

# Reproducible Research, Open Science, Open Data, . . .

## Why ? Who ? How ?

Arnaud Legrand  
CNRS, Inria, University of Grenoble

May 23rd, 2018 – Rennes, Inria DGD-T days

# Who am I?

Application Domain Large scale computing infrastructures



## Research Themes

- Optimization (scheduling, game theory)
- Performance Evaluation (modeling, simulation, analysis)

## Proselytism in Scientific Methodology

- Reproducible Research/Open Science
- Design of Experiments

# Outline

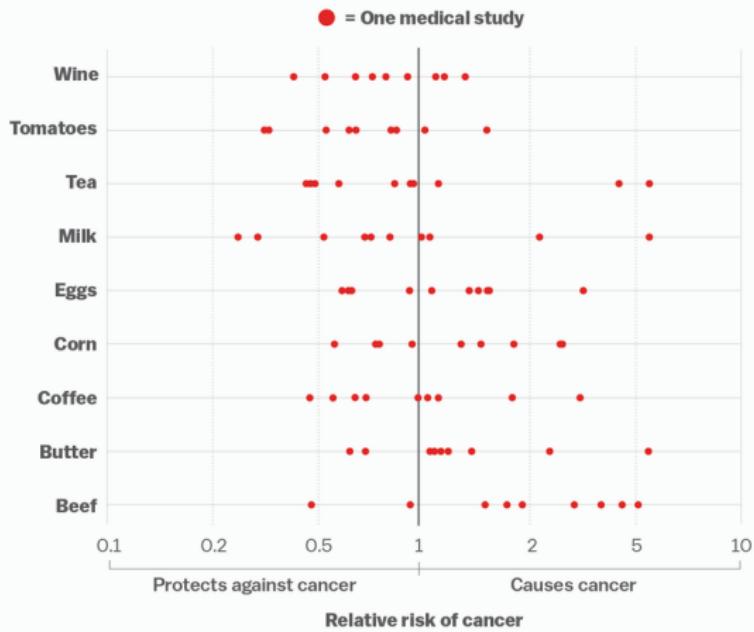
- ① Science crisis ?
- ② Is Computer Science Really Concerned With This?
- ③ Reproducible Research/Open Science in a Nutshell
- ④ Illustrating Nice Ideas Through Different Tools
  - Taking notes and explaining
  - Controlling your environment/experiments
  - Making data/code available
  - Changing evaluation practices
- ⑤ What can Computer Scientists do ?

## Inconsistencies

Is everything we eat associated with cancer? A systematic cookbook review,  
Schoenfeld and Ioannidis, *Amer. Jour. of Clinical Nutrition*, 2013.

# Inconsistencies

Everything we eat both causes and prevents cancer



SOURCE: Schoenfeld and Ioannidis, *American Journal of Clinical Nutrition*

Vox

# 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

Los Angeles Times | BUSINESS

LOCAL U.S. WORLD BUSINESS SPORTS ENTERTAINMENT HEALTH STYLE TRAVEL

## 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.

Science AAAS NEWS SCIENCE JOURNALS CAREERS MULTIMEDIA COLLECTIONS

Science Home Current Issue Previous Issues Science Express Science Products My Science About the Journal

Issue 17 January 2014 Vol. 343 no. 6168 p. 229 DOI: 10.1126/science.1250475

Article Views Article Summary Full Text PDF Article Tools Save to My Folders Download Citation Alert Me When Article is Cited Post to Circulate E-mail This Page Rights & Permissions Commercial Reprints and E-Prints View PubMed Citation Related Content

Marcia McNutt

Science advances on a foundation of trusted approaches that scientists use to gain confidence. Community trust was shaken by reports that a result was not reproducible. Because confidence in results is critical to the scientific community, we are announcing new initiatives. For preclinical studies (one of the top recommendations of the U.S. National Institutes of Health), increasing transparency. Authors will indicate handling (such as how to deal with outliers), what was a sufficient signal-to-noise ratio, whether an experiment was blind to the conduct of the guidelines.

TheScientist EXPLORING LIFE. INSPIRING INNOVATION

## NIH Tackles Irreproducibility

The federal agency speaks out about how to improve the quality of scientific research.

By Jef Akst | January 28, 2014

Announcement: Reducing our irreproducibility : Nature News & Comment

nature.com Sitemap Login Register

## nature International weekly journal of science

Home News & Comment Research Careers & Jobs Current Issue Archive

Archive > Volume 491 > Issue 7448 > Editorial > Article

NATURE | EDITORIAL

### Announcement: Reducing our irreproducibility

24 April 2013

PDF Rights & Permissions

Over the past year, Nature has published a string of articles that highlight the reliability and reproducibility of published research (collected at

The Economist Washington's lawyer surplus How to do a nuclear deal with Iran Investment tips from Nobel economists Junk bonds are back The meaning of Sachin Tendulkar

# HOW SCIENCE GOES WRONG.

nature International weekly journal of science

Menu archive - volume 483 - issue 7391 - editorials - article

NATURE | EDITORIAL

### Must try harder

Nature 483, 509 (29 March 2012) doi:10.1038/483509a Published online 28 March 2012

PDF Citation Reprints Rights & permissions Article metrics

Too many sloppy mistakes are creeping into scientific papers. Lab heads must look more rigorously at the data — and at themselves.

Courtesy V. Stodden, SC, 2015

# 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

Los Angeles Times | BUSINESS



Last Week Tonight with John Oliver:  
Scientific Studies (HBO), May 2016

Courtesy V. Stodden, SC, 2015

# Austerity in Fiscal Policy

2010 "gross debt [...] exceeding 90 percent of the economy has a significant negative effect on economic growth"

– Reinhart et Rogoff: *Growth in a Time of Debt*

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

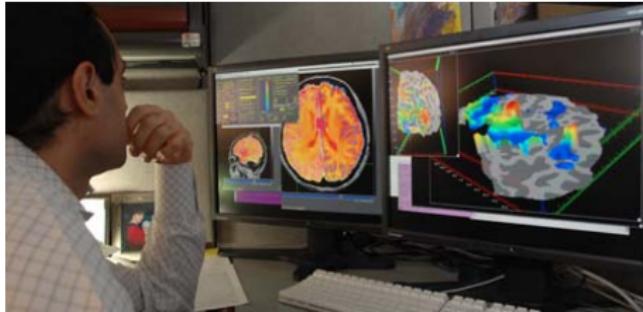
– Herndon, Ash and Pollin

combining data across centuries, exchange rate regimes, public and private debt, and debt denominated in foreign currency as well as domestic currency

– Wray

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

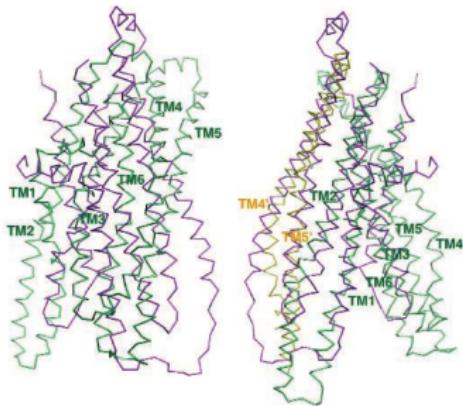
Yet, a scientific debate has at least been possible.



- 2010: Bennett et al. and the dead salmon 😊
- 2016: Eklund, Nichols, and Knutsson. A bug in fmri software could invalidate 15 years of brain research (*40,000 articles*, although it is a bit more subtle than this).
- 2016: Nichols.  $\approx 3\,600$  articles may have to be revisited for confirmation.

These article do not necessarily invalidate everything but force the community to improve their practice.

# Geoffrey Chang's incorrect protein structures



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 computational biology

# A Reproducibility Crisis? What are the Consequences ?

## The Duke University scandal with scientific misconduct on lung cancer

- *Nature Medicine* - 12, 1294 - 1300 (2006) **Genomic signatures to guide the use of chemotherapeutics**, by Anil Potti and 16 other researchers from Duke University and University of South Florida
- Major commercial labs licensed it and were about to start using it before two statisticians discovered and publicized its faults

Dr. Baggerly and Dr. Coombes found errors almost immediately. Some seemed careless — moving a row or a column over by one in a giant spreadsheet — while others seemed inexplicable. The Duke team shrugged them off as “clerical errors.”

The Duke researchers continued to publish papers on their genomic signatures in prestigious journals. Meanwhile, they started three trials using the work to decide which drugs to give patients.

- Retractions: January 2011. Ten papers that Potti coauthored in prestigious journals were retracted for varying reasons

# Well... Stronger and Stronger Consequences

A recent scandal In 2013, *Dong-Pyou Han*, a former assistant professor of biomedical sciences at Iowa State University was disgraced:

- Falsified blood results to make it appear as though a vaccine he was working on had exhibited anti-HIV activity
- Han and his team received  $\approx \$19$  million from NIH
- Retraction and resignation of university

Han was sentenced in 2015 to 57 months imprisonment for fabricating and falsifying data in HIV vaccine trials. He was also fined US \$7.2 million!

## We should avoid witch-hunt

- August 5, 2014, Yoshiki Sasai (stem cell, considered for Nobel Prize) hanged in his laboratory at the RIKEN (Japan). Fraud suspicion...
- In 1986, a young postdoctoral fellow at MIT accused her director, Thereza Imanishi-Kari, of falsifying the results of a study published in Cell and co-signed by the Nobel laureate David Baltimore. [...] Declared guilty, Univ. presidency resignation, and finally cleared. This put the careers of two outstanding researchers on hold for ten years based on unfounded accusations.

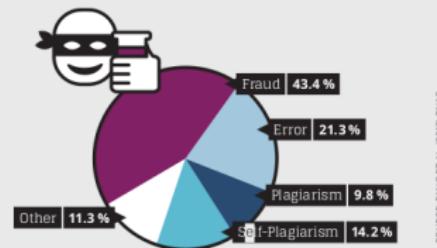
Scientific fraud is bad but let's be careful Have a look at the wikipedia *list of academic scandals*.

# Is Fraud a new phenomenon?

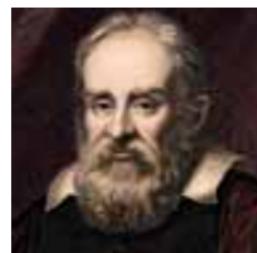
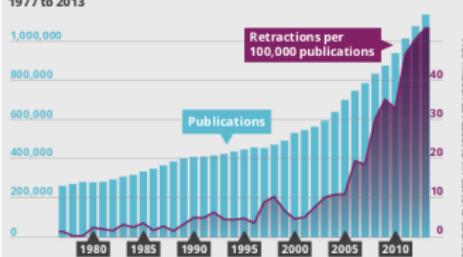
## *The Battle against Scientific Fraud in the CNRS International Magazine*

### Biomedical fraud in figures

Cause of retraction 1977 to 2012



Number of publications and retractions  
1977 to 2013



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

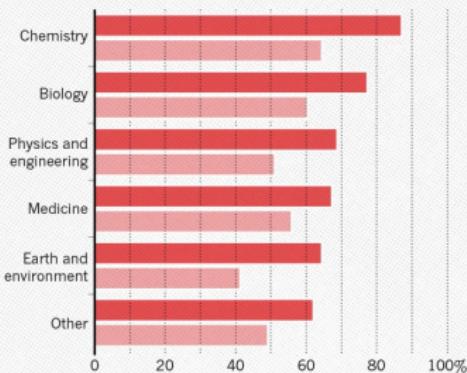
Is it only a matter of Fraud ?

# Why are scientific studies so difficult to reproduce?

## HAVE YOU FAILED TO REPRODUCE AN EXPERIMENT?

Most scientists have experienced failure to reproduce results.

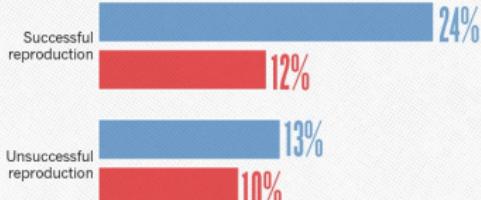
- Someone else's
- My own



## 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.

- Published
- Failed to publish



Number of respondents from each discipline:  
Biology 703, Chemistry 106, Earth and environmental 95.

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

## Wrap-up

- Oncology : "more than half studies published in prestigious journals cannot be reproduced in industrial labs"
- Psychology : "replicating a hundred of major articles: only one third of coherent results"



Whistle blowers, sick institutions, broken system, ?..

No. Questionning previous work is part of the scientific process

Just like honesty, rigor and transparency...

Risks science credibility put into question. No more difference with crooks!

