

# REPRODUCIBILITY CRISIS, OPEN SCIENCE,... AND COMPUTER SCIENCE

---

Arnaud Legrand



Séminaire des élèves, ENS Lyon  
October 2022



# SCIENTIFIC CONSENSUS VS. DEMOCRACY AND FREEDOM OF SPEECH



# PUBLIC EVIDENCE FOR A LACK OF REPRODUCIBILITY

- J.P. Ioannidis. *Why Most Published Research Findings Are False* PLoS Med. 2005.
- *Lies, Damned Lies, and Medical Science*, The Atlantic. Nov, 2010
- *Reproducibility: A tragedy of errors*, Nature, Feb 2016.
- Steen RG, *Retractions in the scientific literature: is the incidence of research fraud increasing?*, J. Med. Ethics 37, 2011



Science has lost its way, at a big cost to humanity

Researchers are rewarded for splashy findings, not for double-checking accuracy. So many scientists looking for cures to diseases have been building on ideas that aren't even true.

The screenshot shows the Science journal website. At the top, there are links for LOCAL, U.S., WORLD, BUSINESS, SPORTS, ENTERTAINMENT, HEALTH, STYLE, TRAVEL, and more. Below this, a large headline reads "Science has lost its way, at a big cost to humanity". Underneath, a sub-headline says "Researchers are rewarded for splashy findings, not for double-checking accuracy. So many scientists looking for cures to diseases have been building on ideas that aren't even true." The main content area features an article by Marcia McNutt titled "Reproducibility". The article summary states: "Science advances on a foundation of trusted data. But a recent study found that the approach that scientists used to gain confidence in their results was failing. The scientific community was shaken by reports that a troubling number of studies could not be reproduced." The article continues with details about the lack of reproducibility in science and the need for transparency and rigor. The author's name, Marcia McNutt, is mentioned as editor-in-chief of Science.

The screenshot shows the Nature journal website. At the top, there is a banner with the text "Announcement: Reducing our irreproducibility: Nature News & Comment". Below this, the main navigation menu includes Home, News & Comment, Research, Careers & Jobs, Current Issue, and Archive. The main content area features an announcement titled "Announcement: Reducing our irreproducibility" dated 24 April 2013. The announcement discusses the issue of irreproducibility in science and how Nature is addressing it through editorial policies and guidelines.



The screenshot shows the Nature journal website. At the top, there is a banner with the text "Announcement: Reducing our irreproducibility: Nature News & Comment". Below this, the main navigation menu includes Home, News & Comment, Research, Careers & Jobs, Current Issue, and Archive. The main content area features an announcement titled "Announcement: Reducing our irreproducibility" dated 24 April 2013. The announcement discusses the issue of irreproducibility in science and how Nature is addressing it through editorial policies and guidelines.

The screenshot shows the The Scientist magazine website. The main heading is "TheScientist EXPLORING LIFE. INSPIRING INNOVATION". Below this, a sub-headline reads "NIH Tackles Irreproducibility". The text states: "The federal agency speaks out about how to improve the quality of scientific research." At the bottom, it says "By Jef Akst | January 28, 2014". The footer of the page includes "Courtesy V. Stodden, SC, 2015".

# NEWSWORTHY STORIES ABOUT SCIENTIFIC MISCONDUCT

**Dong-Pyou Han** Assistant professor, Biomedical sciences, Iowa State University, 2013

*Falsified blood results to make it appear as though a vaccine exhibited anti-HIV activity*

- Han and his team received  $\approx$  \$19 million from NIH
- 1 retracted publication and resignation of university. Sentenced in 2015 to 57 months imprisonment for fabricating and falsifying data in HIV vaccine trials. \$7.2 million!

# NEWSWORTHY STORIES ABOUT SCIENTIFIC MISCONDUCT

**Dong-Pyou Han** Assistant professor, Biomedical sciences, Iowa State University, 2013

*Falsified blood results to make it appear as though a vaccine exhibited anti-HIV activity*

- Han and his team received  $\approx$  \$19 million from NIH
- 1 retracted publication and resignation of university. Sentenced in 2015 to 57 months imprisonment for fabricating and falsifying data in HIV vaccine trials. \$7.2 million!

**Diederik Stapel** Professor, Social Psychology, Univ. Tilburg, 2011

*I failed as a scientist. I adapted research data and fabricated research. Not once, but several times, not for a short period, but over a longer period of time. [...] I am aware of the suffering and sorrow that I caused to my colleagues... I did not withstand the pressure to score, to publish, the pressure to get better in time. I wanted too much, too fast. In a system where there are few checks and balances, where people work alone, I took the wrong turn.*

58 retracted publications

# NEWSWORTHY STORIES ABOUT SCIENTIFIC MISCONDUCT

**Dong-Pyou Han** Assistant professor, Biomedical sciences, Iowa State University, 2013

*Falsified blood results to make it appear as though a vaccine exhibited anti-HIV activity*

- Han and his team received  $\approx$  \$19 million from NIH
- 1 retracted publication and resignation of university. Sentenced in 2015 to 57 months imprisonment for fabricating and falsifying data in HIV vaccine trials. \$7.2 million!

**Diederik Stapel** Professor, Social Psychology, Univ. Tilburg, 2011

*I failed as a scientist. I adapted research data and fabricated research. Not once, but several times, not for a short period, but over a longer period of time. [...] I am aware of the suffering and sorrow that I caused to my colleagues... I did not withstand the pressure to score, to publish, the pressure to get better in time. I wanted too much, too fast. In a system where there are few checks and balances, where people work alone, I took the wrong turn.*

58 retracted publications

**Brian Wansink** Professor, Psychological Nutrition, Cornell, 2016

*I gave her a data set of a self-funded, failed study which had null results. I said "This cost us a lot of time and our own money to collect. There's got to be something here we can salvage because it's a cool (rich & unique) data set." I told her what the analyses should be. [...] Every day she came back with puzzling new results, and every day we would scratch our heads, ask "Why," and come up with another way to reanalyze the data with yet another set of plausible hypotheses*

17 retracted publications

# SCIENTIFIC MISCONDUCT? WHAT ARE THE CONSEQUENCES ?

**Reinhart and Rogoff** Professors of Economics at Harvard

*gross debt [...] exceeding 90 percent of the economy has a significant negative effect on economic growth* – Growth in a Time of Debt (2010)

*While using RR's working spreadsheet, we identified coding errors, selective exclusion of available data, and unconventional weighting of summary statistics.* – 2013: Herndon, Ash and Pollin

*For 3 years, austerity was not presented as an option but as a necessity.*

– 2013: Paul Krugman

At least, a scientific debate has been possible.

# SCIENTIFIC MISCONDUCT? WHAT ARE THE CONSEQUENCES ?

**Reinhart and Rogoff** Professors of Economics at Harvard

*gross debt [...] exceeding 90 percent of the economy has a significant negative effect on economic growth* – Growth in a Time of Debt (2010)

*While using RR's working spreadsheet, we identified coding errors, selective exclusion of available data, and unconventional weighting of summary statistics.* – 2013: Herndon, Ash and Pollin

*For 3 years, austerity was not presented as an option but as a necessity.*

– 2013: Paul\_Krugman

At least, a scientific debate has been possible.

## Bad science is deleterious

- It is used to backup stupid politics, it affects people's life, ...
- It blurs the frontier between scientists and crooks

## Media attention inflates conspiracy opinions 😞

- *Scientific result are worthless.*
- *Scientists can't even agree with each others on economy/climate/vaccine/5G/...*
- *Stop the scientific dictatorship/lobby!*

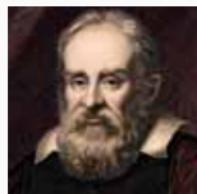
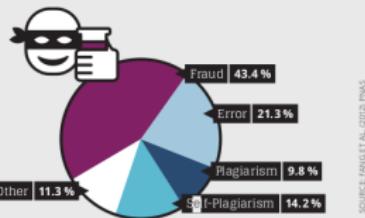
# A CREDIBILITY CRISIS?

How so? Why now? Why is this important? What can we do about it?

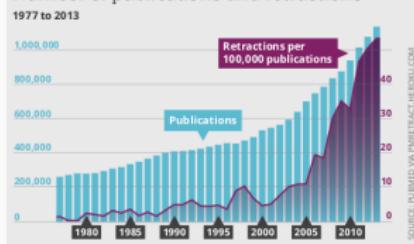
## *The Battle against Scientific Fraud in the CNRS International Magazine*

### Biomedical fraud in figures

Cause of retraction 1977 to 2012



Number of publications and retractions



**Galileo** (data fabrication), **Ptolemy** (plagiarism), **Mendel** (data enhancement), **Pasteur** (rigorous but hid failures), ...

Scientific misconduct is obviously wrong but it's **not new!**

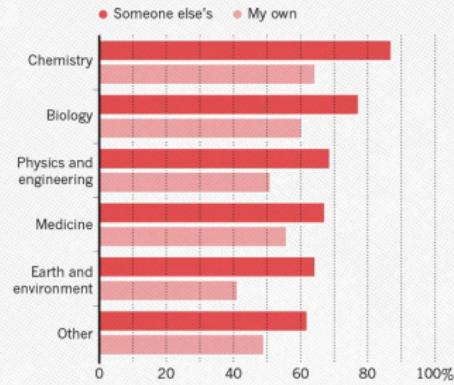
- Every domain has its black sheep

- The publish or perish pressure is a pain

# A REPRODUCIBILITY CRISIS?

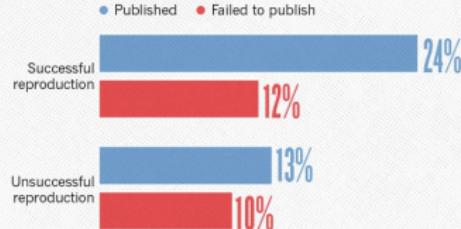
## HAVE YOU FAILED TO REPRODUCE AN EXPERIMENT?

Most scientists have experienced failure to reproduce results.



## HAVE YOU EVER TRIED TO PUBLISH A REPRODUCTION ATTEMPT?

Although only a small proportion of respondents tried to publish replication attempts, many had their papers accepted.



Number of respondents from each discipline:  
Biology 703, Chemistry 106, Earth and environmental 95,  
Medicine 203, Physics and engineering 236, Other 233.

1,500 scientists lift the lid on reproducibility,

Nature, May 2016

## Social causes

- Fraud, conflict of interest (pharmaceutic, ...)
- No incentive to reproduce/check our own work (afap), nor the work of others (big results!), nor to allow others to check (competition)
- Peer review does not scale: 1+ million articles per year!

## Methodological or technical causes

- The many biases (apophenia, confirmation, hindsight, experimenter, ...): bad designs
- Selective reporting, weak analysis (statistics, data manipulation mistakes, computational errors)
- Lack of information, code/raw data unavailable

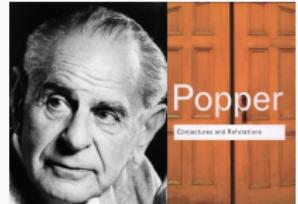
# NO TRANSPARENCY NO CONSENSUS



# REPRODUCIBILITY OF EXPERIMENTAL RESULTS: THE HALLMARK OF SCIENCE

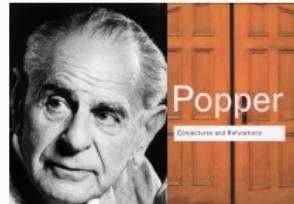
1934: Karl Popper puts the notions of **falsifiability** and **crucial experiment** as the **hallmark of science**

- If no experiment can be set up to **disprove** your theory, it is not science
- Good experiments **discriminate** good theories from bad ones
- Non-reproducible single occurrences are of no significance to science



# REPRODUCIBILITY OF EXPERIMENTAL RESULTS: THE HALLMARK OF SCIENCE

1934: Karl Popper puts the notions of **falsifiability** and **crucial experiment** as the **hallmark of science**



- If no experiment can be set up to **disprove** your theory, it is not science
- Good experiments **discriminate good theories from bad ones**
- Non-reproducible single occurrences are of no significance to science

## An ideal rather than the norm

Popper's proposal works well for Physics from the 18th century but is not so simple for many other domains:

- Theory of evolution
- Biology (every animal does not behave in the same way)
- Spotting a SuperNova
- Anthropology (impact on people from a remote culture)
- Particle Physics (a single LHC)

# REPRODUCIBILITY: A CORE VALUE OF SCIENCE

---

1. Universality: Science aims for objective findings, accessible to anyone

Reproducibility acts as a Universality/Robustness control

2. Incremental: We build on each others work but everybody makes mistakes

Methods, biases, ... How to discriminate sound theories experiments from bad ones? 😊

Reproducibility acts as a Quality control

# REPRODUCIBILITY: A CORE VALUE OF SCIENCE

1. Universality: Science aims for objective findings, accessible to anyone

Reproducibility acts as a Universality/Robustness control

2. Incremental: We build on each others work but everybody makes mistakes

Methods, biases, ... How to discriminate sound theories experiments from bad ones? 😊

Reproducibility acts as a Quality control

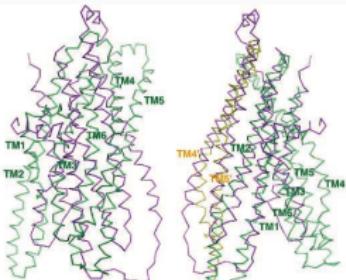
But, scientific practices have greatly evolved, in particular since we rely on computers



How computers broke science – and what we can do about it

– Ben Marwick, The conversation, 2015

# How COMPUTERS BROKE SCIENCE



Geoffrey Chang (Scripps, UCSD) works on crystallography and studies the structure of cell membrane proteins.

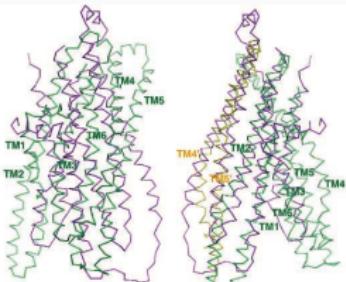
He specialized in structures of **multidrug resistant transporter proteins in bacteria**: MsbA de Escherichia Choli (Science, 2001), Vibrio cholera (Mol. Biology, 2003), Salmonella typhimurium (Science, 2005)

2006: Inconsistencies reveal **a programming mistake**

*A homemade data-analysis program had flipped two columns of data, inverting the electron-density map from which his team had derived the protein structure.*

**5 retractions** that motivate **improved software engineering practices** in comp. biology

# How COMPUTERS BROKE SCIENCE



Geoffrey Chang (Scripps, UCSD) works on crystallography and studies the structure of cell membrane proteins.

