

FORMATION À GRANDE ÉCHELLE À LA RECHERCHE REPRODUCTIBLE

Arnaud Legrand (CNRS/Univ. Grenoble Alpes – LIG)

Réseau de référents science ouverte à la CPU
14 Mars 2023



NO TRANSPARENCY NO CONSENSUS



OPEN SCIENCE

Plan National pour la Science Ouverte (BSN ~ CoSO)

- CNRS, Inria, INRAE, ...
- Many flavors : *Citizen Science*

Main pillars :

1. Open access
2. Open data
3. Open source
 - Open hardware
4. Open methodology (**Reproducible Research**)
 - Open-notebook science
 - Open science infrastructures
5. Open peer review (avoid collusion)
6. Open educational resources



**NO TRANSPARENCY
NO CONSENSUS**



DIFFERENT REPRODUCIBILITY CONCERN IN MODERN SCIENCE

Social Sciences, Oncology, ... methodology, statistics, pre-registration

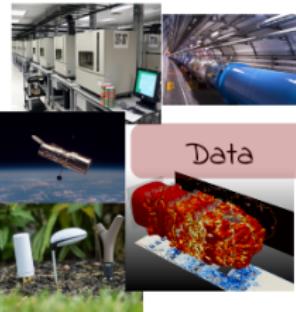
Genomics software engineering, computational reproducibility, provenance

Computational fluid dynamics numerical issues

Artificial Intelligence most of the above

The processing steps between raw observations and findings have gotten increasingly numerous and complex

Authors



Data

DIFFERENT REPRODUCIBILITY CONCERN IN MODERN SCIENCE

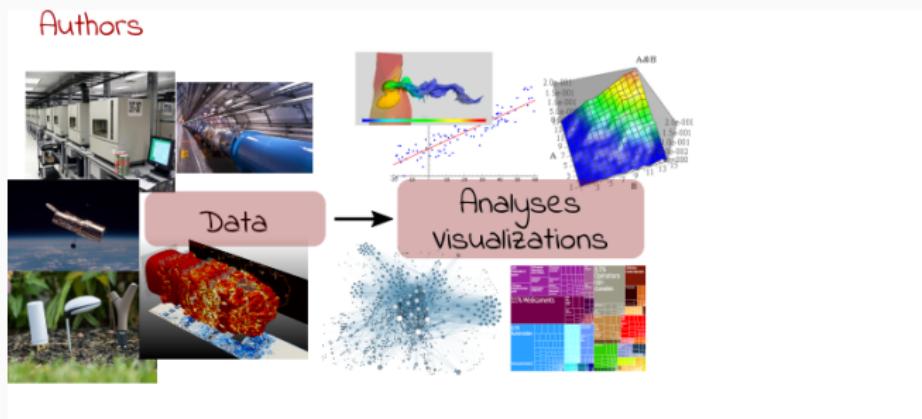
Social Sciences, Oncology, ... methodology, statistics, pre-registration

Genomics software engineering, computational reproducibility, provenance

Computational fluid dynamics numerical issues

Artificial Intelligence most of the above

The processing steps between raw observations and findings have gotten increasingly numerous and complex



DIFFERENT REPRODUCIBILITY CONCERN IN MODERN SCIENCE

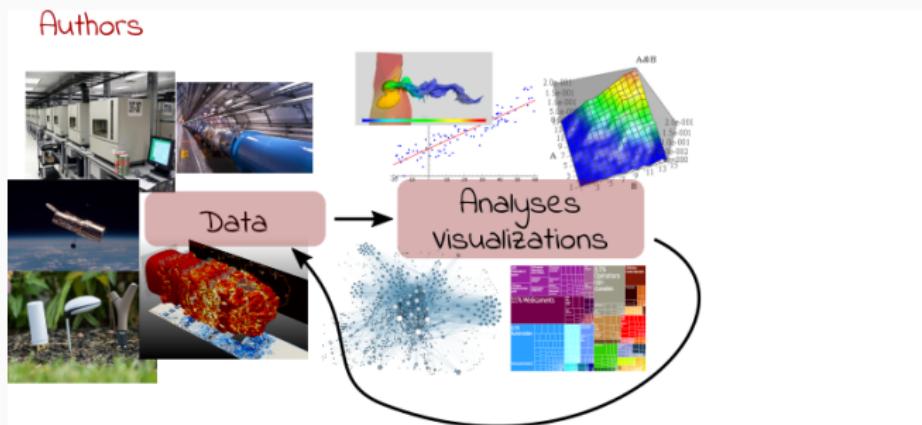
Social Sciences, Oncology, ... methodology, statistics, pre-registration