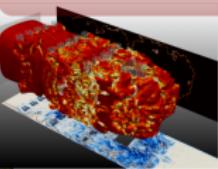
# Outline

- ① Science crisis ?
- ② Is Computer Science Really Concerned With This?
- ③ Reproducible Research/Open Science in a Nutshell
- ④ Illustrating Nice Ideas Through Different Tools
  - Taking notes and explaining
  - Controlling your environment/experiments
  - Making data/code available
  - Changing evaluation practices
- ⑤ What can Computer Scientists do ?

# Computational science!



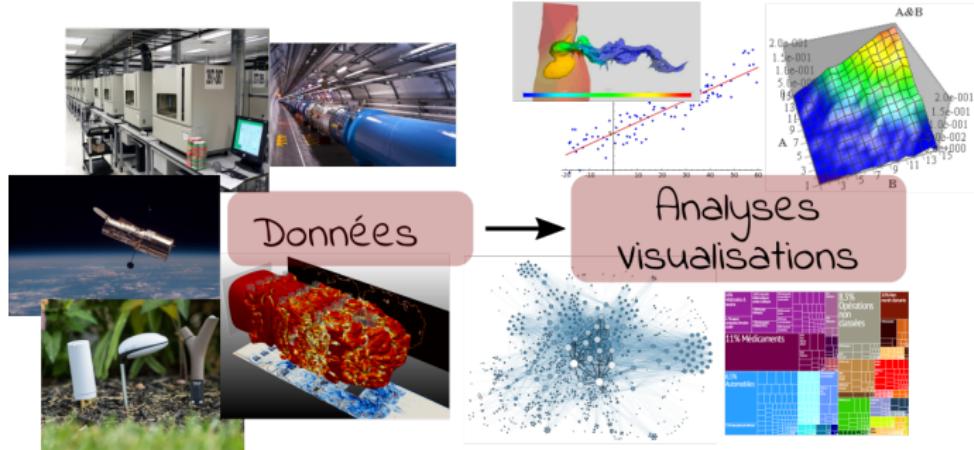
Données



*Today the computer is just as important a tool for chemists as the test tube. Simulations are so realistic that they predict the outcome of traditional experiments...*

*– Nobel Comity (Chemistry), 2013*

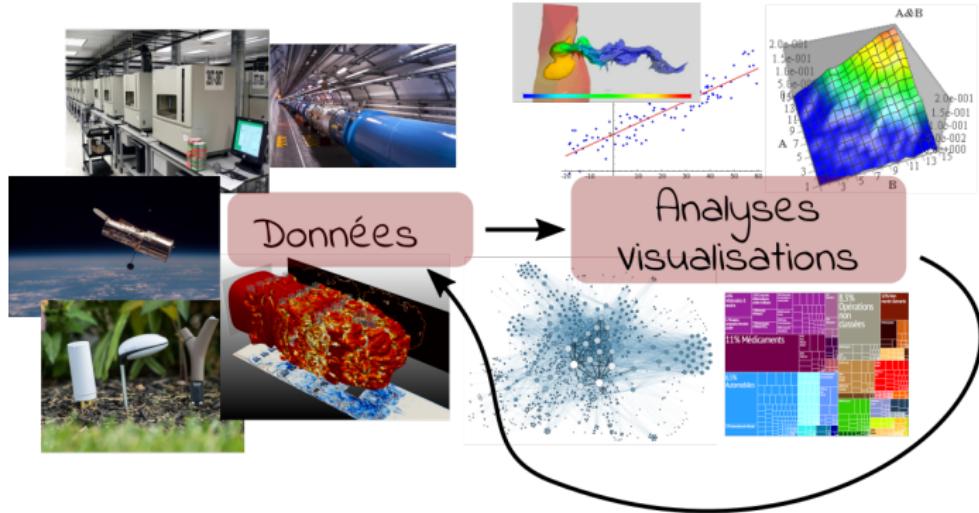
# Computational science!



*Today the computer is just as important a tool for chemists as the test tube. Simulations are so realistic that they predict the outcome of traditional experiments...*

*– Nobel Comity (Chemistry), 2013*

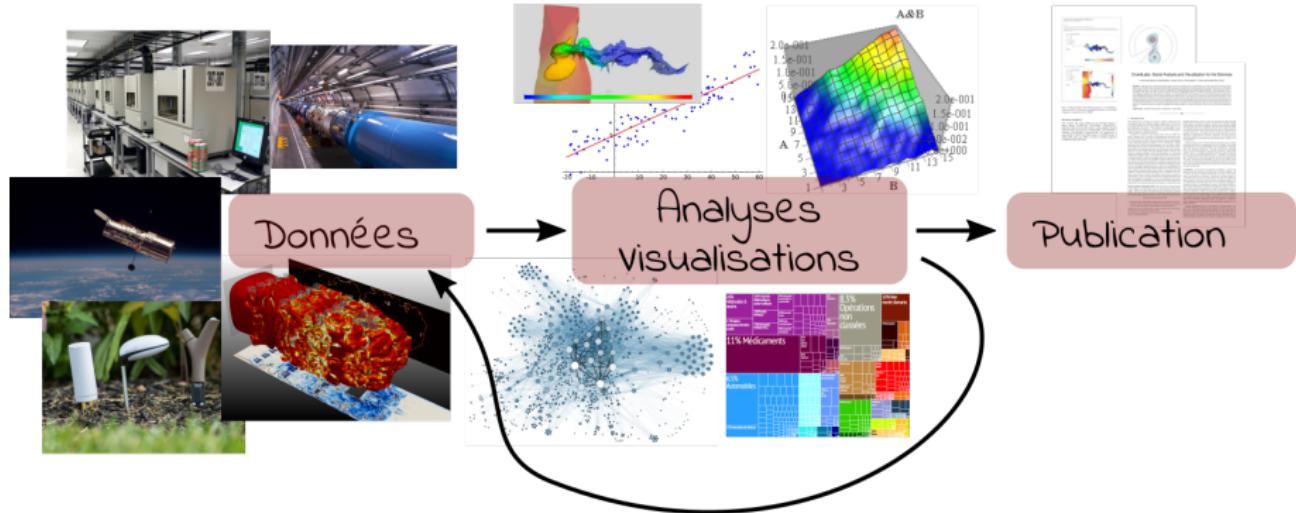
# Computational science!



*Today the computer is just as important a tool for chemists as the test tube. Simulations are so realistic that they predict the outcome of traditional experiments...*

– Nobel Comity (Chemistry), 2013

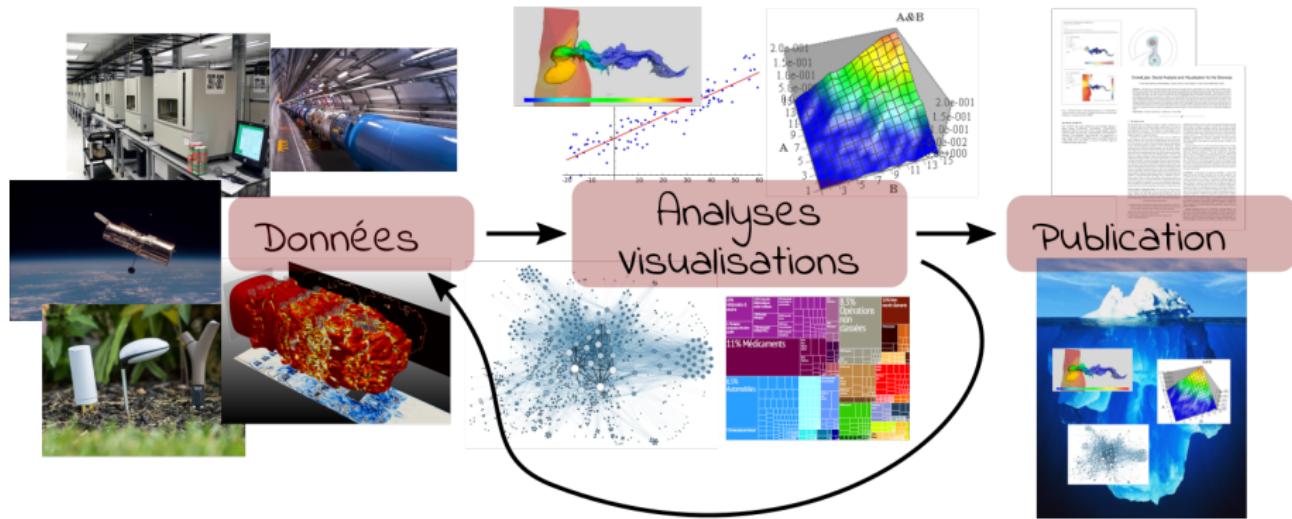
# Computational science!



*Today the computer is just as important a tool for chemists as the test tube. Simulations are so realistic that they predict the outcome of traditional experiments...*

– Nobel Comity (Chemistry), 2013

# Computational science!



*Today the computer is just as important a tool for chemists as the test tube. Simulations are so realistic that they predict the outcome of traditional experiments...*

*– Nobel Comity (Chemistry), 2013*

# Aren't Computers Good for Science ?

How computers broke science – and what we can do to fix it.

- Point and click
- Spreadsheets : programming and data manipulation mistakes
  - Membrane-Associated Ring Finger (C3HC4) 1, E3 Ubiquitin Pro Ligase → MARCH1 → 2016-03-01 → 1456786800
  - 2310009E13 → 2.31E+19
- Complex software stacks : avoid proprietary software as much as possible
- Bugs : *Programming is difficult !*

# All this is about Natural 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 Natural 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...

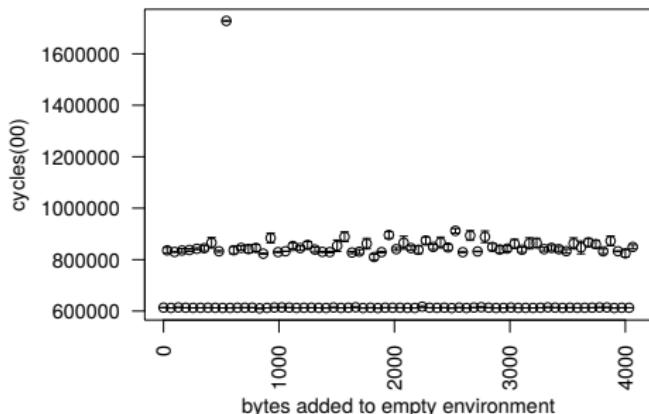
# Experimenting with computers

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

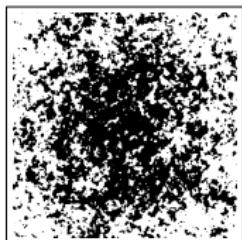
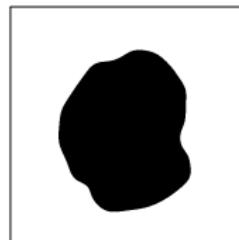
# Computer performance ? Well, I design algorithms!

- "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



# I design Machine Learning Algorithms



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

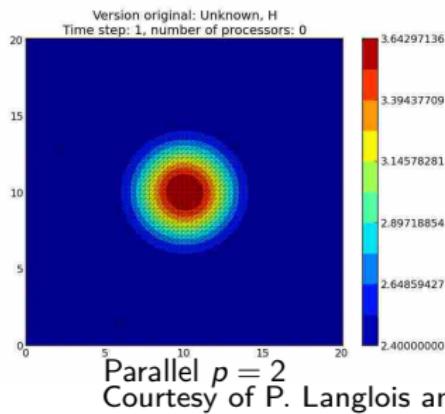
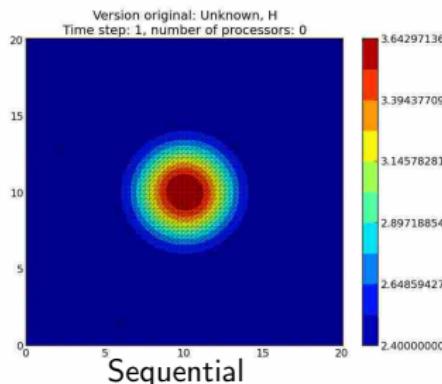
Did I mention we have parallel machines nowadays? 😊

## Telemac2D: the simplest gouttedo simulation

### The gouttedo 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, ...)



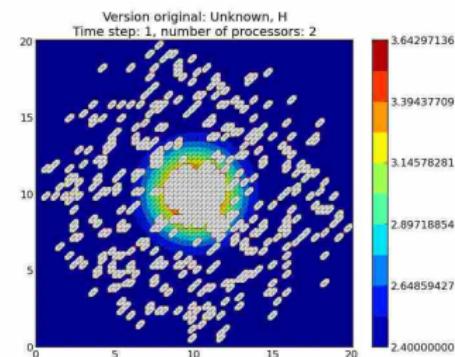
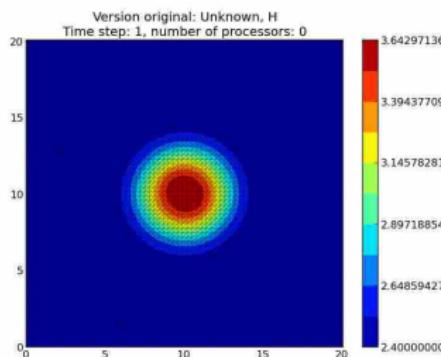
# All I care about is the algorithm output

