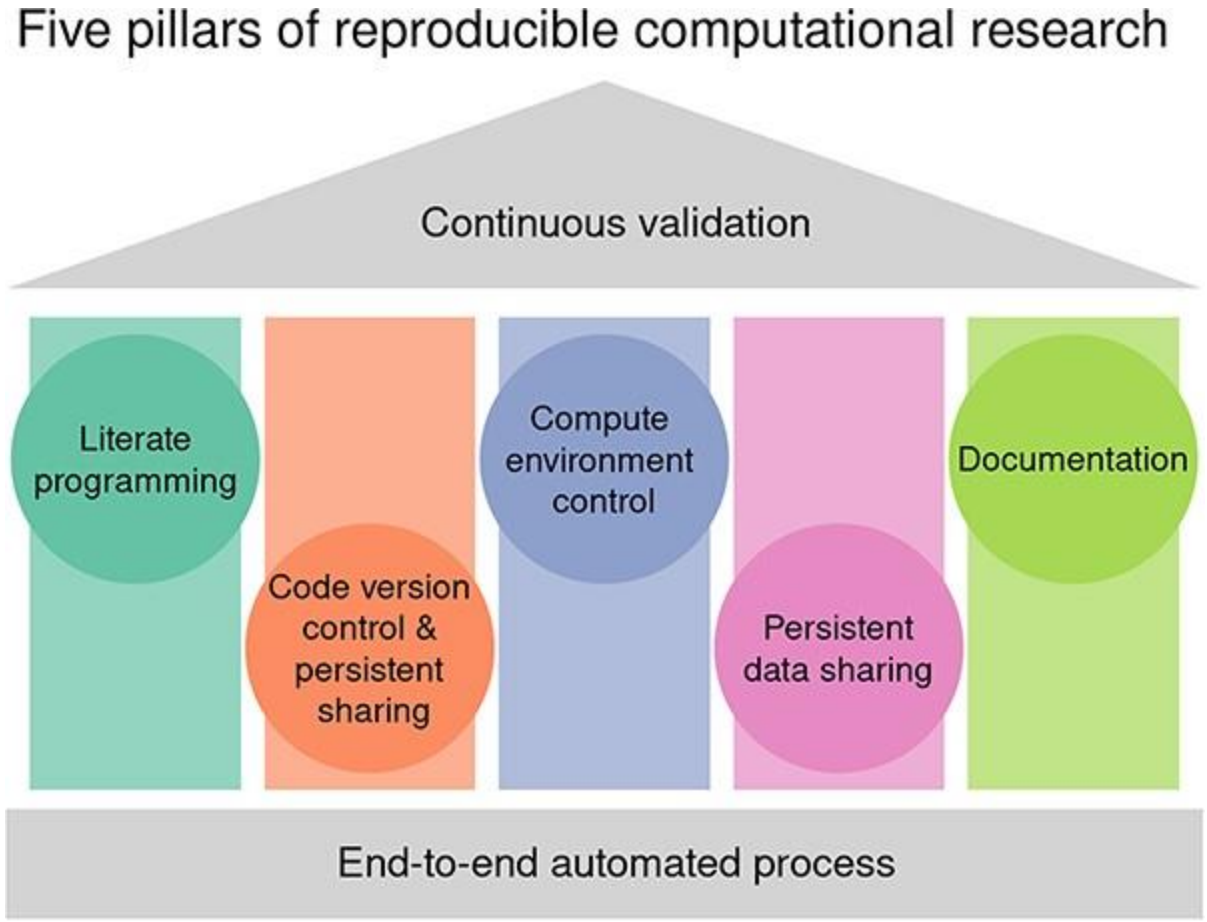
formalize, document and execute data analyses



scripted workflows

Human intervention, Excel and mouse forbidden !!

Free as in Freedom



Key Points

→ Irreproducibility of bioinformatics studies remains a significant and still-relevant problem.

→ We present the five pillars framework, a set of best practices that enable extremely reproducible workflows.

→ Widespread adoption of these principles will enhance research reliability and will speed translation of basic research to tangible benefits.

Mark Ziemann, Pierre Poulain and Anusuiya Bora (2023)