Genomics software engineering, computational reproducibility, provenance

Computational fluid dynamics numerical issues

Artificial Intelligence most of the above

The processing steps between raw observations and findings have gotten increasingly numerous and complex



DIFFERENT REPRODUCIBILITY CONCERN IN MODERN SCIENCE

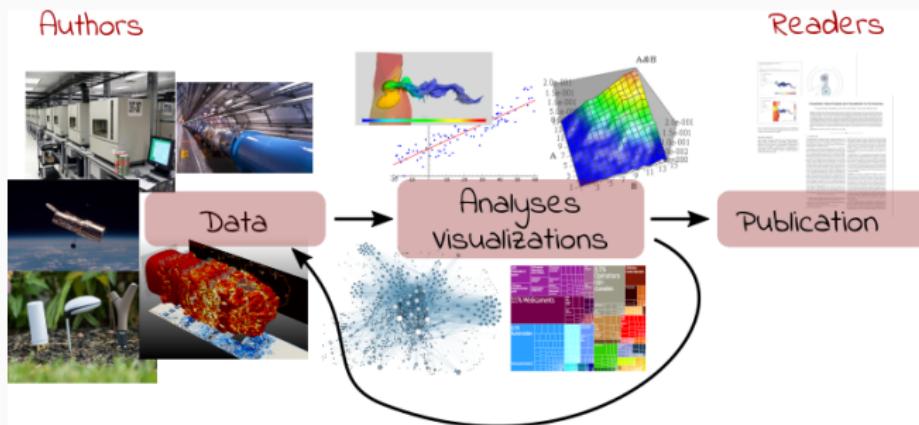
Social Sciences, Oncology, ... methodology, statistics, pre-registration

Genomics software engineering, computational reproducibility, provenance

Computational fluid dynamics numerical issues

Artificial Intelligence most of the above

The processing steps between raw observations and findings have gotten increasingly numerous and complex



DIFFERENT REPRODUCIBILITY CONCERN IN MODERN SCIENCE

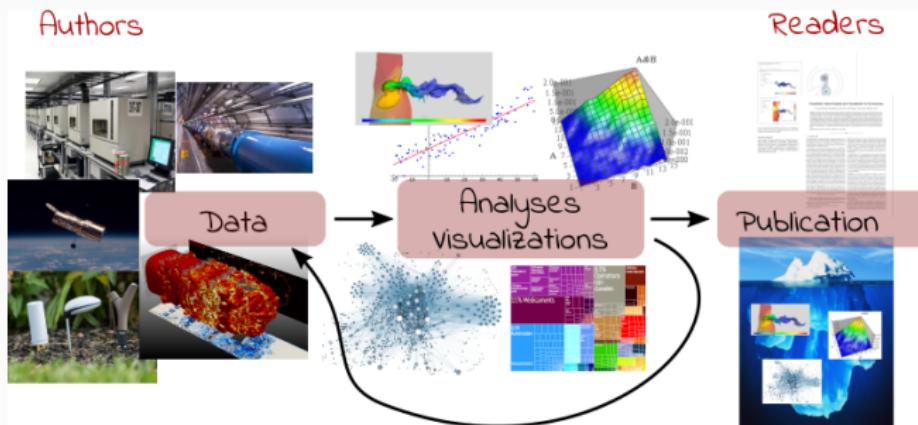
Social Sciences, Oncology, ... methodology, statistics, pre-registration

Genomics software engineering, computational reproducibility, provenance

Computational fluid dynamics numerical issues

Artificial Intelligence most of the above

The processing steps between raw observations and findings have gotten increasingly numerous and complex



DIFFERENT REPRODUCIBILITY CONCERN IN MODERN SCIENCE

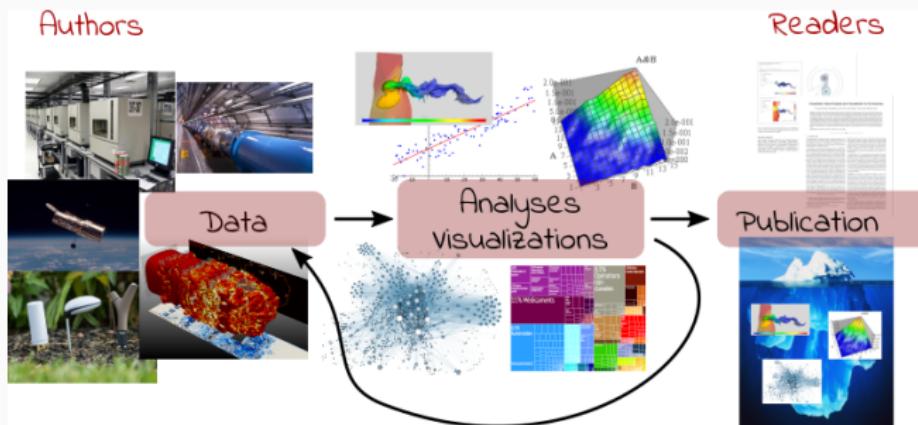
Social Sciences, Oncology, ... methodology, statistics, pre-registration

