Reproducible Research for Computer Scientists

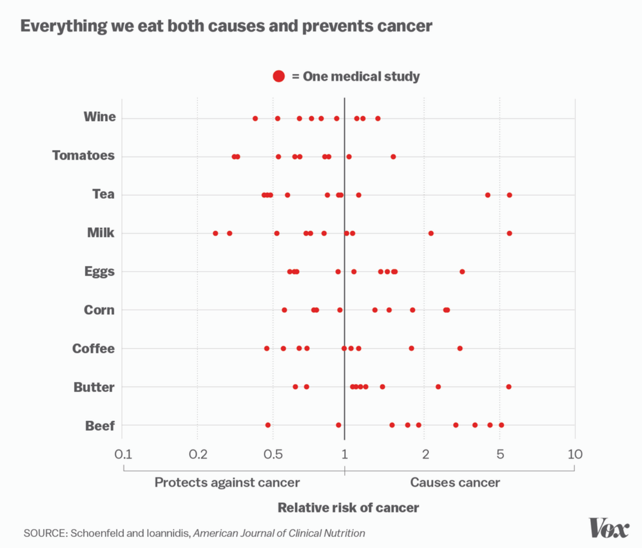
Lucas M. Schnorr

Comp. Syst. Perf. Analysis

Inconsistencies

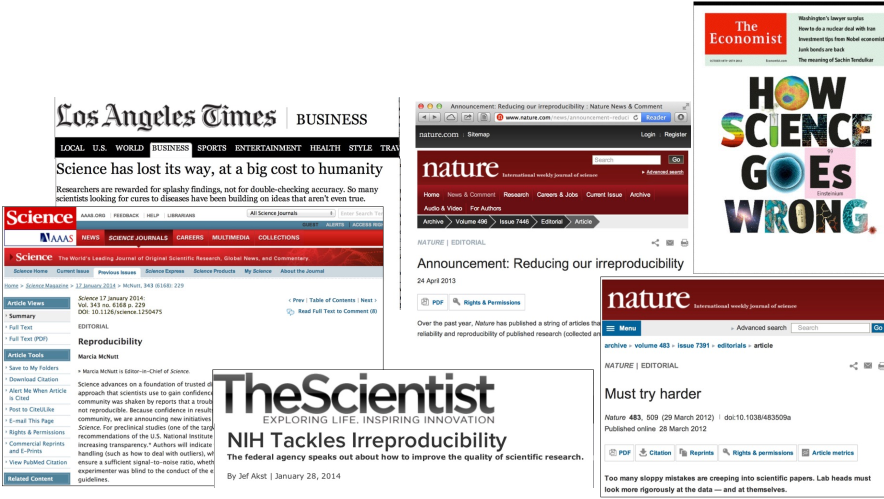
Is everything we eat associated with cancer? A systematic cookbook review, Schoenfeld and Ioannidis, *Amer. Jour. of Clinical Nutrition*, 2013.

Inconsistencies



Public evidence for a Lack of Reproducibility

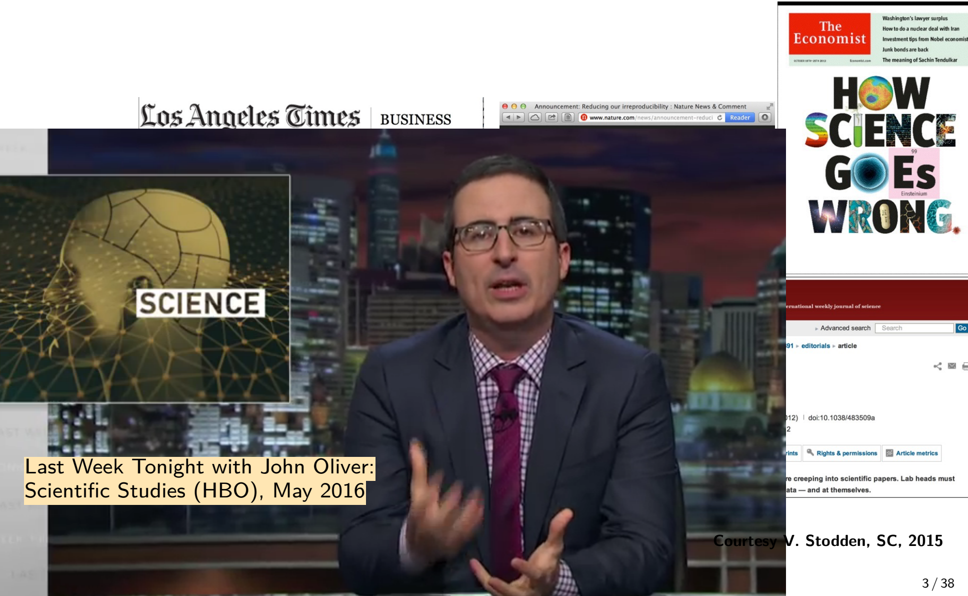
- J.P. Ioannidis. *Why Most Published Research Findings Are False* PLoS Med. 2005.
- *Lies, Damned Lies, and Medical Science*, The Atlantic. Nov, 2010



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



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

Courtesy V. Stodden, SC, 2015

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.*

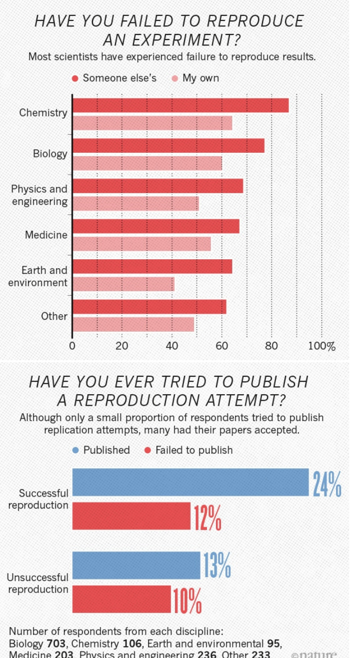
– Herndon, Ash and Pollin

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.



