

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

Arnaud Legrand



MVA, ENS Paris-Saclay

March 2023



WHAT IS SCIENCE?

Question 1: In less than 5 lines give a definition of "Science"

WHAT IS SCIENCE?

Question 1: In less than 5 lines give a definition of "Science"

Dictionary of science and technology

1. the study of the physical and natural world and phenomena, especially by using systematic observation and experiment
2. a particular area of study or knowledge of the physical world
3. a systematically organized body of knowledge about a particular subject

New Oxford Dictionary the intellectual and practical activity encompassing the systematic study of the structure and behavior of the physical and natural world through observation and experiment : the world of science and technology.

1. a particular area of this : veterinary science | the agricultural sciences.
2. a systematically organized body of knowledge on a particular subject : the science of criminology.
3. archaic knowledge of any kind.

WHAT IS SCIENCE?

Question 1: In less than 5 lines give a definition of "Science"

Dictionary of science and technology

1. the study of the physical and natural world and phenomena, especially by using systematic observation and experiment
2. a particular area of study or knowledge of the physical world
3. a systematically organized body of knowledge about a particular subject

New Oxford Dictionary the intellectual and practical activity encompassing the systematic study of the structure and behavior of the physical and natural world through observation and experiment : the world of science and technology.

1. a particular area of this : veterinary science | the agricultural sciences.
2. a systematically organized body of knowledge on a particular subject : the science of criminology.
3. archaic knowledge of any kind.

Building Reliable Knowledge

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.

A screenshot of the Science journal website. The top navigation bar includes links for AAAS-ORG, FEEDBACK, HELP, LIBRARIANS, All Science Journals, and a search bar. Below this is a red banner for AAAS NEWS, SCIENCE JOURNALS, CAREERS, MULTIMEDIA, and COLLECTIONS. The main content area features an article titled "Science has lost its way, at a big cost to humanity" by Marcia McNutt. The article discusses the lack of reproducibility in science. At the bottom left, there are "Article Views" and "Full Text" links, along with a sidebar for "Article Tools" and "Related Content".

A screenshot of the Nature journal website. The top navigation bar includes links for Home, News & Comment, Research, Careers & Jobs, Current Issue, and Archive. Below this is a search bar and a "nature" logo. The main content area features an announcement titled "Announcement: Reducing our irreproducibility" dated April 24, 2013. The announcement discusses the journal's commitment to improving reproducibility.

A screenshot of the Nature journal website. The top navigation bar includes links for Menu, Advanced search, and a search bar. Below this is a large graphic with the text "HOW SCIENCE GOES WRONG." in colorful, stylized letters. The main content area features an announcement titled "Announcement: Reducing our irreproducibility" dated April 24, 2013. The announcement discusses the journal's commitment to improving reproducibility.

A screenshot of The Scientist magazine website. The top navigation bar includes links for Home, News, Features, Methods, Techniques, Columns, and Supplements. Below this is a search bar and a "the scientist" logo. The main content area features an article titled "NIH Tackles Irreproducibility" by Jef Akst. The article discusses the National Institutes of Health's efforts to improve reproducibility in scientific research.

Courtesy V. Stodden, SC, 2015

A screenshot of the Nature journal website. The top navigation bar includes links for Menu, Advanced search, and a search bar. Below this is a large graphic with the text "HOW SCIENCE GOES WRONG." in colorful, stylized letters. The main content area features an announcement titled "Announcement: Reducing our irreproducibility" dated April 24, 2013. The announcement discusses the journal's commitment to improving reproducibility.

2/100

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.

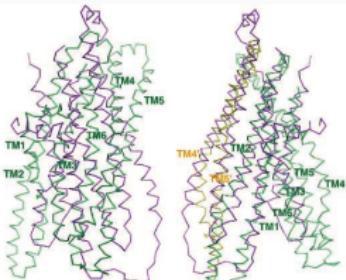
Bad science is deleterious

- 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!*

How COMPUTERS BROKE SCIENCE



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

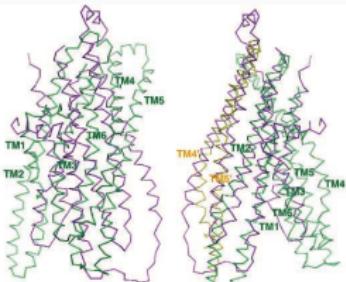
He specialized in structures of **multidrug resistant transporter proteins in bacteria**: MsbA de Escherichia Choli (Science, 2001), Vibrio cholera (Mol. Biology, 2003), Salmonella typhimurium (Science, 2005)

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

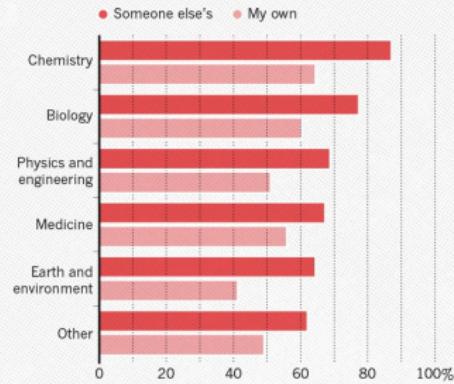
There is **worse!**

- 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

A REPRODUCIBILITY CRISIS?

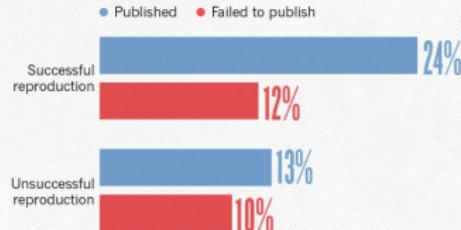
HAVE YOU FAILED TO REPRODUCE AN EXPERIMENT?

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.



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



DIFFERENT REPRODUCIBILITY CONCERN IN MODERN SCIENCE

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

Genomics software engineering, computational reproducibility, provenance

Computational fluid dynamics numerical issues

Artificial Intelligence most of the above

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

Authors



Data

DIFFERENT REPRODUCIBILITY CONCERN IN MODERN SCIENCE

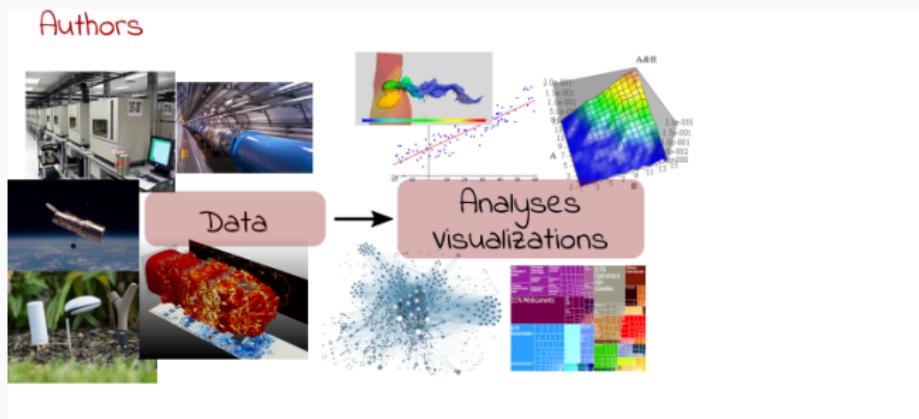
Social Sciences, Oncology, ... methodology, statistics, pre-registration