Genomics software engineering, computational reproducibility, provenance

Computational fluid dynamics numerical issues

Artificial Intelligence most of the above

The processing steps between raw observations and findings have gotten increasingly numerous and complex

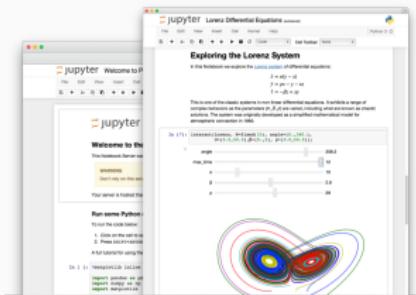


Reproducible Research = Bridging the Gap by working Transparently

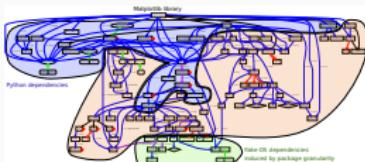
EXISTING TOOLS, EMERGING STANDARDS

Reproducibility issues induced by computers

Notebooks and workflows



Software environments



Sharing platforms



RECHERCHE REPRODUCTIBLE : PRINCIPES MÉTHODOLOGIQUES POUR UNE SCIENCE TRANSPARENTE



Recherche reproductible : principes méthodologiques pour une science transparente (sur FUN MOOC)

- Réflexions : début 2017

Session	Période	Participants	GitLab actifs	Attestations
1	Oct. – Dec. 2018	3416	601	290
2	Apr. – June 2019	2103	283	135
3	Mar.2020 – ...	14625		1552

ÉQUIPE ENSEIGNANTE

Contenu, slides, exercices, ... (CNRS)



Konrad Hinsen
Physique/Bio-chimie
Orléans



Arnaud Legrand
Informatique
Grenoble



Christophe Pouzat
Neuro-bio/Stats.
Paris Strasbourg

Réalisation, animation, ... (Inria)



Laurence Farhi
Pédagogique



Marie-Hélène Comte
Pédagogique



Aurélie Bayle
Pédagogique



Benoît Rospars
Informatique

Relecture et animation Marie-Gabrielle Dondon (INSERM)

Et plus tard : Alexandre Hocquet, Sabrina Granger, etc.

PÉRIMÈTRE DU COURS

Sujet "technique" mais besoin de s'adresser au plus grand nombre

Public visé pour la Recherche Reproductible

- Scientifiques (physique, biologie, SHS, maths...)
- Doctorants, post-docs, ingénieurs, enseignants-chercheurs

Pré-requis

- Démarche scientifique dans son propre domaine
- Utilisation *de base* d'un ordinateur
- Programmation *de base* en R ou Python

Ce qu'on ne couvre pas

- Les bases de stats., programmation/algorithme, ...
- Les points trop techniques (images docker, branches git...)

Bonnes pratiques logiciels libres et matures, format texte

STRUCTURE

Pédagogie

- les concepts avant la technique
- des exercices pratiques pour monter en compétence
- de la documentation et un forum

Modules (≈ 1 par semaine) :

1. Cahier de notes, cahier de labo
2. Le document computationnel
3. Analyse intelligible et répllicable
4. La réalité du terrain

*gitlab, markdown
jupyter|Rstudio|OrgMode
Peer evaluation
Les enfers*

Organisation Parties communes + 3 parcours :

- Jupyter (Python/R) : sur nos serveurs
- Rstudio (R) : sur leur machine
- Org-Mode (Python/R/...) : pour les plus téméraires