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

- 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, ?..

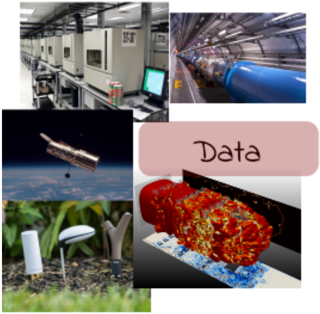
No. Questionning previous work is part of the scientific process

Just like honesty, rigor and transparency...

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

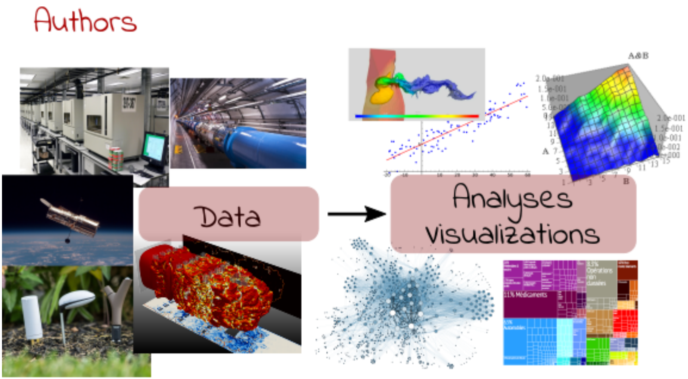
Outline

Authors



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...

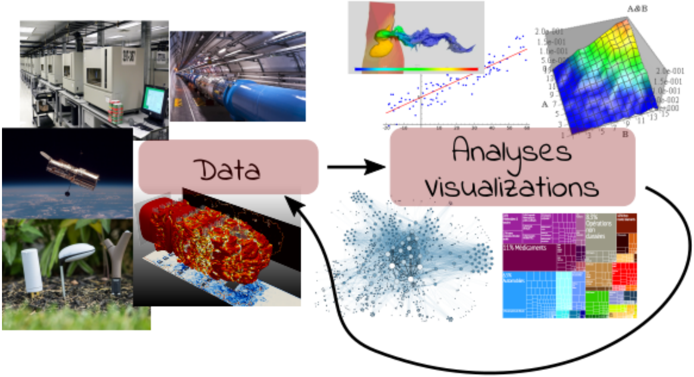
– Nobel Comity (Chemistry), 2013



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...

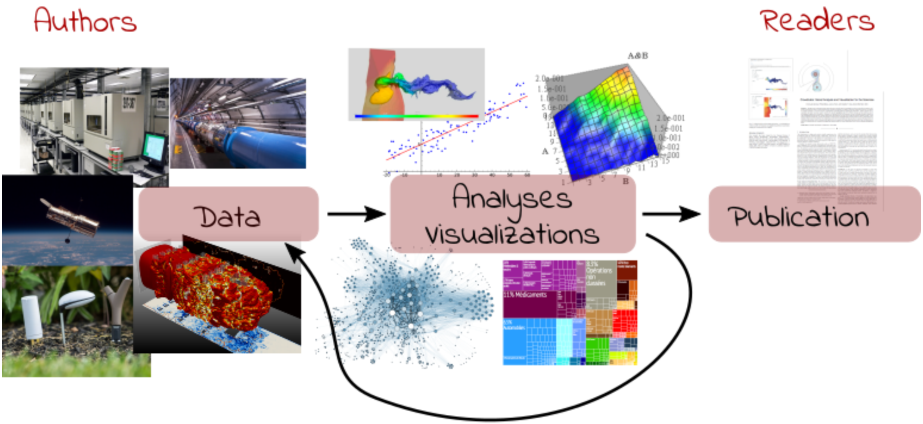
– Nobel Comity (Chemistry), 2013

Authors



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...

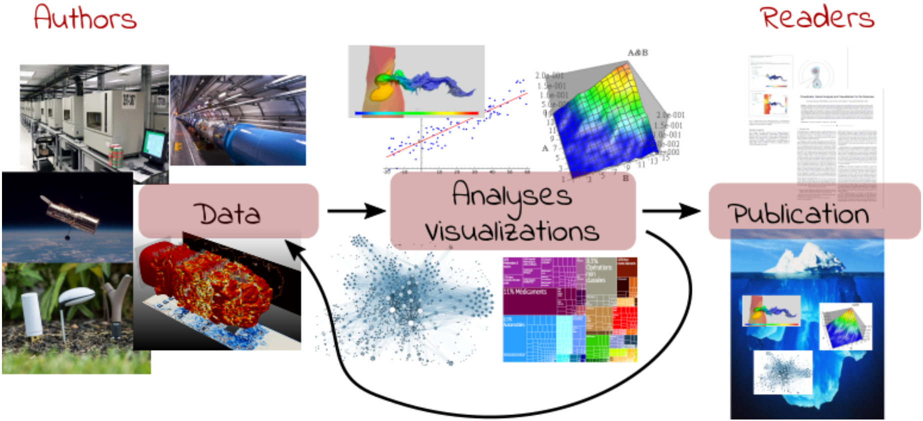
– Nobel Comity (Chemistry), 2013



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).

- 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)

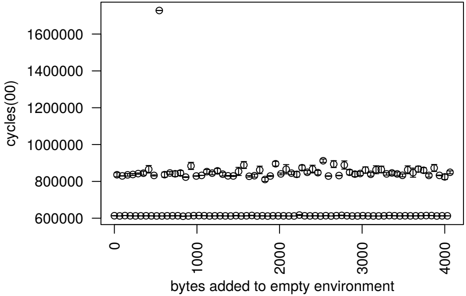
Right, why should we care about computers? They are deterministic machines after all, right? 😉

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

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!

- "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

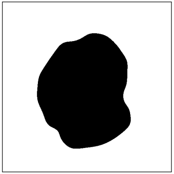
Computer performance ? Well, I design algorithms!