Genomics software engineering, computational reproducibility, provenance

Computational fluid dynamics numerical issues

Artificial Intelligence most of the above

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



DIFFERENT REPRODUCIBILITY CONCERN IN MODERN SCIENCE

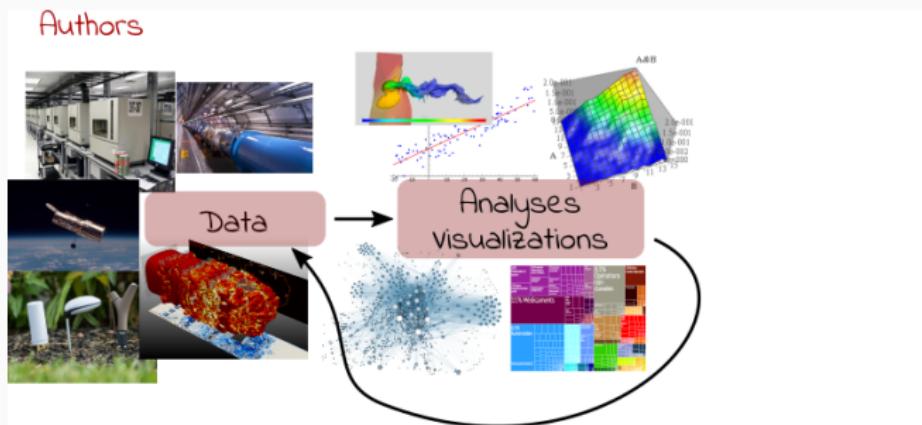
Social Sciences, Oncology, ... methodology, statistics, pre-registration

Genomics software engineering, computational reproducibility, provenance

Computational fluid dynamics numerical issues

Artificial Intelligence most of the above

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



DIFFERENT REPRODUCIBILITY CONCERN IN MODERN SCIENCE

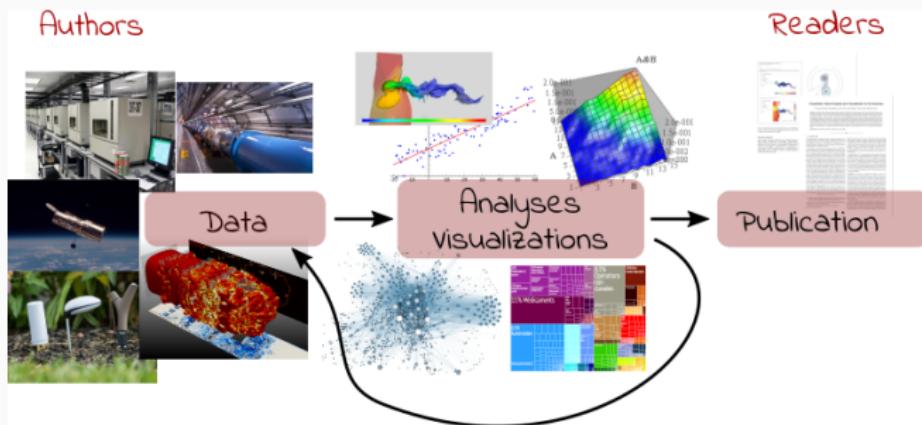
Social Sciences, Oncology, ... methodology, statistics, pre-registration

Genomics software engineering, computational reproducibility, provenance

Computational fluid dynamics numerical issues

Artificial Intelligence most of the above

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



DIFFERENT REPRODUCIBILITY CONCERN IN MODERN SCIENCE

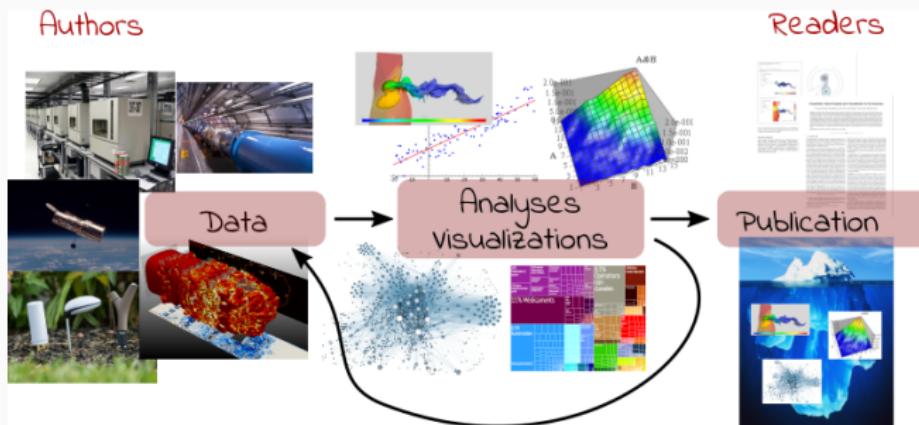
Social Sciences, Oncology, ... methodology, statistics, pre-registration

Genomics software engineering, computational reproducibility, provenance

Computational fluid dynamics numerical issues

Artificial Intelligence most of the above

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



DIFFERENT REPRODUCIBILITY CONCERN IN MODERN SCIENCE

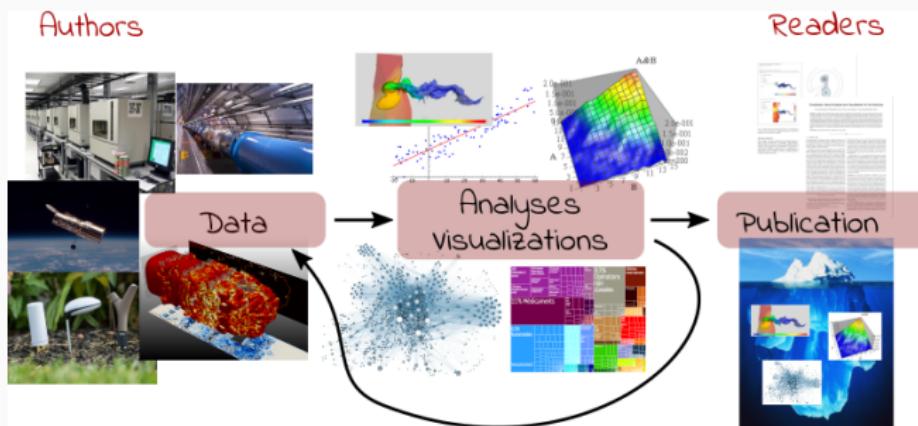
Social Sciences, Oncology, ... methodology, statistics, pre-registration

Genomics software engineering, computational reproducibility, provenance

Computational fluid dynamics numerical issues

Artificial Intelligence most of the above

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



Reproducible Research = Bridging the Gap by working Transparently 8/100

REPRODUCIBLE RESEARCH PRACTICES

REPRODUCIBILITY (GLOSSARY MAY VARY)

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

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

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

¹Up-to measurement variability and precision

²Independently from (random) sampling variability (fight bias)

³Bitwise

"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 level 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).