Did I mention we have parallel machines nowadays? 😊

A white plot displays a non-reproducible value

Numerical reproducibility?

time step = 1



Courtesy of P. Langlois and R. Nheili

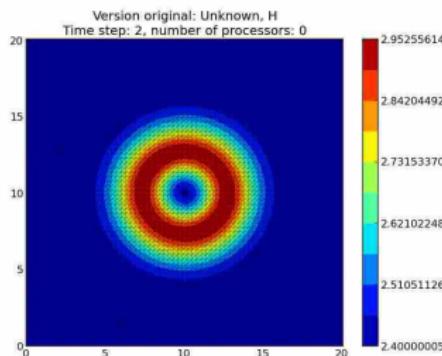
# All I care about is the algorithm output

Did I mention we have parallel machines nowadays? 😊

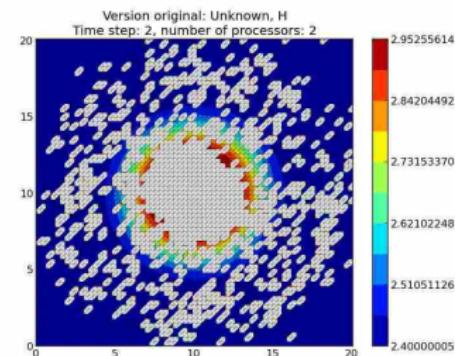
A white plot displays a non-reproducible value

Numerical reproducibility?

time step = 2



Sequential



Parallel  $p = 2$

Courtesy of P. Langlois and R. Nheili

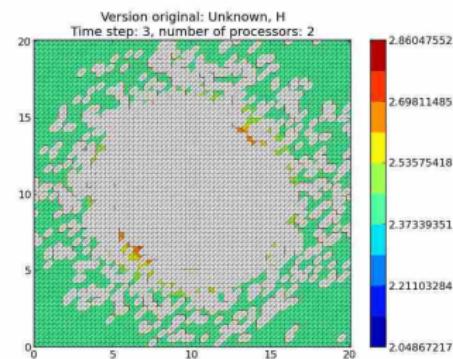
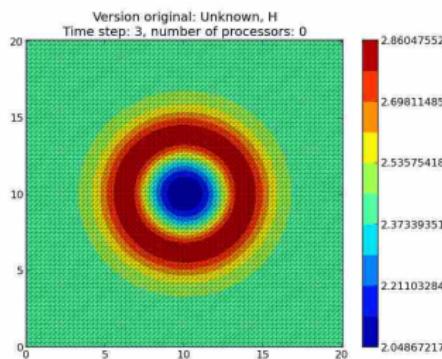
# All I care about is the algorithm output

Did I mention we have parallel machines nowadays? 😊

A white plot displays a non-reproducible value

Numerical reproducibility?

time step = 3



Courtesy of P. Langlois and R. Nheili

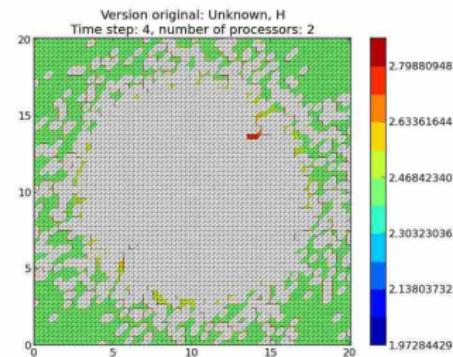
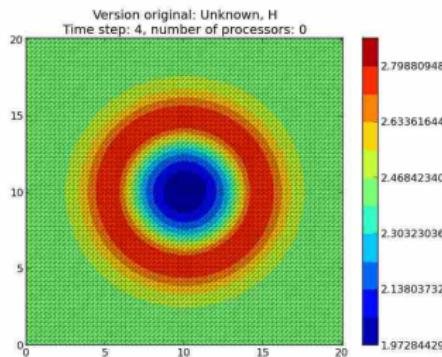
# All I care about is the algorithm output

Did I mention we have parallel machines nowadays? 😊

A white plot displays a non-reproducible value

Numerical reproducibility?

time step = 4



Courtesy of P. Langlois and R. Nheili

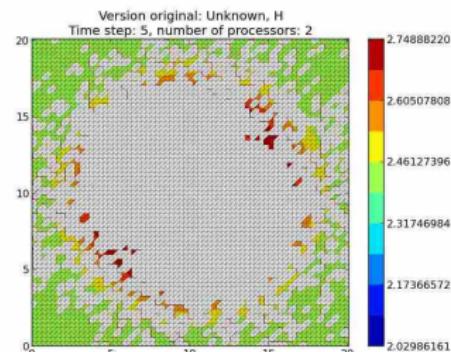
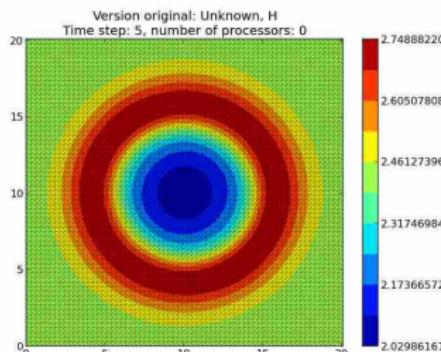
# All I care about is the algorithm output

Did I mention we have parallel machines nowadays? 😊

A white plot displays a non-reproducible value

Numerical reproducibility?

time step = 5



Courtesy of P. Langlois and R. Nheili

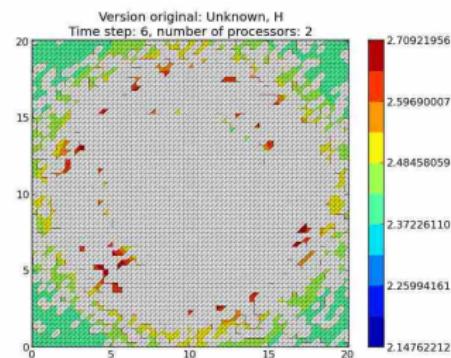
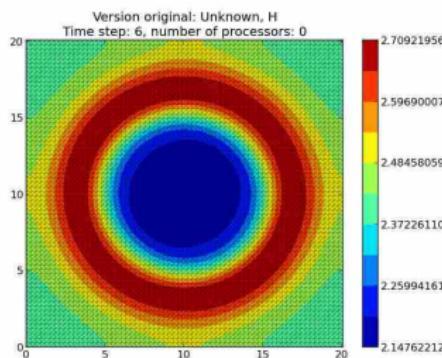
# All I care about is the algorithm output

Did I mention we have parallel machines nowadays? 😊

A white plot displays a non-reproducible value

Numerical reproducibility?

time step = 6



Courtesy of P. Langlois and R. Nheili

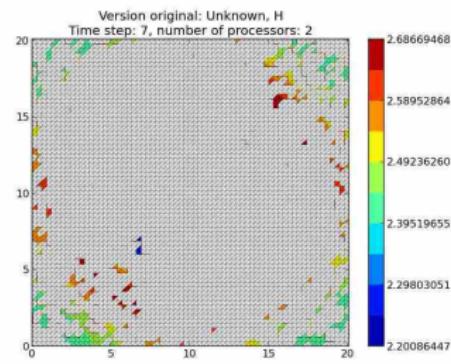
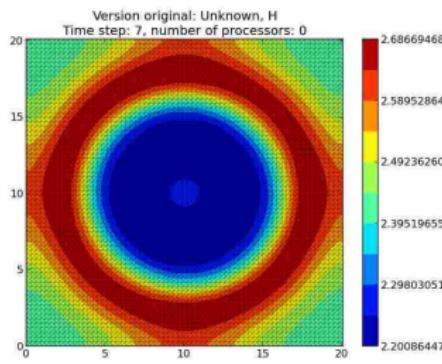
# All I care about is the algorithm output

Did I mention we have parallel machines nowadays? 😊

A white plot displays a non-reproducible value

Numerical reproducibility?

time step = 7



Courtesy of P. Langlois and R. Nheili

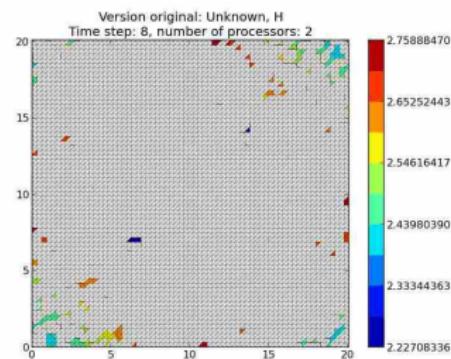
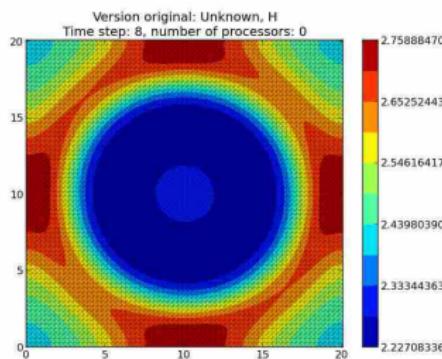
# All I care about is the algorithm output

Did I mention we have parallel machines nowadays? 😊

A white plot displays a non-reproducible value

Numerical reproducibility?

time step = 8



Courtesy of P. Langlois and R. Nheili

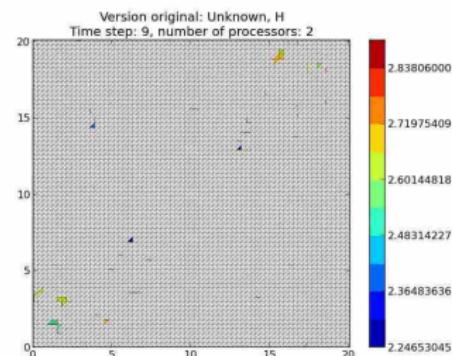
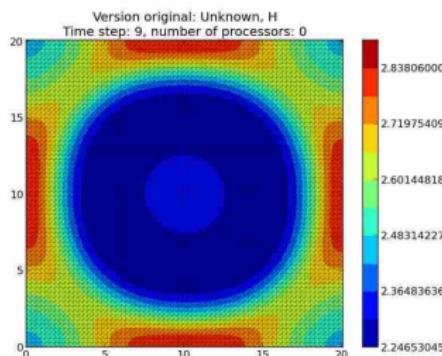
# All I care about is the algorithm output

Did I mention we have parallel machines nowadays? 😊

A white plot displays a non-reproducible value

Numerical reproducibility?

time step = 9



Courtesy of P. Langlois and R. Nheili

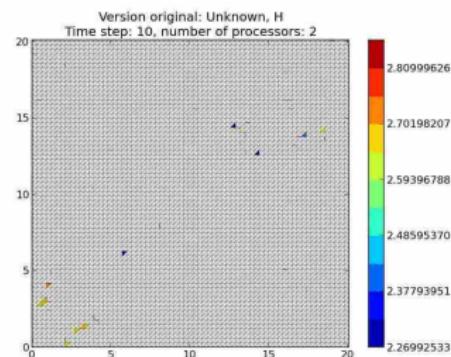
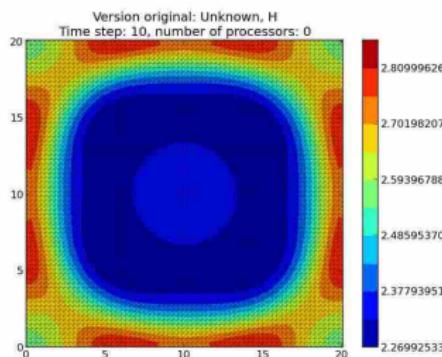
# All I care about is the algorithm output

Did I mention we have parallel machines nowadays? 😊

A white plot displays a non-reproducible value

Numerical reproducibility?