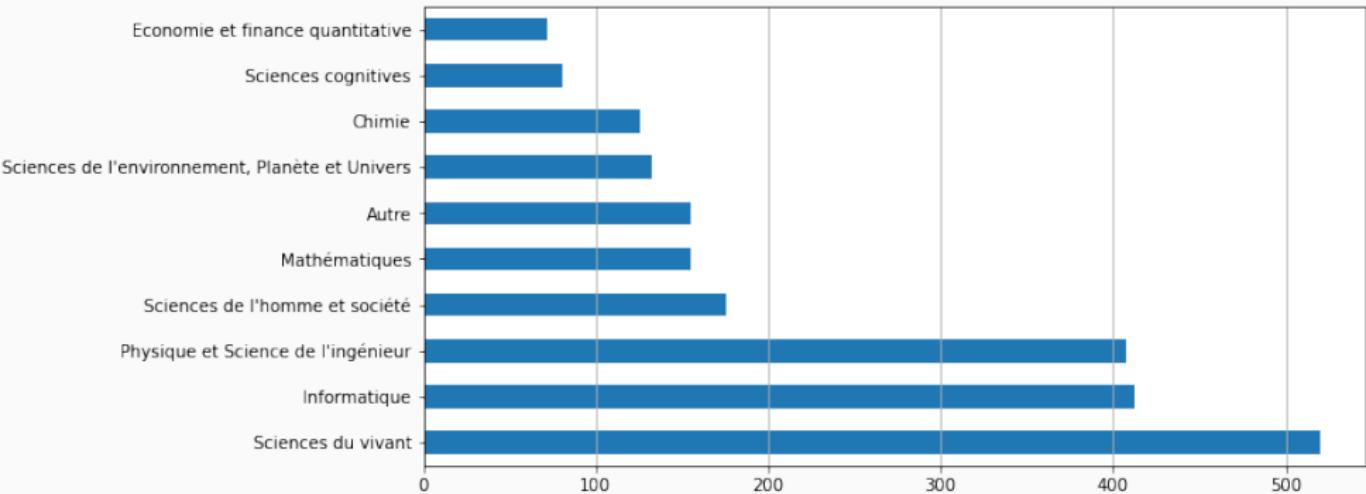


PROFIL SOCIO-DÉMOGRAPHIQUE

Public jeune 66% entre 19 et 35 ans | 59% d'hommes

- 87% résidant en France

Situation professionnelle 50% doctorants, 13% salariés du public , 10% salariés du privé, 10% étudiants, 5% enseignant-chercheur



Disciplines SVT, Environnement, Planète et Univers, ... : 34%

- Informatique : 27%

CONTRAINTE DU MOOC

Les MOOCs sur France Université Numérique

- Substitut à l'école/université
Public éduqué, formation complémentaire sur sujets pointus
- Audience majoritairement francophone
~~ Français/[English](#)
- Pas de MOOC à l'étranger sur le sujet

Attestation de suivi ≠ diplôme

- Inscription gratuite ~~ inscriptions nombreuses mais
efficacité de l'enseignement délicate à évaluer
- ADUM (~ 24 heures de cours)
- Open badges

Satisfaction 96% très satisfaits et satisfaits

- Mise en pratique : 48% tout à fait, 42% probablement

91% de "pas beaucoup" vont conseiller le MOOC 😊

Écoles doctorales Un relai moyennement efficace

- "SPAM" régulier
- Le MOOC RR bien moins suivi que les MOOCs
 - « *Ethique de la recherche* » (avril – juin 2020 : 6 577 inscrits, 2 664 attestations)
 - « *Intégrité scientifique dans les métiers de la recherche* » (oct. 2019 – sep. 2020 : 8 171 inscrits, 2 349 attestations)

Outils présentés et exercices (1/2)

Module 1 Prise de note

- Balisage léger avec **Markdown**
- Gestion de version avec **GitLab**
- Annotation et indexation

The screenshot shows a GitLab interface for a repository named 'LASCON_2018'. The sidebar on the left lists various repository sections: Overview, Repository (selected), Commits (highlighted in blue), Files, Branches, Tags, Contributors, Graph, Compare, Charts, Locked Files, Registry, Issues (0), and Merge Requests (0). The main content area displays a list of commits:

- ff927c0b2ce3e5c8ad7cafc696653299894f5d19 · 1: MarquesPages · 18 Jan, 2018 1 commit
- Start git section · Christophe Pouzat authored 5 · 12 Jan, 2018 1 commit
- Evolving notes up to wiki · Christophe Pouzat authored 6 · 11 Jan, 2018 1 commit
- Lightweight markup language · Christophe Pouzat authored a · 10 Jan, 2018 5 commits
- Historical part done. · Christophe Pouzat authored a · 10 Jan, 2018 5 commits
- Placcius' closet again. · Christophe Pouzat authored a · 10 Jan, 2018 5 commits
- Leishus done. · Christophe Pouzat authored a · 10 Jan, 2018 5 commits
- Codex done. · Christophe Pouzat authored a · 10 Jan, 2018 5 commits
- Moring's work · 12 Jan, 2018 1 commit

At the bottom of the sidebar, there is a link to 'Collapse sidebar'.