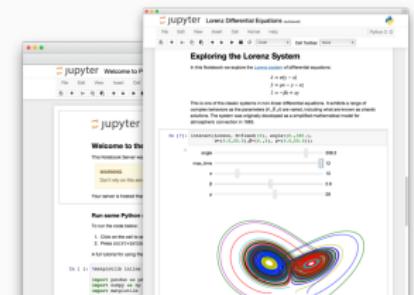
- make incremental improvements in electronic-document software
- seek partners for broadening standards (and making incremental improvements).

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.

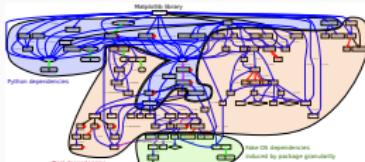
Now that we have begun using CD-ROM publication, we can go much further. Every figure caption contains a pushbutton that jumps to the appropriate science directory (folder) and initiates a figure rebuild command and then displays the figure, possibly as a movie or interactive program. We normally display 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^AT_EX word processing systems. Naturally some software includes licensed software, but with the exception

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 π vaut approximativement

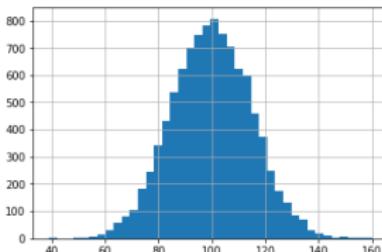
3.141592653589793

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 π (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 the following details:

- Title:** # Un document computationnel
- In [1]:** A code cell containing:

```
from math import *
print(pi)
3.141592653589793
```
- Out [1]:** The output 3.141592653589793, followed by a note: "Mais calculé avec la [méthode des aiguilles de Buffon](#) (https://fr.wikipedia.org/wiki/Aiguille_de_Buffon), on obtientrait comme approximation :".
- In [2]:** A code cell containing:

```
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=np.pi/2)
2/(sum((x+np.sin(theta))>1))/N
```
- Out [2]:** The output 3.14371986944998765, followed by a note: "On peut inclure des formules mathématiques comme $\sqrt{2/\pi} \exp(-x^2/2)$ et des dessins qui n'ont rien à voir avec π (si ce n'est une constante de normalisation...)."
- In [3]:** A code cell containing:

```
%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()
```
- Out [3]:** A histogram showing a bell-shaped distribution centered around 100.

Document final

Un document computationnel

Mon ordinateur m'indique que π vaut approximativement

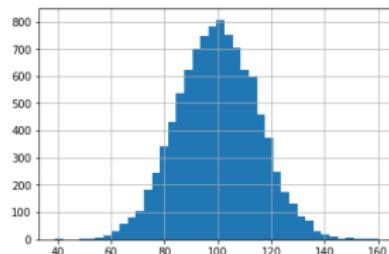
3.141592653589793

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=np.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 π (si ce n'est une constante de normalisation...).



TOOL 1: COMPUTATIONAL NOTEBOOKS/LITTERATE PROGRAMMING

Document initial dans son environnement

```
# 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 aussi comme approximation : 3.141592653589793

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

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()
```

Document final

Un document computationnel

Mon ordinateur m'indique que π vaut approximativement

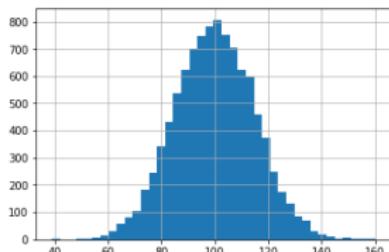
3.141592653589793

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=np.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 π (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'. The notebook contains three code cells:

- In [1]:** Prints the value of pi calculated using theBuffon's needle method. It includes a note from Wikipedia about the method.
- In [2]:** Calculates pi again using a different approach involving uniform random numbers and sine values.
- In [3]:** Plots a histogram of 100,000 random numbers between 0 and 160, showing a bell-shaped distribution centered around 100.

Document final

Un document computationnel

Mon ordinateur m'indique que π vaut approximativement

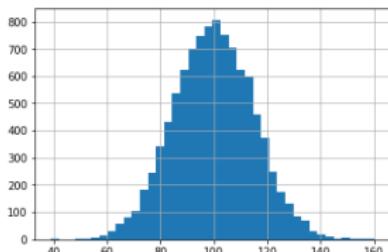
3.141592653589793

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 π (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), 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]: 3,1437198694098765
```

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 π (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()
```

Document final

Un document computationnel

Mon ordinateur m'indique que π vaut approximativement

3.141592653589793

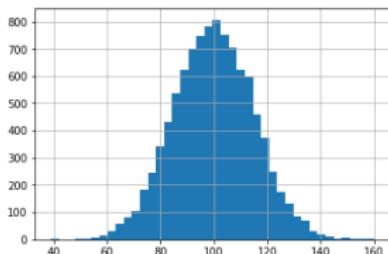
Mais calculé avec la méthode des aiguilles de Buffon, on obtiendrait comme approximation :

```
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 π (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 notebook has three cells:

- In [1]:** Prints the value of pi: `# Un document computationnel`
3.141592653589793
- In [2]:** Prints the value of pi again and calculates the ratio of points inside a unit square to the total number of points to estimate pi. Includes a note about theBuffon's needle method.
3.1437198694098765
- In [3]:** Plots a histogram of 100,000 random points between 0 and 100. The plot shows a bell-shaped curve centered around 100.

Document final

Un document computationnel

Mon ordinateur m'indique que π vaut approximativement

3.141592653589793

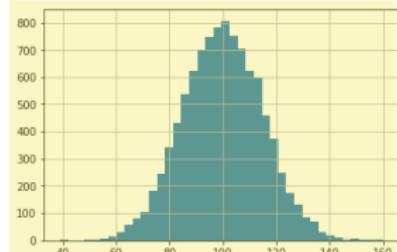
Mais calculé avec la **méthode des aiguilles de Buffon**, on obtiendrait comme approximation :

```
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 π (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]:** Prints π as 3.141592653589793.
- In [2]:** Generates random points and calculates the ratio of points inside a unit circle to total points to approximate π .
- In [3]:** Plots a histogram of random numbers between 40 and 160, showing a bell-shaped distribution centered around 100.

