

# FORMATION À GRANDE ÉCHELLE À LA RECHERCHE REPRODUCTIBLE

---

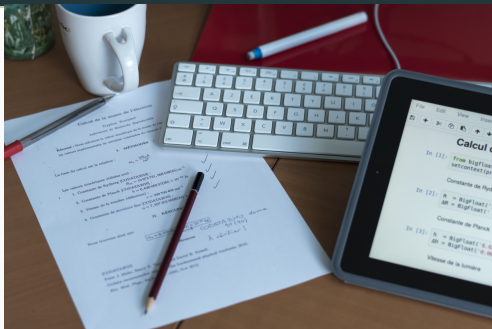
Arnaud Legrand (CNRS/Univ. Grenoble Alpes – LIG)

Recherche reproductible : état des lieux  
9 Mars 2023



RECHERCHE REPRODUCTIBLE :  
PRINCIPES MÉTHODOLOGIQUES  
POUR UNE SCIENCE TRANSPARENTE

---



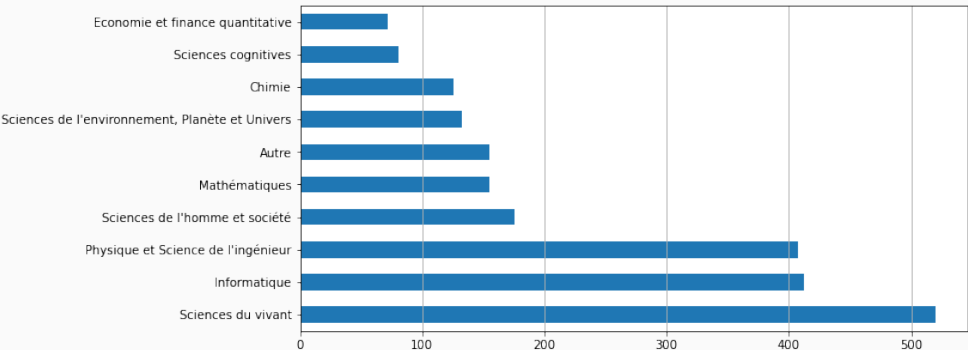
## Recherche reproductible : principes méthodologiques pour une science transparente (sur FUN MOOC)

- **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 – ...	13328		1552

# 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



- SVT, Environnement, Planète et Univers, ... : 34%
- Informatique : 27%

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



Konrad Hinsén  
Physique/Bio-chimie



Arnaud Legrand  
Informatique



Christophe Pouzat  
Neuro-bio/Stats.

## Réalisation, animation, ... (Inria)



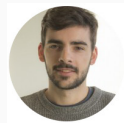
Laurence Farhi  
Pédagogique



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.*

## Les MOOCs sur France Université Numérique

- Pas un 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

## Complément aux écoles doctorales

- Un relai moyennement efficace
- 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)

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

## **Public visé pour la Recherche Reproductible**

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

## **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/algorithmique, ...
- Les points trop techniques (images docker, branches git...)

**Bonnes pratiques** logiciels libres et matures, format texte

**Troisième édition** tentative de **dissémination vers LSH** mais difficile

## 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 *gitlab, markdown*
2. Le document computationnel *jupyter/Rstudio/OrgMode*
3. Analyse intelligible et répliquable *Peer evaluation*
4. La réalité du terrain *Les enfers*

## Organisation Parties communes + 3 parcours :

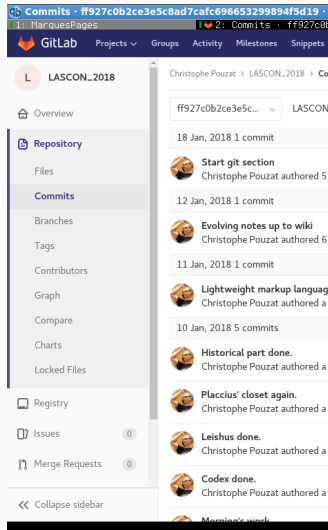
- Jupyter (Python/R) : sur nos serveurs
- Rstudio (R) : sur leur machine
- Org-Mode (Python/R/...) : pour les plus téméraires





## Module 1 Prise de note

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



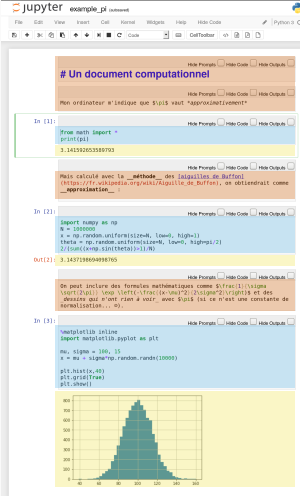
# OUTILS PRÉSENTÉS ET EXERCICES (1/2)

## Module 1 Prise de note

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

## Module 2 Document computationnel

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



The screenshot displays a Jupyter Notebook window titled "exemple\_pi". The notebook content includes:

- A title cell: "# Un document computationnel"
- A text cell: "Mon ordinateur m'indique que  $\pi$  a pour valeur 'approximativement'..."
- An input cell (In [1]):

```
from math import *\nprint(pi)\n3.141592653589793
```
- A text cell: "Mais calculé avec la `__methode__` des aiguilles de Buffon ([https://fr.wikipedia.org/wiki/Aiguille\\_de\\_Buffon](https://fr.wikipedia.org/wiki/Aiguille_de_Buffon)), on obtiendrait comme `__approximation__` :"
- An input cell (In [2]):

```
import numpy as np\nN = 100000\nx = np.random.uniform(size=N, low=0, high=1)\nyes = np.random.uniform(size=N, low=0, high=0.2)\n2/(sum((x*np.sin(theta))>1)/N)
```
- An output cell (Out [2]):

```
3.1457198604998705
```
- A text cell: "On peut inclure des formules mathématiques comme  $\frac{1}{\sqrt{2}}$  via `sqrt(2)/2` ou  $\frac{1}{\sqrt{2}}$  via `sqrt(2)/2` et des données qui n'ont rien à voir, avec `rand` (si ce n'est une constante de normalisation...)"
- An input cell (In [3]):

```
import matplotlib\nplt.plot(x)\nplt.grid(True)\nplt.show()
```
- An output cell (Out [3]): A plot showing a distribution of data points, likely representing the results of a simulation or experiment.

### Module 3 Analyse répliquable

- Explication univoque de : provenance, transformation, analyse statistique, ...
- **Évaluation par les pairs** : 7 sujets
  1. Concentration de CO<sub>2</sub> dans l'atmosphère depuis 1958
  2. Pouvoir d'achat des ouvriers anglais du 16<sup>ème</sup> au 19<sup>ème</sup> siècle
  3. L'épidémie de choléra à Londres en 1854
  4. Estimation de la latence et de la capacité d'une connexion réseau

...

## Module 3 Analyse répliquable

- Explicitation univoque de : provenance, transformation, analyse statistique, ...
- **Évaluation par les pairs** : 7 sujets
  1. Concentration de CO<sub>2</sub> dans l'atmosphère depuis 1958
  2. Pouvoir d'achat des ouvriers anglais du 16<sup>ème</sup> au 19<sup>ème</sup> siècle
  3. L'épidémie de choléra à Londres en 1854
  4. Estimation de la latence et de la capacité d'une connexion réseau

...

## 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
- 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