

Reproducible Research: Where Do We Stand?

Arnaud Legrand
CNRS, Inria, University of Grenoble

November 9th, 2017 – LIRIS, Lyon

Outline

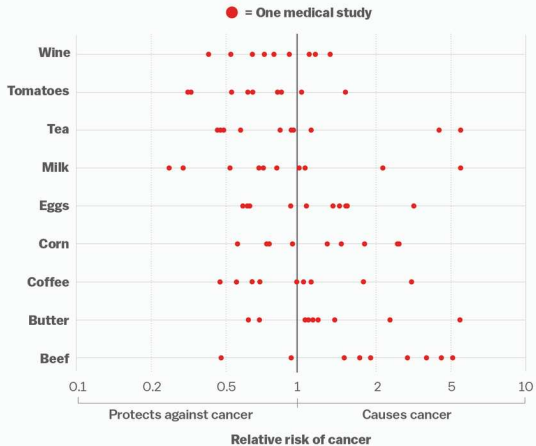
- ① Science crisis ?
- ② How is CS Concerned Really With This?
- ③ Reproducible Research/Open Science in a Nutshell
- ④ Illustrating Nice Ideas Through Different Tools
- ⑤ And In Practice?
- ⑥ What can Computer Scientists do ?

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.ORG | FEEDBACK | HELP | LIBRARIES | All Science Journals | Search | Alerts | Access | RSS

NEWS | SCIENCE JOURNALS | CAREERS | MULTIMEDIA | COLLECTIONS

The World's Leading Journal of Original Scientific Research, Global News, and Commentary

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

Home > Science Magazine > 32 January 2014 > Mchurt, 343 (6160): 229

Article Views: Science 17 January 2014; Vol. 343 no. 6168 p. 229 DOI: 10.1126/science.1250475

Summary | Full Text | Full Text (PDF)

EDITORIAL

Reproducibility

Marcia McHurt

Marcia McHurt is Editor-in-Chief of Science.

Science advances on a foundation of trusted data. An approach that scientists use to gain confidence in results was shaken by reports that a troubling number of research findings are not reproducible. Because confidence in results is the foundation of science, we are announcing new initiatives to ensure the quality of scientific research. For preclinical studies (one of the largest areas of research), we will require authors to provide recommendations of the U.S. National Institute of Standards and Technology (NIST) on how to ensure a sufficient signal-to-noise ratio, which will ensure that the results are not the product of the experimenter's bias or the conduct of the experiment.

Save to My Folders | Download Citation | Alert Me When Article is Cited | Post to CiteLike | E-mail This Page | Rights & Permissions | Commercial Reprints and E-Prints | View PubMed Citation | Related Content

TheScientist
EXPLORING LIFE. INSPIRING INNOVATION

NIH Tackles Irreproducibility

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

By Jeff Akst | January 28, 2014

Announcement: Reducing our irreproducibility - Nature News & Comment

nature.com | Stamp | Login | Register

nature International weekly journal of science

Home | News & Comment | Research | Careers & Jobs | Current Issue | Archive | Audio & Video | For Authors

Archive > Volume 496 > Issue 7460 > 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 have highlighted the reliability and reproducibility of published research (collected and

The Economist

Washington's longer 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.

Endless

nature International weekly journal of science

Menu | Advanced search | Search

archive - volume 483 - issue 7391 - editorials - article

NATURE | EDITORIAL

Must try harder

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

Announcement: Reducing our irreproducibility: Nature News & Comment
www.nature.com/news/announcement-reduc
nature.com | Signup Login | Register



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.

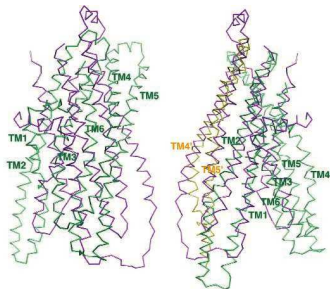
Yet, a scientific debate has at least been possible.



- 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 Escheria 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 retracts 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.

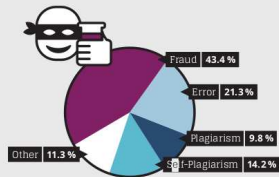
Scientific fraud is bad but let's be careful Have a look at the wikipedia *list of academic scandals*.

Is Fraud a new phenomenon?

The Battle against Scientific Fraud in the CNRS International Magazine

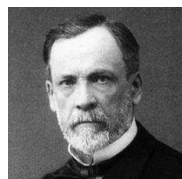
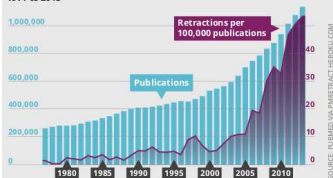
Biomedical fraud in figures

Cause of retraction 1977 to 2012



Number of publications and retractions

1977 to 2013



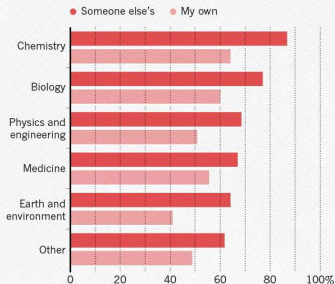
Galileo (data fabrication), Ptolemy (plagiarism), Mendel (data enhancement), **Pasteur** (rigorous but hid failures), ...

Is it only a matter of Fraud ?

Why are scientific studies so difficult to reproduce?

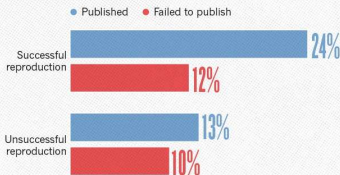
HAVE YOU FAILED TO REPRODUCE AN EXPERIMENT?

Most scientists have experienced failure to reproduce results.



HAVE YOU EVER TRIED TO PUBLISH A REPRODUCTION ATTEMPT?

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



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

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, ?..

Questioning previous work is part of the scientific process

Just like honesty, rigor and transparency. . .

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

Outline

- ① Science crisis ?
- ② How is CS Concerned Really With This?
- ③ Reproducible Research/Open Science in a Nutshell
- ④ Illustrating Nice Ideas Through Different Tools
- ⑤ And In Practice?
- ⑥ What can Computer Scientists do ?

Computational science!

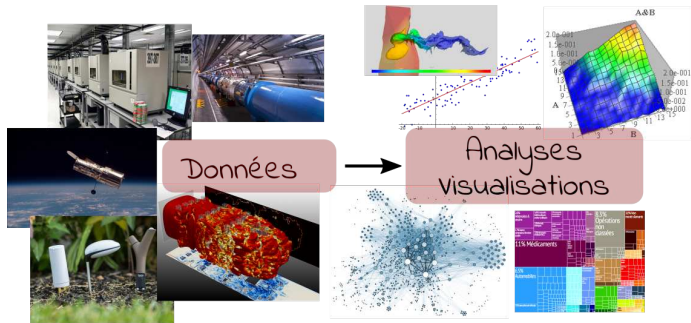


Données

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

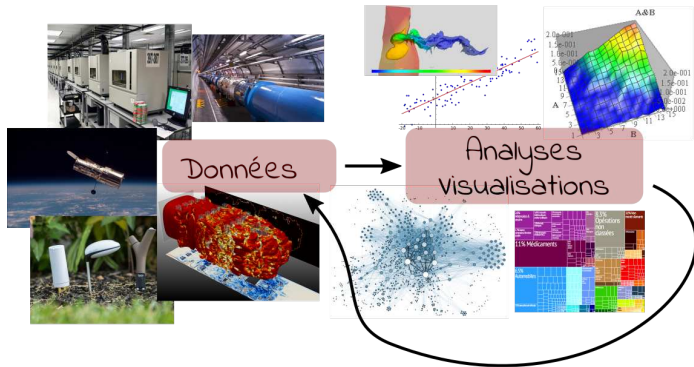
Computational science!