time step = 10



Courtesy of P. Langlois and R. Nheili

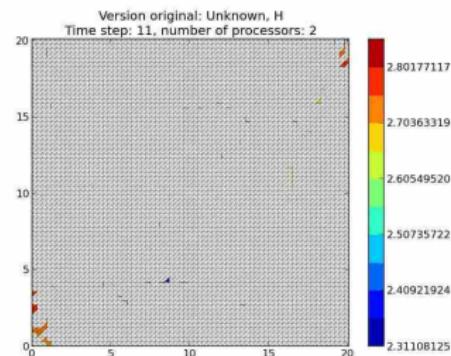
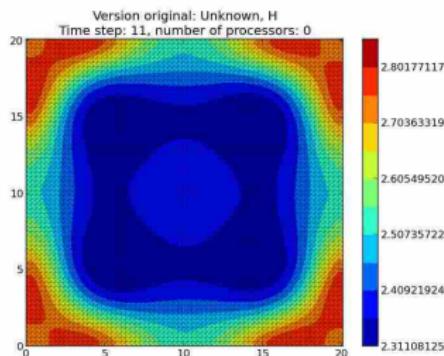
# All I care about is the algorithm output

Did I mention we have parallel machines nowadays? 😊

A white plot displays a non-reproducible value

Numerical reproducibility?

time step = 11



Courtesy of P. Langlois and R. Nheili

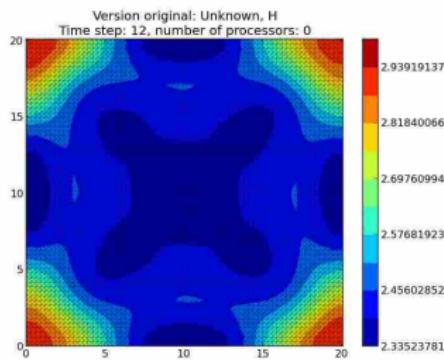
# All I care about is the algorithm output

Did I mention we have parallel machines nowadays? 😊

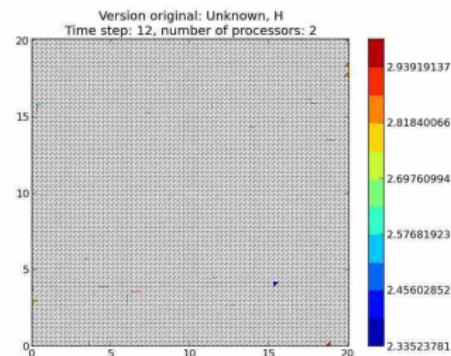
A white plot displays a non-reproducible value

Numerical reproducibility?

time step = 12



Sequential



Parallel  $p = 2$

Courtesy of P. Langlois and R. Nheili

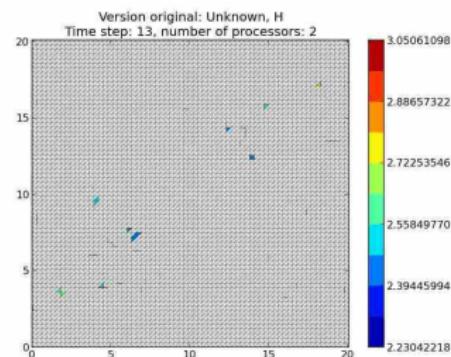
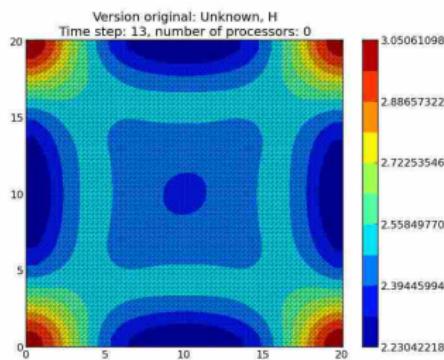
# All I care about is the algorithm output

Did I mention we have parallel machines nowadays? 😊

A white plot displays a non-reproducible value

Numerical reproducibility?

time step = 13



Courtesy of P. Langlois and R. Nheili

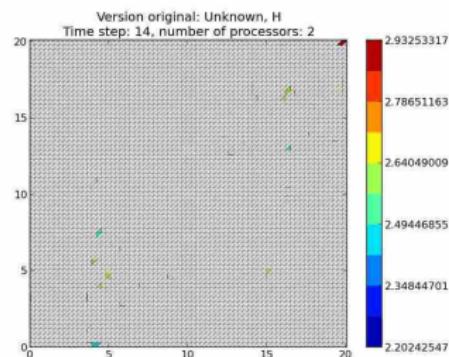
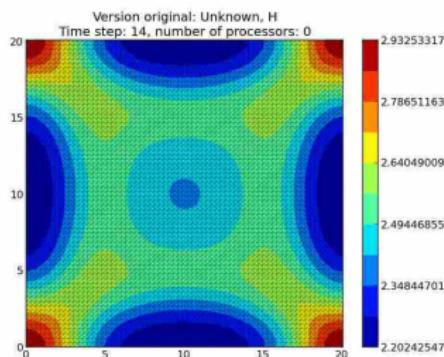
# All I care about is the algorithm output

Did I mention we have parallel machines nowadays? 😊

A white plot displays a non-reproducible value

Numerical reproducibility?

time step = 14



Courtesy of P. Langlois and R. Nheili

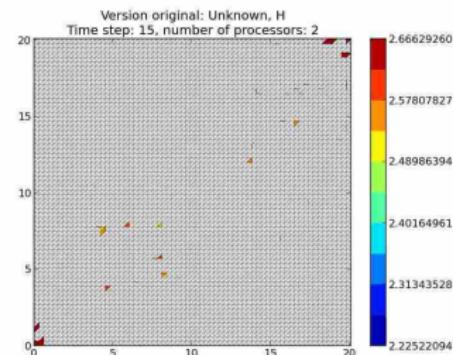
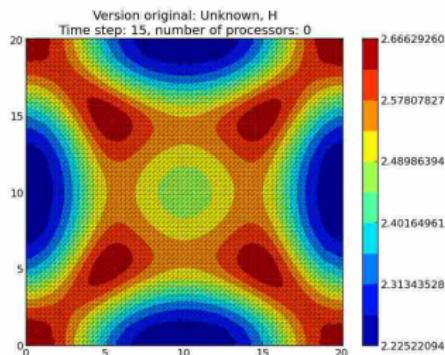
# All I care about is the algorithm output

Did I mention we have parallel machines nowadays? 😊

A white plot displays a non-reproducible value

NO numerical reproducibility!

time step = 15



Courtesy of P. Langlois and R. Nheili

# All I care about is the algorithm output

Did I mention we have **parallel machines** nowadays? 😊

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

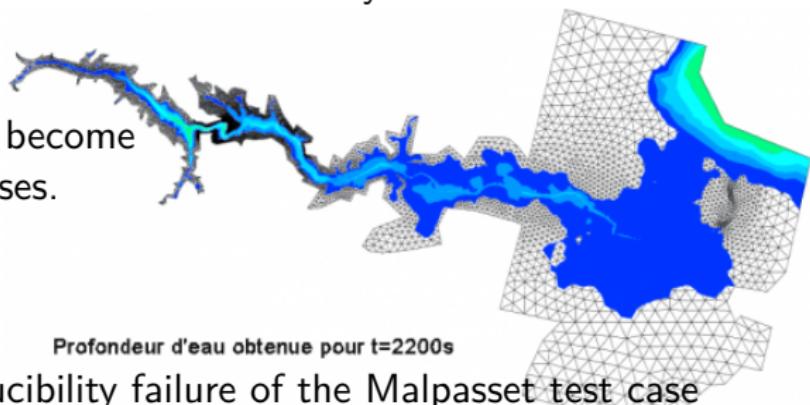


TABLE 1.1: Reproducibility failure of the Malpasset test case

	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.

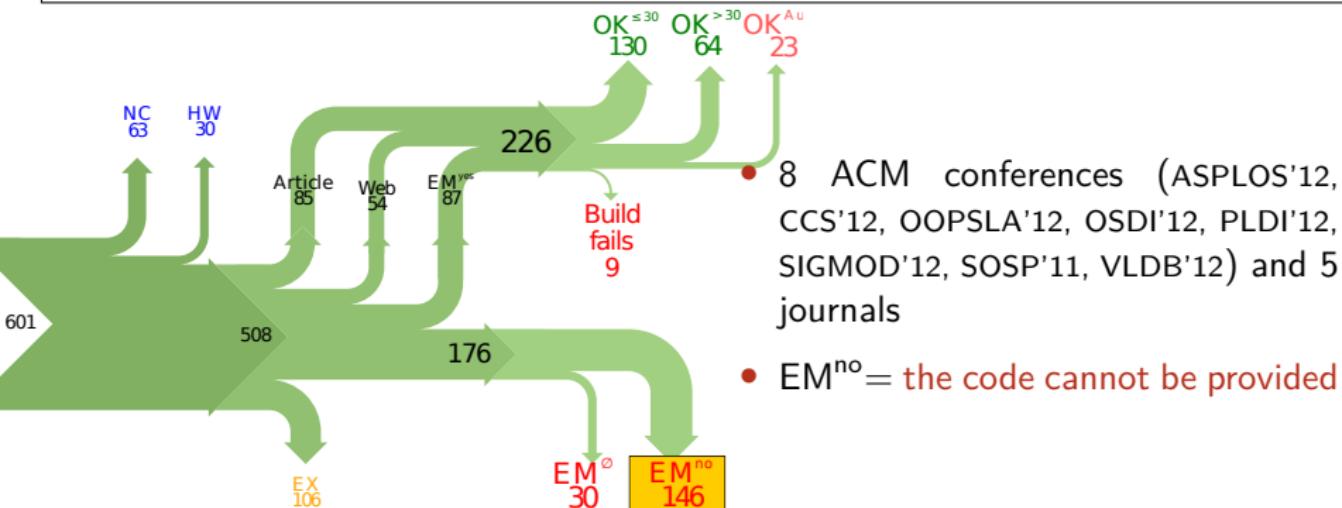
Courtesy of P. Langlois and R. Nheili

# Outline

- ① Science crisis ?
- ② Is Computer Science Really Concerned With This?
- ③ Reproducible Research/Open Science in a Nutshell
- ④ Illustrating Nice Ideas Through Different Tools
  - Taking notes and explaining
  - Controlling your environment/experiments
  - Making data/code available
  - Changing evaluation practices
- ⑤ What can Computer Scientists do ?

# Lack of reproducibility is foremost a cultural matter

Collberg, Christian et Al., *Measuring Reproducibility in Computer Systems Research*, <http://reproducibility.cs.arizona.edu/>



# The Dog Ate my Homework !!!

- Versioning Problems

*Thanks for your interest in the implementation of our paper. The good news is that I was able to find some code. I am just hoping that it is a stable working version of the code, and matches the implementation we finally used for the paper. Unfortunately, I have lost some data when my laptop was stolen last year. The bad news is that the code is not commented and/or clean.*

*Attached is the ⟨system⟩ source code of our algorithm. I'm not very sure whether it is the final version of the code used in our paper, but it should be at least 99% close. Hope it will help.*

# The Dog Ate my Homework !!!

- Versioning Problems
- Bad Backup Practices

*Unfortunately, the server in which my implementation was stored had a disk crash in April and three disks crashed simultaneously. While the help desk made significant effort to save the data, my entire implementation for this paper was not found.*

# The Dog Ate my Homework !!!

- Versioning Problems
- Bad Backup Practices
- Code Will be Available Soon

*Unfortunately the current system is **not mature enough at the moment**, so it's not yet publicly available. We are actively working on a number of extensions and **things are somewhat volatile**. However, once things stabilize we plan to release it to outside users. At that point, we would be happy to send you a copy.*

# The Dog Ate my Homework !!!

- Versioning Problems
- Bad Backup Practices
- Code Will be Available Soon
- No Intention to Release

*I am afraid that the source code was never released. The code was **never** intended to be released so is not in any shape for general use.*

# The Dog Ate my Homework !!!

- Versioning Problems
- Bad Backup Practices
- Code Will be Available Soon
- No Intention to Release
- Programmer Left

*⟨STUDENT⟩ was a graduate student in our program but he left while back so I am responding instead. For the paper we used a prototype that included many moving pieces that only ⟨STUDENT⟩ knew how to operate and we did not have the time to integrate them in a ready-to-share implementation before he left. Still, I hope you can build on the ideas/technique of the paper.*

*Unfortunately, the author who has done most of the coding for this paper has passed away and the code is no longer maintained.*

# The Dog Ate my Homework !!!

- Versioning Problems
- Bad Backup Practices
- Code Will be Available Soon
- No Intention to Release
- Programmer Left
- Commercial Code

*Since this work has been done at <COMPANY> we don't open-source code unless there is a compelling business reason to do so. So unfortunately I don't think we'll be able to share it with you.*

*The code owned by <COMPANY>, and AFAIK the code is not open-source. Your best bet is to reimplement :( Sorry.*

# The Dog Ate my Homework !!!

- Versioning Problems
- Bad Backup Practices
- Code Will be Available Soon
- No Intention to Release
- Programmer Left
- Commercial Code
- Proprietary Academic Code

