

# RECHERCHE REPRODUCTIBLE ET IA

---

Arnaud Legrand



2èmes Journées du Réseau National de la Recherche Repro  
Table ronde  
Mars 2024



**Alexis Arnaud** (UGA/GRICAD) coordinateur de la Cellule Data de l'UGA, confidentialité des données

**Ninon Burgos** (CNRS/Institut du Cerveau/ARAMIS) IA pour aide à la décision médicale

**Ambre Davat** (GRESEC/ chaire éthique et IA)

**Jean-Luc Parouty** (CNRS/SIMaP) coordinateur de la formation FIDLE

**Didier Schwab** (UGA/LIG, GETALP) traitement automatisé des langues et de la parole

## DIFFERENT REPRODUCIBILITY CONCERNS IN MODERN SCIENCE

Many different **terms** and **issues** depending on the **domain**

- Reproduce, Replicate, Repeat, Rerun, Redo, Reuse, Register, Report

## DIFFERENT REPRODUCIBILITY CONCERNS IN MODERN SCIENCE

Many different **terms** and **issues** depending on the **domain**

- Reproduce, Replicate, Repeat, Rerun, Redo, Reuse, Register, Report

**Biology, Oncology** sample provenance, clinical trials  $\rightsquigarrow$  standardized protocols

**Psychology, Nutrition** HARKING, p-hacking  $\rightsquigarrow$  pre-registration

## DIFFERENT REPRODUCIBILITY CONCERNS IN MODERN SCIENCE

Many different **terms** and **issues** depending on the **domain**

- Reproduce, Replicate, Repeat, Rerun, Redo, Reuse, Register, Report

**Biology, Oncology** sample provenance, clinical trials  $\rightsquigarrow$  standardized protocols

**Psychology, Nutrition** HARKING, p-hacking  $\rightsquigarrow$  pre-registration

**Genomics** software engineering, comput. reproducibility, provenance

**Computational fluid dynamics** numerical chaos, parallel architectures

# DIFFERENT REPRODUCIBILITY CONCERNS IN MODERN SCIENCE

Many different **terms** and **issues** depending on the **domain**

- Reproduce, Replicate, Repeat, Rerun, Redo, Reuse, Register, Report

**Biology, Oncology** sample provenance, clinical trials  $\rightsquigarrow$  standardized protocols

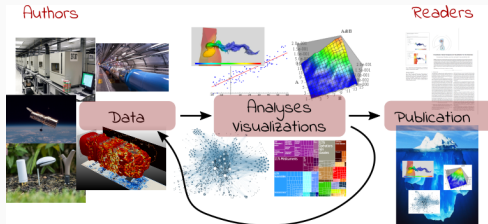
**Psychology, Nutrition** HARKING, p-hacking  $\rightsquigarrow$  pre-registration

**Genomics** software engineering, comput. reproducibility, provenance

**Computational fluid dynamics** numerical chaos, parallel architectures

**Artificial Intelligence** most of the above 😊

AFAIC, I care about **transparency**





**Machine Learning:** **Trouble at the lab**, The Economist 2013

*According to some estimates, three-quarters of published scientific papers in the field of machine learning are bunk because of this "overfitting".*  
– Alex "Sandy" Pentland

**The Reproducibility Crisis in ML-based science** (Princeton workshop 2022)

*Reproducibility failures in ML-based science are systemic. We found 20 reviews across 17 scientific fields (medicine, neuroimaging, autism diagnosis, genomics, computer security, ...) that find errors in a total of 329 papers that use ML-based science and in some cases leading to wildly overoptimistic conclusion. [...] **complex ML models don't perform substantively better than decades-old LR models.***

***Data leakage:** spurious relationship between the independent variables and the target variable that arises as an artifact of the data collection, sampling, or pre-processing strategy.*

– **S. Kapoor and A. Narayanan**

## Same Old, Same Old...

- Software profusion and chaos
- Complex process

## Big Sensitive Data Large volume of data and computation

- Different from climate-change/astrophysics/... research ?
- Sensitive data and privacy issues

## "Autonomy" of Science GAFAM/NATU, OpenAI/Mistral/Anthropic/... and Academia. Same game ?

- Jean-Zay  $\approx$  40 PetaFlops, Meta's AI supercomputer  $\approx$  5,000 PetaFlops
- A moving target
- Reproducibility vs. Market shares

AI is now driving science 😞 (e.g., astrophysics)

## Scientific Integrity ChatGPT/LLM/ML4everything disruption

- Scientific integrity, Bias, Opacity, ...  $\rightsquigarrow$  lost of trust