Document final

Un document computationnel

Mon ordinateur m'indique que π vaut approximativement

3.141592653589793

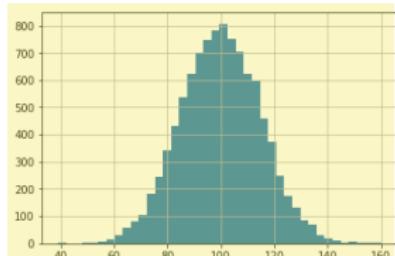
Mais calculé avec la **méthode des aiguilles de Buffon**, on obtiendrait comme approximation :

```
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)
```

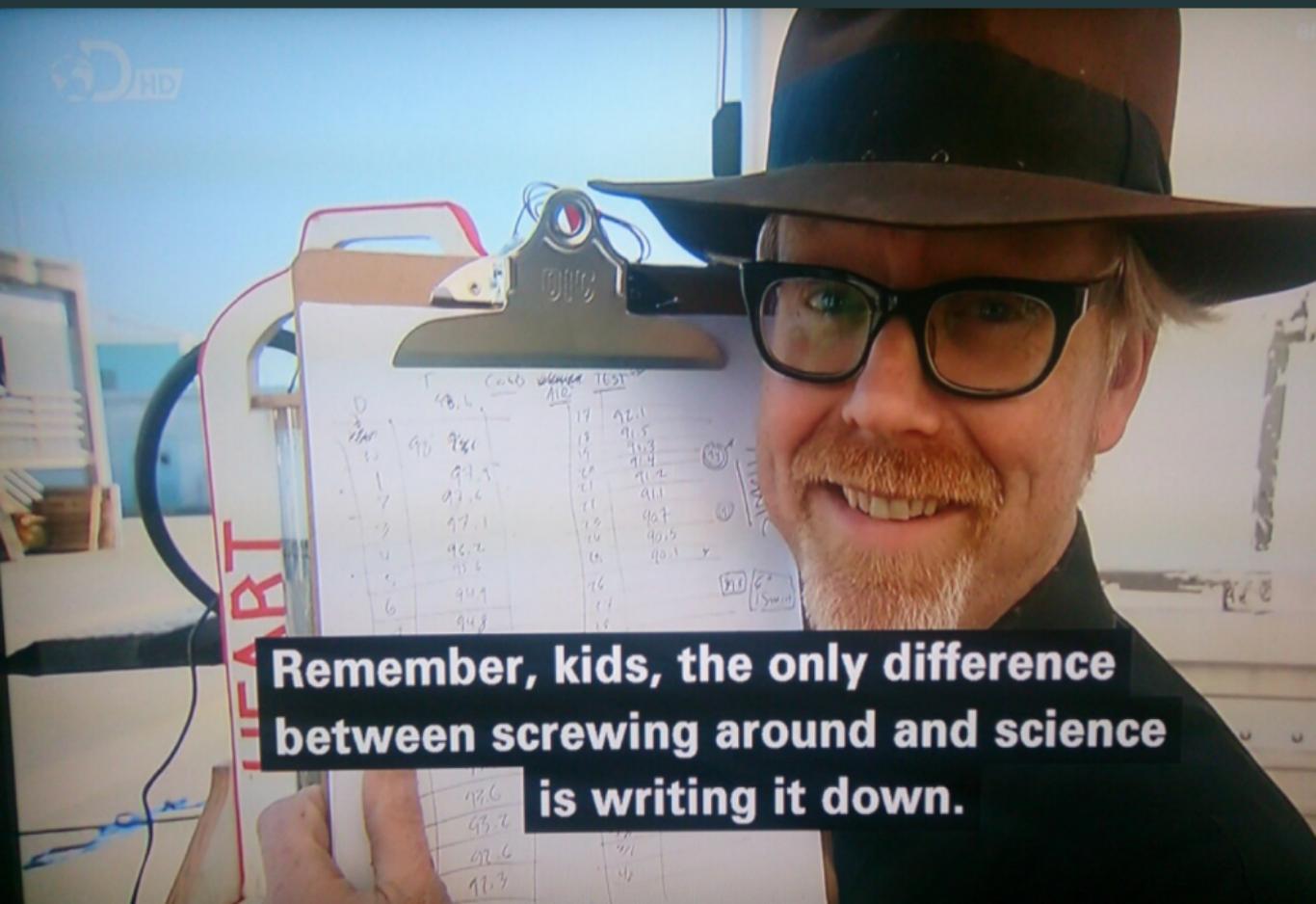
3.1437198694098765

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 π (si ce n'est une constante de normalisation... ☺).



TOOL 1 BIS: LABORATORY NOTEBOOKS, COMPUTATIONAL DOCUMENTS



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

TOOL 1 TER: 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 π vaut "approximativement"

In [1]:

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

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

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)  
if (x**2 + np.sin(theta)**2) < 1/n
```

Out[2]: 3.143719869495785

On peut inclure des formules mathématiques comme $\frac{1}{\sqrt{\pi}}$ via `mathop`.

In [3]:

```
%matplotlib inline  
import matplotlib.pyplot as plt  
  
n, sigma = 100, 33  
x = np.random.normal(0, sigma, n)  
  