*Unfortunately, the  $\langle SYSTEM \rangle$  sources are not meant to be opensource (the code is partially property of  $\langle UNIVERSITY 1 \rangle$ ,  $\langle UNIVERSITY 2 \rangle$  and  $\langle UNIVERSITY 3 \rangle$ .)*

*If this will change I will let you know, albeit I do not think there is an intention to make the  $\langle SYSTEM \rangle$  sources opensource in the near future.*

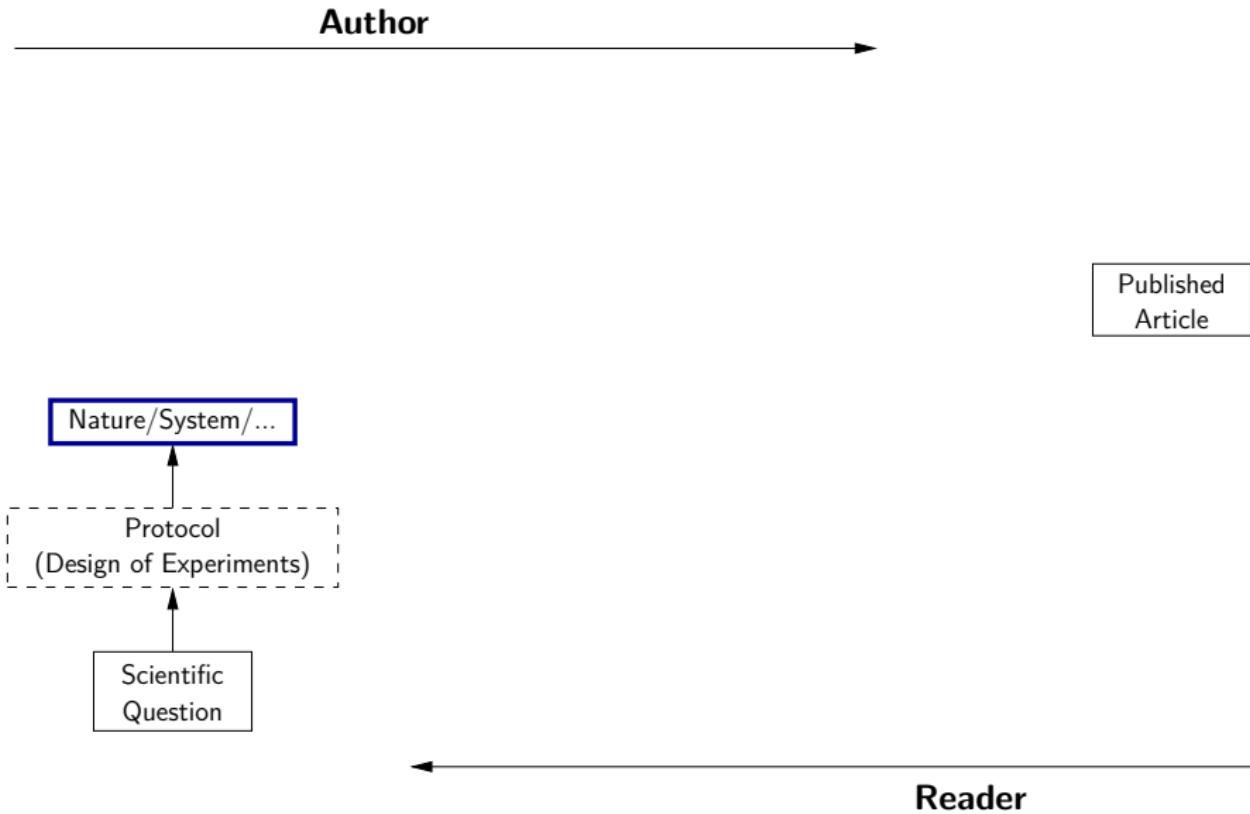
*If you're interested in obtaining the code, we only ask for a description of the research project that the code will be used in (which may lead to some joint research), and we also have a software license agreement that the University would need to sign.*

# The Dog Ate my Homework !!!

- Versioning Problems
- Bad Backup Practices
- Code Will be Available Soon
- No Intention to Release
- Programmer Left
- Commercial Code
- Proprietary Academic Code
- Research vs. Sharing
- ...
- ...

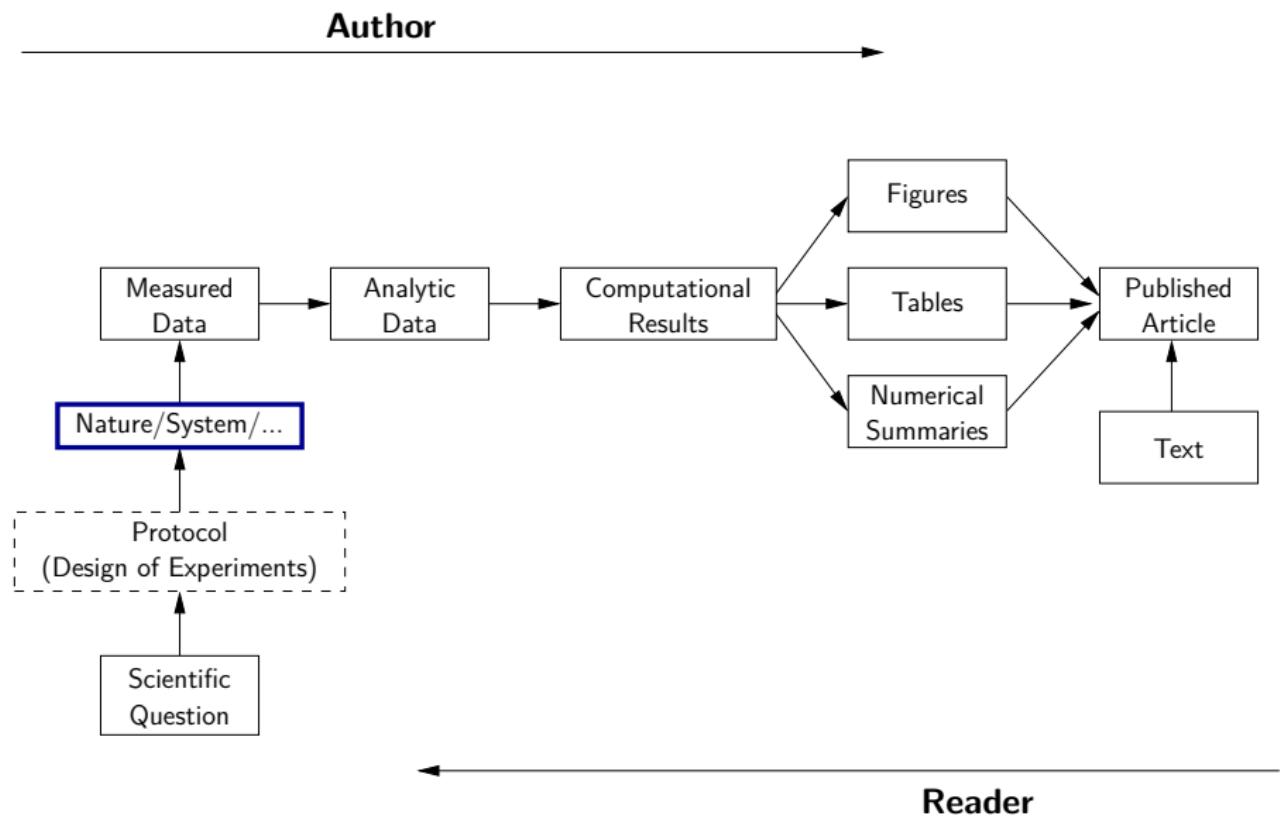
*In the past when we attempted to share it, we found ourselves spending more time getting outsiders up to speed than on our own research. So I finally had to establish the policy that we will not provide the source code outside the group.*

# Reproducible Research: Trying to Bridge the Gap



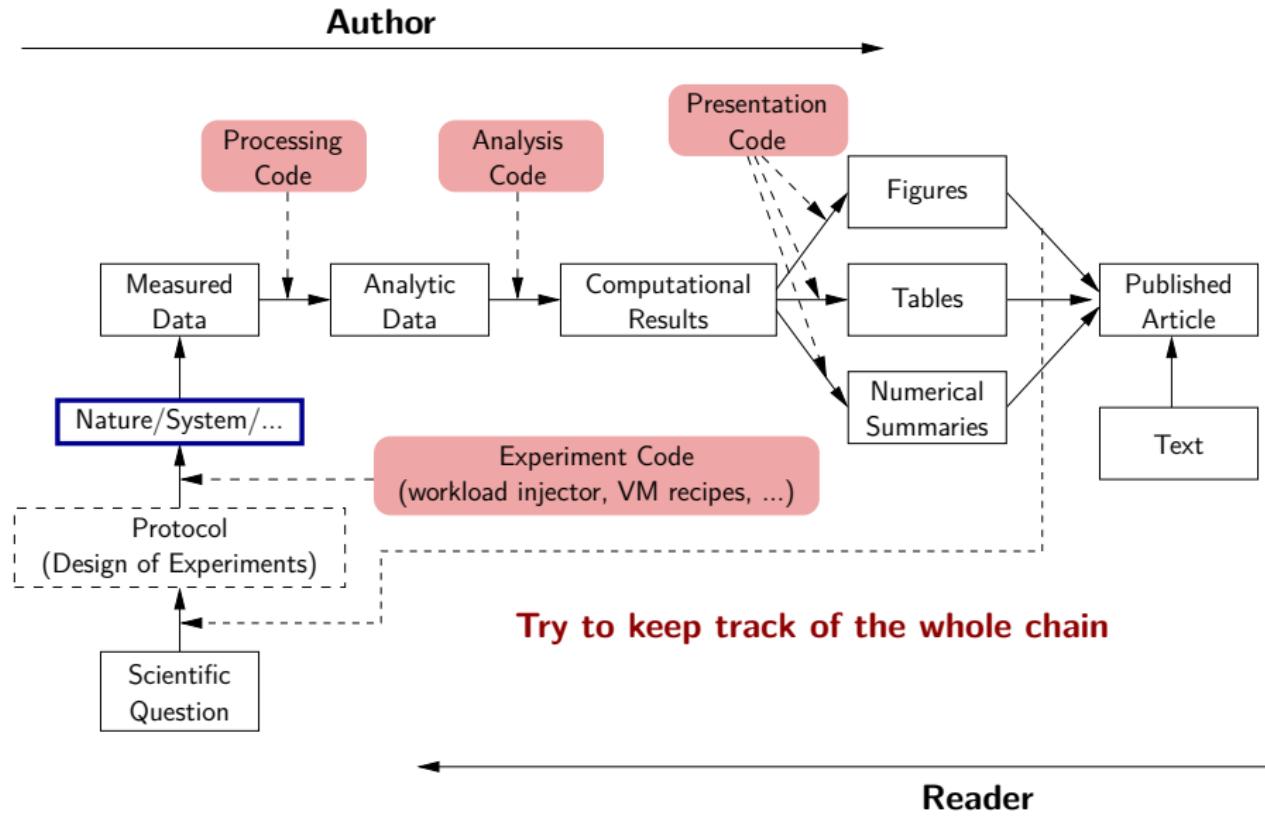
Inspired by Roger D. Peng's lecture on reproducible research, May 2014