He specialized in structures of **multidrug resistant transporter proteins in bacteria**: MsbA de Escherichia Choli (Science, 2001), Vibrio cholera (Mol. Biology, 2003), Salmonella typhimurium (Science, 2005)

2006: Inconsistencies reveal **a programming mistake**

*A homemade data-analysis program had flipped two columns of data, inverting the electron-density map from which his team had derived the protein structure.*

5 retractions that motivate **improved software engineering practices** in comp. biology

There is **worse!**

- The generalized and intensive use of **spreadsheets** (**COVID tracing**)
- Relying on **black box** statistical methods is infinitely easier than understanding them  
(Learning and Data Analytics frameworks = nuke)
- Numerical errors and software environment unawareness

# 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



# DIFFERENT REPRODUCIBILITY CONCERN IN MODERN SCIENCE

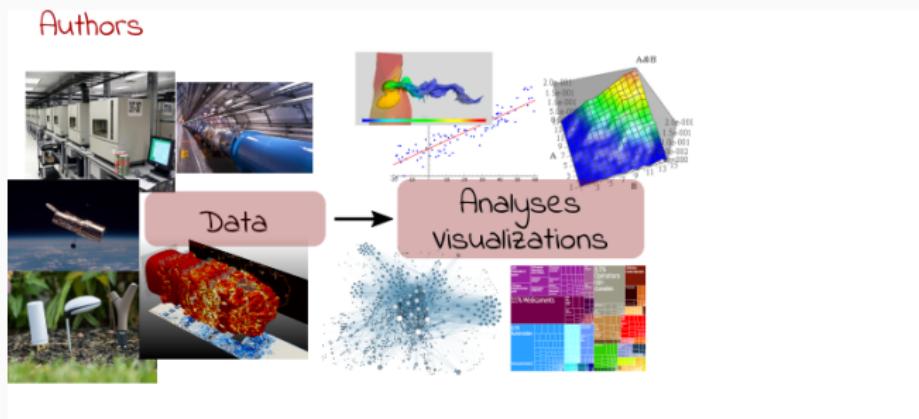
**Social Sciences, Oncology, ...** methodology, statistics, pre-registration

**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

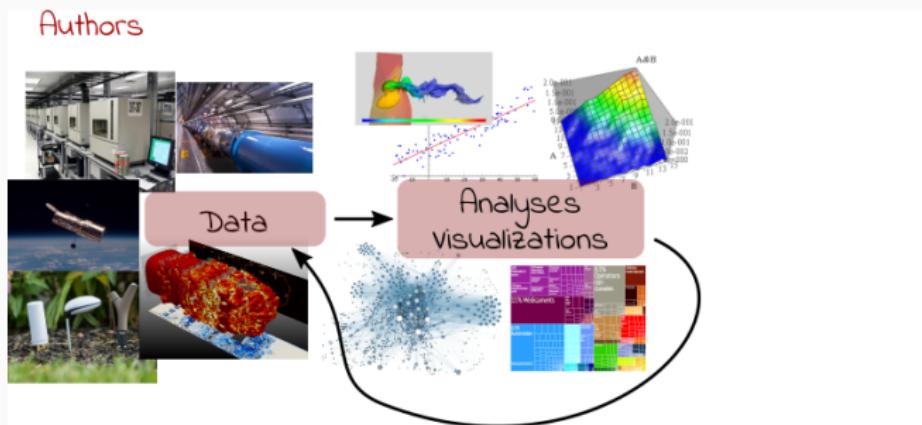
**Social Sciences, Oncology, ...** methodology, statistics, pre-registration

**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

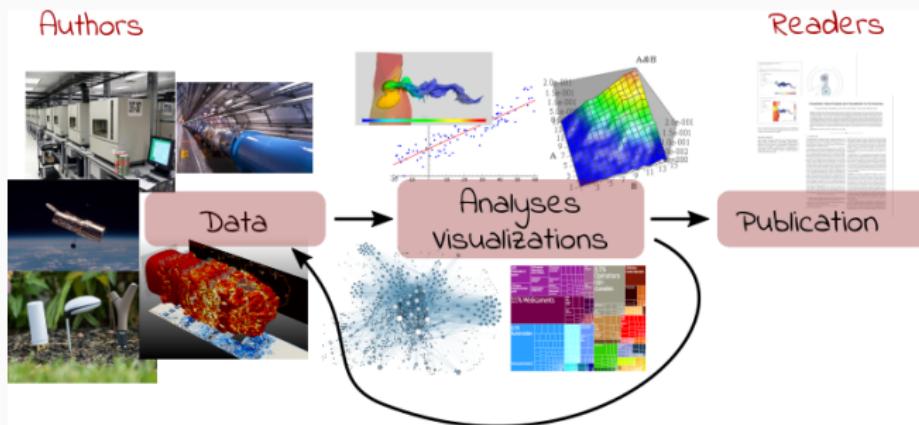
**Social Sciences, Oncology, ...** methodology, statistics, pre-registration

**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

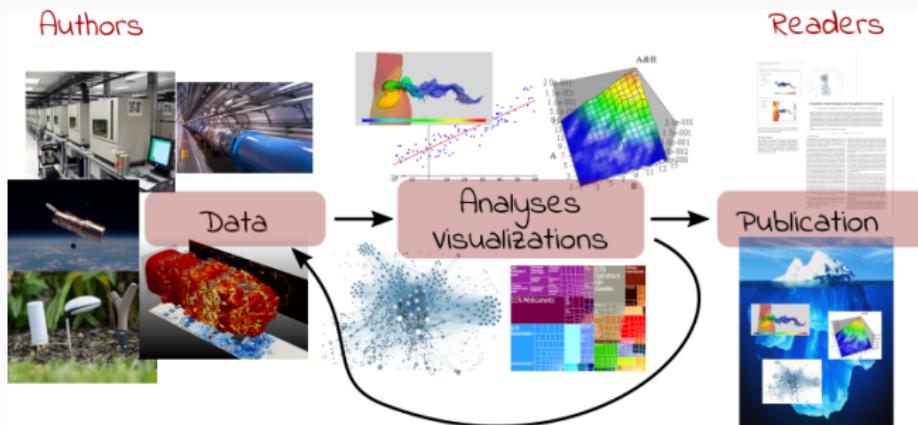
**Social Sciences, Oncology, ...** methodology, statistics, pre-registration

**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

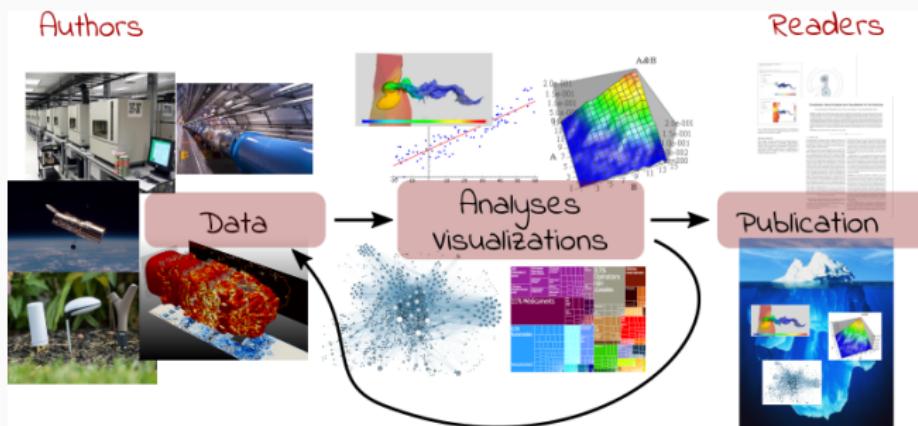
**Social Sciences, Oncology, ...** methodology, statistics, pre-registration

**Genomics** software engineering, computational reproducibility, provenance

**Computational fluid dynamics** numerical issues

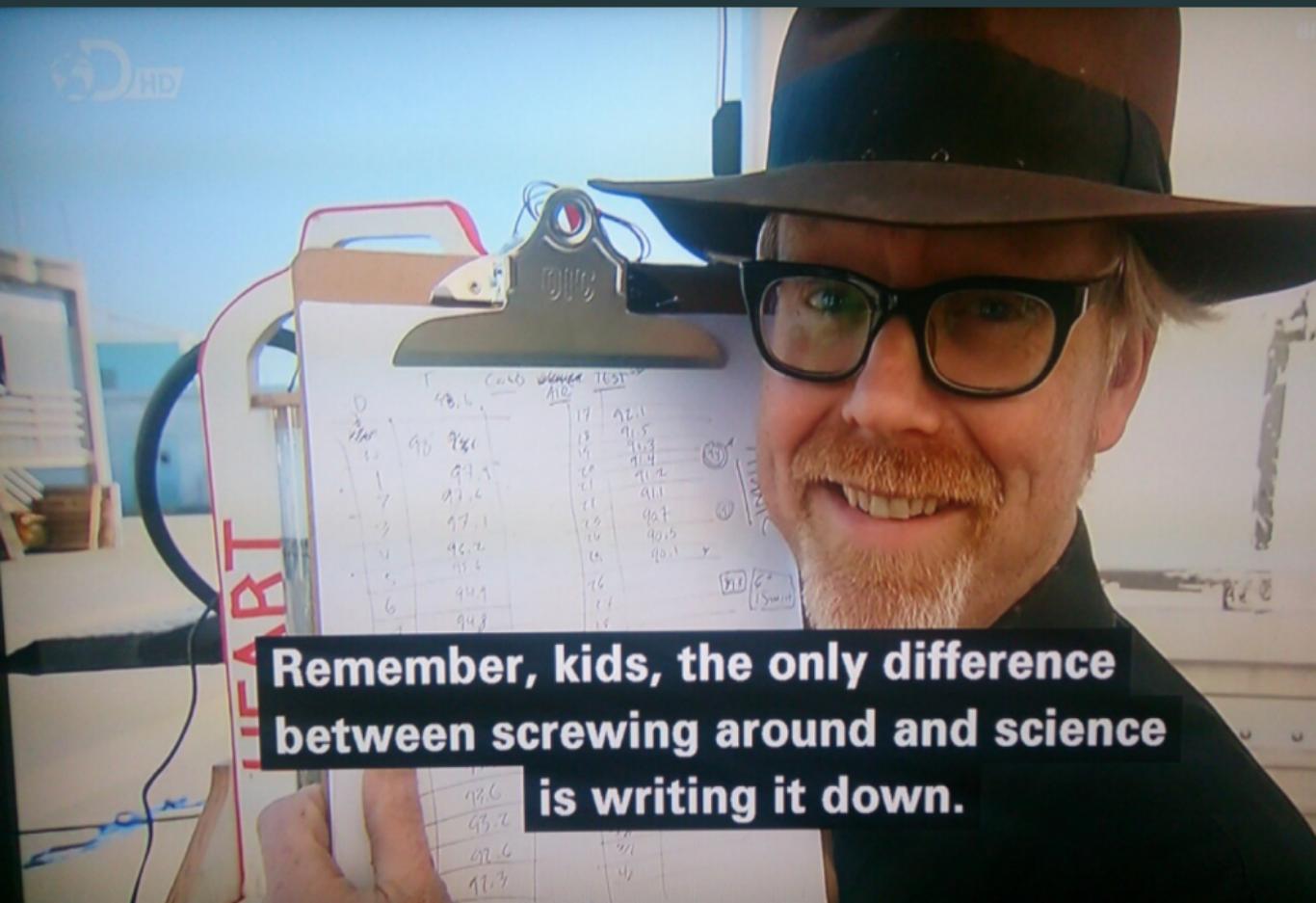
**Artificial Intelligence** most of the above

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



Reproducible Research = Bridging the Gap by working Transparently 11/45

# MYTHBUSTERS: SCIENCE VS. SCREWING AROUND



**Remember, kids, the only difference  
between screwing around and science  
is writing it down.**

# REPRODUCIBLE RESEARCH PRACTICES

---

# "REPRODUCIBLE RESEARCH": FIRST APPEARANCE

Claerbout & Karrenbach, meeting of the Society of Exploration Geophysics, 1992

## Electronic Documents Give Reproducible Research a New Meaning

RE1.3

Jon F. Claerbout and Martin Karrenbach, Stanford Univ.

### SUMMARY

A revolution in education and technology transfer follows from the marriage of word processing and software command scripts. In this marriage an author attaches to every figure caption a pushbutton or a name tag usable to recalculate the figure from all its data, parameters, and programs. This provides a new meaning of reproducibility in computer documents.

In 1990, we set this sequence of goals:

- Learn how to merge a publication with its underlying computational analysis.
- Teach researchers how to prepare a document in a form where they themselves can reproduce their own research results a year or more later by "pressing a single button".
- Learn how to leave finished work in a condition where coworkers can reproduce the calculation including the final illustration by pressing a button in its caption.
- Prepare a complete copy of our local software environment so that graduating students can take their work away with them to other sites, press a button, and reproduce their Stanford work.
- Merge electronic documents written by multiple authors (SEP reports).

- make incremental improvements in electronic-document software
- seek partners for broadening standards (and making incremental improvements).

Our basic goal is reproducible research. The electronic document is our means to this end. In principle, reproducibility in research can be achieved without electronic documents and that is how we started. Our first nonelectronic reproducible document was a textbook in which the paper document contained the name of a program script in every figure caption. The program scripts were organized by book chapter and section so they could be correlated to an accompanying magnetic tape dump of the file system. The magnetic tape also contained all the necessary data to feed the program script.