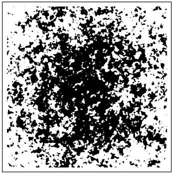
- "Real" problems are all NP-hard, Log-APX, etc.
- Real workload = ~~NP-completeness proof widgets~~, regularities and properties (difficult to formally state but that should be exploited)

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

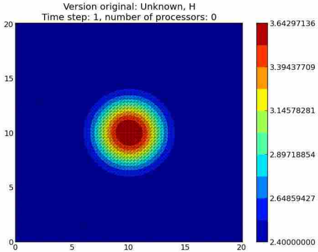
Did I mention we have parallel machines nowadays? ☺

Telemac2D: the simplest gouttedo simulation

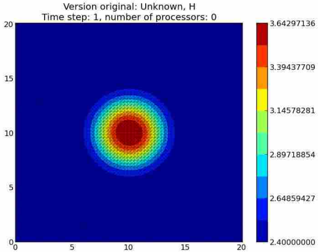
The gouttedo test case

- 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, ...)



Sequential



Parallel $p = 2$

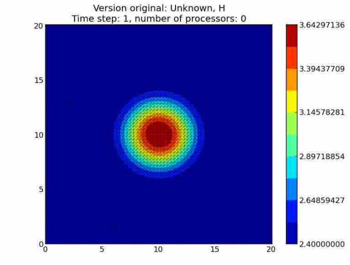
Courtesy of P. Langlois and R. Nheili

Did I mention we have parallel machines nowadays? ☺

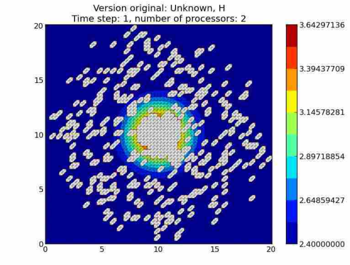
A white plot displays a non-reproducible value

Numerical reproducibility?

time step = 1



Sequential



Parallel $p = 2$

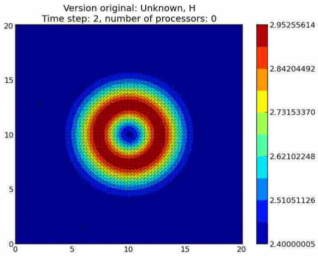
Courtesy of P. Langlois and R. Nheili

Did I mention we have parallel machines nowadays? ☺

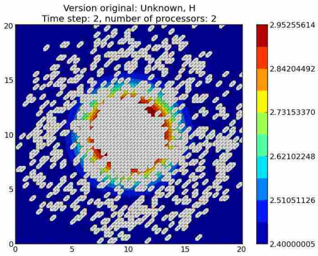
A white plot displays a non-reproducible value

Numerical reproducibility?

time step = 2



Sequential



Parallel $p = 2$

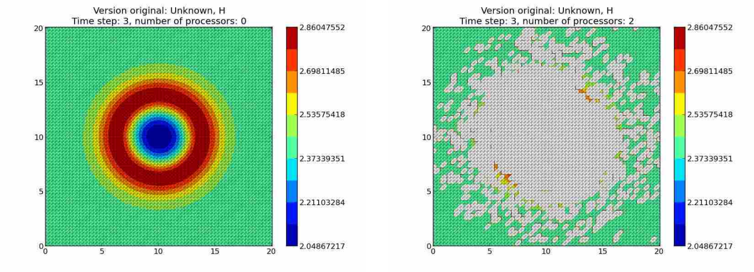
Courtesy of P. Langlois and R. Nheili

Did I mention we have parallel machines nowadays? ☺

A white plot displays a non-reproducible value

Numerical reproducibility?

time step = 3



Sequential | Parallel $p = 2$

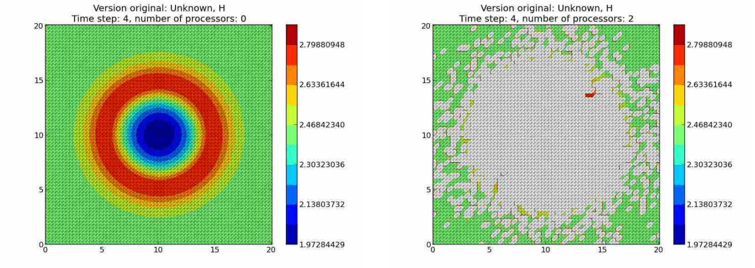
Courtesy of P. Langlois and R. Nheili

Did I mention we have parallel machines nowadays? ☺

A white plot displays a non-reproducible value

Numerical reproducibility?

time step = 4



Sequential

Parallel $p = 2$

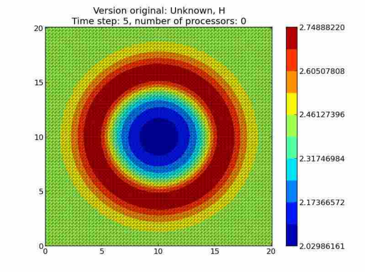
Courtesy of P. Langlois and R. Nheili

Did I mention we have parallel machines nowadays? ☺

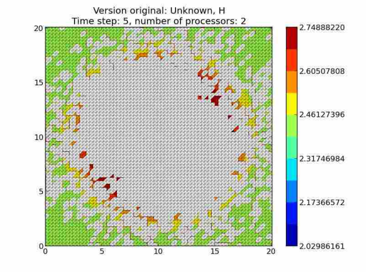
A white plot displays a non-reproducible value

Numerical reproducibility?

time step $= 5$



Sequential



Parallel $p = 2$

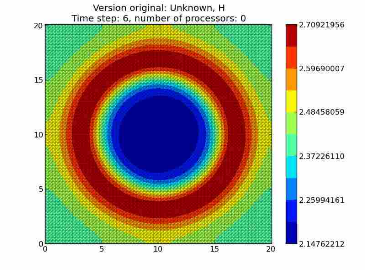
Courtesy of P. Langlois and R. Nheili

Did I mention we have parallel machines nowadays? ☺

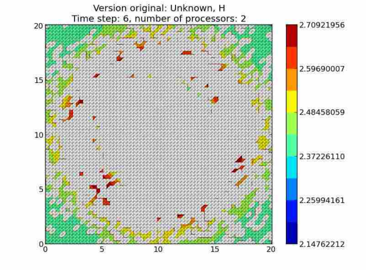
A white plot displays a non-reproducible value