# Reproducible Research: Trying to Bridge the Gap



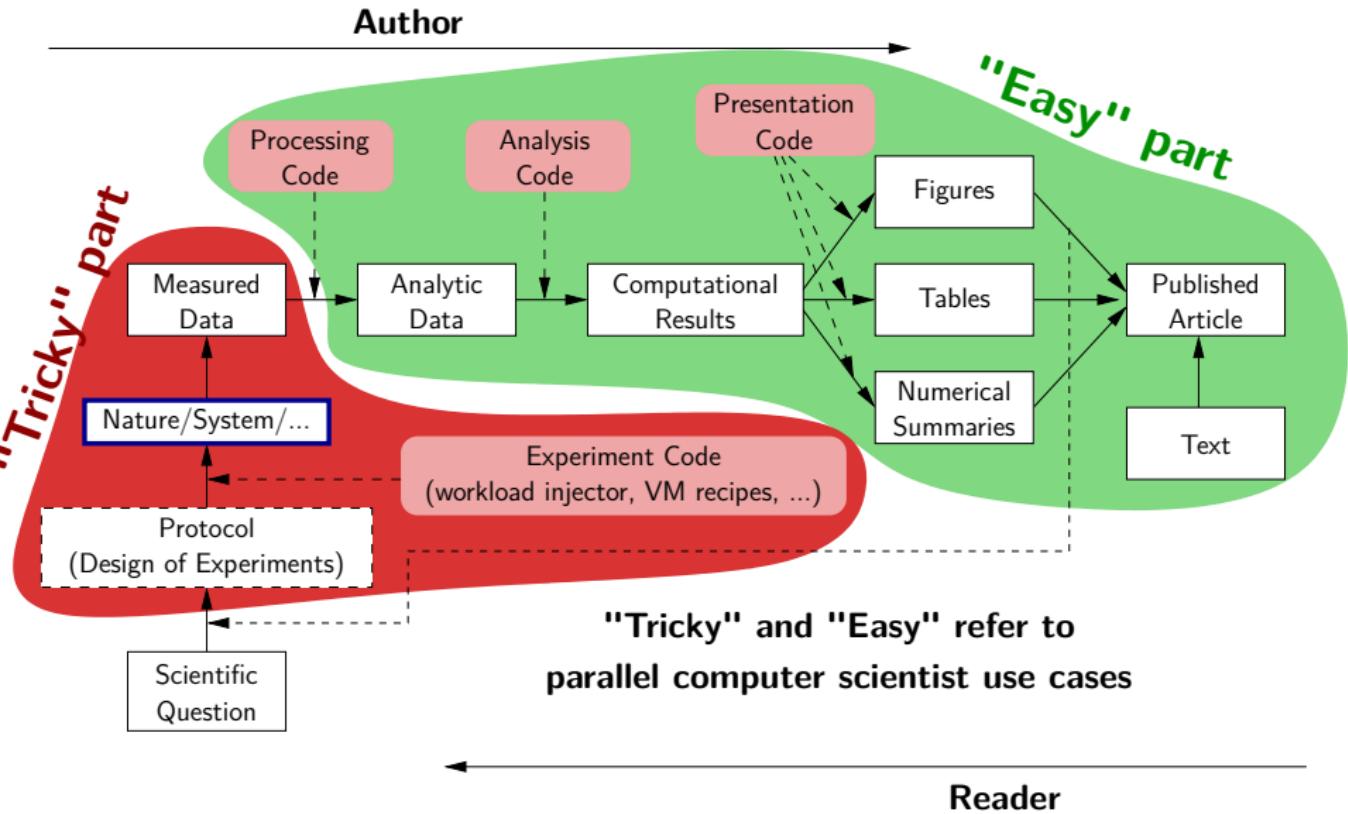
Inspired by Roger D. Peng's lecture on reproducible research, May 2014

# Reproducible Research: Trying to Bridge the Gap



Inspired by Roger D. Peng's lecture on reproducible research, May 2014

# Reproducible Research: Trying to Bridge the Gap



Inspired by Roger D. Peng's lecture on reproducible research, May 2014

# Paradigm Shift

- ① Lack of information, data access
  - ~~ *The right tracking/sharing tools*
- ② Lack of technical and scientific rigor
  - ~~ *Better methodology (e.g., training in stats)*
- ③ Computation/programming mistakes
  - ~~ *Continuous integration, code review, environment control, numerical analysis, ...*
- ④ Little incentive to care/share/...
  - ~~ *The contribution evaluation processs (publication, recruiting, ...) needs to evolve*



in  
c $\leftrightarrow$ de  
we  
trust;

Transparency increases the chances of finding mistakes  
and getting rid of them

# Outline

- ① Science crisis ?
- ② Is Computer Science Really Concerned With This?
- ③ Reproducible Research/Open Science in a Nutshell
- ④ Illustrating Nice Ideas Through Different Tools
  - Taking notes and explaining
  - Controlling your environment/experiments
  - Making data/code available
  - Changing evaluation practices
- ⑤ What can Computer Scientists do ?

# Outline

- ① Science crisis ?
- ② Is Computer Science Really Concerned With This?
- ③ Reproducible Research/Open Science in a Nutshell
- ④ Illustrating Nice Ideas Through Different Tools
  - Taking notes and explaining
  - Controlling your environment/experiments
  - Making data/code available
  - Changing evaluation practices
- ⑤ What can Computer Scientists do ?

# Ipython/Jupyter Notebook

Web app: create and share documents that contain live code, equations, visualizations, and explanatory text

The image shows a screenshot of the Jupyter Notebook interface. On the right, a window titled "jupyter Lorenz Differential Equations" is open, displaying a notebook page about the Lorenz system. The page includes text, mathematical equations, and a plot. On the left, a separate browser window also displays a "jupyter Welcome to P" page, which includes a warning message and a "Run some Python" section.

**jupyter** Lorenz Differential Equations (autosaved)

File Edit View Insert Cell Kernel Help

Code Cell Toolbar: None

## Exploring the Lorenz System

In this Notebook we explore the [Lorenz system](#) of differential equations:

$$\begin{aligned}\dot{x} &= \sigma(y - x) \\ \dot{y} &= \rho x - y - xz \\ \dot{z} &= -\beta z + xy\end{aligned}$$

This is one of the classic systems in non-linear differential equations. It exhibits a range of complex behaviors as the parameters ( $\sigma$ ,  $\beta$ ,  $\rho$ ) are varied, including what are known as chaotic solutions. The system was originally developed as a simplified mathematical model for atmospheric convection in 1963.

In [7]: `interact(Lorenz, N=fixed(10), angle=(0.,360.),  
sigma=(0.0,50.0),beta=(0.,5), rho=(0.0,50.0));`

angle: 308.2  
max\_time: 12  
 $\sigma$ : 10  
 $\beta$ : 2.6  
 $\rho$ : 28

A full tutorial for using the

In [ ]: `#matplotlib inline`

```
import pandas as pd
import numpy as np
import matplotlib
```

# Computational Document (Jupyter Notebook)

## Un document computationnel

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

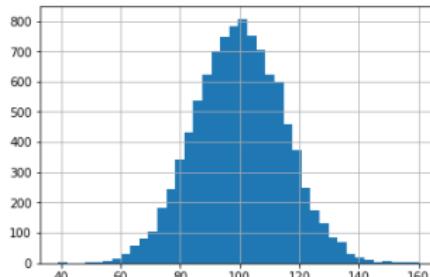
3.141592653589793

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... ☺).



# Computational Document (Jupyter Notebook)

Document initial dans son environnement

jupyter example\_pi (automated)

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

# Un document computationnel

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

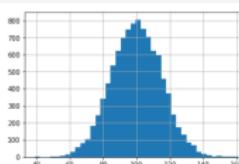
```
In [1]: from math import * print(pi) 3.141592653589793
```

Mais calculé avec la [méthode des aiguilles de Buffon](https://fr.wikipedia.org/wiki/Algouille_de_Buffon) ([https://fr.wikipedia.org/wiki/Algouille\\_de\\_Buffon](https://fr.wikipedia.org/wiki/Algouille_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  $\sqrt{2/\pi} \exp(-\frac{(x-\mu)^2}{2\sigma^2})$  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  $\pi$  vaut *approximativement*

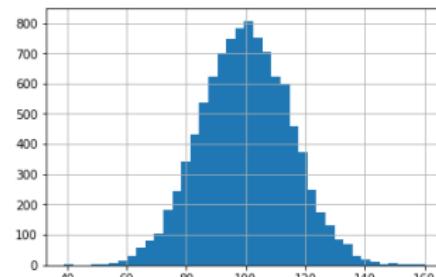
3.141592653589793

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... ☺).



# Computational Document (Jupyter Notebook)

Document initial dans son environnement



```
# Un document computationnel
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 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)

Out[2]: 3.1437198694098765

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

Mark Down

Document final

## Un document computationnel

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

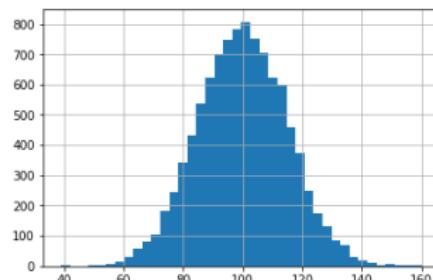
3.141592653589793

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

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...).



# Computational Document (Jupyter Notebook)

Document initial dans son environnement

jupyter example\_pi [untrusted]

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

# 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](https://fr.wikipedia.org/wiki/Aiguille_de_Buffon)), on obtiendrait comme approximation :

In [2]:

```
import numpy as np
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)
```

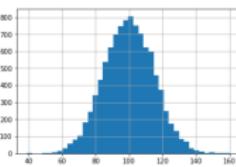
Out[2]: 3.1437198694098765

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  $\pi$  (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  $\pi$  vaut *approximativement*

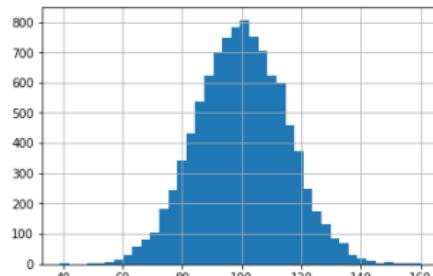
3.141592653589793

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... ☺).



# Computational Document (Jupyter Notebook)

Document initial dans son environnement

jupyter example\_pi [untrusted]

# 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](https://fr.wikipedia.org/wiki/Aiguille_de_Buffon)), on obtiendrait comme approximation :

In [2]:

```
import numpy as np
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)
```

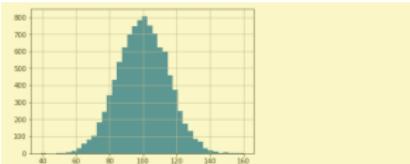
Out [2]: 3.1437198694098765

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  $\pi$  (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  $\pi$  vaut *approximativement*

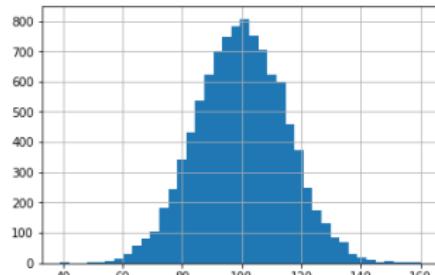
3.141592653589793

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... ☺).



Résultats

# Computational Document (Jupyter Notebook)

Document initial dans son environnement

jupyter example\_pi [untrusted]

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

# 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](#), on obtiendrait comme approximation :

In [2]:

```
import numpy as np
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]:

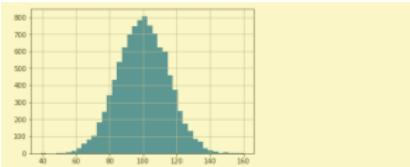
3.1437198694098765

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  $\pi$  (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  $\pi$  vaut *approximativement*