plt.hist(x, 40)  
plt.title("Histogramme")  
plt.show()
```



TOOL 1 TER: 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 l'adresse www.inserm.fr. Nous les disposons sous forme d'un fichier CSV (Comma Separated Values) qui contient des données brutes et nécessite toujours une analyse préliminaire pour prendre de l'avis sur elles. Le présent type de fichier CSV est un tableau de données.

Voici quelques lignes extraites du fichier d'origine :

Nom du tableau : Liste des cas de grippe	
cas	Nombre total de cas (10 400)
cas_id	Identifiant unique de chaque cas
cas_n	Numéro de l'individu dans la liste de cas
cas_sexe	Sexe de l'individu dans la liste de cas
cas_âge	Âge de l'individu dans la liste de cas
cas_covid	Nombre de personnes atteintes de COVID-19 dans la liste de cas
cas_risque	Nombre de tous individus en risque de contamination (en cas de grippe saisonnière)
cas_risque_grippe	Nombre de personnes atteintes de grippe dans la liste de cas
cas_risque_grippe_grippe	Nombre de personnes atteintes de grippe et de COVID-19 dans la liste de cas
cas_risque_grippe_grippe_grippe	Nombre de personnes atteintes de grippe, de COVID-19 et de grippe saisonnière dans la liste de cas
cas_risque_grippe_grippe_grippe_grippe	Nombre de personnes atteintes de grippe, de COVID-19 et de grippe saisonnière et de grippe grippe dans la liste de cas

In [2]: `# Import des librairies nécessaires pour l'analyse de données.
import pandas as pd
import numpy as np`

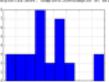
Notre tableau de données est maintenant chargé dans la variable `df`.

cas	cas_id	cas_n	cas_sexe	cas_âge	cas_covid	cas_risque	cas_risque_grippe	cas_risque_grippe_grippe	cas_risque_grippe_grippe_grippe	cas_risque_grippe_grippe_grippe_grippe
1	20000000	1	Homme	20	0	20	20	0	0	0
2	20000001	2	Homme	20	0	20	20	0	0	0
3	20000002	3	Homme	20	0	20	20	0	0	0
4	20000003	4	Homme	20	0	20	20	0	0	0
5	20000004	5	Homme	20	0	20	20	0	0	0
6	20000005	6	Homme	20	0	20	20	0	0	0
7	20000006	7	Homme	20	0	20	20	0	0	0
8	20000007	8	Homme	20	0	20	20	0	0	0
9	20000008	9	Homme	20	0	20	20	0	0	0
10	20000009	10	Homme	20	0	20	20	0	0	0
11	20000010	11	Homme	20	0	20	20	0	0	0
12	20000011	12	Homme	20	0	20	20	0	0	0
13	20000012	13	Homme	20	0	20	20	0	0	0
14	20000013	14	Homme	20	0	20	20	0	0	0
15	20000014	15	Homme	20	0	20	20	0	0	0
16	20000015	16	Homme	20	0	20	20	0	0	0
17	20000016	17	Homme	20	0	20	20	0	0	0
18	20000017	18	Homme	20	0	20	20	0	0	0
19	20000018	19	Homme	20	0	20	20	0	0	0
20	20000019	20	Homme	20	0	20	20	0	0	0
21	20000020	21	Homme	20	0	20	20	0	0	0
22	20000021	22	Homme	20	0	20	20	0	0	0
23	20000022	23	Homme	20	0	20	20	0	0	0
24	20000023	24	Homme	20	0	20	20	0	0	0
25	20000024	25	Homme	20	0	20	20	0	0	0
26	20000025	26	Homme	20	0	20	20	0	0	0
27	20000026	27	Homme	20	0	20	20	0	0	0
28	20000027	28	Homme	20	0	20	20	0	0	0
29	20000028	29	Homme	20	0	20	20	0	0	0
30	20000029	30	Homme	20	0	20	20	0	0	0
31	20000030	31	Homme	20	0	20	20	0	0	0
32	20000031	32	Homme	20	0	20	20	0	0	0
33	20000032	33	Homme	20	0	20	20	0	0	0
34	20000033	34	Homme	20	0	20	20	0	0	0
35	20000034	35	Homme	20	0	20	20	0	0	0
36	20000035	36	Homme	20	0	20	20	0	0	0
37	20000036	37	Homme	20	0	20	20	0	0	0
38	20000037	38	Homme	20	0	20	20	0	0	0
39	20000038	39	Homme	20	0	20	20	0	0	0
40	20000039	40	Homme	20	0	20	20	0	0	0
41	20000040	41	Homme	20	0	20	20	0	0	0
42	20000041	42	Homme	20	0	20	20	0	0	0
43	20000042	43	Homme	20	0	20	20	0	0	0
44	20000043	44	Homme	20	0	20	20	0	0	0
45	20000044	45	Homme	20	0	20	20	0	0	0
46	20000045	46	Homme	20	0	20	20	0	0	0
47	20000046	47	Homme	20	0	20	20	0	0	0
48	20000047	48	Homme	20	0	20	20	0	0	0
49	20000048	49	Homme	20	0	20	20	0	0	0
50	20000049	50	Homme	20	0	20	20	0	0	0
51	20000050	51	Homme	20	0	20	20	0	0	0
52	20000051	52	Homme	20	0	20	20	0	0	0
53	20000052	53	Homme	20	0	20	20	0	0	0
54	20000053	54	Homme	20	0	20	20	0	0	0
55	20000054	55	Homme	20	0	20	20	0	0	0
56	20000055	56	Homme	20	0	20	20	0	0	0
57	20000056	57	Homme	20	0	20	20	0	0	0
58	20000057	58	Homme	20	0	20	20	0	0	0
59	20000058	59	Homme	20	0	20	20	0	0	0
60	20000059	60	Homme	20	0	20	20	0	0	0
61	20000060	61	Homme	20	0	20	20	0	0	0
62	20000061	62	Homme	20	0	20	20	0	0	0
63	20000062	63	Homme	20	0	20	20	0	0	0
64	20000063	64	Homme	20	0	20	20	0	0	0
65	20000064	65	Homme	20	0	20	20	0	0	0
66	20000065	66	Homme	20	0	20	20	0	0	0
67	20000066	67	Homme	20	0	20	20	0	0	0
68	20000067	68	Homme	20	0	20	20	0	0	0
69	20000068	69	Homme	20	0	20	20	0	0	0
70	20000069	70	Homme	20	0	20	20	0	0	0
71	20000070	71	Homme	20	0	20	20	0	0	0
72	20000071	72	Homme	20	0	20	20	0	0	0
73	20000072	73	Homme	20	0	20	20	0	0	0
74	20000073	74	Homme	20	0	20	20	0	0	0
75	20000074	75	Homme	20	0	20	20	0	0	0
76	20000075	76	Homme	20	0	20	20	0	0	0
77	20000076	77	Homme	20	0	20	20	0	0	0
78	20000077	78	Homme	20	0	20	20	0	0	0
79	20000078	79	Homme	20	0	20	20	0	0	0
80	20000079	80	Homme	20	0	20	20	0	0	0
81	20000080	81	Homme	20	0	20	20	0	0	0
82	20000081	82	Homme	20	0	20	20	0	0	0
83	20000082	83	Homme	20	0	20	20	0	0	0
84	20000083	84	Homme	20	0	20	20	0	0	0
85	20000084	85	Homme	20	0	20	20	0	0	0
86	20000085	86	Homme	20	0	20	20	0	0	0
87	20000086	87	Homme	20	0	20	20	0	0	0
88	20000087	88	Homme	20	0	20	20	0	0	0
89	20000088	89	Homme	20	0	20	20	0	0	0
90	20000089	90	Homme	20	0	20	20	0	0	0
91	20000090	91	Homme	20	0	20	20	0	0	0
92	20000091	92	Homme	20	0	20	20	0	0	0
93	20000092	93	Homme	20	0	20	20	0	0	0
94	20000093	94	Homme	20	0	20	20	0	0	0
95	20000094	95	Homme	20	0	20	20	0	0	0
96	20000095	96	Homme	20	0	20	20	0	0	0
97	20000096	97	Homme	20	0	20	20	0	0	0
98	20000097	98	Homme	20	0	20	20	0	0	0
99	20000098	99	Homme	20	0	20	20	0	0	0
100	20000099	100	Homme	20	0	20	20	0	0	0

On voit que les 100 derniers numéros de cas, qui touchent environ 20% de la population française, sont assez bons (0 à 20).

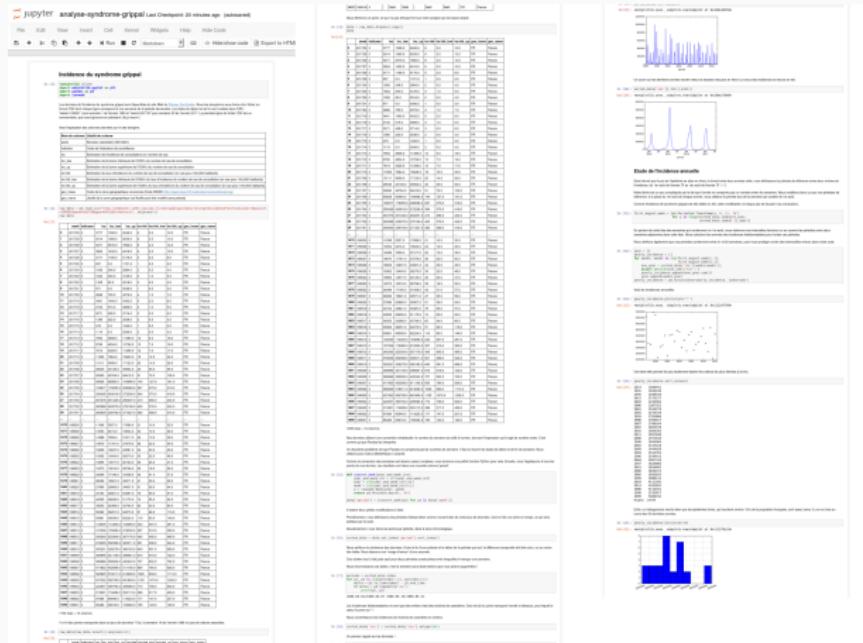
In [3]: `# Import des bibliothèques nécessaires`

In [4]: `# Analyse des données avec les bibliothèques pandas et matplotlib`



TOOL 1 TER: WORKFLOWS

Notebooks are no panacea and do not help developing clean code



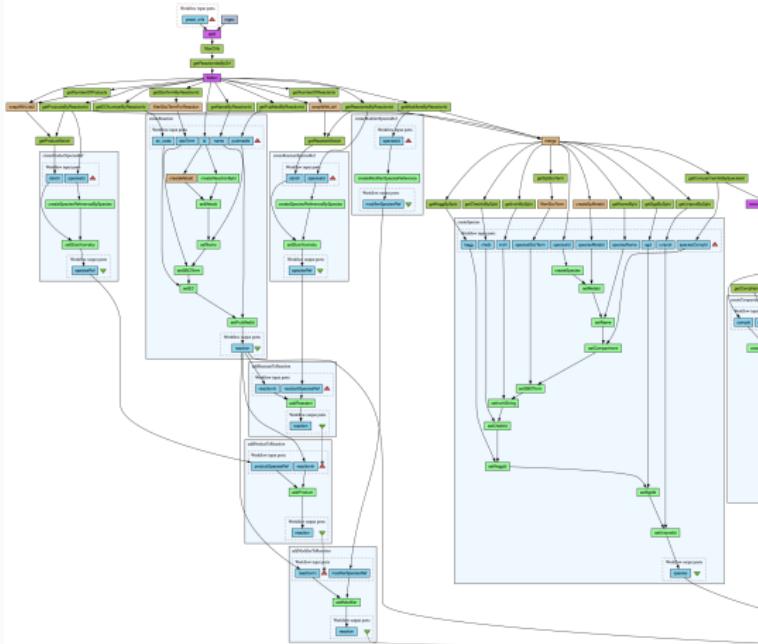
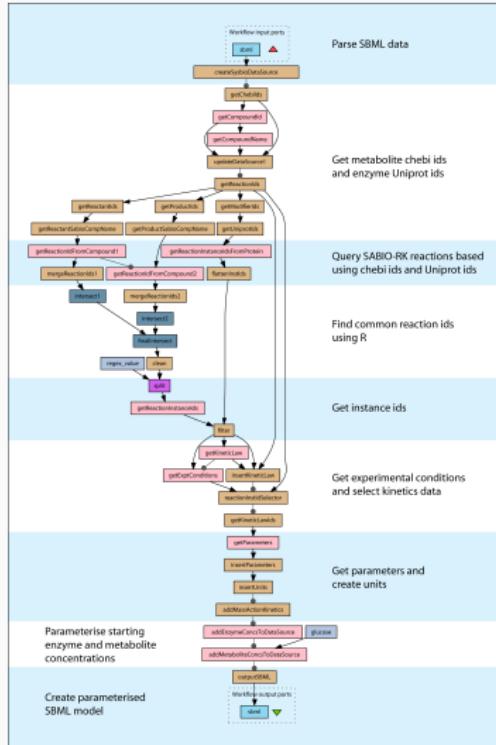