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

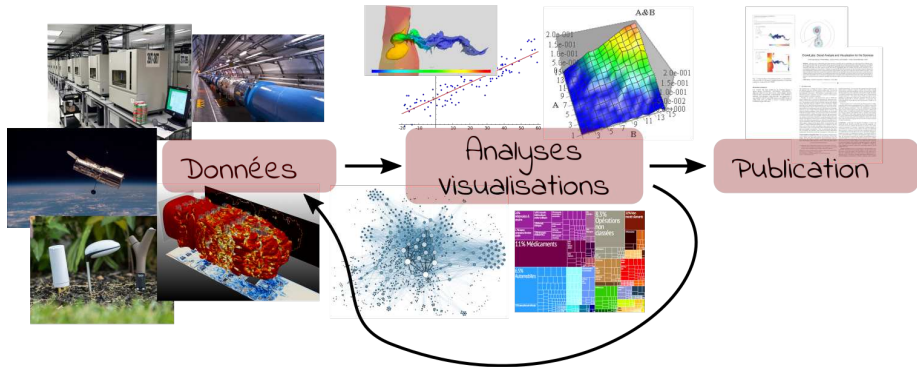
Computational science!



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

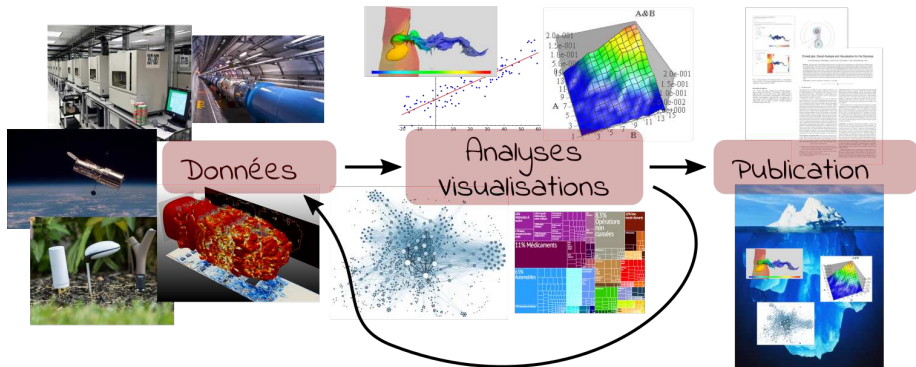
Computational science!



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 Protein 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).

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

– Dijkstra

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

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

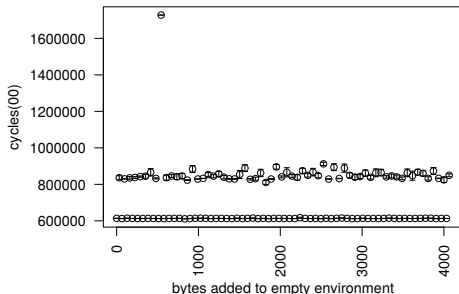
Experimenting with computers

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

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)

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

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)

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

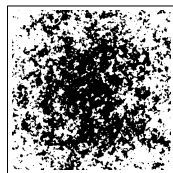
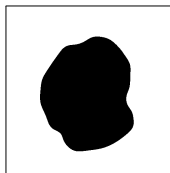
Machine Learning: Trouble at the lab, The Economist 2013



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)

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



All I care about is the algorithm output

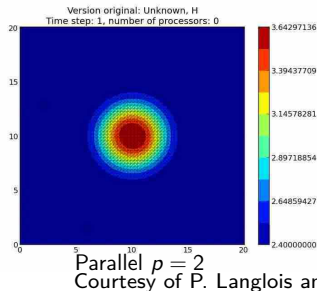
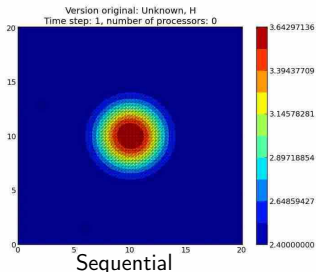
Did I mention we have **parallel machines** nowadays? 😊

Telemac2D: the simplest goutedo simulation

The goutedo 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, ...)



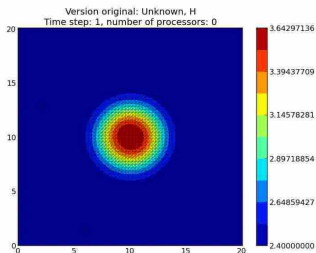
All I care about is the algorithm output

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

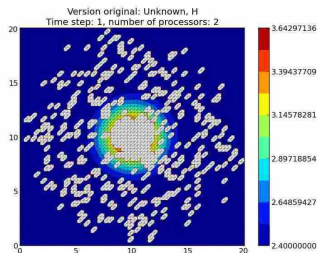
A white plot displays a non-reproducible value

Numerical reproducibility?

time step = 1



Sequential



Parallel $p = 2$

Courtesy of P. Langlois and R. Nheili

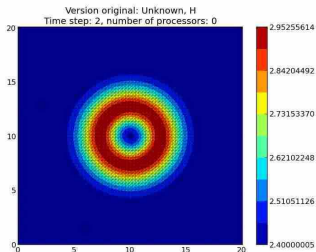
All I care about is the algorithm output

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

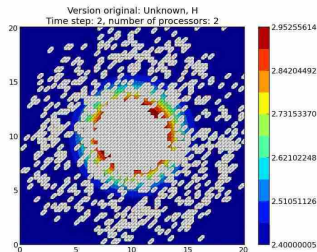
A white plot displays a non-reproducible value

Numerical reproducibility?

time step = 2



Sequential



Parallel $p = 2$

Courtesy of P. Langlois and R. Nheili

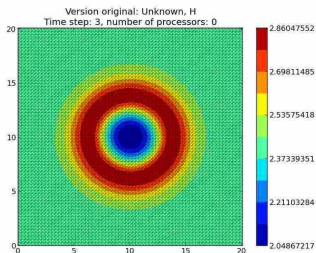
All I care about is the algorithm output

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

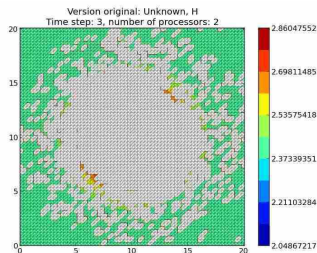
A white plot displays a non-reproducible value

Numerical reproducibility?

time step = 3



Sequential



Parallel $p = 2$

Courtesy of P. Langlois and R. Nheili

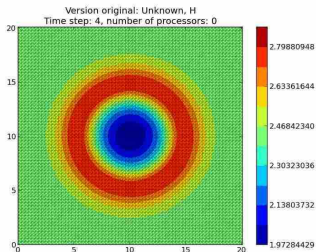
All I care about is the algorithm output

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

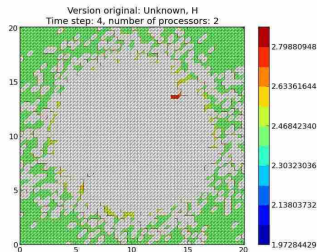
A white plot displays a non-reproducible value

Numerical reproducibility?

time step = 4



Sequential



Parallel $p = 2$

Courtesy of P. Langlois and R. Nheili

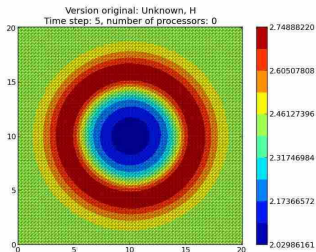
All I care about is the algorithm output

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

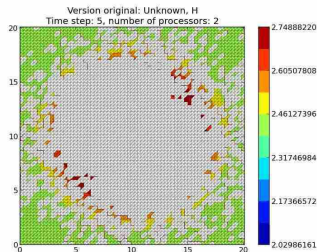
A white plot displays a non-reproducible value

Numerical reproducibility?

time step = 5



Sequential



Parallel $p = 2$

Courtesy of P. Langlois and R. Nheili

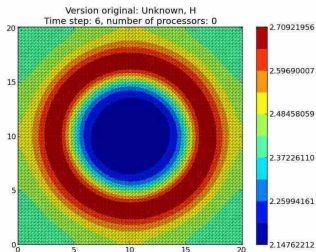
All I care about is the algorithm output

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

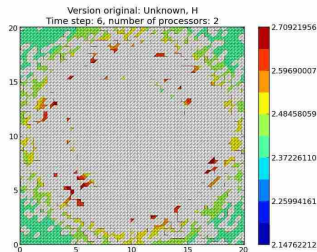
A white plot displays a non-reproducible value

Numerical reproducibility?