Numerical reproducibility?

time step $= 6$



Sequential



Parallel $p = 2$

Courtesy of P. Langlois and R. Nheili

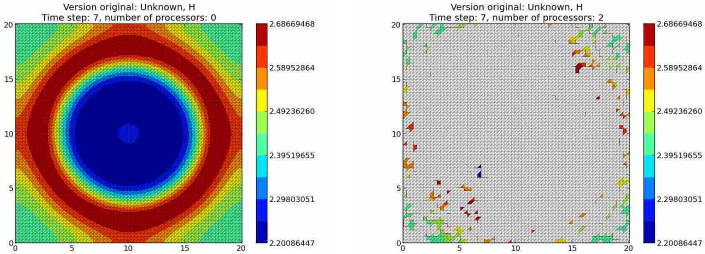
All I care about is the algorithm output

Did I mention we have parallel machines nowadays? ☺

A white plot displays a non-reproducible value

Numerical reproducibility?

time step $= 7$



Sequential

Parallel $p = 2$

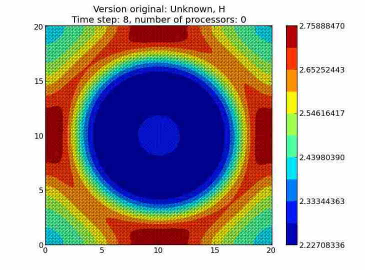
Courtesy of P. Langlois and R. Nheili

Did I mention we have parallel machines nowadays? ☺

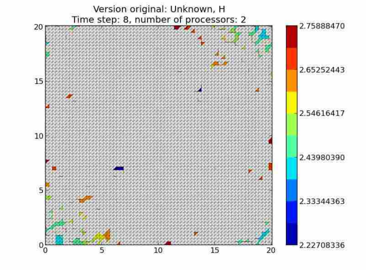
A white plot displays a non-reproducible value

Numerical reproducibility?

time step $= 8$



Sequential



Parallel $p = 2$

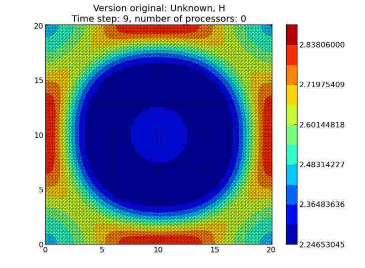
Courtesy of P. Langlois and R. Nheili

Did I mention we have parallel machines nowadays? ☺

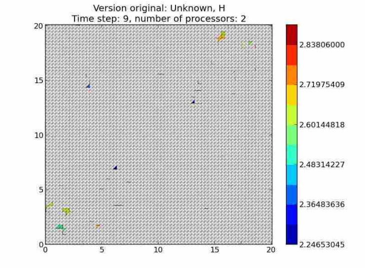
A white plot displays a non-reproducible value

Numerical reproducibility?

time step = 9



Sequential



Parallel $p = 2$

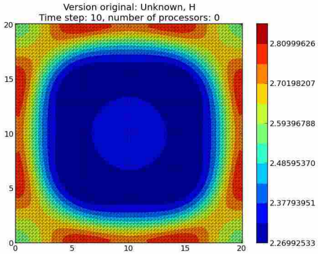
Courtesy of P. Langlois and R. Nheili

Did I mention we have parallel machines nowadays? ☺

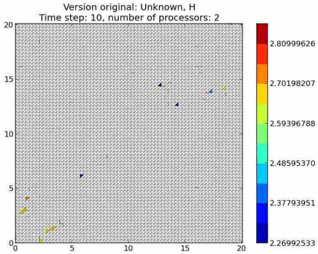
A white plot displays a non-reproducible value

Numerical reproducibility?

time step $= 10$



Sequential



Parallel $p = 2$

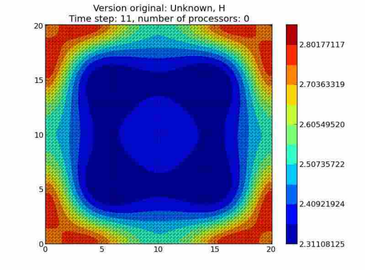
Courtesy of P. Langlois and R. Nheili

Did I mention we have parallel machines nowadays? ☺

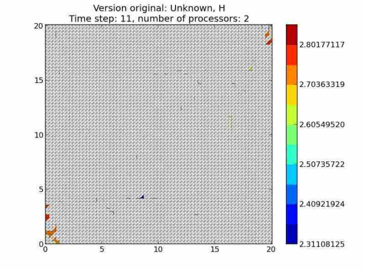
A white plot displays a non-reproducible value

Numerical reproducibility?

time step $= 11$



Sequential



Parallel $p = 2$

Courtesy of P. Langlois and R. Nheili

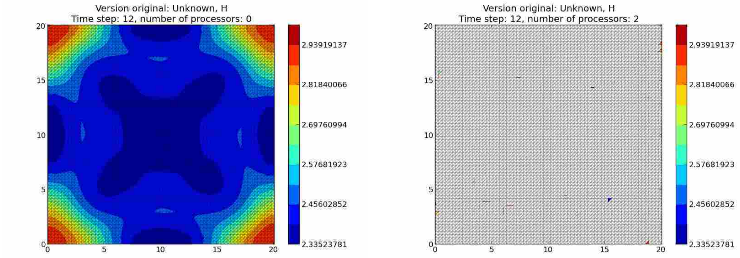
All I care about is the algorithm output

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

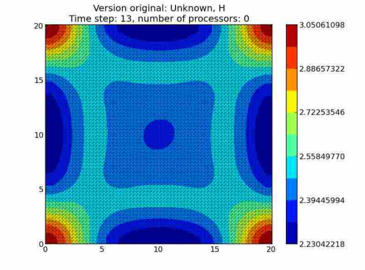
All I care about is the algorithm output