TOOL 1 TER: WORKFLOWS

Notebooks are no panacea and do not help developing clean code

The image displays a 4x3 grid of Jupyter Notebook screenshots, each illustrating a different step or aspect of a workflow. The notebooks cover topics such as estimating color names from web images, performing k-means clustering, visualizing color distributions, and analyzing machine learning model results.

- Row 1:**
 - Estimating Color Names by Web Image**: A notebook cell showing code to extract colors from a URL and calculate the most frequent colors.
 - Dimensionality reduction and k-means**: A notebook cell showing the use of PCA and k-means clustering on a dataset of 1000 images.
 - Dimensionality reduction and k-means results**: A heatmap titled "Color space" showing the distribution of colors in a 2D space.
- Row 2:**
 - Dimensionality reduction**: A notebook cell showing PCA and t-SNE dimensionality reduction on a dataset of 1000 images.
 - Dimensionality reduction and k-means**: A notebook cell showing the results of dimensionality reduction followed by k-means clustering.
 - Dimensionality reduction and k-means results**: A heatmap titled "Color space" showing the results of the dimensionality reduction and k-means clustering process.
- Row 3:**
 - PCA**: A notebook cell showing PCA on a dataset of 1000 images.
 - Dimensionality reduction and k-means**: A notebook cell showing the combined process of PCA and k-means clustering.
 - Analysis**: A notebook cell showing the results of the dimensionality reduction and k-means clustering process.
- Row 4:**
 - Preprocessing**: A notebook cell showing preprocessing steps like scaling and standardization.
 - Dimensionality reduction and k-means**: A notebook cell showing the dimensionality reduction and k-means clustering process.
 - Prediction**: A notebook cell showing the results of the dimensionality reduction and k-means clustering process.
- Row 5:**
 - Dimensionality reduction and k-means**: A notebook cell showing the dimensionality reduction and k-means clustering process.
 - Dimensionality reduction and k-means**: A notebook cell showing the dimensionality reduction and k-means clustering process.
 - Analysis**: A notebook cell showing the results of the dimensionality reduction and k-means clustering process.
- Row 6:**
 - Dimensionality reduction and k-means**: A notebook cell showing the dimensionality reduction and k-means clustering process.
 - Dimensionality reduction and k-means**: A notebook cell showing the dimensionality reduction and k-means clustering process.
 - Analysis**: A notebook cell showing the results of the dimensionality reduction and k-means clustering process.

TOOL 1 TER: WORKFLOWS



TOOL 1 TER: WORKFLOWS

Workflows:

- Clearer high-level view
- **Explicit** composition of codes and data movement
- 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: `make`, dask, drake, swift, `snakemake`, ...
- Hybrids: SOS-notebook, ...

GOOD PRACTICE #2

CONTROLLING SOFTWARE ENVIRONMENT

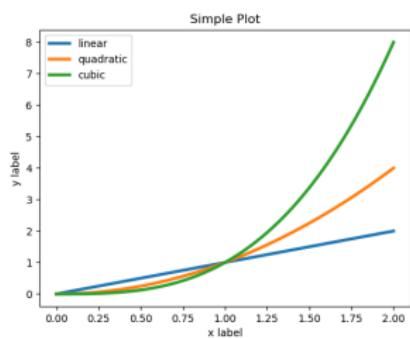
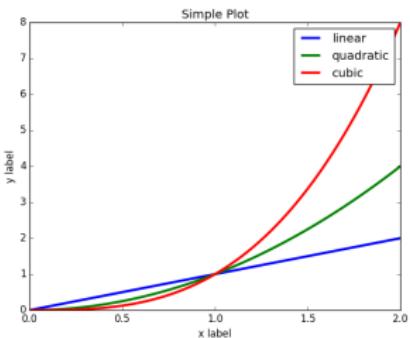
ARGH... DAMNED COMPUTERS

- 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 😊
 1. No one will ever run/use your code if it isn't easy to install
 2. No one will ever manage to run your code if you don't document how to run it
 3. Others (even you) are unlikely to get the same results unless you control and share your software environment

SOFTWARE DEPENDENCIES: HORROR STORIES

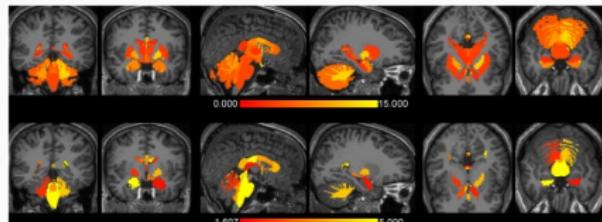
SOFTWARE DEPENDENCIES: HORROR STORIES

- Software environment evolution



SOFTWARE DEPENDENCIES: HORROR STORIES

- Software environment evolution
- 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:

- volume: $8.8 \pm 6.6\%$ (range 1.3-**64.0%**)
- cortical thickness: $2.8 \pm 1.3\%$ (range 1.1-7.7%)

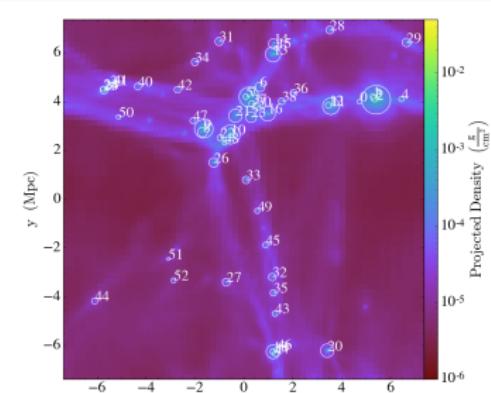
About a factor two smaller differences were found between the Mac and HP workstations and between Mac OSX 10.5 and OSX 10.6.

In the context of an ongoing study, users are discouraged to update to a new major release of either FreeSurfer or operating system.

Formal assessment of the accuracy of FreeSurfer is desirable.

SOFTWARE DEPENDENCIES: HORROR STORIES

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

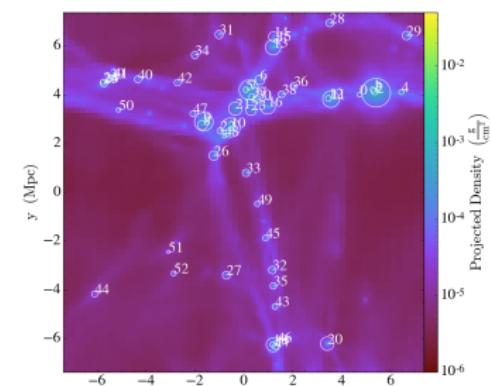


Assessing Reproducibility: An Astrophysical Example of Computational Uncertainty in the HPC Context (ResCuE-HPC, 2018)

Compiler	Optim.	Largest Halo Avg Mass.	Std. Err	Walltime
gcc@6.2.0	None	2.273E 46	1.069E 44	22h

SOFTWARE DEPENDENCIES: HORROR STORIES

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

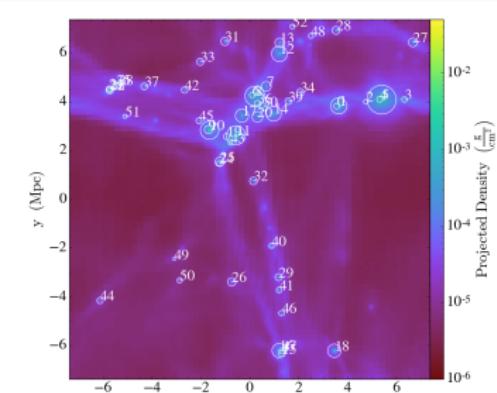


Assessing Reproducibility: An Astrophysical Example of Computational Uncertainty in the HPC Context (ResCuE-HPC, 2018)

Compiler	Optim.	Largest Halo Avg Mass.	Std. Err	Walltime
gcc@6.2.0	None	2.273E 46	1.069E 44	22h

SOFTWARE DEPENDENCIES: HORROR STORIES

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

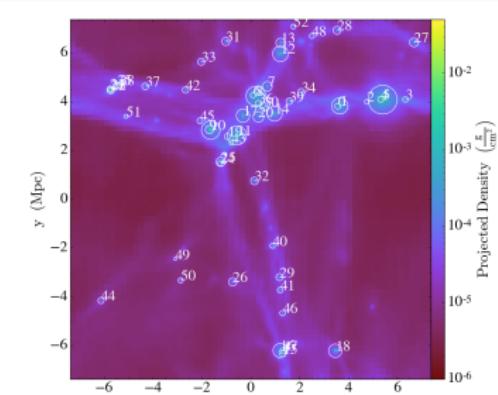


Assessing Reproducibility: An Astrophysical Example of Computational Uncertainty in the HPC Context (ResCuE-HPC, 2018)

Compiler	Optim.	Largest Halo		Walltime
		Avg Mass.	Std. Err	
gcc@6.2.0	None	2.273E 46	1.069E 44	22h
gcc@6.2.0	Normal	2.266E 46	1.218E 44	10h
gcc@6.2.0	High	2.275E 46	1.199E 44	9h

SOFTWARE DEPENDENCIES: HORROR STORIES

- Software environment evolution
 - 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	2.273E 46	1.069E 44	22h
gcc@6.2.0	Normal	2.266E 46	1.218E 44	10h
gcc@6.2.0	High	2.275E 46	1.199E 44	9h
intel@16.0.3	None	22.71 E 46	1.587E 44	39h
intel@16.0.3	Normal	43.30 E 46	1.248E 44	7h
intel@16.0.3	High	2.268E 46	1.414E 44	6h
cce@8.5.5	Low	43.11 E 46	1.353E 44	16h
cce@8.5.5	Normal	2.271E 46	1.261E 44	6h
cce@8.5.5	High	2.272E 46	1.341E 44	5h

COMPLEX ECOSYSTEMS

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

3.5.1

COMPLEX ECOSYSTEMS

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

3.5.1