Now that we have begun using CD-ROM publication, we can go much further. Every figure caption contains a pushbutton that jumps to the appropriate science directory (folder) and initiates a figure rebuild command and then displays the figure, possibly as a movie or interactive program. We normally display seismic images of the earth's interior, but to reach wider audiences, Figure 1 shows a satellite weather picture which the pushbutton will animate as seen on commercial television. We include all our plot software as well as freely available software from many sources, including compilers and the L<sup>A</sup>T<sub>E</sub>X word processing systems. Naturally some software includes licensed software, but with the exception

# REPRODUCIBILITY (GLOSSARY MAY VARY)

Many **definitions** (*replicability, repeatability, reproducibility*), sometimes conflicting  
(*new data, same person, independent researcher*)

<b>experimental</b> reproducibility	similar input (data) + similar experimental protocol	→	<b>similar results</b> <sup>1</sup>
<b>statistical</b> reproducibility	different input (data) + same analysis	→	<b>same conclusions</b> <sup>2</sup>
<b>computational</b> reproducibility	similar input (data) + same code/software + same software environment	→	<b>exact same results</b> <sup>3</sup>

Reproducible Research = A way of doing science so that scientific experiments, discoveries, results, etc. can be easily reproduced (done again), to be confirmed, or to be built on for the next study.

– Courtesy G. Durrif, 2021

---

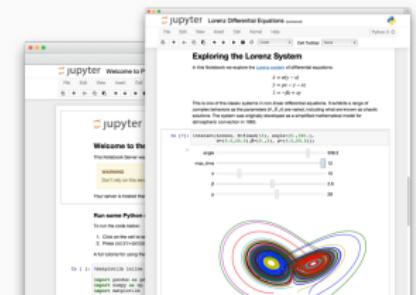
<sup>1</sup>Up-to measurement variability and precision

<sup>2</sup>Independently from (random) sampling variability (fight bias)

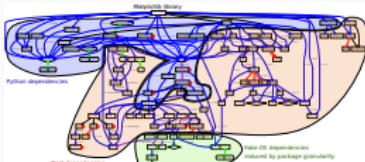
<sup>3</sup>Bitwise

# EXISTING TOOLS, EMERGING STANDARDS

## Notebooks and workflows



## Software environments



## Sharing platforms



## GOOD PRACTICE #1

### TAKING NOTES AND DOCUMENTING

---



## Author

- I thought I used the same parameters but **I'm getting different results!**
- The new student wants to compare with **the method I proposed last year**
- My advisor asked me whether I took care of setting this or this but **I can't remember**
- The damned fourth reviewer asked for a major revision and wants me to change **Figure 3. Which code and which data set did I use?**
- **It worked yesterday!** 6 months later: Why did I do that?

## Reviewer

- As usual, there is **no confidence interval**, I wonder about the variability and whether the difference is **significant** or not
- That can't be true, I'm sure **they removed some points**
- Why is this graph in logscale? **How would it look like otherwise?** I'm not even **sure** of what this value means. If only I could access the generation script

# TOOL 1: COMPUTATIONAL NOTEBOOKS/LITTERATE PROGRAMMING

## Un document computationnel

Mon ordinateur m'indique que  $\pi$  vaut approximativement

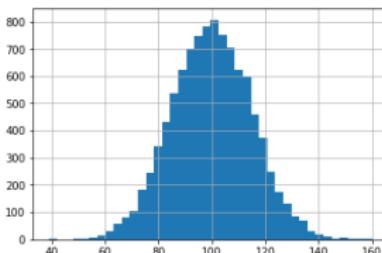
3.141592653589793

Mais calculé avec la méthode des [aiguilles de Buffon](#), on obtiendrait comme approximation :

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

3.1437198694098765

On peut inclure des formules mathématiques comme  $\frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$  et des dessins qui n'ont rien à voir avec  $\pi$  (si ce n'est une constante de normalisation... ☺).



# TOOL 1: COMPUTATIONAL NOTEBOOKS/LITTERATE PROGRAMMING

Document initial dans son environnement

The screenshot shows a Jupyter Notebook interface with several code cells:

- In [1]:**

```
from math import *
print(pi)
```

 Output: 3.141592653589793
- In [2]:**

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

 Output: 3.14371986944998765
- In [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()
```

 Output: A histogram showing a normal distribution centered around 100.

Document final

## Un document computationnel

Mon ordinateur m'indique que  $\pi$  vaut approximativement

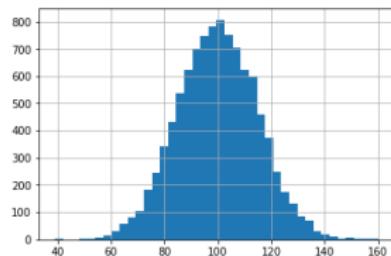
3.141592653589793

Mais calculé avec la [méthode des aiguilles de Buffon](#), on obtiendrait comme approximation :

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

3.14371986944998765

On peut inclure des formules mathématiques comme  $\frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$  et des dessins qui n'ont rien à voir avec  $\pi$  (si ce n'est une constante de normalisation... ☺).



# TOOL 1: COMPUTATIONAL NOTEBOOKS/LITTERATE PROGRAMMING

Document initial dans son environnement

A screenshot of a Jupyter Notebook interface. The title bar says "jupyter example\_pi (modified)". The main area shows a cell with the following content:

```
# Un document computationnel

Mon ordinateur m'indique que $\pi$ vaut "approximativement"

In [1]: from math import * print(pi)
3.141592653589793

Mais calculé avec la méthode des aiguilles de Buffon (https://fr.wikipedia.org/wiki/Aiguille\_de\_Buffon), on obtient comme approximation :
```

Another cell below it shows:

```
In [2]: import numpy as np N = 1000000 x = np.random.uniform(size=N, low=0, high=1) theta = np.random.uniform(size=N, low=0, high=pi/2) 2*(sum((x*np.sin(theta))>1))/N
```

The output of this cell is:

```
Out[2]: 3.14371986944998765
```

Below this, another cell contains:

```
In [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()
```

And its output is a histogram of a normal distribution centered at 100.

Document final

## Un document computationnel

Mon ordinateur m'indique que  $\pi$  vaut approximativement

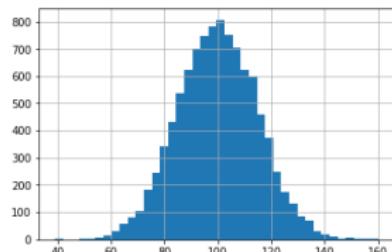
3.141592653589793

Mais calculé avec la [méthode des aiguilles de Buffon](#), on obtient comme approximation :

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

3.14371986944998765

On peut inclure des formules mathématiques comme  $\frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$  et des dessins qui n'ont rien à voir avec  $\pi$  (si ce n'est une constante de normalisation... ☺).



# TOOL 1: COMPUTATIONAL NOTEBOOKS/LITTERATE PROGRAMMING

Document initial dans son environnement

The screenshot shows a Jupyter Notebook interface with three code cells:

- In [1]:** Prints the value of pi (3.141592653589793) and includes a note about calculating pi using Buffon's needle method.
- In [2]:** Calculates a numerical approximation of pi using a Monte Carlo simulation with 1000000 points.
- In [3]:** Plots a histogram of 100 random numbers between 0 and 100, showing a bell-shaped distribution centered around 100.

Annotations with red arrows point from the text "Code" in the first section to the code blocks in the notebook, and from the text "Document initial dans son environnement" to the top of the screenshot.

Document final

## Un document computationnel

Mon ordinateur m'indique que  $\pi$  vaut approximativement

3.141592653589793

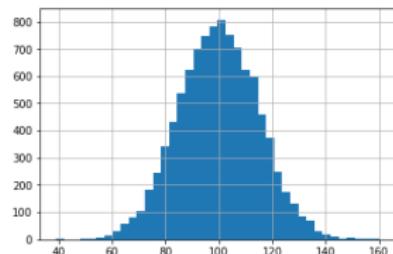
Mais calculé avec la méthode des [aiguilles de Buffon](#), on obtiendrait comme approximation :

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

3.14371986949098765

On peut inclure des formules mathématiques comme  $\frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$  et

des dessins qui n'ont rien à voir avec  $\pi$  (si ce n'est une constante de normalisation... ☺).



# TOOL 1: COMPUTATIONAL NOTEBOOKS/LITTERATE PROGRAMMING

Document initial dans son environnement

# Un document computationnel

```
In [1]:  
from math import *  
print(pi)  
3,141592653589793
```

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 :

```
In [2]:  
import numpy as np  
N = 1000000  
x = np.random.uniform(size=N, low=0, high=1)  
theta = np.random.uniform(size=N, low=0, high=pi/2)  
2/(sum((x+np.sin(theta))>1))/N  
Out[2]: 3,143719869498765
```

On peut inclure des formules mathématiques comme  $\frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$  et des dessins qui n'ont rien à voir avec  $\pi$  (si ce n'est une constante de normalisation... ☺).

```
In [3]:  
%matplotlib inline  
import matplotlib.pyplot as plt  
  
mu, sigma = 100, 15  
x = mu + sigma*np.random.randn(10000)  
  
plt.hist(x, 99)  
plt.grid(True)  
plt.show()
```

Résultats

Document final

## Un document computationnel

Mon ordinateur m'indique que  $\pi$  vaut approximativement

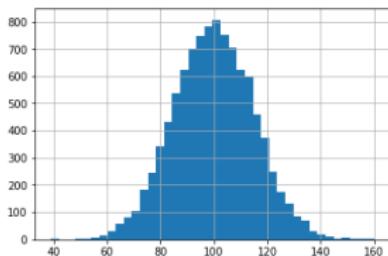
3.141592653589793

Mais calculé avec la **méthode des aiguilles de Buffon**, on obtiendrait comme approximation :

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

3.143719869498765

On peut inclure des formules mathématiques comme  $\frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$  et des dessins qui n'ont rien à voir avec  $\pi$  (si ce n'est une constante de normalisation... ☺).



# TOOL 1: COMPUTATIONAL NOTEBOOKS/LITTERATE PROGRAMMING

Document initial dans son environnement

A screenshot of a Jupyter Notebook interface. The top bar shows 'jupyter example\_pi' and 'Python 3'. Below is a toolbar with various icons. The main area contains three code cells:

- In [1]:** Prints  $\pi$  to the console.
- In [2]:** Calculates the value of  $\pi$  using Buffon's needle method. It includes a note from Wikipedia about the formula and its derivation.
- In [3]:** Generates a histogram of random numbers between 0 and 100, showing a bell-shaped distribution.

Red arrows point from the text "Document initial dans son environnement" to the notebook interface and from the word "Export" to the histogram.

Document final

## Un document computationnel

Mon ordinateur m'indique que  $\pi$  vaut approximativement

3.141592653589793

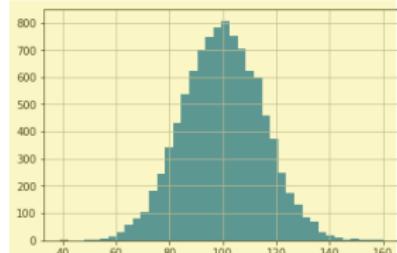
Mais calculé avec la **méthode des aiguilles de Buffon**, on obtiendrait comme approximation :

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

3.1437198694098765

Export →

On peut inclure des formules mathématiques comme  $\frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$  et des dessins qui n'ont rien à voir avec  $\pi$  (si ce n'est une constante de normalisation... ☺).



# TOOL 1: COMPUTATIONAL NOTEBOOKS/LITTERATE PROGRAMMING

Document initial dans son environnement

The screenshot shows a Jupyter Notebook interface with three code cells:

- In [1]:** Prints the value of  $\pi$  (3.141592653589793) and includes a note about calculating it with the Buffon's needle method.
- In [2]:** Generates random points and calculates the ratio of points below a line to the total number of points, which is used to approximate  $\pi$ .
- In [3]:** Plots a histogram of 100,000 random numbers between 0 and 1, showing a bell-shaped distribution centered at 0.5.

Document final

## Un document computationnel

Mon ordinateur m'indique que  $\pi$  vaut approximativement

3.141592653589793

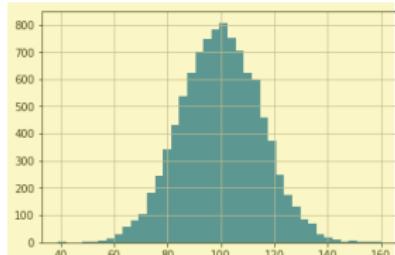
Mais calculé avec la méthode des [aiguilles de Buffon](#), on obtiendrait comme approximation :

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

3.1437198694098765

Export

On peut inclure des formules mathématiques comme  $\frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$  et des dessins qui n'ont rien à voir avec  $\pi$  (si ce n'est une constante de normalisation... ☺).



# JOURNAL AND REPRODUCIBLE ARTICLE DEMO

Document your:

- **Hypotheses**: keep track of your ideas/line of thoughts
- **Experiments**: details on how and why an experiment was run, including failed or ambiguous attempts
- **Initial analysis or interpretation of these experiments**: was the outcome conform to the expectation or not? does it (in)validate the hypothesis? **why** did you do this or that ?
- **Organization**: keep track of things to do/fix/test/improve

Write for the future you

I have a very intense usage of my journal and I can **demo this today**

- Experiment results are better **structured by dates (add tags)**
- Final rendering of results (figures, tables, article, presentation) should be reproducible
- Use plain text and lightweight markup languages (e.g., **LATEX** or Markdown)

# TOOL 1 BIS: WORKFLOWS

Notebooks are no panacea and do not help developing clean code

jupyter example\_pi [segment]

File Edit View Insert Cell Kernel Widgets Help Hide Code Hide Outputs Python 3 Cell Toolbar

# Un document computationnel

Mon ordinateur m'indique que  $\pi$  vaut "approximativement"

In [1]:

```
from math import *  
print(pi)  
3.141592653589793
```

Mais calculé avec la `__method__` des (ajoutées de Buffet) `__approximation__`, on obtiendrait comme approximation :

In [2]:

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

Out[2]:

On peut inclure des formules mathématiques comme  $\frac{1}{\sqrt{2}}$  via `math` ou `sympy`. On peut également faire des choses un peu plus élaborées, comme dessiner ce qu'il faut pour avoir une loi normale (ce qui devrait donner ce qu'il faut pour avoir une loi normale de normalisation...).

In [3]:

```
%matplotlib inline  
import matplotlib.pyplot as plt  
  
N = 1000000  
x = np.random.uniform(0, high=1)  
plt.hist(x, 40)  
plt.title("Normal distribution")  
plt.show()
```

# TOOL 1 BIS: WORKFLOWS

Notebooks are no panacea and do not help developing clean code

jupyter analyse-syndrome-grippal Last Checkpoint 20 minutes ago (autosaved)

File Edit View Insert Cell Kernel Help Hide Code Export to HTML

In [1]: `#!/usr/bin/python3  
# Import des librairies nécessaires pour l'analyse de données.  
import pandas as pd  
import numpy as np`

Les données de l'Institut de Santé Publique peuvent être trouvées à ce lien : <https://www.inserm.fr/>. Nous les disposons sous forme d'un fichier "syndrome\_grippal" pour servir à la démonstration. Téléchargez-le pour accéder au fichier CSV sur un autre ordinateur.

Voici quelques lignes extraites du fichier CSV :

Nom de colonne	Type de colonne
date	Date en format YYYY-MM-DD
id	Identifiant unique de la consultation
age	Âge de la personne consultée
sex	Sexe de la personne consultée
nb_cold	Nombre de personnes atteintes de syndrome grippal
nb_cold_1000	Nombre de personnes atteintes de syndrome grippal pour 1000 personnes
nb_cold_10000	Nombre de personnes atteintes de syndrome grippal pour 10000 personnes
nb_cold_100000	Nombre de personnes atteintes de syndrome grippal pour 100000 personnes
geo_zone	Zone de la carte géographique concernée (Code INSEE) : <a href="https://www.insee.fr/fr/methodes-referencement/codes-insee.htm">https://www.insee.fr/fr/methodes-referencement/codes-insee.htm</a>
nb_cold_max	Nombre maximum de personnes atteintes de syndrome grippal pour 100000 personnes
nb_cold_min	Nombre minimum de personnes atteintes de syndrome grippal pour 100000 personnes
nb_cold_mean	Moyenne des personnes atteintes de syndrome grippal pour 100000 personnes
nb_cold_std	Écart-type des personnes atteintes de syndrome grippal pour 100000 personnes

In [2]: `# On va voir si le rapport entre nb_cold_max et nb_cold_min est constant ou non.  
df.groupby('geo_zone').nb_cold_min.nunique()`

Out[2]:

geo_zone	nb_cold_min
000	280.00
001	207.00
002	217.00
003	212.00
004	212.00
005	212.00
006	212.00
007	212.00
008	212.00
009	212.00
010	212.00
011	212.00
012	212.00
013	212.00
014	212.00
015	212.00
016	212.00
017	212.00
018	212.00
019	212.00
020	212.00
021	212.00
022	212.00
023	212.00
024	212.00
025	212.00
026	212.00
027	212.00
028	212.00
029	212.00
030	212.00
031	212.00
032	212.00
033	212.00
034	212.00
035	212.00
036	212.00
037	212.00
038	212.00
039	212.00
040	212.00
041	212.00
042	212.00
043	212.00
044	212.00
045	212.00
046	212.00
047	212.00
048	212.00
049	212.00
050	212.00
051	212.00
052	212.00
053	212.00
054	212.00
055	212.00
056	212.00
057	212.00
058	212.00
059	212.00
060	212.00
061	212.00
062	212.00
063	212.00
064	212.00
065	212.00
066	212.00
067	212.00
068	212.00
069	212.00
070	212.00
071	212.00
072	212.00
073	212.00
074	212.00
075	212.00
076	212.00
077	212.00
078	212.00
079	212.00
080	212.00
081	212.00
082	212.00
083	212.00
084	212.00
085	212.00
086	212.00
087	212.00
088	212.00
089	212.00
090	212.00
091	212.00
092	212.00
093	212.00
094	212.00
095	212.00
096	212.00
097	212.00
098	212.00
099	212.00
100	212.00
101	212.00
102	212.00
103	212.00
104	212.00
105	212.00
106	212.00
107	212.00
108	212.00
109	212.00
110	212.00
111	212.00
112	212.00
113	212.00
114	212.00
115	212.00
116	212.00
117	212.00
118	212.00
119	212.00
120	212.00
121	212.00
122	212.00
123	212.00
124	212.00
125	212.00
126	212.00
127	212.00
128	212.00
129	212.00
130	212.00
131	212.00
132	212.00
133	212.00
134	212.00
135	212.00
136	212.00
137	212.00
138	212.00
139	212.00
140	212.00
141	212.00
142	212.00
143	212.00
144	212.00
145	212.00
146	212.00
147	212.00
148	212.00
149	212.00
150	212.00
151	212.00
152	212.00
153	212.00
154	212.00
155	212.00
156	212.00
157	212.00
158	212.00
159	212.00
160	212.00
161	212.00
162	212.00
163	212.00
164	212.00
165	212.00
166	212.00
167	212.00
168	212.00
169	212.00
170	212.00
171	212.00
172	212.00
173	212.00
174	212.00
175	212.00
176	212.00
177	212.00
178	212.00
179	212.00
180	212.00
181	212.00
182	212.00
183	212.00
184	212.00
185	212.00
186	212.00
187	212.00
188	212.00
189	212.00
190	212.00
191	212.00
192	212.00
193	212.00
194	212.00
195	212.00
196	212.00
197	212.00
198	212.00
199	212.00
200	212.00
201	212.00
202	212.00
203	212.00
204	212.00
205	212.00
206	212.00
207	212.00
208	212.00
209	212.00
210	212.00
211	212.00
212	212.00
213	212.00
214	212.00
215	212.00
216	212.00
217	212.00
218	212.00
219	212.00
220	212.00
221	212.00
222	212.00
223	212.00
224	212.00
225	212.00
226	212.00
227	212.00
228	212.00
229	212.00
230	212.00
231	212.00
232	212.00
233	212.00
234	212.00
235	212.00
236	212.00
237	212.00
238	212.00
239	212.00
240	212.00
241	212.00
242	212.00
243	212.00
244	212.00
245	212.00
246	212.00
247	212.00
248	212.00
249	212.00
250	212.00
251	212.00
252	212.00
253	212.00
254	212.00
255	212.00
256	212.00
257	212.00
258	212.00
259	212.00
260	212.00
261	212.00
262	212.00
263	212.00
264	212.00
265	212.00
266	212.00
267	212.00
268	212.00
269	212.00
270	212.00
271	212.00
272	212.00
273	212.00
274	212.00
275	212.00
276	212.00
277	212.00
278	212.00
279	212.00
280	212.00
281	212.00
282	212.00
283	212.00
284	212.00
285	212.00
286	212.00
287	212.00
288	212.00
289	212.00
290	212.00
291	212.00
292	212.00
293	212.00
294	212.00
295	212.00
296	212.00
297	212.00
298	212.00
299	212.00
300	212.00
301	212.00
302	212.00
303	212.00
304	212.00
305	212.00
306	212.00
307	212.00
308	212.00
309	212.00
310	212.00
311	212.00
312	212.00
313	212.00
314	212.00
315	212.00
316	212.00
317	212.00
318	212.00
319	212.00
320	212.00
321	212.00
322	212.00
323	212.00
324	212.00
325	212.00
326	212.00
327	212.00
328	212.00
329	212.00
330	212.00
331	212.00
332	212.00
333	212.00
334	212.00
335	212.00
336	212.00
337	212.00
338	212.00
339	212.00
340	212.00
341	212.00
342	212.00
343	212.00
344	212.00
345	212.00
346	212.00
347	212.00
348	212.00
349	212.00
350	212.00
351	212.00
352	212.00
353	212.00
354	212.00
355	212.00
356	212.00
357	212.00
358	212.00
359	212.00
360	212.00
361	212.00
362	212.00
363	212.00
364	212.00
365	212.00
366	212.00
367	212.00
368	212.00
369	212.00
370	212.00
371	212.00
372	212.00
373	212.00
374	212.00
375	212.00
376	212.00
377	212.00
378	212.00
379	212.00
380	212.00
381	212.00
382	212.00
383	212.00
384	212.00
385	212.00
386	212.00
387	212.00
388	212.00
389	212.00
390	212.00
391	212.00
392	212.00
393	212.00
394	212.00
395	212.00
396	212.00
397	212.00
398	212.00
399	212.00
400	212.00
401	212.00
402	212.00
403	212.00
404	212.00
405	212.00
406	212.00
407	212.00
408	212.00
409	212.00
410	212.00
411	212.00
412	212.00
413	212.00
414	212.00
415	212.00
416	212.00
417	212.00
418	212.00
419	212.00
420	212.00
421	212.00
422	212.00
423	212.00
424	212.00
425	212.00
426	212.00
427	212.00
428	212.00
429	212.00
430	212.00
431	212.00
432	212.00
433	212.00
434	212.00
435	212.00
436	212.00
437	212.00
438	212.00
439	212.00
440	212.00
441	212.00
442	212.00
443	212.00
444	212.00
445	212.00
446	212.00
447	212.00
448	212.00
449	212.00
450	212.00
451	212.00
452	212.00
453	212.00
454	212.00
455	212.00
456	212.00
457	212.00
458	212.00
459	212.00
460	212.00
461	212.00
462	212.00
463	212.00
464	212.00
465	212.00
466	212.00
467	212.00
468	212.00
469	212.00
470	212.00
471	212.00
472	212.00
473	212.00
474	212.00
475	212.00
476	212.00
477	212.00
478	212.00
479	212.00
480	212.00
481	212.00
482	212.00
483	212.00
484	212.00
485	212.00
486	212.00
487	212.00
488	212.00
489	212.00
490	212.00
491	212.00
492	212.00
493	212.00
494	212.00
495	212.00
496	212.00
497	212.00
498	212.00</td

## TOOL 1 BIS: WORKFLOWS

Notebooks are no panacea and do not help developing clean code

The figure shows a Jupyter Notebook interface with several code cells, data tables, and plots.

- Title:** analyse-syndrome-grippal (Last Checkpoint 20 minutes ago - Published)
- Data Tables:** Two large tables are displayed, likely representing daily or weekly counts of influenza-like illness cases across different regions or categories.
- Plots:** Several plots are shown:
  - A line plot of 'Nombre de cas par jour' (Number of cases per day) from January 1 to March 15, 2013, showing high variability with multiple peaks.
  - A histogram of 'Nombre de cas par jour' (Number of cases per day) for the same period.
  - A scatter plot of 'Nombre de cas par jour' (Number of cases per day) versus 'Date' (Date).
  - A histogram of 'Nombre de cas par semaine' (Number of cases per week) for the period from Jan 1 to Mar 15, 2013.
- Code Cells:** Numerous code cells are present, including:
  - Import statements: `import pandas as pd` and `import numpy as np`.
  - Data loading: `df = pd.read\_csv('data.csv')` and `df2 = pd.read\_csv('data2.csv')`.
  - Data processing: `df['Date'] = pd.to\_datetime(df['Date'])` and `df['Semaine'] = df['Date'].dt.isocalendar().week`.
  - Statistical calculations: `df['Cas'].sum()` and `df['Cas'].mean()`.
  - Plotting: `df['Cas'].plot()` and `df['Cas'].hist()`.
  - Conditional logic: `if len(df) > 0` and `if len(df2) > 0`.
  - Looping: `for i in range(1, len(df))` and `for i in range(1, len(df2))`.

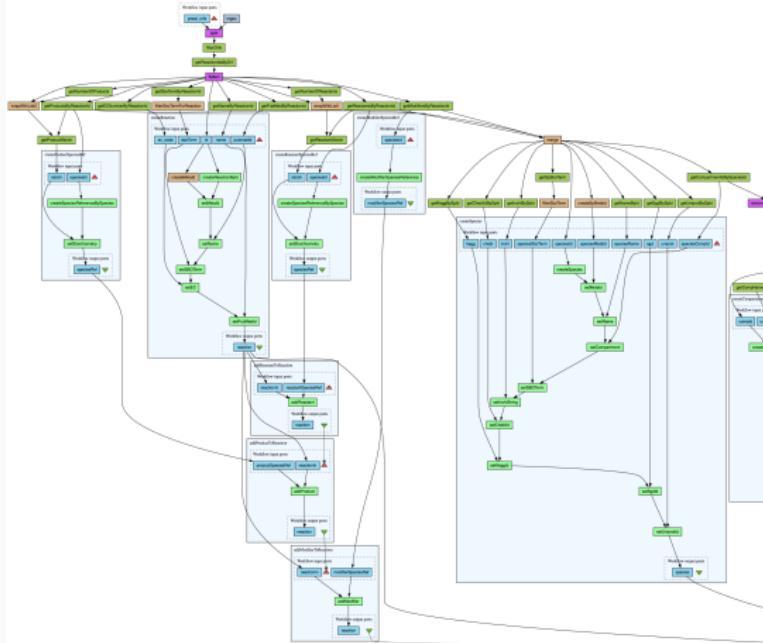
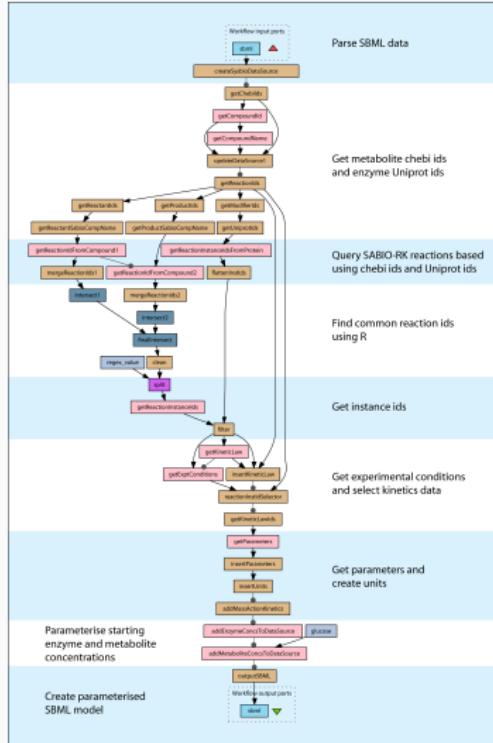
# TOOL 1 BIS: WORKFLOWS

Notebooks are no panacea and do not help developing clean code

The image displays a 4x3 grid of Jupyter Notebook screenshots, each showing a different step in a machine learning project. The steps include:

- Gathering Color Names by Web Image Searchers**: A cell with code to extract color names from web images.
- Model training architecture**: A cell with code for a neural network architecture.
- Model fitting**: A cell with code for training a model with data.
- Model evaluation**: A cell with code for evaluating model results.
- Modeling the training data**: A cell with code for creating a scatter plot of training data.
- Modeling the test data**: A cell with code for creating a scatter plot of test data.
- Modeling the results**: A cell with code for creating a heatmap of model results.
- Modeling the error**: A cell with code for creating a scatter plot of prediction error vs training sample variance.
- Modeling the classifier**: A cell with code for creating a scatter plot of classifier results.
- Modeling the classifier**: Another cell with code for creating a scatter plot of classifier results.
- Modeling the classifier**: A third cell with code for creating a scatter plot of classifier results.
- Modeling the classifier**: A fourth cell with code for creating a scatter plot of classifier results.

# TOOL 1 BIS: WORKFLOWS



### Workflows:

- Clearer high-level view
- Composition of codes and data movement made explicit
- Safer sharing, reusing, and execution
- Notebooks are a variant that is both impoverished and richer
- No simple/mature path from a notebook to a workflow

### Examples:

- Galaxy, Kepler, Taverna, Pegasus, Collective Knowledge, VisTrails
- Light-weight: dask, drake, swift, snakemake, ...
- Hybrids: SOS-notebook, ...

## GOOD PRACTICE #2

### CONTROLLING SOFTWARE ENVIRONMENT

---

# ARGH... DAMNED COMPUTERS

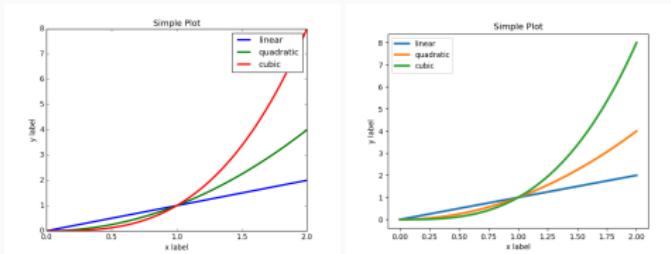
---

- Alice: I got 3.123123 Bob: I got segfault
- Damned! It used to work!!! Whenever I upgrade my computer, things break so I try to stay away from this 😞
- Whenever trying the code of my colleague, I had to install **libFoo-1.5c** but I broke everything and now neither his code nor mine works! 😞
- But hey! Here is my code. It's on GitHub so feel free to play with it! I'm doing open science 😊

Seriously ? It's 21st century. How come all this is so painful ?

# BACKWARDS COMPATIBILITY

- Software environment evolution



## BACKWARDS COMPATIBILITY

---

- Software environment evolution
- Software evolution and OS heterogeneity

The Effects of FreeSurfer Version, Workstation Type, and Macintosh Operating System Version on Anatomical Volume and Cortical Thickness Measurements (PLOS ONE, 2012)

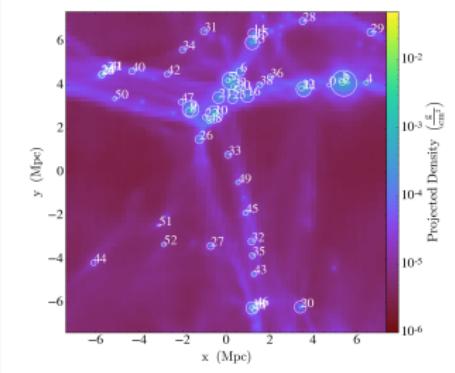
*Significant differences in volume and cortical thickness were revealed across FreeSurfer versions. In addition, less pronounced differences were found between the Mac and HP workstations and between Mac OSX 10.5 and OSX 10.6.*

# BACKWARDS COMPATIBILITY

- Software environment evolution
- Software evolution and OS heterogeneity
- Impact of the compiler

Assessing Reproducibility: An Astrophysical Example of Computational Uncertainty in the HPC Context  
(ResCuE-HPC, 2018)

Compiler	Optim.	Largest Halo Avg Mass.	Std. Err	Walltime
gcc@6.2.0	None	2.273E46	1.069E44	22h
gcc@6.2.0	Normal	2.266E46	1.218E44	10h
gcc@6.2.0	High	2.275E46	1.199E44	9h
intel@16.0.3	None	2.271E45	1.587E44	39h
intel@16.0.3	Normal	4.330(45)	1.248E44	7h
intel@16.0.3	High	2.268E46	1.414E44	6h
cce@8.5.5	Low	4.311(45)	1.353E44	16h
cce@8.5.5	Normal	2.271E46	1.261E44	6h
cce@8.5.5	High	2.272E46	1.341E44	5h

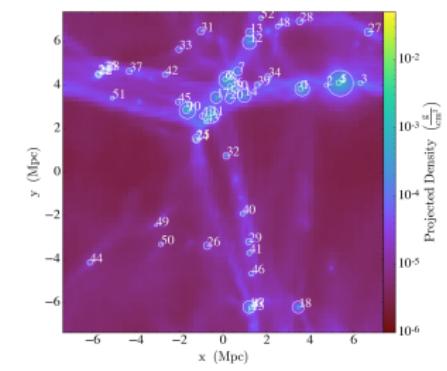


# BACKWARDS COMPATIBILITY

- Software environment evolution
- Software evolution and OS heterogeneity
- Impact of the compiler

Assessing Reproducibility: An Astrophysical Example of Computational Uncertainty in the HPC Context  
(ResCuE-HPC, 2018)

Compiler	Optim.	Largest Halo Avg Mass.	Std. Err	Walltime
gcc@6.2.0	None	2.273E46	1.069E44	22h
gcc@6.2.0	Normal	2.266E46	1.218E44	10h
gcc@6.2.0	High	2.275E46	1.199E44	9h
intel@16.0.3	None	2.271E45	1.587E44	39h
intel@16.0.3	Normal	4.330(45)	1.248E44	7h
intel@16.0.3	High	2.268E46	1.414E44	6h
cce@8.5.5	Low	4.311(45)	1.353E44	16h
cce@8.5.5	Normal	2.271E46	1.261E44	6h
cce@8.5.5	High	2.272E46	1.341E44	5h

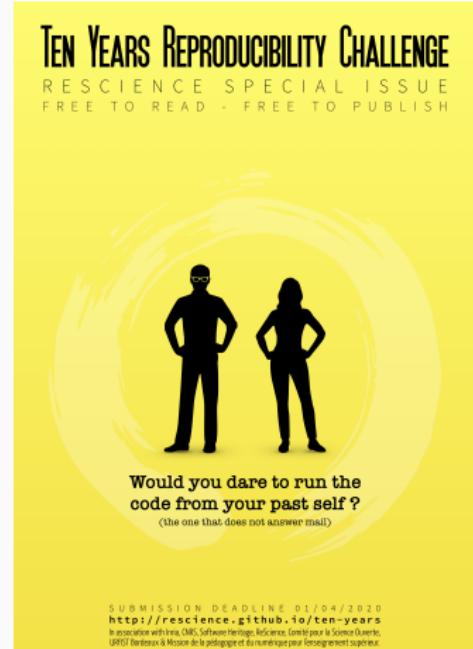


# BACKWARDS COMPATIBILITY

- Software environment evolution
- Software evolution and OS heterogeneity
- Impact of the compiler

Assessing Reproducibility: An Astrophysical Example of Computational Uncertainty in the HPC Context  
(ResCuE-HPC, 2018)

Compiler	Optim.	Largest Halo Avg Mass.	Std. Err	Walltime
gcc@6.2.0	None	2.273E46	1.069E44	22h
gcc@6.2.0	Normal	2.266E46	1.218E44	10h
gcc@6.2.0	High	2.275E46	1.199E44	9h
intel@16.0.3	None	2.271E45	1.587E44	39h
intel@16.0.3	Normal	4.330(45)	1.248E44	7h
intel@16.0.3	High	2.268E46	1.414E44	6h
cce@8.5.5	Low	4.311(45)	1.353E44	16h
cce@8.5.5	Normal	2.271E46	1.261E44	6h
cce@8.5.5	High	2.272E46	1.341E44	5h



<http://rescience.github.io/ten-years/>

# COMPLEX ECOSYSTEMS