Did I mention we have parallel machines nowadays? ☺

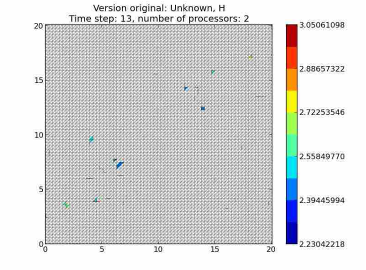
A white plot displays a non-reproducible value

Numerical reproducibility?

time step $= 13$



Sequential



Parallel $p = 2$

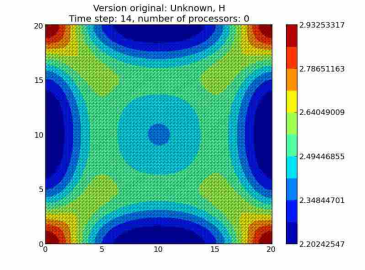
Courtesy of P. Langlois and R. Nheili

Did I mention we have parallel machines nowadays? ☺

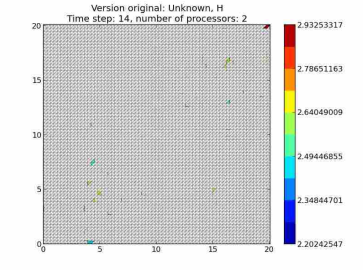
A white plot displays a non-reproducible value

Numerical reproducibility?

time step $= 14$



Sequential



Parallel $p = 2$

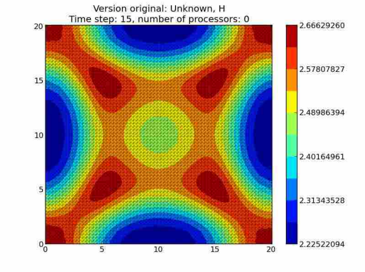
Courtesy of P. Langlois and R. Nheili

Did I mention we have parallel machines nowadays? 😉

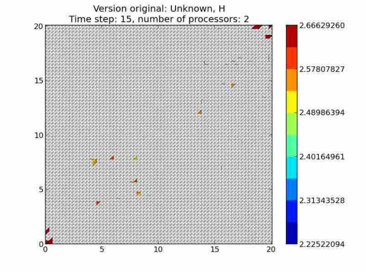
A white plot displays a non-reproducible value

NO numerical reproducibility!

time step $= 15$



Sequential



Parallel $p = 2$

Courtesy of P. Langlois and R. Nheili

Did I mention we have parallel machines nowadays? ☺

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

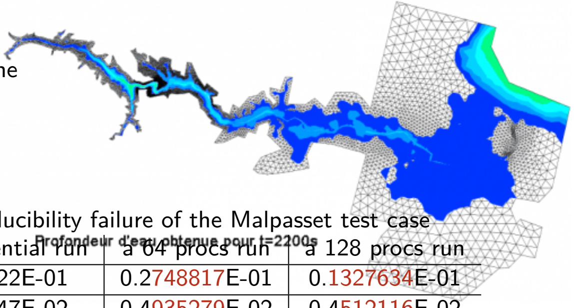


TABLE 1.1: Reproducibility failure of the Malpasset test case

	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 in the algorithm, in its implementation, in its execution.

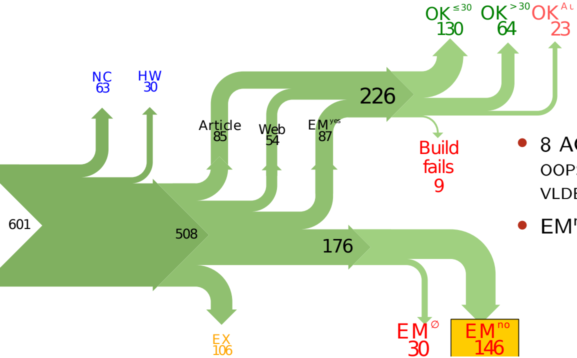
The whole chain needs to be revisited.