```
1 apt show python3-matplotlib
```

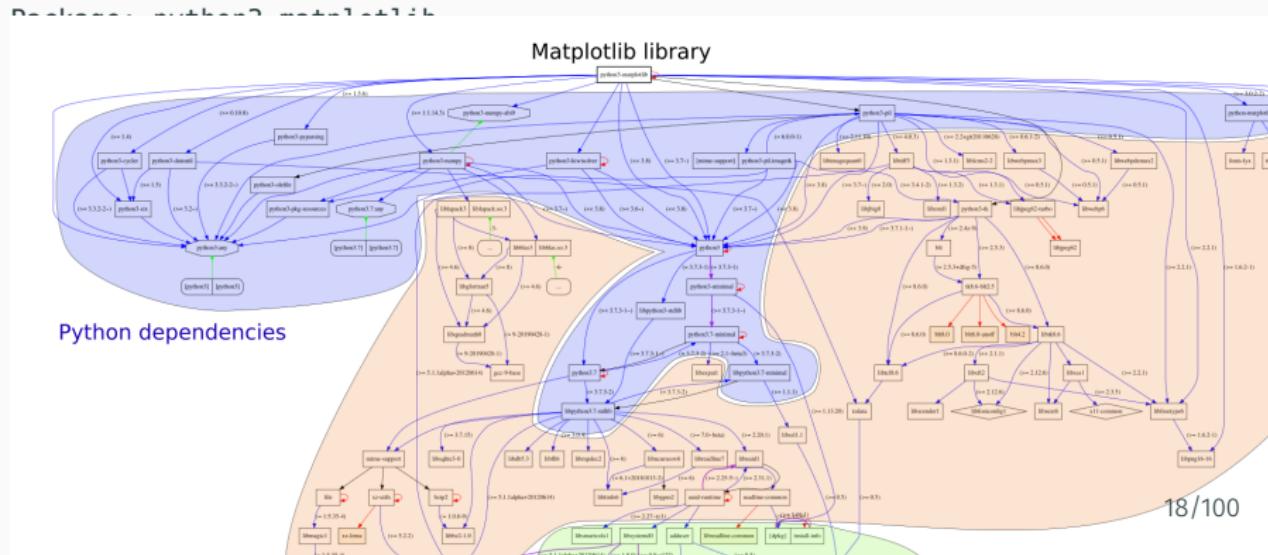
Package: python3-matplotlib
Version: 3.5.1-2+b1
Source: matplotlib (3.5.1-2)
Maintainer: Sandro Tosi <morph@debian.org>
Installed-Size: 27.6 MB
Depends: libjs-jquery, libjs-jquery-ui, python-matplotlib-data (>= 3.5.1),
 python3-dateutil, python3-pil.imagetk, python3-pyparsing (>= 1.5.6),
 python3-six (>= 1.4), python3-numpy (>= 1:1.20.0), python3-numpy-
 abi9,
 python3 (<< 3.11), python3 (>= 3.9~), python3-cycler (>= 0.10.0),
 python3-fonttools, python3-kiwisolver, python3-packaging, python3-
 pil,
 python3:any, libc6 (>= 2.29), libfreetype6 (>= 2.2.1),
 libgcc-s1 (>= 3.3.1), libqhull-r8.0 (>= 2020.1), libstdc++6 (>= 11)
Recommends: python3-tk
Suggests: dvipng, ffmpeg, fonts-staypuft, ghostscript, gir1.2-gtk-3.0, inkscape,

COMPLEX ECOSYSTEMS

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

3.5.1