time step = 6



Sequential



Parallel $p = 2$

Courtesy of P. Langlois and R. Nheili

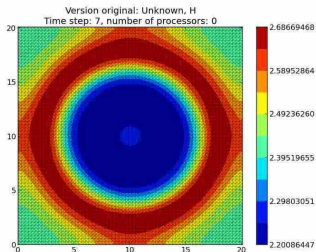
All I care about is the algorithm output

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

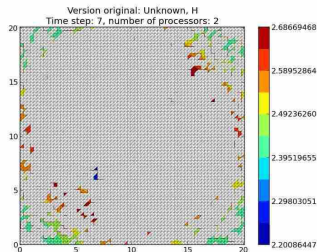
A white plot displays a non-reproducible value

Numerical reproducibility?

time step = 7



Sequential



Parallel $p = 2$

Courtesy of P. Langlois and R. Nheili

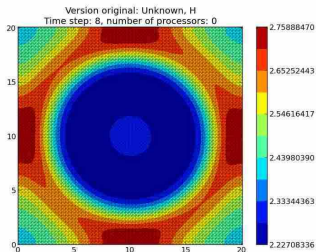
All I care about is the algorithm output

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

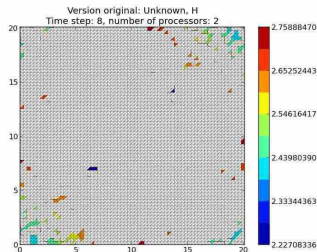
A white plot displays a non-reproducible value

Numerical reproducibility?

time step = 8



Sequential



Parallel $p = 2$

Courtesy of P. Langlois and R. Nheili

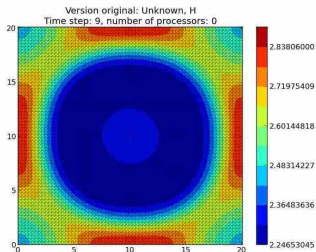
All I care about is the algorithm output

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

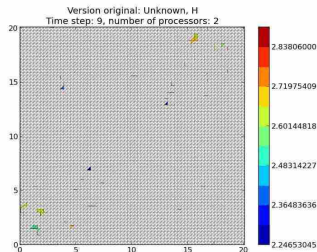
A white plot displays a non-reproducible value

Numerical reproducibility?

time step = 9



Sequential



Parallel $p = 2$

Courtesy of P. Langlois and R. Nheili

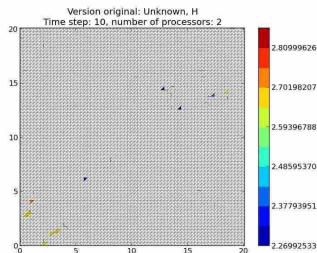
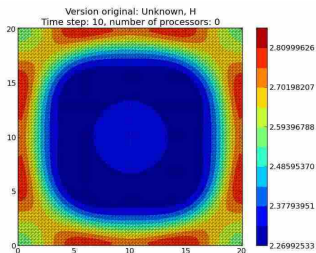
All I care about is the algorithm output

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

A white plot displays a non-reproducible value

Numerical reproducibility?

time step = 10



Courtesy of P. Langlois and R. Nheili

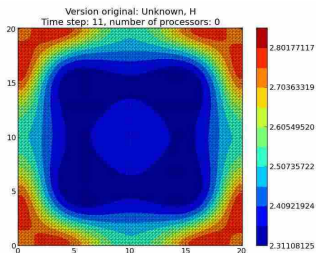
All I care about is the algorithm output

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

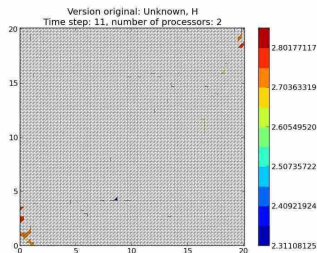
A white plot displays a non-reproducible value

Numerical reproducibility?

time step = 11



Sequential



Parallel $p = 2$

Courtesy of P. Langlois and R. Nheili

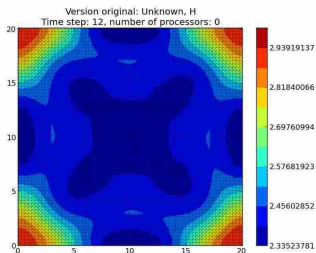
All I care about is the algorithm output

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

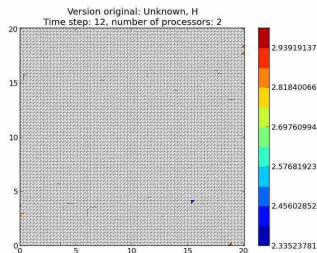
A white plot displays a non-reproducible value

Numerical reproducibility?

time step = 12



Sequential



Parallel $p = 2$

Courtesy of P. Langlois and R. Nheili

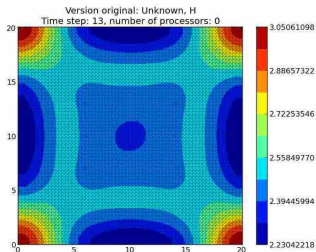
All I care about is the algorithm output

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

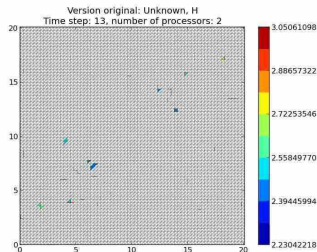
A white plot displays a non-reproducible value

Numerical reproducibility?

time step = 13



Sequential



Parallel $p = 2$

Courtesy of P. Langlois and R. Nheili

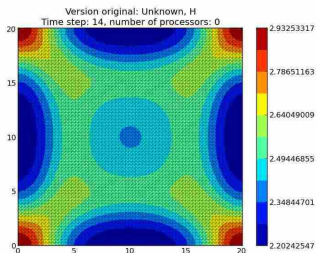
All I care about is the algorithm output

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

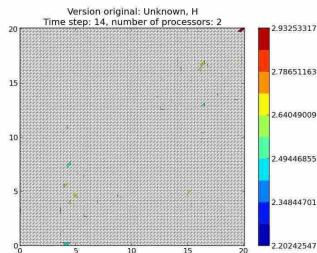
A white plot displays a non-reproducible value

Numerical reproducibility?

time step = 14



Sequential



Parallel $p = 2$

Courtesy of P. Langlois and R. Nheili

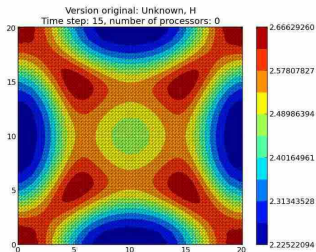
All I care about is the algorithm output

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

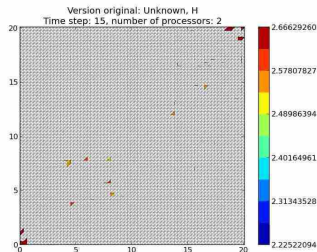
A white plot displays a non-reproducible value

NO numerical reproducibility!

time step = 15



Sequential



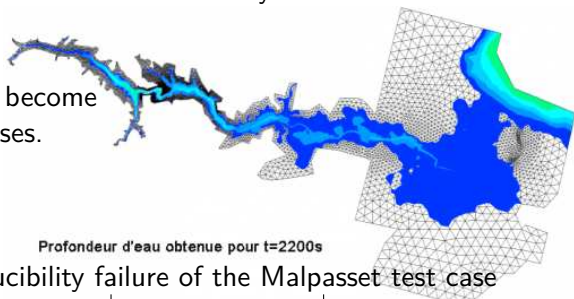
Parallel $p = 2$

Courtesy of P. Langlois and R. Nheili

All I care about is the algorithm output

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

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



Profondeur d'eau obtenue pour t=2200s

TABLE 1.1: Reproducibility failure of the Malpasset test case

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

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

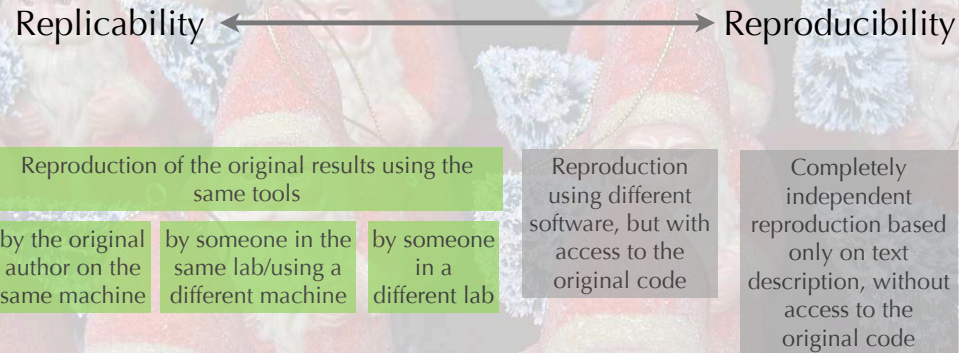
The whole chain needs to be revisited.