Courtesy of P. Langlois and R. Nheili

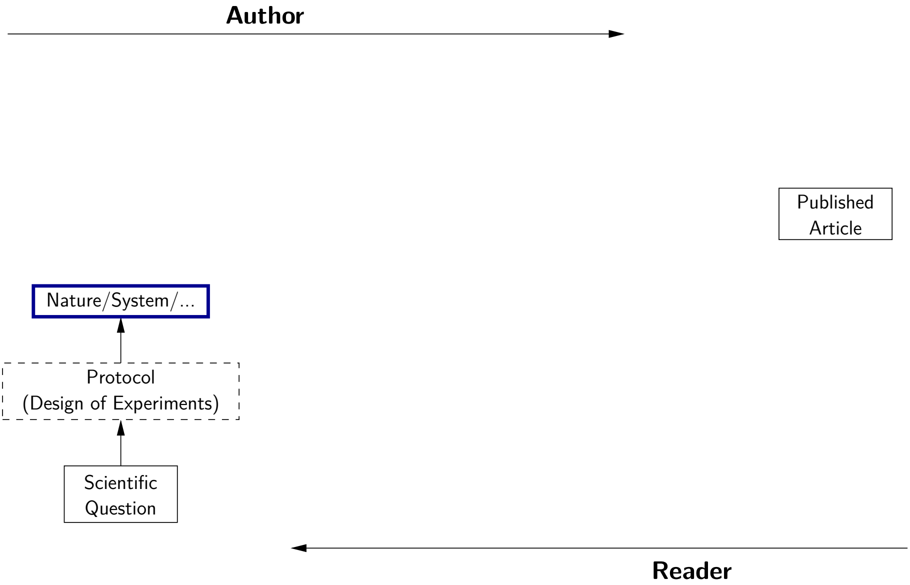
Outline

Lack of reproducibility is foremost a cultural matter

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

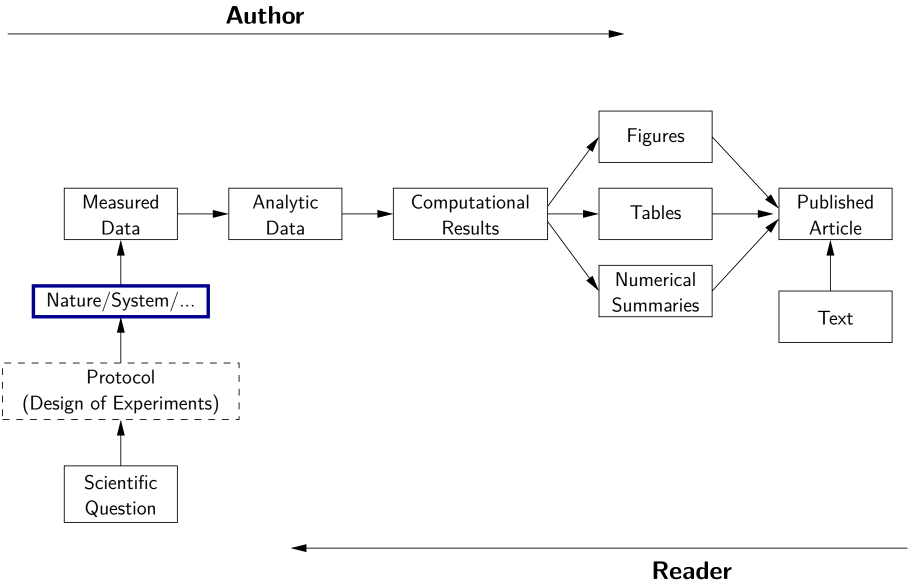


- 8 ACM conferences (ASPLOS'12, CCS'12, OOPSLA'12, OSDI'12, PLDI'12, SIGMOD'12, SOSP'11, VLDB'12) and 5 journals
- EM^{no} = the code cannot be provided



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

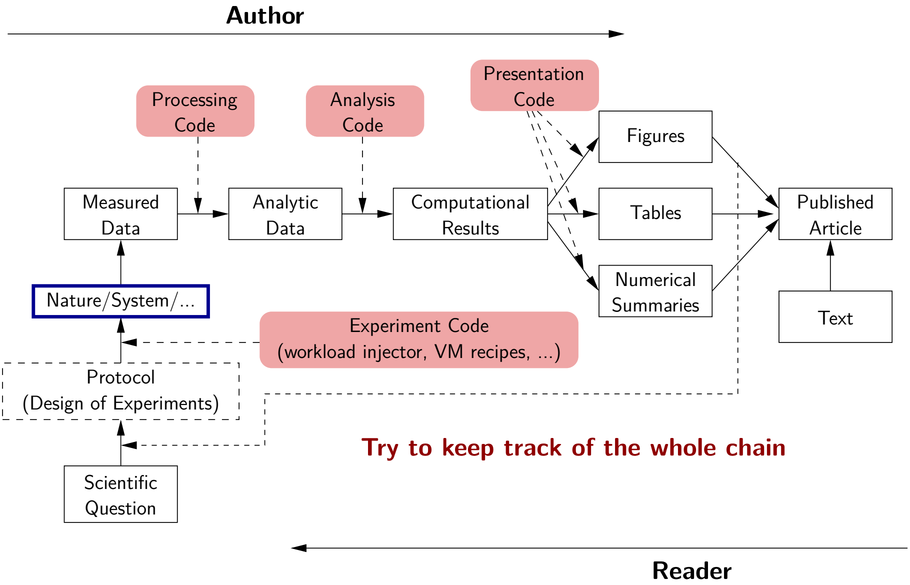
In this series of lectures, we'll go from right to left and see how we can improve.



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

In this series of lectures, we'll go from right to left and see how we can improve.

Reproducible Research: Trying to Bridge the Gap

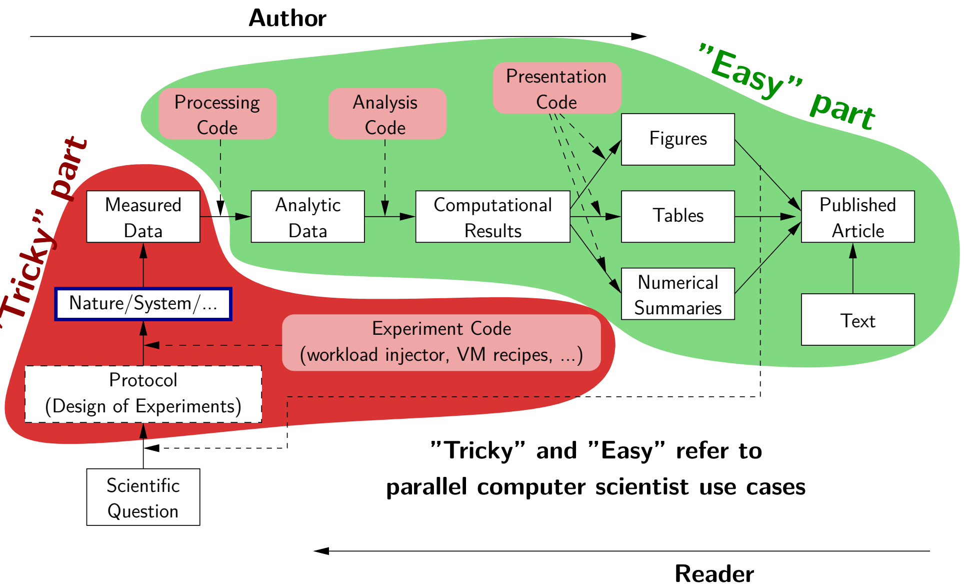


Try to keep track of the whole chain

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

In this series of lectures, we'll go from right to left and see how we can improve.

Reproducible Research: Trying to Bridge the Gap



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

In this series of lectures, we'll go from right to left and see how we can improve.