3.141592653589793

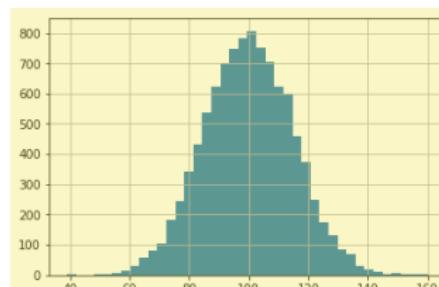
Mais calculé avec la [méthode des aiguilles de Buffon](#), on obtiendrait comme approximation :

```
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... ☺).



# Vistrails: a Workflow Engine for Provenance Tracking

## An *Provenance-Rich* Paper: ALPS2.0

The ALPS project release 2.0:  
Open source software for strongly correlated  
systems

B. Bauer<sup>1</sup> L. D. Carr<sup>2</sup> H.G. Evertz<sup>3</sup> A. Feiguin<sup>4</sup> J. Freire<sup>5</sup>  
S. Fuchs<sup>6</sup> L. Gamper<sup>1</sup> J. Gukelberger<sup>1</sup> E. Gulf<sup>7</sup> S. Guertler<sup>8</sup>  
A. Hehn<sup>1</sup> R. Igarashi<sup>9,10</sup> S.V. Isakov<sup>1</sup> D. Koop<sup>5</sup> P.N. Ma<sup>1</sup>  
P. Mates<sup>1,5</sup> H. Matsuo<sup>11</sup> O. Parcollet<sup>12</sup> G. Pawłowski<sup>13</sup>  
J.D. Picom<sup>14</sup> L. Pollet<sup>1,15</sup> E. Santos<sup>9</sup> V.W. Scarola<sup>16</sup>  
U. Schollwöck<sup>17</sup> C. Silva<sup>3</sup> B. Sturer<sup>1</sup> S. Todo<sup>10,11</sup> S. Trebst<sup>18</sup>  
M. Troyer<sup>1</sup> M. L. Wall<sup>19</sup> P. Werner<sup>1</sup> S. Wessel<sup>19,20</sup>

<sup>1</sup>Theoretische Physik, ETH Zurich, 8093 Zurich, Switzerland

<sup>2</sup>Department of Physics, Colorado School of Mines, Golden, CO 80401, USA

<sup>3</sup>Institut für Theoretische Physik, Technische Universität Graz, A-8010 Graz, Austria

<sup>4</sup>Department of Physics and Astronomy, University of Wyoming, Laramie, Wyoming 82071, USA

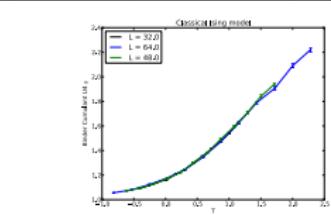
<sup>5</sup>Scientific Computing and Imaging Institute, University of Utah, Salt Lake City, Utah 84112, USA

<sup>6</sup>Institute for Theoretical Physics, Georg-August-Universität Göttingen, Göttingen, Germany

<sup>7</sup>Columbia University, New York, NY 10027, USA

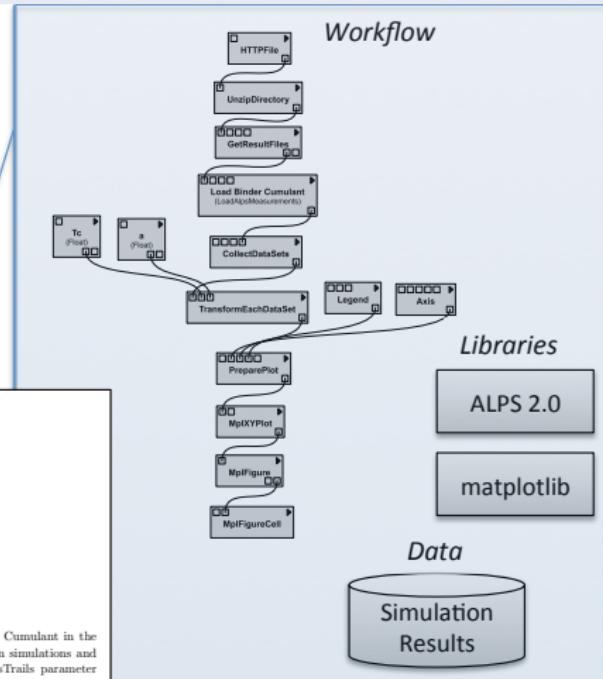
<sup>8</sup>Bethe Center for Theoretical Physics, Universität Bonn, Nussallee 12, 53115 Bonn, Germany

arXiv:1101.2646v4 [cond-mat.str-el] 23 May 2011



† Correspond-

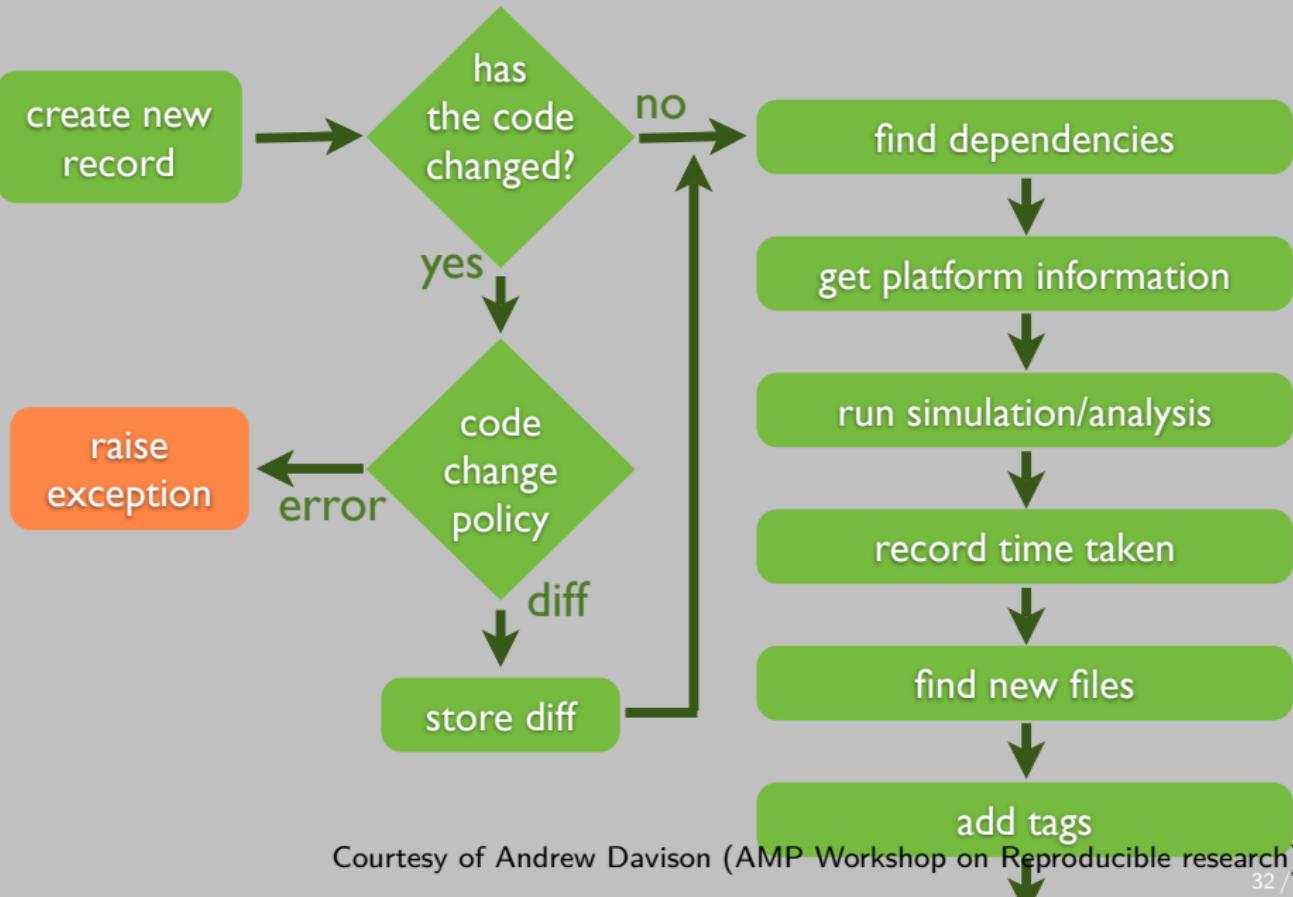
Figure 3. In this example we show a data collapse of the Binder Cumulant in the classical Ising model. The data has been produced by remotely run simulations and the critical exponent has been obtained with the help of the VisTrails parameter exploration functionality.



Courtesy of Juliana Freire (AMP Workshop on Reproducible research)

[Reynier et al., ISTAT 2011]

# Sumatra: an "experiment engine" that helps taking notes



Courtesy of Andrew Davison (AMP Workshop on Reproducible research)

# Sumatra: an "experiment engine" that helps taking notes

```
$ smt comment 20110713-174949 "Eureka! Nobel prize  
here we come."
```

# Sumatra: an "experiment engine" that helps taking notes

```
$ smt tag "Figure 6"
```

# Sumatra: an "experiment engine" that helps taking notes

Sumatra: TestProject: List of records

TestProject: List of records

Delete Include data	Label	Reason	Outcome	Duration	Processes	Simulator		Script			Date	Time	Tags
						Name	Version	Repository	Main file	Version			
<input type="checkbox"/>	<a href="#">20100709-154255</a>		'Eureka! Nobel prize here we come.'	0.59 s		Python	2.5.2	/Users/andrew/tmp/SumatraTest	main.py	396c2020ca50	09/07/2010	15:42:55	
<input type="checkbox"/>	<a href="#">20100709-154309</a>			0.59 s		Python	2.5.2	/Users/andrew/tmp/SumatraTest	main.py	396c2020ca50	09/07/2010	15:43:09	
<input type="checkbox"/>	<a href="#">haggling</a>	'determine whether the gourd is worth 3 or 4 shekels'	'apparently, it is worth NaN shekels.'	0.59 s		Python	2.5.2	/Users/andrew/tmp/SumatraTest	main.py	396c2020ca50	09/07/2010	15:43:20	<a href="#">foobar</a>
<input type="checkbox"/>	<a href="#">20100709-154338</a>	'test effect of a smaller time constant'		0.59 s		Python	2.5.2	/Users/andrew/tmp/SumatraTest	main.py	396c2020ca50	09/07/2010	15:43:38	
<input type="checkbox"/>	<a href="#">haggling_repeat</a>	Repeat experiment haggling	The new record exactly matches the original.	0.58 s		Python	2.5.2	/Users/andrew/tmp/SumatraTest	main.py	396c2020ca50	09/07/2010	15:43:47	

# My daily solution: Emacs/Org-mode

## My journal a single file

- Chronological organization with TAGS
- Heterogenous information
- Code execution is convenient

## My students' laboratory notebooks

- Conventions
- ≈ Chronological organization
- Track info about experiments (who, what, when, how, **why**) and data analysis

## Replicable articles

- Easily built on the previous documents
- Many hidden sections (data download, perl/R scripts)
- Sources connect figures with data and meta-data

# Outline

- ① Science crisis ?
- ② Is Computer Science Really Concerned With This?
- ③ Reproducible Research/Open Science in a Nutshell
- ④ Illustrating Nice Ideas Through Different Tools
  - Taking notes and explaining
  - Controlling your environment/experiments
  - Making data/code available
  - Changing evaluation practices
- ⑤ What can Computer Scientists do ?

# Complex Environments

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

```
1 2.1.1
```

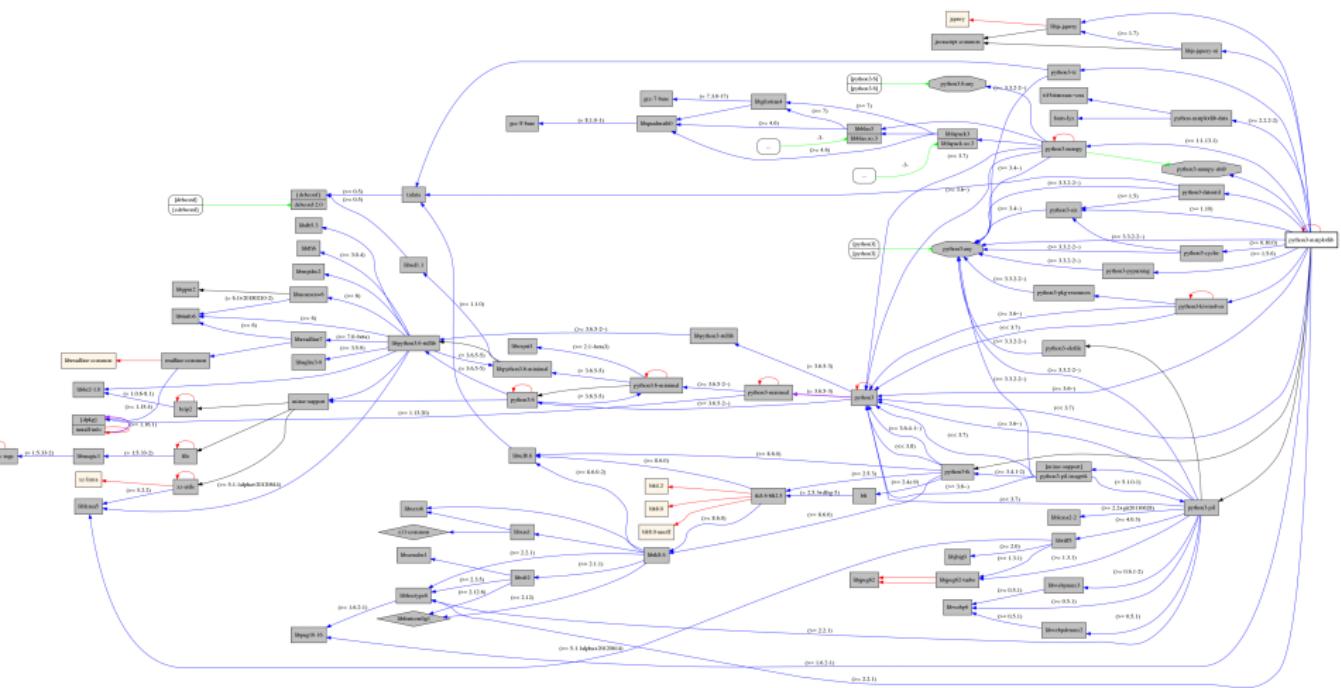
# Complex Environments

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

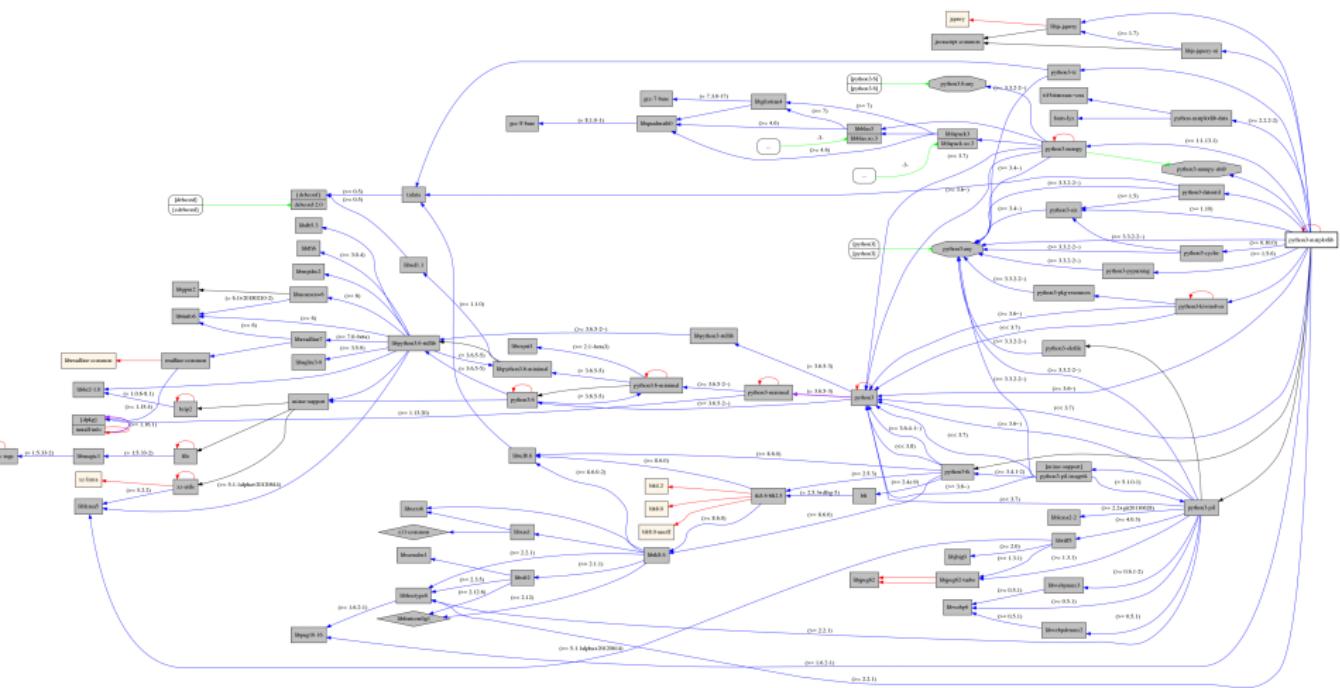
## 2.1.1