Courtesy of P. Langlois and R. Nheili

Outline

- ① Science crisis ?
- ② How is CS Concerned Really With This?
- ③ Reproducible Research/Open Science in a Nutshell
- ④ Illustrating Nice Ideas Through Different Tools
- ⑤ And In Practice?
- ⑥ What can Computer Scientists do ?

Reproducibility: What Are We Talking About?



Courtesy of Andrew Davison (AMP Workshop on Reproducible research)

Reproducible Research: Trying to Bridge the Gap

Author

Published
Article

Nature/System/...

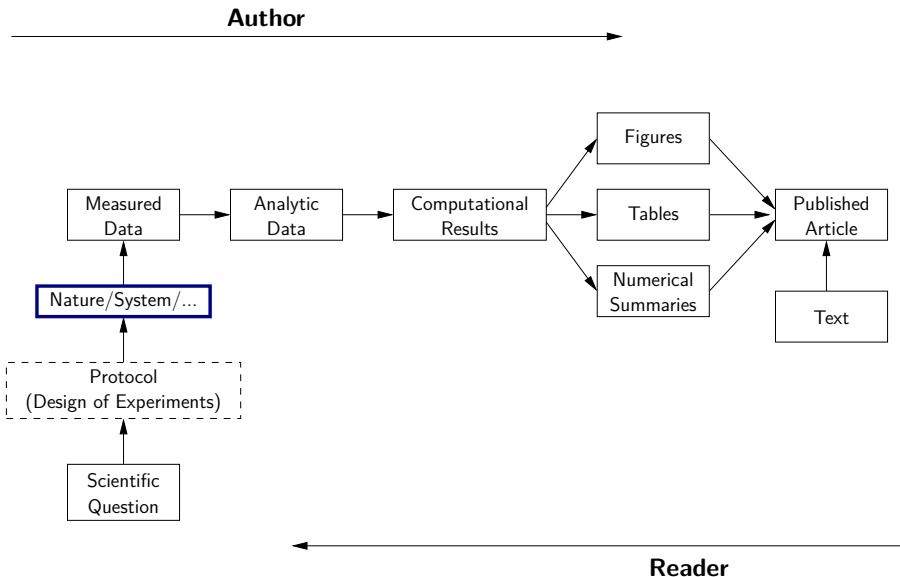
Protocol
(Design of Experiments)

Scientific
Question

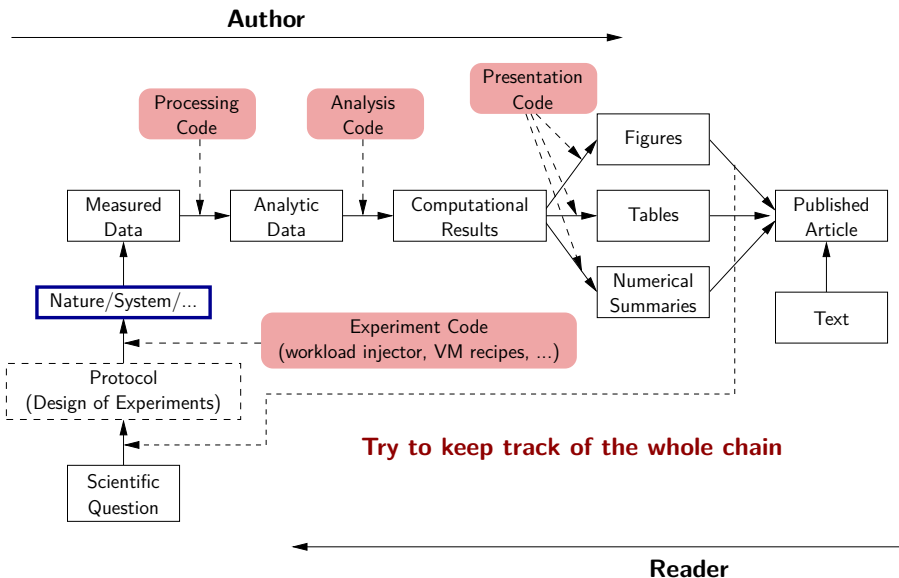
Reader

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

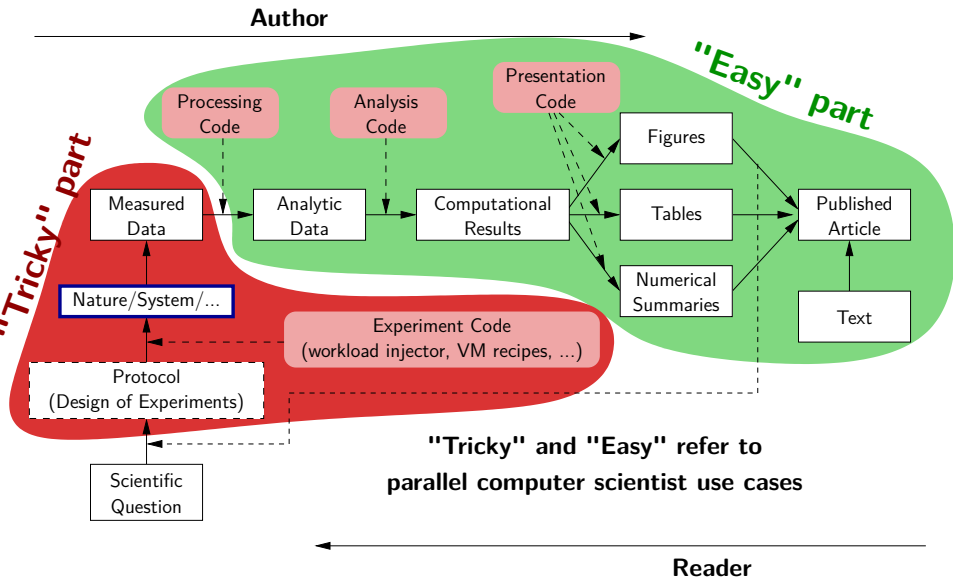
Reproducible Research: Trying to Bridge the Gap



Reproducible Research: Trying to Bridge the Gap



Reproducible Research: Trying to Bridge the Gap



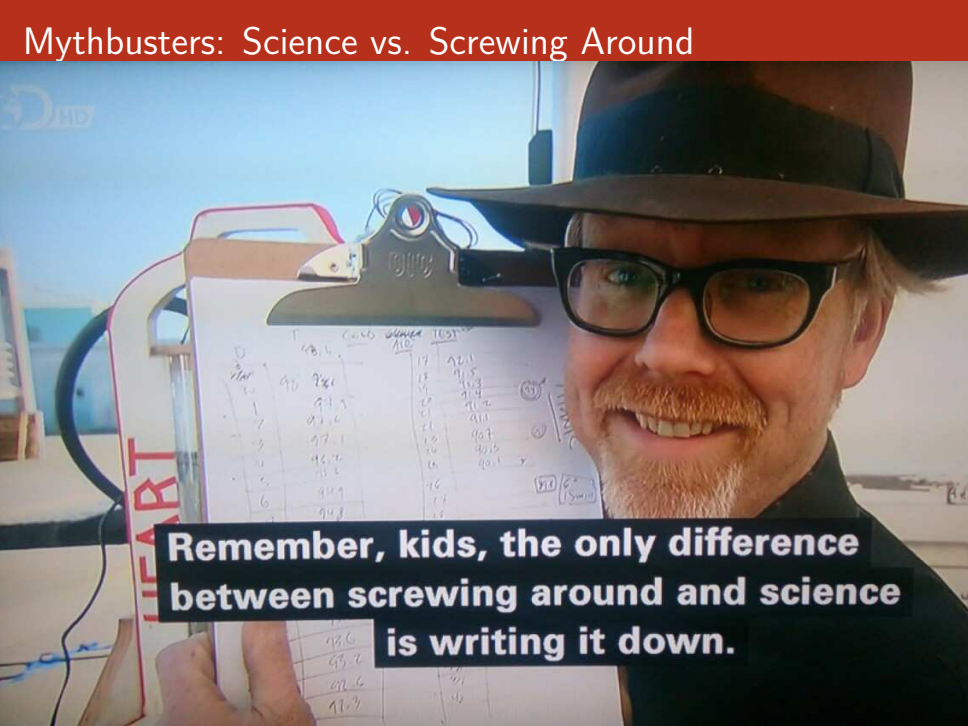
Paradigm Shift

- ❶ Lack of information, data access
- ❷ Computation/programming/statistics mistakes
- ❸ Lack of technical and scientific rigor