Paradigm Shift

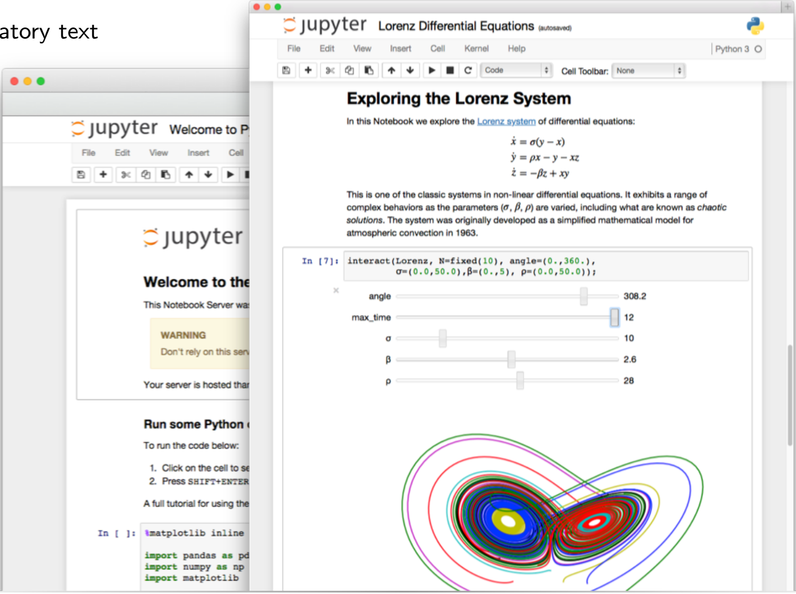
1. Lack of information, data access
   ⇝ *The right tracking/sharing tools*
2. Lack of technical and scientific rigor
   ⇝ *Better methodology (e.g., training in stats)*
3. Computation/programming mistakes
   ⇝ *Continuous integration, code review, environment control, numerical analysis, …*
4. Little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

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

Outline

Outline

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



Un document computationnel

Mon ordinateur m'indique que π vaut *approximativement*

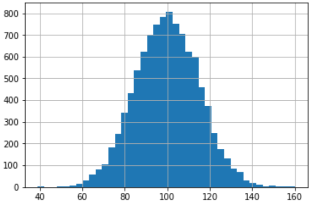
```
3.141592653589793
```

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

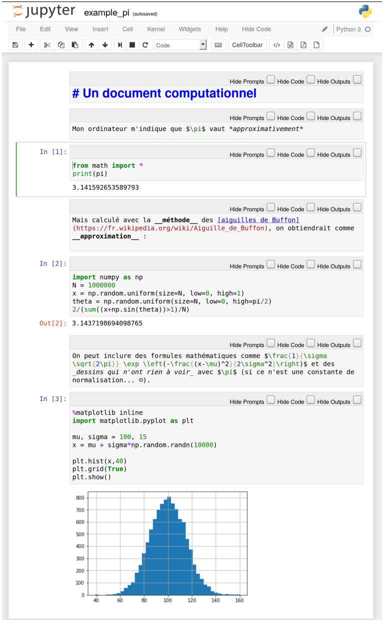
```
import numpy as np
N = 1000000
x = np.random.uniform(size=N, low=0, high=1)
theta = np.random.uniform(size=N, low=0, high=pi/2)
2/(sum((x+np.sin(theta))>1)/N)
```

```
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... ☺).



Document initial dans son environnement



Document final

Un document computationnel

Mon ordinateur m'indique que π vaut *approximativement*

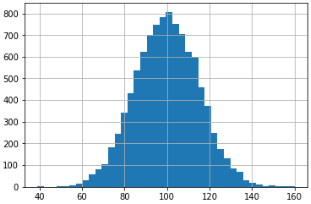
```
3.141592653589793
```

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

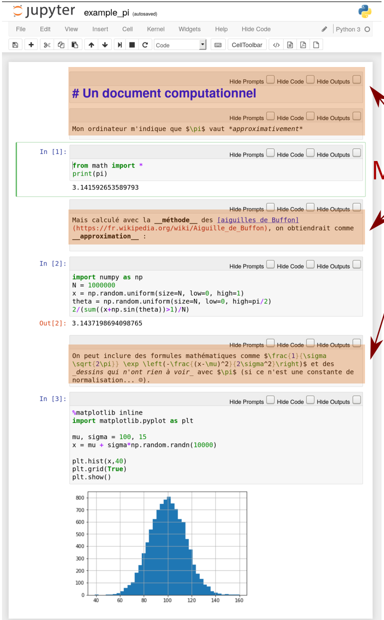
```python
import numpy as np
N = 1000000
x = np.random.uniform(size=N, low=0, high=1)
theta = np.random.uniform(size=N, low=0, high=pi/2)
2/(sum((x+np.sin(theta))>1)/N)
```

```
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... ☺).



Document initial dans son environnement



MarkDown

Document final

Un document computationnel

Mon ordinateur m'indique que π vaut *approximativement*

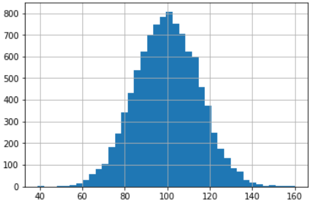
3.141592653589793

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)
```

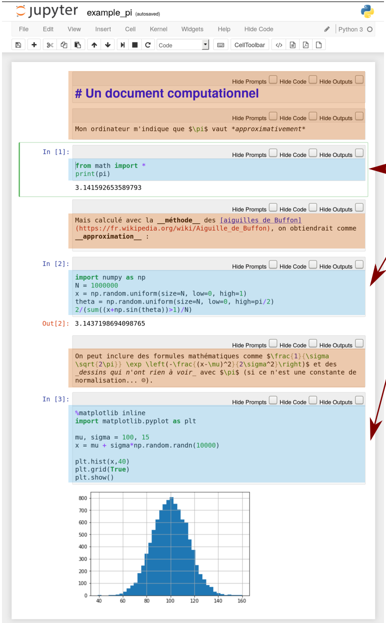
3.1437198694098765

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... ☺).



Computational Document (Jupyter Notebook)

Document initial dans son environnement



Code

Document final

Un document computationnel

Mon ordinateur m'indique que π vaut *approximativement*

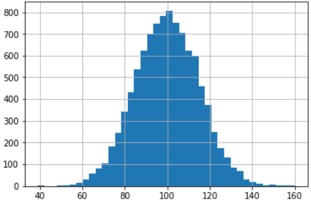
3.141592653589793

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