Outils présentés et exercices (1/2)

Module 1 Prise de note

- Balisage léger avec **Markdown**
- Gestion de version avec **GitLab**
- Annotation et indexation

The screenshot shows a Jupyter Notebook window titled "example_ipynb". It has a toolbar with File, Edit, View, Insert, Cell, Kernel, Widgets, Help, Hide Code, and Python 3. The main area contains two code cells and one figure cell.

In [1]:

```
# Un document computationnel
# Mon ordinateur m'indique que π vaut "approximativement"
In [1]:  
from math import *  
print(pi)  
3.141592653589793
```

In [2]:

```
Mais calculé avec la _méthode_ des _mâtelles du Buffon_
(https://fr.wikipedia.org/wiki/La\_guillotine\_de\_Buffon), on obtient donc
représentation...
```

In [3]:

```
import numpy as np  
N = 1000000  
x = np.random.uniform(size=N, low=0, high=1)  
y = np.random.uniform(size=N, low=0, high=pi/2)  
2/(pi*(np.sin(theta)+1))
```

Out[2]:

```
3.1437198694998765
```

In [3]:

```
# On peut inclure des formules mathématiques comme $1/\sqrt{1-x^2}$  
# pour faire une intégration de $x\sqrt{1-x^2}$ sur l'intervalle $[0,1]$.  
# Voici un exemple : trac([x*sqrt(1-x**2)]*xsign(x**2))**right) et des  
# résultats qui sont bien à voir, avec $0.15$ (si ce n'est une constante de  
# normalisation...).
```

Out[3]:

```
matplotlib inline  
import matplotlib.pyplot as plt  
mu, sigma = 100, 15  
x = mu + sigma*np.random.randn(10000)  
plt.hist(x, 40)  
plt.grid(True)  
plt.show()
```

A histogram is displayed, showing a bell-shaped curve centered at approximately 100, with a peak density of about 800.

Module 2 Document computationnel

- Jupyter / R-Markdown / Org-Mode
- Analyse des données de challenger
 - régression logistique
 - données "écartées"

Module 3 Analyse répliable

- Explication univoque de : provenance, transformation, analyse statistique, ...
- **Évaluation par les pairs** : 7 sujets
 1. Concentration de CO₂ dans l'atmosphère depuis 1958
 2. Pouvoir d'achat des ouvriers anglais du 16^{ème} au 19^{ème} siècle
 3. L'épidémie de choléra à Londres en 1854
 4. Analyse des dialogues dans l'avare de Molière

...

Outils présentés et exercices (2/2)

Module 3 Analyse répliable

- Explication univoque de : provenance, transformation, analyse statistique, ...
 - **Évaluation par les pairs** : 7 sujets
 1. Concentration de CO₂ dans l'atmosphère depuis 1958
 2. Pouvoir d'achat des ouvriers anglais du 16^{ème} au 19^{ème} siècle
 3. L'épidémie de choléra à Londres en 1854
 4. Analyse des dialogues dans l'avare de Molière
- ...

Module 4 Les enfers de la recherche reproductible

- HDF5, workflows, contrôle d'environnements, instabilité numérique
- **Reproduction de l'étude originale de Challenger**
- Articles de ReScience

REPRODUCIBLE RESEARCH II :

PRACTICES AND TOOLS FOR

MANAGING COMPUTATIONS AND DATA

Planning

- En réflexion depuis début 2020
- En **anglais**
- Prévu pour 2021, 2022, Nov. 2023!

Objectif rendre accessible les sujets plus techniques

- Logiciels libres, matures, et spécialisés

Pré-requis sous Linux

- *Familiarité avec la ligne de commande*

PÉRIMÈTRE ET STRUCTURE DU COURS

Planning

- En réflexion depuis début 2020
- En **anglais**
- Prévu pour 2021, 2022, Nov. 2023!

Objectif rendre accessible les sujets plus techniques

- Logiciels libres, matures, et spécialisés

Pré-requis sous Linux

- *Familiarité* avec la ligne de commande

3 gros modules relativement indépendants

- Managing data (FITS/HDF5, **git annex**, Zenodo, Software Heritage)
- Software environment control (**docker**, **singularity**, **guix**)
- Scientific workflow (**make**, **snakemake**)

Fil rouge le décompte des tâches solaires

- 2002 – ... : 28 000 images FITS