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

Mythbusters: Science vs. Screwing Around

A man with a beard, wearing a black hat and glasses, is smiling and holding a clipboard. The clipboard has a silver clip at the top and contains a piece of paper with handwritten data. The data is organized into two columns. The left column has a header 'D' and a sub-header '1000' followed by a list of numbers: 1, 2, 3, 4, 5, 6, 7. The right column has a header '1000' followed by a list of numbers: 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30. The numbers in the right column are written in a cursive script. There are also some circled numbers and a small box with the word 'Screw' inside. The background is a plain wall with a light switch and some wires.

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

Outline

- ① Science crisis ?
- ② How is CS Concerned Really With This?
- ③ Reproducible Research/Open Science in a Nutshell
- ④ Illustrating Nice Ideas Through Different Tools
- ⑤ And In Practice?
- ⑥ What can Computer Scientists do ?

Un document computationnel

Mon ordinateur m'indique que π vaut *approximativement*

3.141592653589793

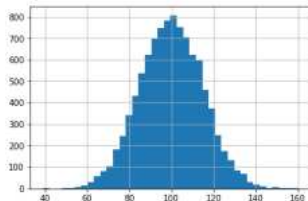
Mais calculé avec la **méthode des aiguilles de Buffon**, on obtiendrait comme **approximation** :

```
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

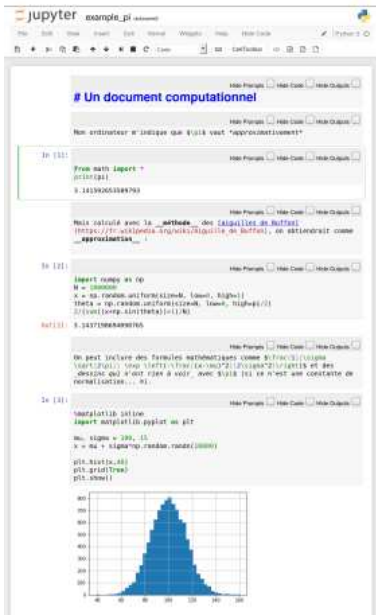
On peut inclure des formules mathématiques comme $\frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$ et

des dessins qui n'ont rien à voir avec π (si ce n'est une constante de normalisation... ☹).



Computational Document

Document initial dans son environnement



The screenshot shows a Jupyter Notebook interface with a title bar "jupyter example_pi.ipynb". The notebook contains four cells:

- Cell 0:** A comment "# Un document computationnel".
- Cell 1:** A comment "Nom attribuer n'indique que si'il s'agit d'un 'approximativement'".
- Cell 2:** Python code to import math and print pi, followed by the output "3.141592653589793".
- Cell 3:** A comment explaining the Buffon's needle method, followed by Python code using numpy to simulate the method, and the output "3.1437198694098765".
- Cell 4:** A comment about including mathematical formulas, followed by code to use matplotlib and pyplot to create a histogram, and the resulting histogram plot.

Document final

Un document computationnel

Mon ordinateur m'indique que π vaut approximativement

3.141592653589793

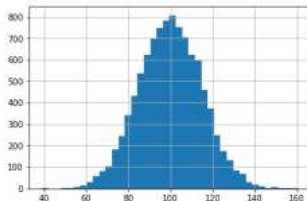
Mais calculé avec la **méthode des aiguilles de Buffon**, on obtiendrait comme **approximation** :

```
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... \propto).



Computational Document

Document initial dans son environnement



Mark Down

Document final

Un document computationnel

Mon ordinateur m'indique que π vaut approximativement

3.141592653589793

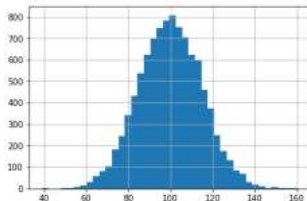
Mais calculé avec la **méthode des aiguilles de Buffon**, on obtiendrait comme **approximation** :

```
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... \odot).



Computational Document

Document initial dans son environnement



The screenshot shows a Jupyter Notebook interface with the following cells:

- Cell 0: A title "# Un document computationnel" and a comment "Mon ordinateur m'indique que π vaut approximativement 3.141592653589793".
- Cell 1: Code to print the value of π : `from math import pi; print(pi)`. The output is 3.141592653589793.
- Cell 2: A comment about the Buffon's needle method: "Puis calculé avec la méthode des aiguilles de Buffon".
- Cell 3: Code to import numpy and generate random numbers: `import numpy as np; N = 1000000; x = np.random.uniform(low=0, high=1); theta = np.random.uniform(low=0, high=pi/2); 2/(sum((x+np.sin(theta))>1)/N)`. The output is 3.143710664995705.
- Cell 4: A comment about plotting: "On peut inclure des formules mathématiques comme $\frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$ ".
- Cell 5: Code to plot a normal distribution: `from matplotlib.pyplot import plt; mu, sigma = 100, 15; x = np.linspace(mu-sigma*4, mu+sigma*4, 10000); plt.hist(x, 100); plt.xlabel('x'); plt.ylabel('N(x)')`. The output is a histogram of a normal distribution.

Code

Document final

Un document computationnel

Mon ordinateur m'indique que π vaut approximativement

3.141592653589793

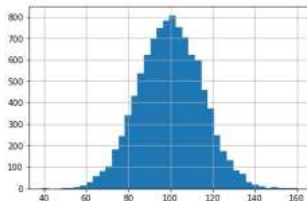
Mais calculé avec la **méthode des aiguilles de Buffon**, on obtiendrait comme approximation :

```
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

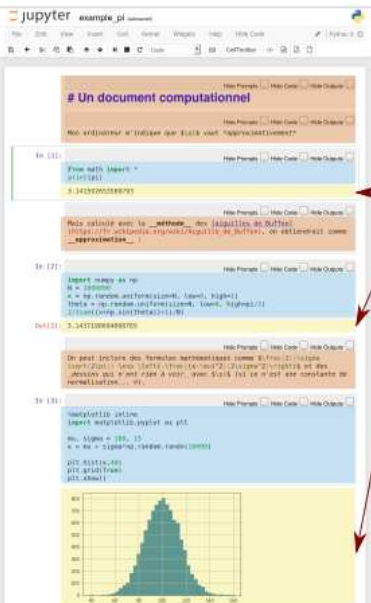
On peut inclure des formules mathématiques comme $\frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$ et

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



Computational Document

Document initial dans son environnement



The screenshot shows a Jupyter Notebook interface with the title "jupyter exemple_01.ipynb". The notebook contains several code cells:

- Cell 0: A comment in French: "# Un document computationnel".
- Cell 1: A comment in French: "Mon ordinateur m'indique que π vaut approximativement 3.141592653589793".
- Cell 2: A comment in French: "Mais calculé avec la méthode des aiguilles de Buffon, on obtiendrait comme approximation :".
- Cell 3: Python code to import numpy and calculate an approximation of pi using the Buffon's needle method. The output is 3.1437198694098765.
- Cell 4: A comment in French: "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... \odot)".
- Cell 5: Python code to create a plot of a normal distribution. The output is a histogram with a normal distribution curve overlaid.

Résultats

Document final

Un document computationnel

Mon ordinateur m'indique que π vaut approximativement

3.141592653589793

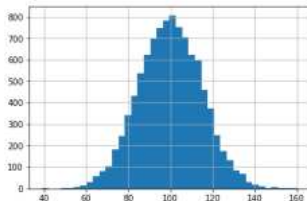
Mais calculé avec la **méthode des aiguilles de Buffon**, on obtiendrait comme approximation :

```
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... \odot).



Computational Document

Document initial dans son environnement

Jupyter example.py

```
# Un document computationnel

Mon ordinateur m'indique que  $\pi$  vaut approximativement
3.141592653589793


In [10]:
from math import *
print(pi)
3.141592653589793

Puis calculé avec la méthode des aiguilles de Buffon
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...  $\propto$ ).