```
1 Version: 2.1.1-2  
2 Depends: python3-dateutil, python-matplotlib-data (>= 2.1.1-2),  
3 python3-pyparsing (>= 1.5.6), python3-six (>= 1.10), python3-tz,  
4 libjs-jquery, libjs-jquery-ui, python3-numpy (>= 1:1.13.1),  
5 python3-numpy-abi9, python3 (<< 3.7), python3 (>= 3.6~),  
6 python3-cycler (>= 0.10.0), python3:any (>= 3.3.2-2~), libc6 (>=  
7 2.14), libfreetype6 (>= 2.2.1), libgcc1 (>= 1:3.0), libpng16-16  
8 (>= 1.6.2-1), libstdc++6 (>= 5.2), zlib1g (>= 1:1.1.4)
```

# Complex Environments



# Complex Environments



*Virtual images or Docker containers* allow to freeze environments

Two approaches: *preserve the mess* vs. *encourage cleanliness*

# Reprozip

Automagically pack your experiment to fight **dependency hell**

## ON THE ORIGINAL MACHINE

```
$ pip install reprozip
$ reprozip trace ./myexperiment -my --options inputs/somefile.csv other_file_here.bin
experiment: 0%... 25%... 50%... 75%... 100%
result: 42.137
Configuration file written in .reprozip/config.yml
Edit that file then run the packer -- use 'reprozip pack -h' for help
$ reprozip pack my_experiment.rpz
[REPROZIP] 17:26:42.588 INFO: Creating pack my_experiment.rpz...
[REPROZIP] 17:26:42.589 INFO: Adding files from package coreutils...
[REPROZIP] 17:26:42.601 INFO: Adding files from package libc6...
[REPROZIP] 17:26:42.906 INFO: Adding other files...
[REPROZIP] 17:26:43.450 INFO: Adding metadata...
```

## ON ANOTHER MACHINE

```
$ pip install reprounzip[all]
$ reprounzip vagrant setup my_experiment.rpz mydirectory
Bringing machine 'default' up with 'virtualbox' provider...
==> default: Importing base box 'remram/debian-7-amd64'...
==> default: Booting VM...
==> default: Machine booted and ready!
==> default: Running provisioner: shell...
$ reprounzip vagrant run mydirectory
experiment: 0%... 25%... 50%... 75%... 100%
result: 42.137
$ reprounzip vagrant upload /tmp/new_config:global-config
$ reprounzip vagrant run mydirectory --cmdline ./myexperiment --other --options
inputs/somefile.csv
experiment: 0%... 25%... 50%... 75%... 100%
result: -17.814
```

## Other options:

CARE, CDE, Reprozip, Pipenv | Docker, Singularity | Spack, GUIX, NIX

# Controlling experiments

- Naive way: sh + ssh + ...
- Better way: use a workflow management system (taverna, galaxy, kepler, vistrails, ...)
- Parallel/distributed experiments require specific experiment engines
  - ▶ **Expo** (2007-, G5K)
  - ▶ **XPflow** (2012-, G5K)
  - ▶ **Execo** (2013-, G5K)
- Parallel/distributed experiments require specific experiment engines
  - ▶ Plush (2006-, Planetlab)
  - ▶ OMF (2009-, Wireless)
  - ▶ Splay (2008), ...

although nothing  
specific to G5K

They differ in the underlying paradigms and the platforms for which they have been designed

A survey of general-purpose experiment management tools for distributed systems, T. Buchert, C. Ruiz, L. Nussbaum, O. Richard, FGCS, 2014

- Control your **numerical results** (random generators, libraries, rounding and non-determinism, ...)

# Outline

- ① Science crisis ?
- ② Is Computer Science Really Concerned With This?
- ③ Reproducible Research/Open Science in a Nutshell
- ④ Illustrating Nice Ideas Through Different Tools
  - Taking notes and explaining
  - Controlling your environment/experiments
  - Making data/code available
  - Changing evaluation practices
- ⑤ What can Computer Scientists do ?

# Hosting platforms

- Your webpage 😞
- Articles: Arxiv, HAL
- Code: Github or Institutional Gitlab
- Figshare, Zenodo 😊
- Companion websites ([elsevier executable paper](#) 😞, [runmycode](#),  
[exec&share](#), [code ocean](#) ...)



This may seem easy but is quite tricky:

- Arbitrary limits can make your life painful
- Perennity ([Roberto Di Cosmo's talks](#))
  - CodeSpaces murdered on Amazon, Google Code termination, Gitorious shutdown, ...
  - Disruption of the web of reference: URLs decay (half-life of 4 years), DOIs have little guarantee, ...



# Outline

- ① Science crisis ?
- ② Is Computer Science Really Concerned With This?
- ③ Reproducible Research/Open Science in a Nutshell
- ④ Illustrating Nice Ideas Through Different Tools
  - Taking notes and explaining
  - Controlling your environment/experiments
  - Making data/code available
  - Changing evaluation practices
- ⑤ What can Computer Scientists do ?

# Publishing initiatives

Artifact evaluation and ACM badges Article appendix (SC)



Platforms Elsevier executable paper, ACM pushes for code ocean  
Double blind vs. Open reviews plus "legal aspects"  
RR-oriented & Open Journals IPOL



# Funding Agencies

**US** *NSF has released a new announcement encouraging submission of research projects relating to **reproducibility and replicability**.*

– Jim Kurose, Oct. 2016

*The NIH is committed to promoting **rigorous and transparent research** in all areas of science supported by a variety of grant programs.*

– NIH: Rigor and Reproducibility guidelines

**Europe** *In 2012, the European Commission encouraged all EU Member States to **put public-funded research results in the public sphere** in order to make science better and strengthen their knowledge-based economy.*

– Open Science (Open Access)

*A key element will be capacity building to link literature and data in order to enable a more transparent evaluation of research and **reproducibility** of results.*

– H2020-EINFRA-2014-2015

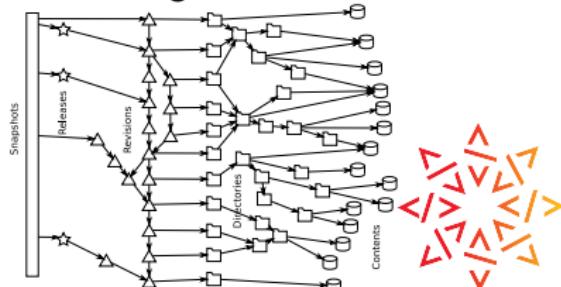
*Encouraging the re-use of research data generated by publicly funded research projects* – Science with and for Society, 27 October 2017

# Outline

- ① Science crisis ?
- ② Is Computer Science Really Concerned With This?
- ③ Reproducible Research/Open Science in a Nutshell
- ④ Illustrating Nice Ideas Through Different Tools
  - Taking notes and explaining
  - Controlling your environment/experiments
  - Making data/code available
  - Changing evaluation practices
- ⑤ What can Computer Scientists do ?

# On the "technical" side (1/2)

- Better documenting what we do: **Laboratory notebooks**
  - Literate programming is great for analysis, and reproducible articles but does not go well yet with conducting experiments and workflows
  - A real adoption of such practice requires more storage and the ability to navigate in such information
- Better software engineering practice: Public releases, **devops approach**, reproducible builds, numerical aspects
  - Moving/evolving technology. Preservation ? Adoption ?
  - Should not slow down research
- Fighting against software/data degradation: **Software Heritage**, zenodo
  - Challenges: multiple! curation, access/privacy, exploitation, navigation, storage, ...



**Software Heritage**  
THE GREAT LIBRARY OF SOURCE CODE

## On the "technical" side (2/2)

- Better experimental practice and platforms: **FIT IoT-lab, G5K** are world leading experimental infrastructures; rely on standard simulators (**SimGrid, NS3**)
  - Maintenance cost, keeping in pace with technology, practices for prototype platforms, control, sharing of experimental conditions with others, experimental engines



- Need for convergence (SILEX) in term of software infrastructure and practice (e.g., security, account management, access, isolation, experiment management, etc.) ?

## On the "social" side

Slight **cultural changes** in our relation to publication and **daily practice**

- Changing our social model to favor adoption of better practice
  - Artifact evaluation, open reviews, ... (e.g., IPOL, ReScience)
  - Alternative evaluation models (publishing, funding, ...)
- **Learning** is the essence of our work. ~Train our researchers and students
  - MOOC on Reproducible Research in Oct. 2018 on FUN
  - Train in statistics, experimental practice, design of experiments

It's up to us. We should care and take the lead

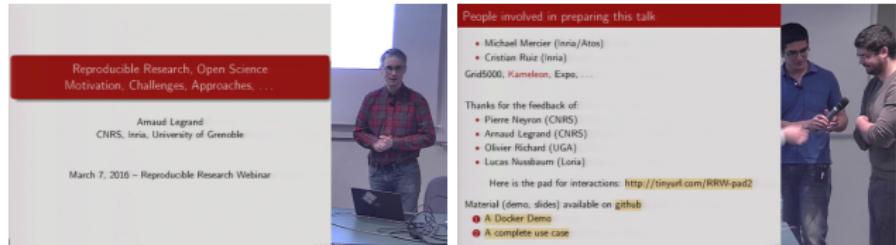
Main benefit:

- **Higher confidence** in our work ~definite competitive advantage
- Our research becomes **sound, deeper, auditable, more visible, reusable**

# Webinars: Learning by Doing

Many different tools/approaches developed in various communities

- ① Replicable article
- ② Logging your activity
- ③ Logging and backing up your data
- ④ Organizing your data
- ⑤ Mastering your environment
- ⑥ Controlling your experiments
- ⑦ Making your data/code/article available
- ⑧ Publication modes
- ⑨ Artifact Evaluation



Literate programming

Controlling your environment



Reproducible Research, Open Science  
Logging and backing up your work  
Git Tips and Tricks, a Scientist Perspective

V. Danjean, A. Legrand, L. Stanisic  
University of Grenoble, CNRS, Inria Bordeaux

June 7, 2016 – Reproducible Research Webinar (Episode IV)

Numerical reproducibility

Logging and backing up

[https://github.com/alegrand/RR\\_webinars](https://github.com/alegrand/RR_webinars)