```
1 apt show python3-matplotlib
```



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 <cmd>`
 - Building images: `docker build -f <Dockerfile>`
 - Sharing through the Docker Hub: `docker pull/push `

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*

```
1   FROM ubuntu:20.04
2   RUN apt-get update
3       && apt-get upgrade -y
4       && 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

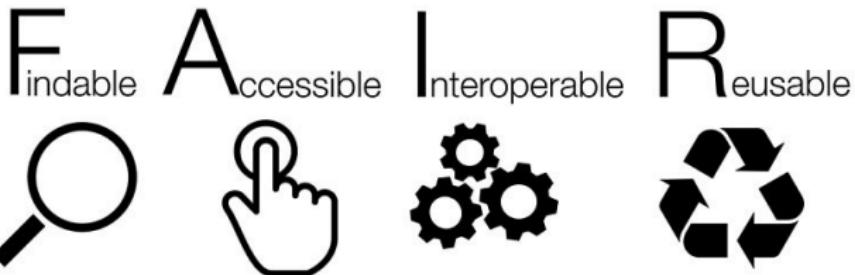
Package managers (the ugly and the good)

- 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/>

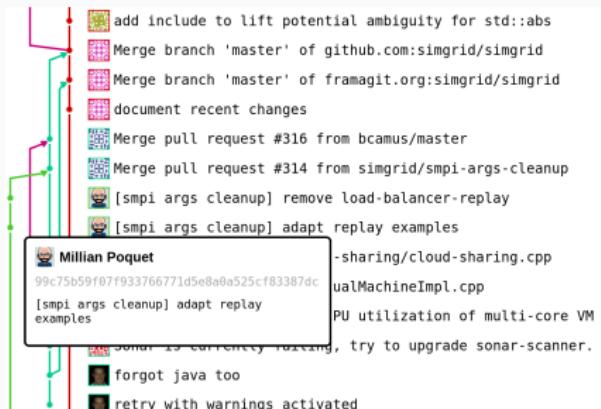
- "*Open as much as possible and close as much as necessary*"
- Management, publication, annotation (metadata), archiving
- Source code = specific data with specific consideration

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



Limitation

- Managing large data: **Git LFS** **Git Annex** (or DataLad)

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

zenodo

Software Archive



Software Heritage

Collect/Preserve/Share

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



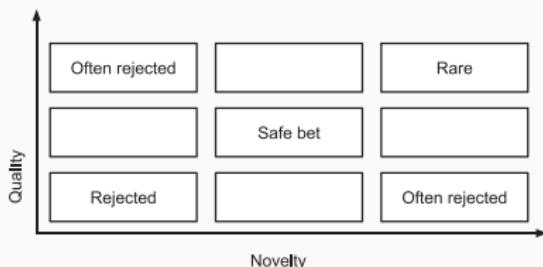
Joelle Pineau @ NeurIPS'18

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

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

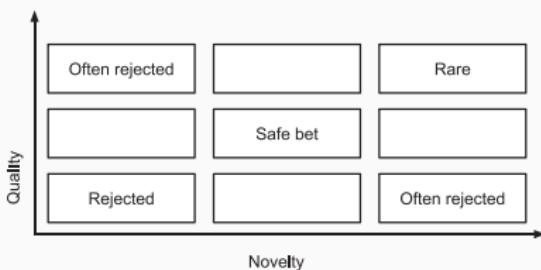
CHANGING ACADEMIC PRACTICES (PUBLISH OR PERISH)

- 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



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

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