In [11]:
import matplotlib.pyplot as plt
mu, sigma = 100, 15
x = mu + sigma*np.random.randn(10000)

plt.hist(x)
plt.xlabel('x')
plt.ylabel('count')
```



A histogram showing the distribution of 10,000 random values generated from a normal distribution with mean 100 and standard deviation 15. The x-axis is labeled 'x' and ranges from 40 to 160. The y-axis is labeled 'count' and ranges from 0 to 800. The distribution is bell-shaped and centered around 100.

Export

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)
```

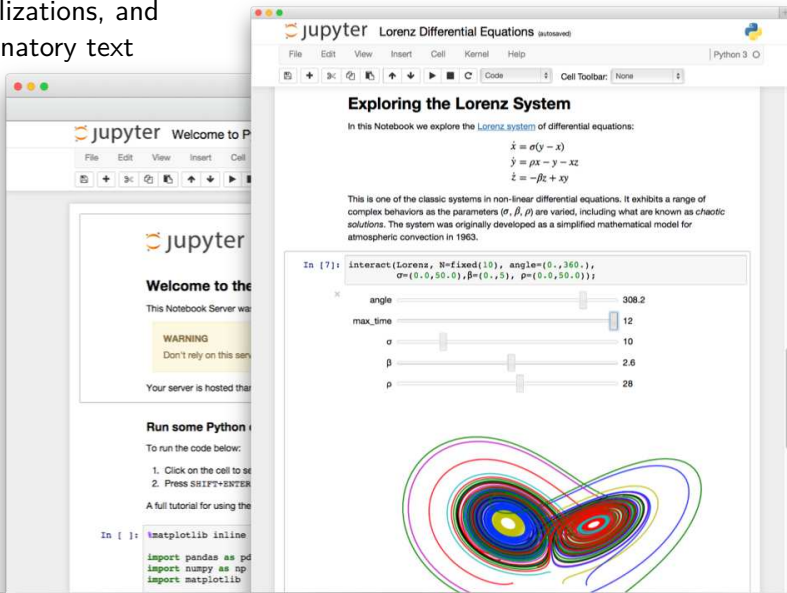
3.1437198694098765

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... \propto).



Ipython/Jupyter Notebook

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



jupyter Lorenz Differential Equations (autosaved)

File Edit View Insert Cell Kernel Help Python 3

Exploring the Lorenz System

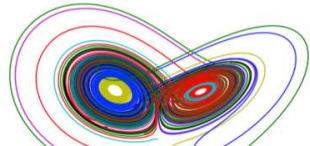
In this Notebook we explore the [Lorenz system](#) of differential equations:

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

This is one of the classic systems in non-linear differential equations. It exhibits a range of complex behaviors as the parameters (σ, β, ρ) 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=fix(10), angle=(0.,360.),
sigma=(0.0,50.0),beta=(0.,5), rho=(0.0,50.0));`

angle 308.2
max_time 12
 σ 10
 β 2.6
 ρ 28



Our Approach: An Infrastructure to Support Provenance-Rich Papers [Koop et al., ICCS 2011]

- ◆ Tools for *authors* to create reproducible papers
 - Specifications that encode the computational processes
 - Package the results
 - Link from publications*Support different approaches*
- ◆ Tools for testers to repeat and validate results
 - Explore different parameters, data sets, algorithms
- ◆ Interfaces for searching, comparing and analyzing experiments and results
 - Can we discover better approaches to a given problem?
 - Or discover relationships among workflows and the problems?
 - How to describe experiments?

An Provenance-Rich Paper: ALPS2.0

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

B. Bauer¹ L. D. Carr² H.G. Evertz³ A. Feiguin⁴ J. Freire⁵
S. Fuchs⁶ L. Gamper¹ J. Gukelberger¹ E. Gull⁷ S. Guertler⁸
A. Hehn¹ R. Igarashi^{9,10} S.V. Isakov¹ D. Koop¹ P.N. Ma¹
P. Mates^{1,2} H. Matsui¹¹ O. Parcollet¹² G. Pawłowski¹³
J.D. Picon¹⁴ L. Pollet¹⁵ E. Santoso¹⁶ V.W. Scarola¹⁶
U. Schollwies¹⁷ C. Silva¹ B. Surer¹ S. Todo^{18,11} S. Trebst¹⁸
M. Troyer^{1,2} M. L. Wall² P. Werner¹ S. Wesel^{19,20}

¹Theoretische Physik, ETH Zurich, 8093 Zurich, Switzerland

²Department of Physics, Colorado School of Mines, Golden, CO 80401, USA

³Institut für Theoretische Physik, Technische Universität Graz, A-8010 Graz, Austria

⁴Department of Physics and Astronomy, University of Wyoming, Laramie, Wyoming 82071, USA

⁵Scientific Computing and Imaging Institute, University of Utah, Salt Lake City, Utah 84112, USA

⁶Institut für Theoretische Physik, Georg-August-Universität Göttingen, Göttingen, Germany

⁷Columbia University, New York, NY 10027, USA

⁸Bethe Center for Theoretical Physics, Universität Bonn, Nussallee 12, 53115 Bonn, Germany

⁹Department of Physics, University of California, Berkeley, CA 94720, USA

¹⁰Department of Physics, University of California, Berkeley, CA 94720, USA

¹¹Department of Physics, University of California, Berkeley, CA 94720, USA

¹²Department of Physics, University of California, Berkeley, CA 94720, USA

¹³Department of Physics, University of California, Berkeley, CA 94720, USA

¹⁴Department of Physics, University of California, Berkeley, CA 94720, USA

¹⁵Department of Physics, University of California, Berkeley, CA 94720, USA

¹⁶Department of Physics, University of California, Berkeley, CA 94720, USA

¹⁷Department of Physics, University of California, Berkeley, CA 94720, USA

¹⁸Department of Physics, University of California, Berkeley, CA 94720, USA

¹⁹Department of Physics, University of California, Berkeley, CA 94720, USA

²⁰Department of Physics, University of California, Berkeley, CA 94720, USA

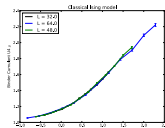
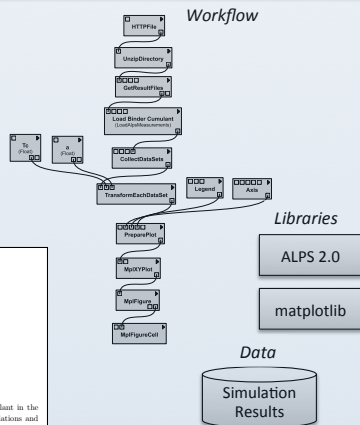


Figure 3. In this example we show a data collapse of the Binder Cumulant in the classical Ising model. The data has been produced by remotely run simulations and the critical exponent has been obtained with the help of the VisTrails parameter exploration functionality.