```
import matplotlib  
print(matplotlib.__version__)
```

3.1.2

# COMPLEX ECOSYSTEMS

```
import matplotlib
print(matplotlib.__version__)
```

## 3.1.2

```
apt show python3-matplotlib
```

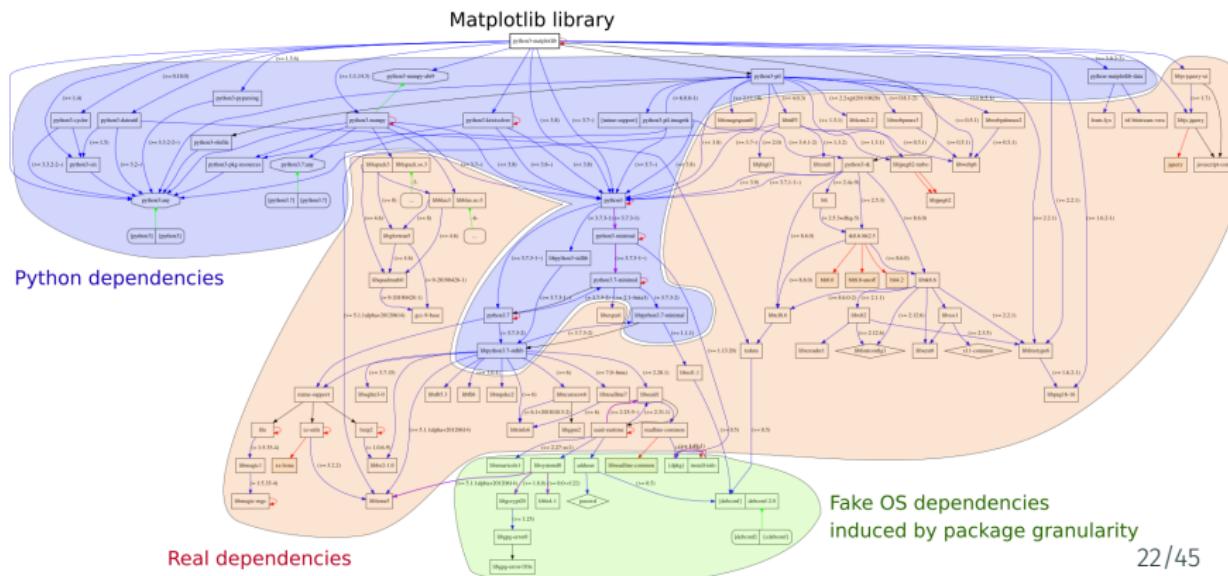
Package: python3-matplotlib  
Version: 3.1.2-2  
Priority: optional  
Section: python  
Source: matplotlib  
Maintainer: Sandro Tosi <morph@debian.org>  
Installed-Size: 15.3 MB  
Depends: python3-dateutil, python-matplotlib-data (>= 3.1.2-2), python3-pyparsing,  
six (>= 1.4), libjs-jquery, libjs-jquery-ui, python3-numpy (>= 1:1.16.0~rc1), py-  
numpy-abi9, python3 (<< 3.9), python3 (>= 3.7~), python3-cycler (>= 0.10.0), py-  
kiwisolver, python3:any, libc6 (>= 2.29), libfreetype6 (>= 2.2.1), libgcc-  
s1 (>= 3.0), libpng16-16 (>= 1.6.2-1), libstdc++6 (>= 5.2)  
Recommends: python3-pil, python3-tk  
Suggests: dvipng, ffmpeg, gir1.2-gtk-3.0, ghostscript, inkscape, ipython3, librs-  
common, python-matplotlib-doc, python3-cairocffi, python3-gi, python3-gi-cairo,  
gobject, python3-nose, python3-pyqt5, python3-scipy, python3-sip, python3-22/45  
tornado, texlive-extra-utils, texlive-latex-extra, ttf-staypuft

# COMPLEX ECOSYSTEMS

```
import matplotlib  
print(matplotlib.__version__)
```

### 3.1.2

```
apt show python3-matplotlib
```



# NON STANDARD ECOSYSTEMS

## No standard

- Linux (`apt`, `rpm`, `yum`), MacOS X (`brew`, `MacPorts`, `Fink`), Windows (?)
- Neither for installation nor for retrieving the information... 😞

```
import sys
print(sys.version)
import matplotlib
print(matplotlib.__version__)
import pandas as pd
print(pd.__version__)

3.7.6 (default, Jan 19 2020, 22:34:52)
[GCC 9.2.1 20200117]
3.1.2
0.25.3
```

```
library(ggplot2)
sessionInfo()

R version 3.6.3 RC (2020-02-21 r77847)
Platform: x86_64-pc-linux-gnu (64-bit)
Running under: Debian GNU/Linux bullseye/sid

Matrix products: default
BLAS:    /usr/lib/x86_64-linux-gnu/atlas/libblas.so.3.10.3
LAPACK:  /usr/lib/x86_64-linux-gnu/atlas/liblapack.so.3.10.3

locale:
[1] C

attached base packages:
[1] stats      graphics   grDevices utils      datasets   methods

other attached packages:
[1] ggplot2_3.2.1

loaded via a namespace (and not attached):
 [1] Rcpp_1.0.3        withr_2.1.2       crayon_1.3.4      dplyr_
 [5] assertthat_0.2.1  grid_3.6.3       R6_2.4.1         life_
 [9] gtable_0.3.0      magrittr_1.5     scales_1.1.0      pillar_
[13] rlang_0.4.4       lazyeval_0.2.2    glue_1.3.1       purrr_
[17] munsell_0.5.0     compiler_3.6.3   pkgconfig_2.0.3   colorspace_
[21] tidyselect_1.0.0   tibble_2.1.3    23/45
```

# ARGH... DAMNED COMPUTERS

---

- Whenever I upgrade my computer, things break so I try to stay away from this 😞
- Whenever trying the code of my colleague, I had to install `libFoo-1.5` but I broke everything and now neither his code nor mine works! 😞
- But hey! Here is my code, feel free to play with it! I'm doing open science 😊

## Are you really aware of your dependencies ?

- No one will ever run/use your code if it isn't easy to install
- No one will ever manage to run your code if you don't document how to run it
- Others (even you) are unlikely to get the same results unless you automate the execution

## TOOL 2: CONTAINERS AND PACKAGE MANAGERS

The good



The bad



The ugly



Automatic tracking

## TOOL 2: CONTAINERS AND PACKAGE MANAGERS

The good



The bad



The ugly



Automatic tracking

Containers

- Pros: Lightweight, Good isolation, Easy to use
  - Running as easy as `docker run <img> <cmd>`
  - Building images: `docker build -f <Dockerfile>`
  - Sharing through the Docker Hub: `docker pull/push <img>`

## TOOL 2: CONTAINERS AND PACKAGE MANAGERS

The good



The bad



The ugly



Automatic tracking

### Containers

- **Pros:** Lightweight, Good isolation, Easy to use
- **Cons:** Opaque, Container build is generally not reproducible
  - Recipes rarely follow *reproducible good practices*