VCR: A Universal Identifier for Computational Results

Chronicling computations in real-time

VCR computation platform Plugin = Computation recorder

Regular program code

```
figure1 = plot(x)
save(figure1, 'figure1.eps')
```

```
> file /home/figure1.eps saved
>
```

VCR: A Universal Identifier for Computational Results

Chronicling computations in real-time

VCR computation platform Plugin = Computation recorder

Program code with VCR plugin

```
repository vcr.nature.com  
verifiable figure1 = plot(x)
```

```
> vcr.nature.com approved:  
> access figure1 at https://vcr.nature.com/ffaaffb148d7
```


VCR: A Universal Identifier for Computational Results

Word-processor plugin App

LaTeX source

```
\includegraphics{figure1.eps}
```

LaTeX source with VCR package

```
\includerresult{vcr.thelancet.com/ffaaffb148d7}
```

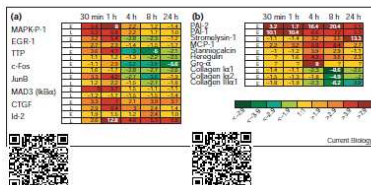
Permanently bind printed graphics to underlying result content

VCR: A Universal Identifier for Computational Results

Research Paper Analysis of replicative senescence Shelton *et al.* 943

Figure 3

Time course of serum stimulation. (a) Early passage (E: PD30) or late passage (L: PD89) BJ cultures were held in 0.5% serum for 2 days, then stimulated with 10% FBS. RNA levels from cultures at the indicated time points (Cy5 channel) were compared with the uninduced starting culture (Cy3 channel). Positive values indicate higher expression in induced cells; negative values indicate lower expression in induced cells. Question marks indicate that there was insufficient signal for detection. A complete listing of serum-responsive genes from this analysis is provided in Supplementary material. (b) The serum-responsiveness of select senescence-regulated genes in early passage (PD30) BJ fibroblasts.



senescence response appears to overlap substantially with gene expression patterns observed in activated fibroblasts during wound healing [24–26]. MCP-1, Gro- α , IL-1 β and IL-15 are strong effectors of macrophage and neutrophil recruitment and activation [27,28]. The upregulation of Toll (Tlr-4) in senescent fibroblasts confirms the overall immune response behavior at senescence. Tlr-4 is an IL-1 receptor homolog and is implicated in the activation of the gene regulatory protein NF- κ B, a function proposed to be part of the innate immune response [29]. The induction of IL-15 at senescence is also consistent with an innate immune response, as IL-15 can be induced by NF- κ B-dependent transcription [30] and also participates in inflammatory disease processes [28].

Deficiencies in the response of senescent cells to serum stimulation have been reported, and include an inability to induce the expression of *c-fos* mRNA [31] and markers of late G1 and S phase [32]. In response to serum, expression of inflammatory chemokines, matrix-degrading proteases and their modulators is induced in early-passage dermal fibroblasts, and expression of matrix collagens is reduced.

This transient burst of activity may represent the natural senescence response to growth factor withdrawal. The markers of the markers associated with senescence in dermal fibro-

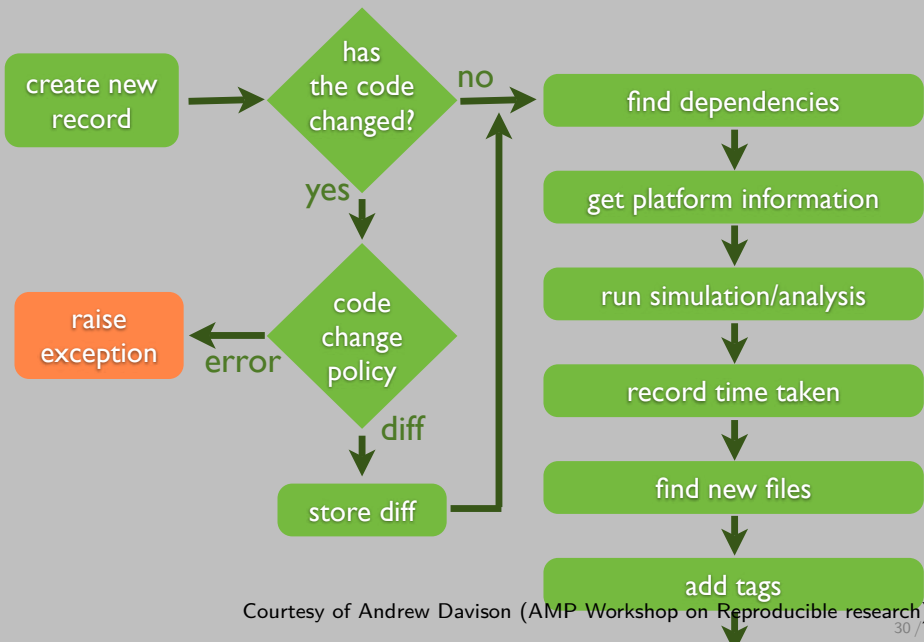
states overlap substantially with those in telomere-induced senescence (W.F., D.N.S., R. Allsopp, S. Lowe, and G. Ferbeyre, unpublished observations) and thus are likely to use many of the same activation processes.

The pattern of gene expression at senescence varies substantially in different cell types. Although the expression of matrix and structural proteins, such as the collagens, keratins and auxiliary factors, is repressed in RPE cells, inflammatory regulators are not induced, in contrast to dermal fibroblasts. Physiologically, this would make sense, as an acute inflammatory response in a tissue critical for normal vision would be likely to have deleterious consequences. However, as the RPE layer has a central role in the deposition and maintenance of extracellular matrix in the retina, decrements in the ability of senescent RPE cells to maintain appropriate expression patterns, as evidenced by decreased expression of collagens, keratins, aggrecan, transglutaminase and so on, would be predicted to have adverse effects on retinal architecture. Dysfunction of the RPE cell layer is considered to be a substantial factor in the development of age-related macular degeneration [36].

Surprisingly, early passage BJ cells do not express many of the markers associated with senescence in dermal fibro-

Courtesy of Marian Gavish and David Donoho (AMP Workshop on Reproducible research)

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"/>	20100709-154255		'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"/>	20100709-154309			0.59 s		Python	2.5.2	/Users/andrew/tmp/SumatraTest	main.py	396c2020ca50	09/07/2010	15:43:09	
<input type="checkbox"/>	haggling	'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	foobar
<input type="checkbox"/>	20100709-154338	'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"/>	haggling_repeat	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	


Courtesy of Andrew Davison (AMP Workshop on Reproducible research)