```
FROM ubuntu:20.04
RUN apt-get update
    && apt-get upgrade -y
    && apt-get install -y ...
```

- Choose a stable image (and the smallest possible)
- Include only the necessary libraries (e.g. no graphics libs)
- Avoid system updates (instead freeze sources)

## TOOL 2: CONTAINERS AND PACKAGE MANAGERS

The good



The bad



The ugly



Automatic tracking

### Containers

- Pros: Lightweight, Good isolation, Easy to use
- Cons: Opaque, Container build is generally not reproducible

### Package managers

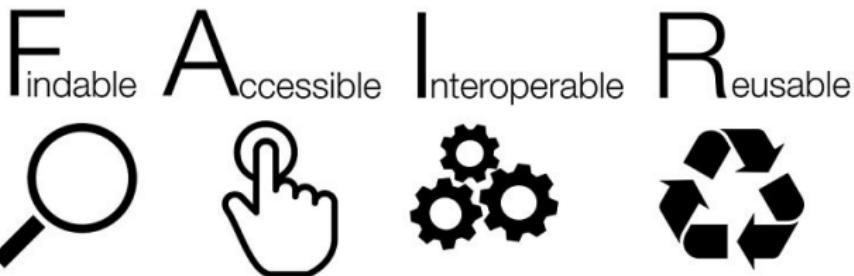
- Language specific: `pip/pipenv/virtualenv`, `conda`, `CRAN/Bioconductor`
  - Limits: version management, durability, permeable, language centric
- **GUIX/NiX** = Full-fledged functional package manager
  - Native support for environment (*à la git*)
  - Isolation through `--pure`
  - Recompile from source (cache recommended)

## GOOD PRACTICE #3

## VERSION CONTROL AND ARCHIVING

---

# FAIR PRINCIPLES



<https://www.go-fair.org/fair-principles/>

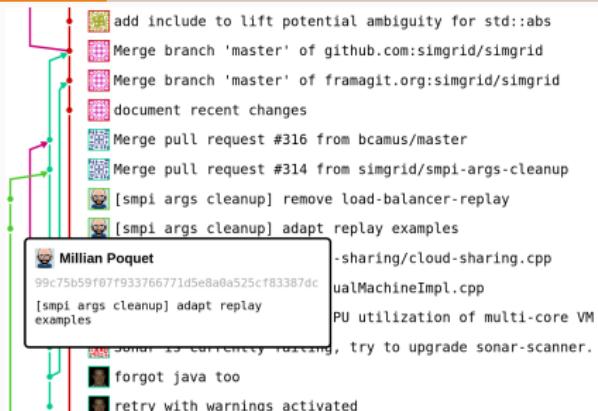
- "*Open as much as possible and close as much as necessary*"
- Management, publication, annotation (metadata), archiving
- Source code = specific data with specific consideration

Let's go beyond general principles!

# TOOL 3: VERSION CONTROL AND FORGE

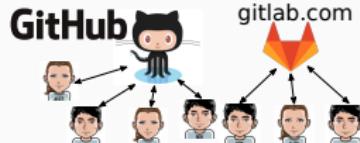
## Git = version control

- Developed in 2005 by Linus Torvalds for the kernel development
- Local and efficient rollbacks
- Distributed: everyone has a full copy of the history



## GitHub, GitLab, and Co

- Free hosting of public projects, social network
- Web interfaces (browsing, preview, online editing)
- User management (read/write, public/private)
- Issues, Continuous Integration, ...



## TOOL 3BIS: FIGHTING INFORMATION LOSS WITH ARCHIVES



or



= awesome collaborations ( $\neq$  archive)

- D. Spinellis. *The Decay and Failures of URL References*. CACM, 46(1), 2003  
*The half-life of a referenced URL is approximately 4 years from its publication date.*
- P. Habibzadeh. *Decay of References to Web sites in Articles Published in General Medical Journals: Mainstream vs Small Journals*. Applied Clinical Informatics. 4 (4), 2013  
*half life ranged from 2.2 years in EMHJ to 5.3 years in BMJ*
- Discontinued forges: Code Space, Gitorious, Google code, Inria Gforge

## TOOL 3BIS: FIGHTING INFORMATION LOSS WITH ARCHIVES



or



= awesome collaborations ( $\neq$  archive)

- D. Spinellis. *The Decay and Failures of URL References*. CACM, 46(1), 2003  
*The half-life of a referenced URL is approximately 4 years from its publication date.*
- P. Habibzadeh. *Decay of References to Web sites in Articles Published in General Medical Journals: Mainstream vs Small Journals*. Applied Clinical Informatics. 4 (4), 2013  
*half life ranged from 2.2 years in EMHJ to 5.3 years in BMJ*
- Discontinued forges: Code Space, Gitorious, Google code, Inria Gforge

Article archives



Data archives



figshare



Software Archive



Software Heritage

Collect/Preserve/Share

## WHAT ABOUT COMPUTER SCIENCE ?

---

# ALL THIS IS ABOUT COMPUTATIONAL SCIENCES. SHOULD WE CARE ?

Computer Science is young and inherits from Mathematics,  
Engineering, Nat. Sciences, Linguistic, ...

Purely theoretical scientists whose practice is close to mathematics  
*may* not be concerned (can't publish a math article without releasing  
the proofs).

- Have a look at talk by Vladimir Voevodsky in 2014 at Princeton 😊

# ALL THIS IS ABOUT COMPUTATIONAL SCIENCES. SHOULD WE CARE ?

Computer Science is young and inherits from Mathematics,  
Engineering, Nat. Sciences, Linguistic, ...

Purely theoretical scientists whose practice is close to mathematics  
may not be concerned (can't publish a math article without releasing  
the proofs).

- Have a look at talk by Vladimir Voevodsky in 2014 at Princeton 😊

*Computer science is not more related to computers than Astronomy to telescopes*

– Dijkstra (mis-attributed)

Right, why should we care about computers? They are **deterministic**  
machines after all, right? 😊

**Model ≠ Reality.** Although designed and built by human beings,  
computer systems are **so complex** that mistakes easily slip in...

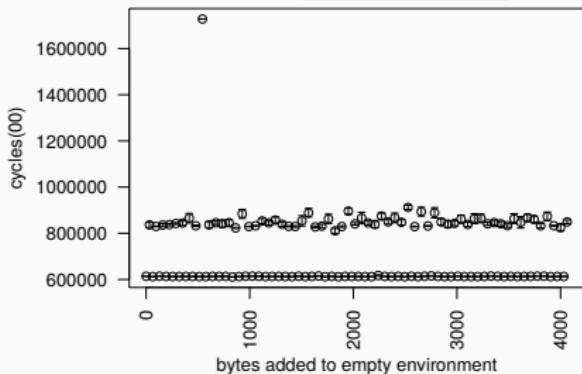
# EXPERIMENTING WITH COMPUTERS

Machines are real!



Brendan Gregg: Shouting in the data center

Machines are complicated



Mytkowicz et al. *Producing wrong data without doing anything obviously wrong!*  
ACM SIGPLAN Not. 44(3), March 2009

Our reality evolves!!! The hardware keeps evolving so most results on old platforms quickly become obsolete (although, we keep building on such results 😊).

We need to regularly revisit and allow others to build on our work!

# COMPUTER PERFORMANCE ? WELL, I DESIGN ALGORITHMS!

- "Real" problems are all NP-hard, Log-APX, etc.
- Real workload = ~~NP-completeness proof widgets~~, regularities and properties (difficult to formally state but that should be exploited)

Algorithms are evaluated on particular **workloads** that impact both their running time and the quality of the solutions

# COMPUTER PERFORMANCE ? WELL, I DESIGN ALGORITHMS!

- "Real" problems are all NP-hard, Log-APX, etc.
- Real workload = ~~NP-completeness proof~~ widgets, regularities and properties (difficult to formally state but that should be exploited)

Algorithms are evaluated on particular **workloads** that impact both their running time and the quality of the solutions

**Image Processing:** **True horror stories**, E. Meinhardt-Llopis, CANUM 2016

- *The proposed multigrid algorithm converges to the solution of the problem in  $O(N)$  using biharmonic functions*
- Surprisingly, our naive multi-scale Gauss-Seidel converges much faster

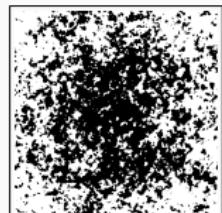
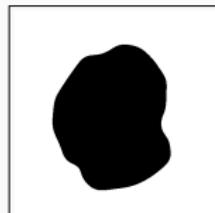
# COMPUTER PERFORMANCE ? WELL, I DESIGN ALGORITHMS!

- "Real" problems are all NP-hard, Log-APX, etc.
- Real workload = ~~NP-completeness proof~~ widgets, regularities and properties (difficult to formally state but that should be exploited)

Algorithms are evaluated on particular **workloads** that impact both their running time and the quality of the solutions

**Image Processing:** True horror stories, E. Meinhardt-Llopis, CANUM 2016

- *The proposed multigrid algorithm converges to the solution of the problem in  $O(N)$  using biharmonic functions*
- Surprisingly, our naive multi-scale Gauss-Seidel converges much faster





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".*

– Sandy Pentland (MIT)

Every month in CACM, there is an article about the ethical consequences of Machine Learning on:

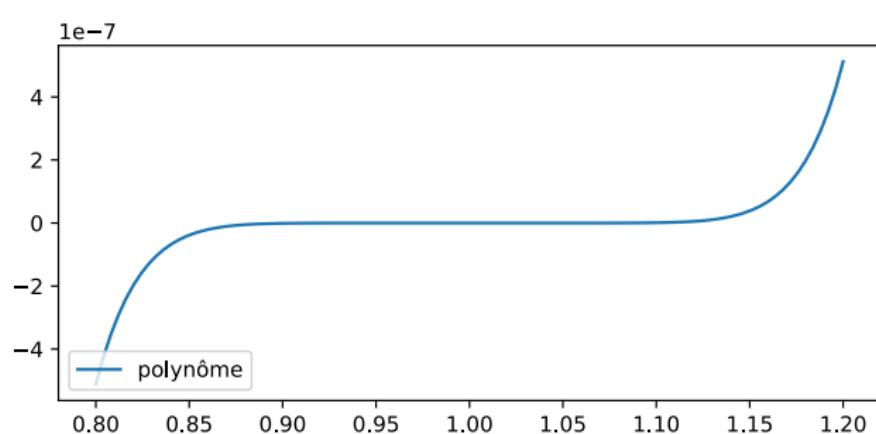
- Car driving, Autonomous guns, Law enforcement (risk assessment, predictive policing), ...

It's Not the Algorithm, It's the Data (CACM, Feb. 2017)

- Advertising, Loan attribution, Selection at University, Organ transplant

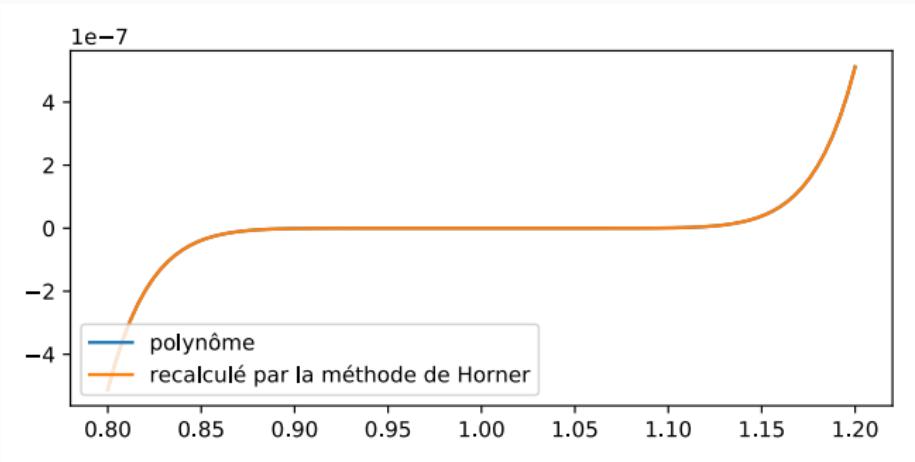
Increasing society concern about **fairness** and **transparency**

# ALL I CARE ABOUT IS THE ALGORITHM OUTPUT (FP)



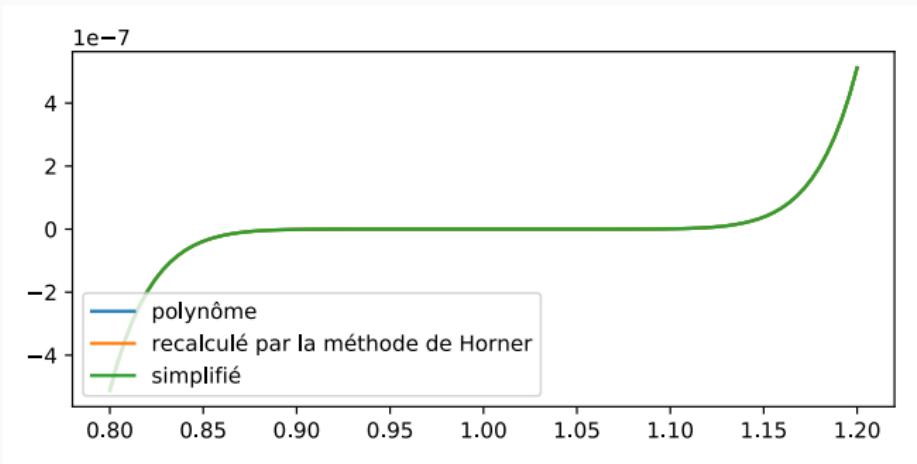
```
def polynome(x):
    return x**9 - 9.*x**8 + 36.*x**7 - 84.*x**6 + 126.*x**5 \
           - 126.*x**4 + 84.*x**3 - 36.*x**2 + 9.*x - 1.
```

# FLOATING-POINT ARITHMETIC



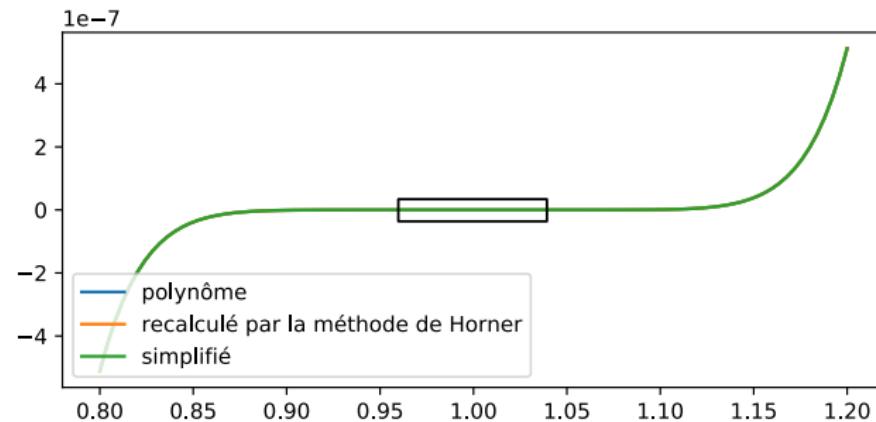
```
def horner(x):
    return x*(x*(x*(x*(x*(x*(x*(x - 9.) + 36.) - 84.) + 126.) \
    - 126.) + 84.) - 36.) + 9.) - 1.
```

# FLOATING-POINT ARITHMETIC

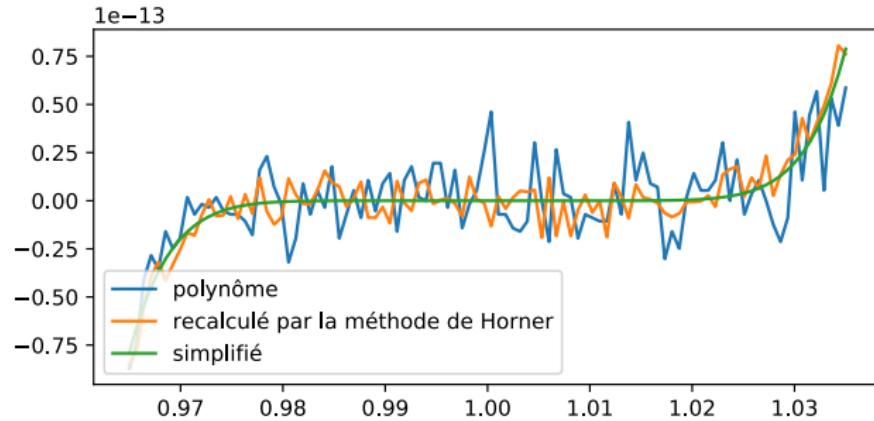


```
def simple(x):
    return (x-1.)**9
# Easy! ;)
```

# FLOATING-POINT ARITHMETIC



# FLOATING-POINT ARITHMETIC



# ROUNDING

---

- Every operation includes implicit rounding.
- $a+b$  is actually `round`( $a+b$ ).
- Unfortunately:

$$\text{round}(\text{round}(a+b)+c) \neq \text{round}(a+\text{round}(b+c)).$$

- Operation order therefore matters.

For a reproducible computation, operation order must be preserved!!!

# HOW TO EXPLAIN IT TO MY COMPILER?

---

To speed up computations, compilers may change operation order, and thus results.

Two options for computing reproducibly:

1. Insist on the preservation of operation order,
  - if the language permits it.
  - Example: Module 'ieee\_arithmetic' in Fortran 2003
2. Make compilation reproducible:
  - Record the precise compiler version
  - Record all compilation options

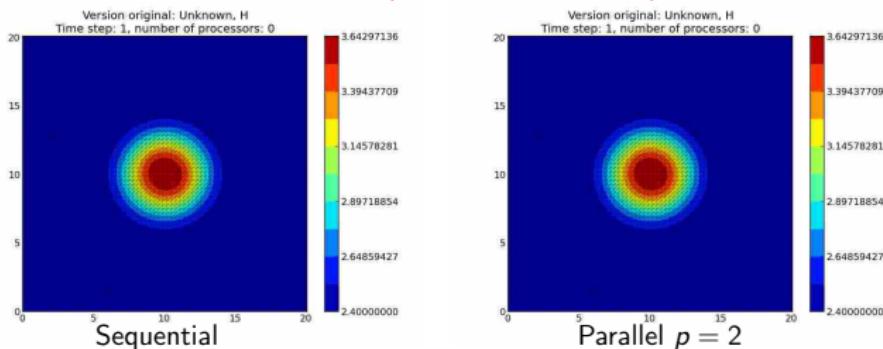
# DID I MENTION WE HAVE PARALLEL MACHINES NOWADAYS?

## Telemac2D: the simplest gouttedeo simulation

### The gouttedeo test case

- 2D-simulation of a water drop fall in a square bassin
- Unknown: water depth for a 0.2 sec time step
- Triangular mesh: 8978 elements and 4624 nodes

Expected numerical reproducibility (time step = 1, 2, ...)



13 / 64

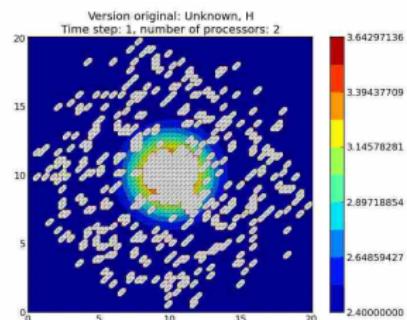
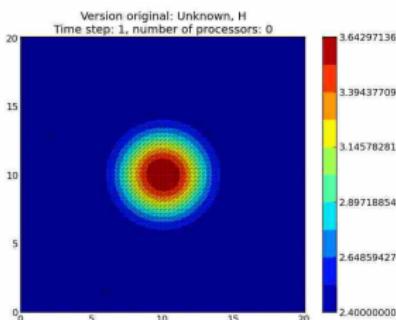
Courtesy of P. Langlois and R. Nheili

# DID I MENTION WE HAVE PARALLEL MACHINES NOWADAYS?

A white plot displays a non-reproducible value

Numerical reproducibility?

time step = 1

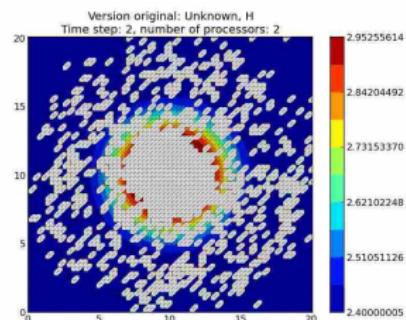
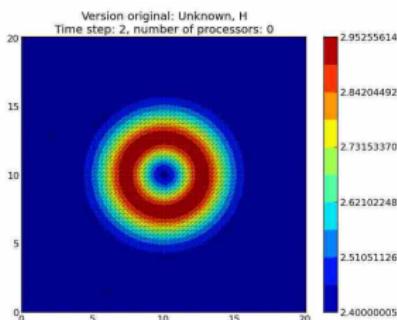


# DID I MENTION WE HAVE PARALLEL MACHINES NOWADAYS?

A white plot displays a non-reproducible value

Numerical reproducibility?

time step = 2

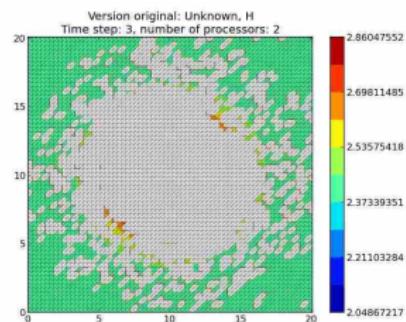
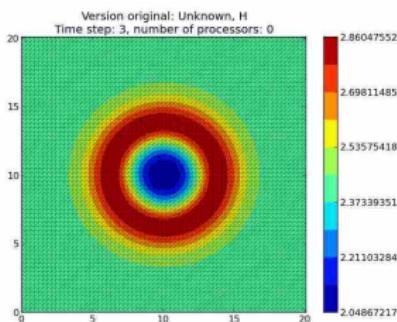


# DID I MENTION WE HAVE PARALLEL MACHINES NOWADAYS?

A white plot displays a non-reproducible value

Numerical reproducibility?

time step = 3

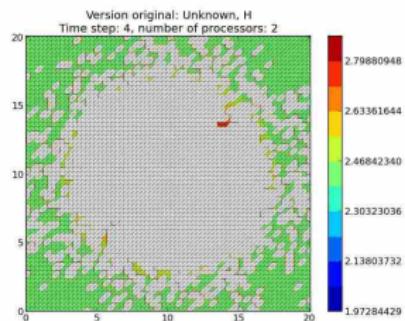
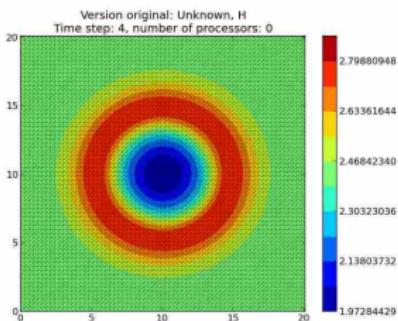


# DID I MENTION WE HAVE PARALLEL MACHINES NOWADAYS?

A white plot displays a non-reproducible value

Numerical reproducibility?

time step = 4

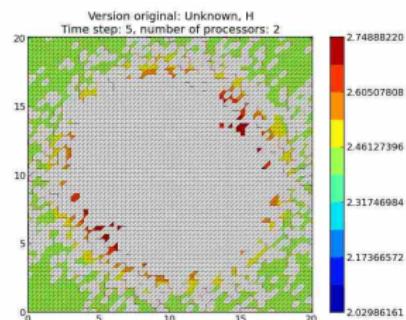
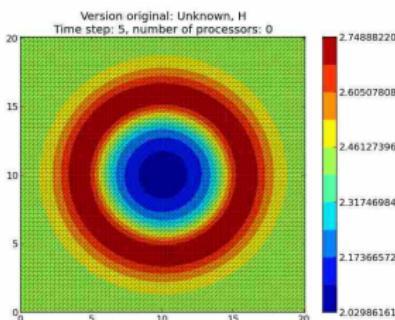


# DID I MENTION WE HAVE PARALLEL MACHINES NOWADAYS?

A white plot displays a non-reproducible value

Numerical reproducibility?

time step = 5

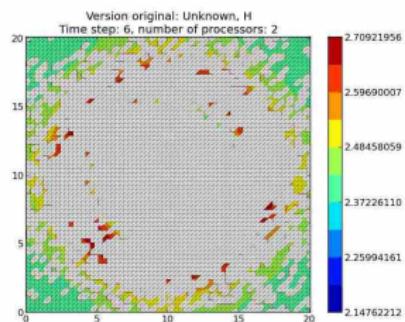
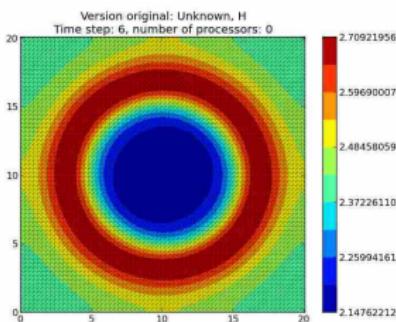


# DID I MENTION WE HAVE PARALLEL MACHINES NOWADAYS?

A white plot displays a non-reproducible value

Numerical reproducibility?

time step = 6

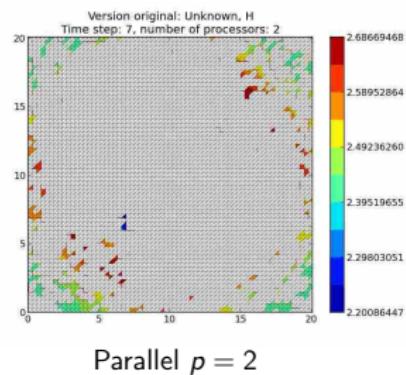
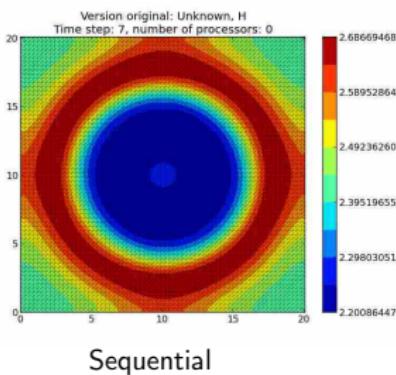


# DID I MENTION WE HAVE PARALLEL MACHINES NOWADAYS?

A white plot displays a non-reproducible value

Numerical reproducibility?

time step = 7

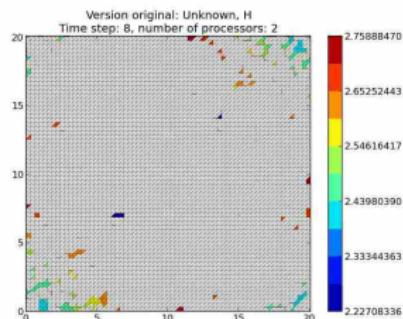
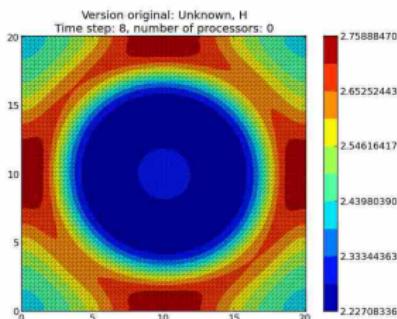


# DID I MENTION WE HAVE PARALLEL MACHINES NOWADAYS?

A white plot displays a non-reproducible value

Numerical reproducibility?

time step = 8

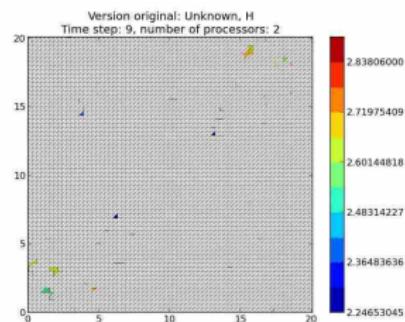
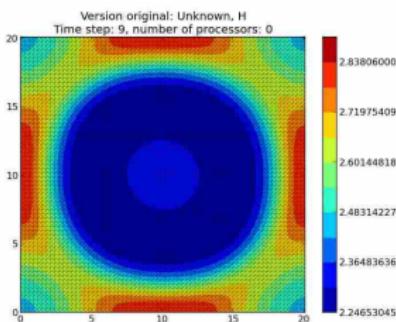


# DID I MENTION WE HAVE PARALLEL MACHINES NOWADAYS?

A white plot displays a non-reproducible value

Numerical reproducibility?

time step = 9

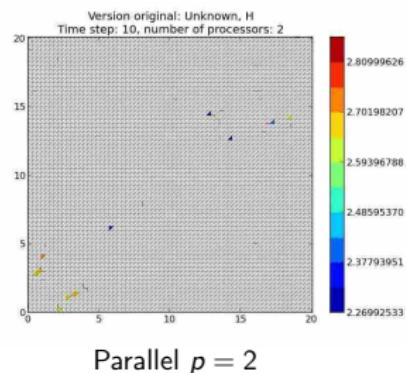
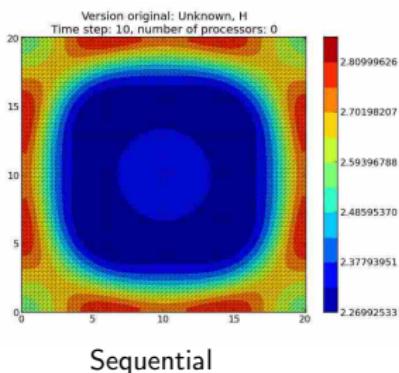


# DID I MENTION WE HAVE PARALLEL MACHINES NOWADAYS?

A white plot displays a non-reproducible value

Numerical reproducibility?

time step = 10

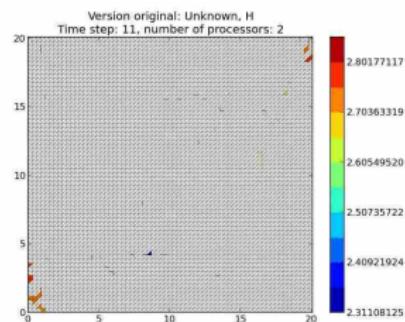
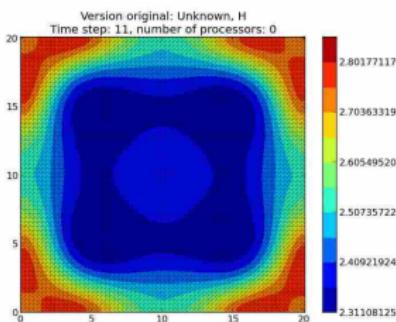


# DID I MENTION WE HAVE PARALLEL MACHINES NOWADAYS?

A white plot displays a non-reproducible value

Numerical reproducibility?

time step = 11

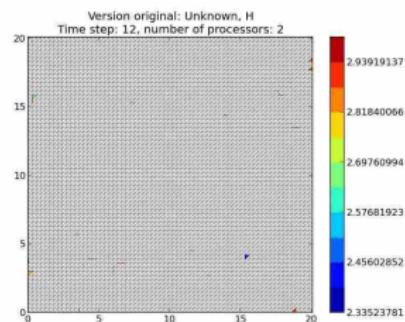
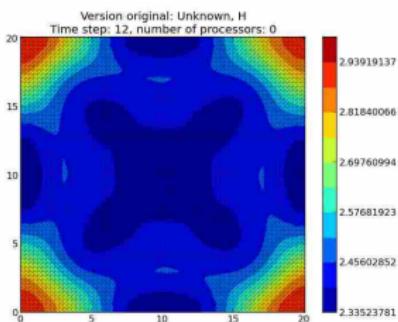


# DID I MENTION WE HAVE PARALLEL MACHINES NOWADAYS?

A white plot displays a non-reproducible value

Numerical reproducibility?

time step = 12

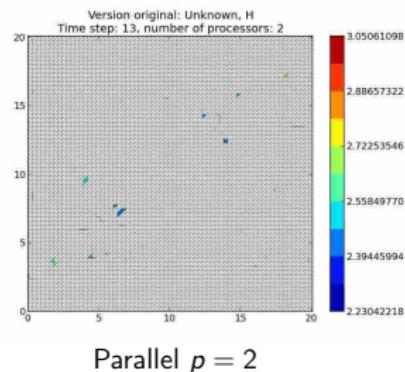
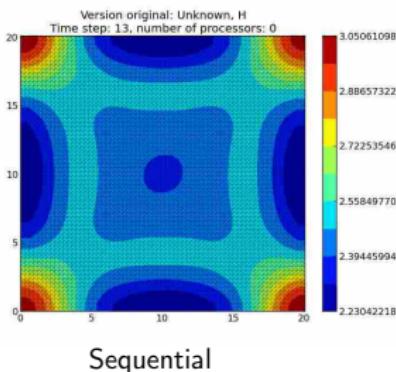


# DID I MENTION WE HAVE PARALLEL MACHINES NOWADAYS?

A white plot displays a non-reproducible value

Numerical reproducibility?

time step = 13

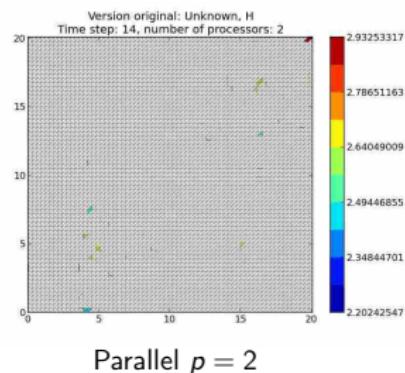
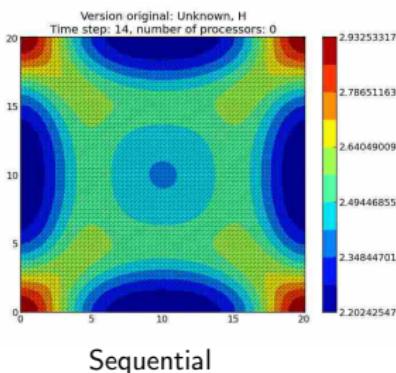


# DID I MENTION WE HAVE PARALLEL MACHINES NOWADAYS?

A white plot displays a non-reproducible value

Numerical reproducibility?

time step = 14