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 reprozip[all]
$ reprozip 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...
$ reprozip vagrant run mydirectory
experiment: 0%... 25%... 50%... 75%... 100%
result: 42.137
$ reprozip vagrant upload /tmp/new_config:global-config
$ reprozip vagrant run mydirectory --cmdline ./myexperiment --other --options
inputs/somefile.csv
experiment: 0%... 25%... 50%... 75%... 100%
result: -17.814
```

New Tools for Computational Reproducibility

- Dissemination Platforms:

ResearchCompendia.org

IPOL

Madagascar

MLOSS.org

thedatahub.org

nanoHUB.org

[Open Science Framework](https://OpenScienceFramework)

[The DataVerse Network](https://TheDataVerseNetwork)

RunMyCode.org

- Workflow Tracking and Research Environments:

VisTrails

Kepler

CDE

Galaxy

GenePattern

Synapse

Sumatra

Taverna

Pegasus

- Embedded Publishing:

Courtesy of Victoria Stodden (UC Davis, Feb 13, 2014)

[Verifiable Computational Research](https://VerifiableComputationalResearch)

Sweave

knitr

[Collage Authoring Environment](https://CollageAuthoringEnvironment)

SHARE

And also: **Org-Mode** 😊, **Figshare**, **Zenodo**, **ActivePapers** 😊, **Elsevier executable paper** 😞, ...

Outline

- ① Science crisis ?
- ② How is CS Concerned Really With This?
- ③ Reproducible Research/Open Science in a Nutshell
- ④ Illustrating Nice Ideas Through Different Tools
- ⑤ And In Practice?
- ⑥ What can Computer Scientists do ?

A Difficult Trade-off

Many different tools/approaches developed in various communities

But mainly two approaches: Automatic vs. Explicit

- **Automatically keeping track of everything**
 - the code that was run (source code, libraries, compilation procedure)
 - processor architecture, OS, machine, date, ...
- **Ensuring others can understand/adapt what was done**
 - Why did I run this? Does it still work when I change this piece of code for this one?

A Difficult Trade-off

Many different tools/approaches developed in various communities

But mainly two approaches: Automatic vs. Explicit

- **Automatically keeping track of everything**

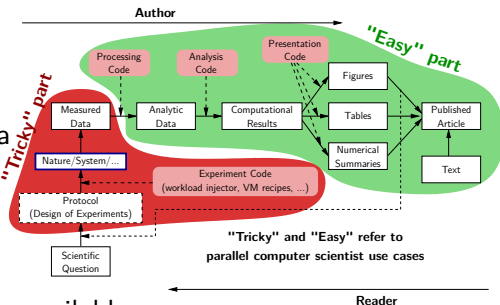
- the code that was run (source code, libraries, compilation procedure)
- processor architecture, OS, machine, date, ...

- **Ensuring others can understand/adapt what was done**

- Why did I run this? Does it still work when I change this piece of code for this one?

And the following key points:

- 1 Replicable article
- 2 Logging your activity
- 3 Logging and backup your data
- 4 Organizing your data
- 5 Mastering your environment
- 6 Controlling your experiments
- 7 Making your data/code/article available



3. Logging and backup your data

What are the options?

- Nothing 😞 (remember the funny examples from the beginning... 😊)
- Incremental backup mechanisms (e.g., time machine)
- The cloud! (e.g., Dropbox and Google Drive 😞 ...)
- Flexible version control systems (e.g. git 😊) where you're in control of what's happening
 - Use a crontab if you really do not want to think about it
 - We have come up with a specific **git branching workflow** for managing experimental results

4. Organizing and managing your data

- Use the machine readable **CSV format**
- Provide **raw** data and **meta** data, not just statistical outputs
- Organization
 - Explain your conventions (e.g., `src/`, `data/`, `script/`, `journal.org`)
 - Git submodules
- **Never** do data manipulation and statistical tests **by hand** or with a spreadsheet 😞
- **Use R**, Python or another free software to read and process raw data.
 - Use a workflow that **documents both data and process**
 - The org-mode tangling mechanism may help

5. Mastering your environment

What are the options?

- Nothing 😊 No, it's not, you have to do something. . .
- Restrict your tools/dependencies to the bare minimum (e.g., python)
 - List them all manually in a README
 - Use **custom shell scripts** or **sosreport** that log all the dependencies you are aware. Ask your friends to check whether this is sufficient. . .
 - Combine everything in **activepapers**, i.e., an HDFS5 file combining datasets and programs working on these datasets in a single package, along with meta data, history, . . .
- Create and distribute your own virtual image (VM, docker, **Singularity**)
- Have tools that **automatically** keep track of dependencies/files and packages up the Code, Data, and Environment
 - **CDE** (Guo et al., 2011) **ReproZip** (Freire et al., 2013), **CARE** (Janin et al., 2014),
 - See **Preserve the Mess or Encourage Cleanliness?** (Thain et al., 2015)
- Use a specific tool to generate customized appliances (kvm, LXC, Virtualbox, iso, . . .): **recipes** with **steps** and **aliases**, execution in **contexts**, **checkpoints**, . . . (**Kameleon**)

6. Controlling your 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)
 - ▶ `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, ...)

7. Making your data/code/article available

- Your webpage 😞
- Figshare, Zenodo 😊, ...
- Companion websites (elsevier executable paper 😞, runmycode, exec&share 😊, ...)
- Github (damn, they're good! 😊), ...

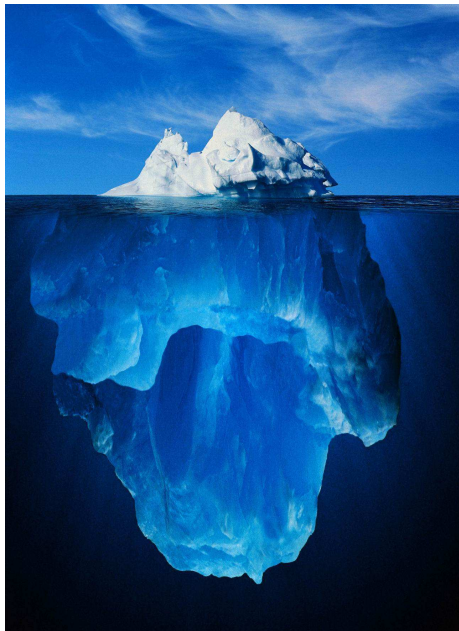
This may seem easy but is more tricky than it looks like:

- Arbitrary limits can make your life painful
- Perennity (Roberto Di Cosmo's talk at R⁴)
 - 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, ...

Many legal aspects about data/code/idea sharing

- I am a civil servant and I strongly believe research is a team sport
- Intellectual property is an important topic we do not want to leave to bureaucrats and lawyers...

Remember the general picture



The article is only the top of the iceberg, we need a way to **dive** and **unveil** what's behind every graphics and number. . .

1. Replicable article (Literate programming)

Donald Knuth: explanation of the program logic in a natural language interspersed with snippets of macros and traditional source code.

I'm way too stupid to program this way 😊 but that's exactly what we need for writing a reproducible article/analysis!

KnitR (a.k.a. Sweave)

For R and Emacs users. Easy replicable articles with a modern IDE (e.g., Rstudio)

Ipynb/Jupyter notebook

Python user → go for Jupyter. Web app, easy to use/setup... Writing replicable article may be tricky though

Org-mode (my favorite! requires Emacs though)

- Org-mode is plain text, very smooth, works both for html, pdf, ...
- Allows to combine all my favorite languages

Note that this generation depends on a computational environment whose preservation is not addressed here (see for example *activepapers*).

A replicable article with Org-Mode

See for example our recent article on the simulation of Multithreaded Sparse Linear Algebra Solvers at ICPADS 2015.

Here are the following important features to exploit:

Structure highly hierarchical

- Sectioning, itemize, enumerate, fonts
- Tags to control what will be exported

Export in several output formats

- Fine control with `#+BEGIN_EXPORT latex`
- Unfortunate need for verbose headers (because of \LaTeX 😞) and black magic in the end of the file (for emacs portability 😞)

Babel (the literate programming part of org-mode). Many possible usage:

- Run babel on export
- Or not. . . and make sure intermediate results are stored (this is how I proceed)
- Dependencies can be expressed
- Caching mechanism
- Side effects are the enemy of reproducibility

2. Logging your activity (Laboratory Notebook)

Do not tie your hands with non-free software like Evernote or OneNote

- Org-mode again!
 - Capture mechanism (notes, todo, ...)
 - Babel favors code reuse, ssh connections in sessions, meta-programming
 - Tagging mechanism to structure the journal
 - Link mechanism, Todo, Calendar views, Tables, ...

I have a very intense usage and so do all my master/PhD students (e.g., [here](#))

- Spending **more than an hour without** at least **writing** what you're working on **is not right**. ... Take a **5 minutes** break and ask yourself what you're doing, what is keeping you busy and where all this is leading you
- While working on something, you will often notice/think about something you should fix/improve but you just don't want to do it now. Take 20 seconds to write a **TODO** entry
- There are moments where you have to **wait for something** (compiling, deployment, ...). It is generally the perfect time for improving your notes (e.g., detail the steps to accomplish a TODO entry)
- **By the end of the day**: daily (and weekly) **review!**
 - Update your lists, decide the next steps, summarize what you did/learnt, ...

Pros and Cons of these three tools

- Ipython notebook:
 - 😊 Easy to set up, user-friendly, machine readable format (JSON), easy sharing on the cloud
 - 😞 Writing an article, JSON, not fully polyglot
- knitr/Rstudio:
 - 😊 Easy to set up, user-friendly, writing articles, easy publishing on **rpubs**
 - 😞 not fully polyglot
- Emacs/Org-mode:
 - 😞 Emacs, steep learning curve
 - 😊 Powerful and versatile, yields control to power users, works both for writing articles and a notebook, good integration on github

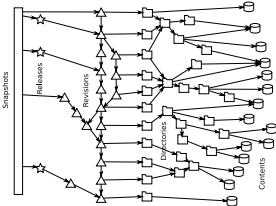
The ultimate tool would combine an engine in an editor that allows collaborative interactive edition

Outline

- ① Science crisis ?
- ② How is CS Concerned Really With This?
- ③ Reproducible Research/Open Science in a Nutshell
- ④ Illustrating Nice Ideas Through Different Tools
- ⑤ And In Practice?
- ⑥ 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 (Sim-Grid, NS3)
 - Maintenance cost, keeping in pace with technology, practices for prototype platforms, control, sharing of experimental conditions with others, experimental engines



- Need for convergence 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)
 - Promote a different model
- **Learning** is the essence of our work. \rightsquigarrow Train our researchers and students
 - Better teaching/understanding of statistics, experimental practice, design of experiments

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

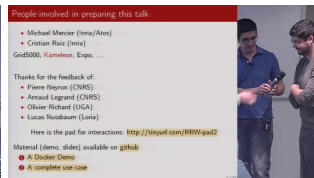
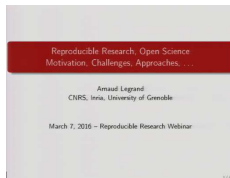
Main benefit:

- **Higher confidence** in our work \rightsquigarrow 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



Numerical reproducibility

Logging and backing up

https://github.com/alegrand/RR_webinars