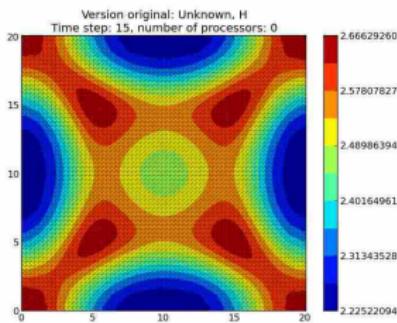


# DID I MENTION WE HAVE PARALLEL MACHINES NOWADAYS?

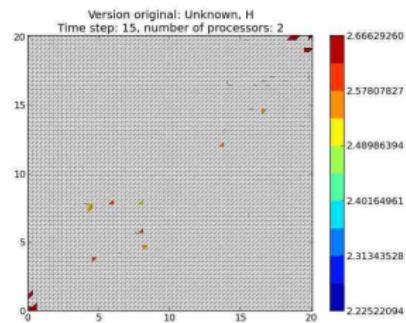
A white plot displays a non-reproducible value

NO numerical reproducibility!

time step = 15



Sequential



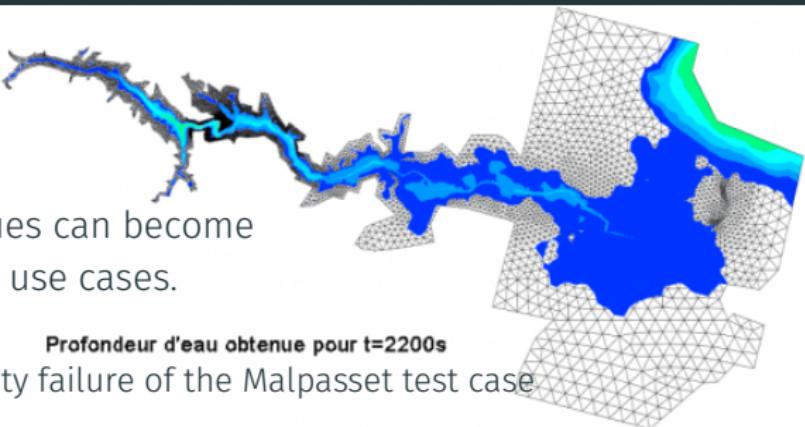
Parallel  $p = 2$

# DID I MENTION WE HAVE PARALLEL MACHINES NOWADAYS?

These numerical issues can become quite harmful in real use cases.

Profondeur d'eau obtenue pour t=2200s

TABLE 1.1: Reproducibility failure of the Malpasset test case



	The sequential run	a 64 procs run	a 128 procs run
depth H	0.3500122E-01	0.2748817E-01	0.1327634E-01
velocity U	0.4029747E-02	0.4935279E-02	0.4512116E-02
velocity V	0.7570773E-02	0.3422730E-02	0.7545233E-02

**Numerical reproducibility:** Approximations in the model, in the algorithm, in its implementation, in its execution.

The whole chain needs to be revisited.

Courtesy of P. Langlois and R. Nheili

WHAT WILL IT TAKE ?

---

# CHANGING RESEARCH PRACTICES

## Soft. Engineering, Statistics, and Reproducible Research in the curricula

**Manifesto:** "*I solemnly pledge*" (**WSSSPE, Lorena Barba, FAIR**)

1. I will teach my graduate students about reproducibility
2. All our research code (and writing) is under version control
3. We will always carry out verification and validation
4. We will share data, plotting script & figure under CC-BY
5. We will upload the preprint to arXiv at the time of submission of a paper
6. We will release code at the time of submission of a paper
7. We will add a "Reproducibility" declaration at the end of each paper
8. I will keep an up-to-date web presence



Learn and Teach using online resources like

- **Software Carpentry, The Turing Way, ...**

# CHANGING PUBLISHING PRACTICES

## Artifact evaluation and ACM badges



## Major conferences

- Supercomputing: Artifact Description (AD) mandatory, Artifact Evaluation (AE) still optional, Double blind vs. RR
- NeurIPS, ICLR: open reviews, reproducibility challenge



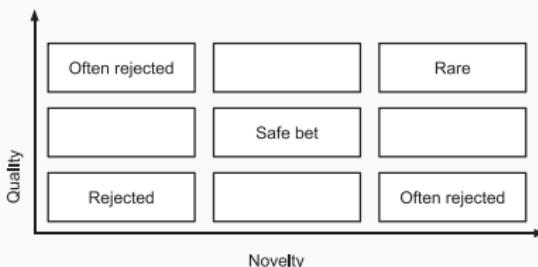
Joelle Pineau @ NeurIPS'18

- ACM SIGMOD 2015-2019, Most Reproducible Paper Award...

Mentalities are evolving people care, make stuff available, errors are found and fixed

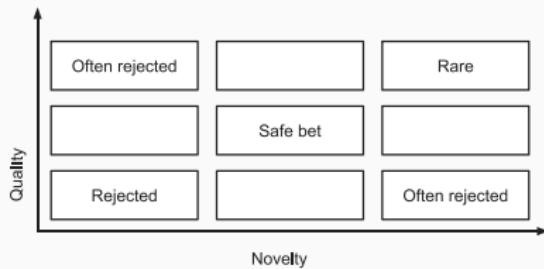
# CHANGING ACADEMIC PRACTICES (PUBLISH OR PERISH)

- Goodhart's Law: Are Academic Metrics Being Gamed?, M. Fire 2019
  - AI: over 1,000 ranked journals ( $\times 10$  in 15 years)
  - Shorter papers with increasing self references
  - More and more papers without any citation
  - Sharp increase in the number of new authors publishing at a much faster rate given their career age
- The Truth, The Whole Truth, and Nothing But the Truth: A Pragmatic, Guide to Assessing Empirical Evaluations, TOPLAS 2016



# CHANGING ACADEMIC PRACTICES (PUBLISH OR PERISH)

- Goodhart's Law: Are Academic Metrics Being Gamed?, M. Fire 2019
  - AI: over 1,000 ranked journals ( $\times 10$  in 15 years)
  - Shorter papers with increasing self references
  - More and more papers without any citation
  - Sharp increase in the number of new authors publishing at a much faster rate given their career age
- The Truth, The Whole Truth, and Nothing But the Truth: A Pragmatic, Guide to Assessing Empirical Evaluations, TOPLAS 2016



- Impact factor abandoned by Dutch university in hiring and promotion, decisions. Nature, June 2021. Faculty and staff members at Utrecht University will be evaluated by their commitment to open science

# WHAT ABOUT 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**

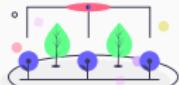


THAT'S ALL FOLKS!

---

# RESOURCES AND ACKNOWLEDGMENTS

Vers une recherche  
reproductible  
Faire évoluer ses pratiques



Loïc Desquillet, Sabrina Granger, Boris Hejblum,  
Arnaud Legrand, Pascal Pernot, Nicolas Rougier

A non-technical introduction to reproducibility issues (in French)

- Loïc Desquillet, Sabrina Granger, Boris Hejblum, Pascal Pernot, Nicolas Rougier

# RESOURCES AND ACKNOWLEDGMENTS



A non-technical introduction to reproducibility issues (in French)

- Loïc Desquillet, Sabrina Granger, Boris Hejblum, Pascal Pernot, Nicolas Rougier

MOOC Reproducible Research: Methodological principles for a transparent science, Learning Lab Inria

- Konrad Hinsen, Christophe Pouzat
- 3rd Edition: March 2020 – March 2023  
(15,000+)



Stay tuned for the MOOC "Advanced RR" planned for 2021 2022 2023

- Software environment control
- Scientific workflow
- Managing data