```python
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)
```

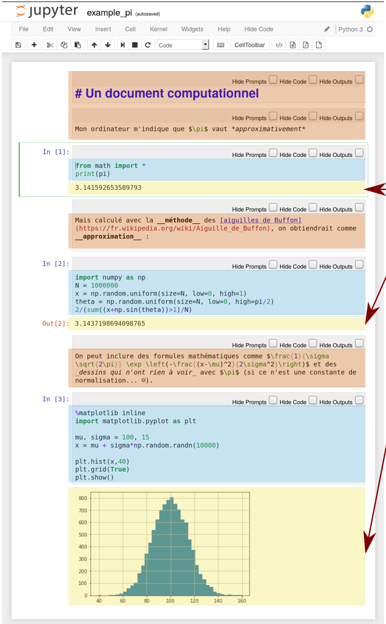
3.1437198694098765

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 (Jupyter Notebook)

Document initial dans son environnement



Résultats

Document final

Un document computationnel

Mon ordinateur m'indique que π vaut *approximativement*

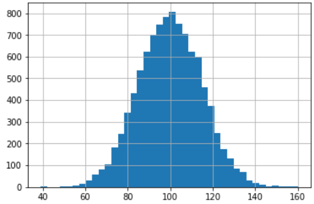
```
3.141592653589793
```

Mais calculé avec la **méthode** des [aiguilles de Buffon](https://fr.wikipedia.org/wiki/Aiguille_de_Buffon), on obtiendrait comme **approximation** :

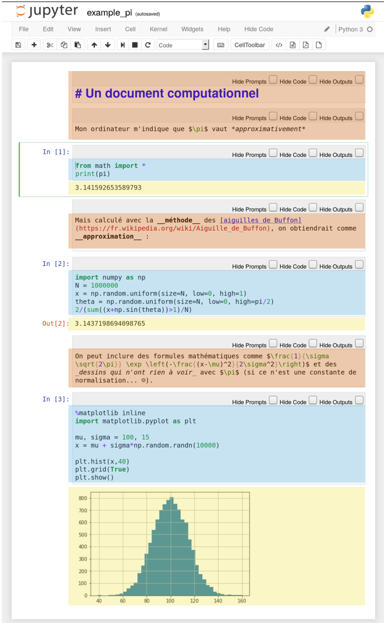
```
import numpy as np
N = 1000000
x = np.random.uniform(size=N, low=0, high=1)
theta = np.random.uniform(size=N, low=0, high=pi/2)
2/(sum((x+np.sin(theta))>1)/N)
```

```
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... ☺).



Document initial dans son environnement



Export

Document final

Un document computationnel

Mon ordinateur m'indique que π vaut *approximativement*

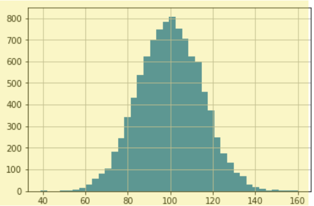
```
3.141592653589793
```

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

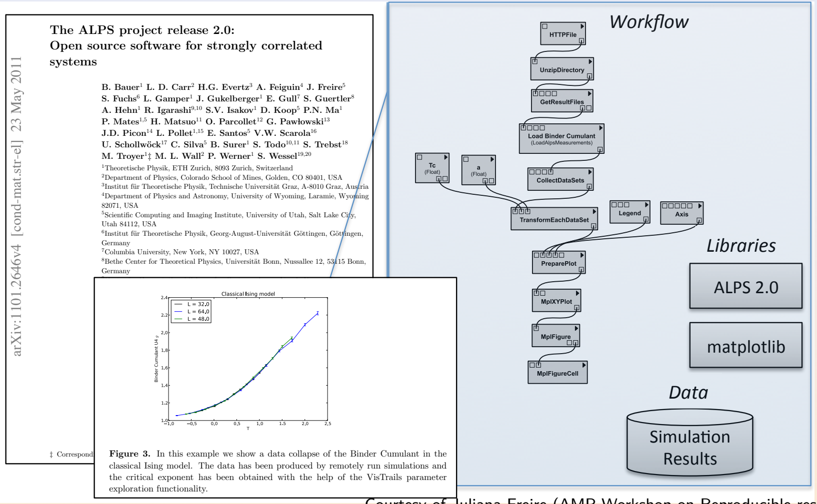
```python
import numpy as np
N = 1000000
x = np.random.uniform(size=N, low=0, high=1)
theta = np.random.uniform(size=N, low=0, high=pi/2)
2/(sum((x+np.sin(theta))>1)/N)
```

```
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... ☺).



An *Provenance-Rich* Paper: ALPS2.0



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

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

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

[1]Theoretische Physik, ETH Zurich, 8093 Zurich, Switzerland
[2]Department of Physics, Colorado School of Mines, Golden, CO 80401, USA
[3]Institut für Theoretische Physik, Technische Universität Graz, A-8010 Graz, Austria
[4]Department of Physics and Astronomy, University of Wyoming, Laramie, Wyoming 82071, USA
[5]Scientific Computing and Imaging Institute, University of Utah, Salt Lake City, Utah 84112, USA
[6]Institut für Theoretische Physik, Georg-August-Universität Göttingen, Göttingen, Germany
[7]Columbia University, New York, NY 10027, USA
[8]Bethe Center for Theoretical Physics, Universität Bonn, Nussallee 12, 53115 Bonn, Germany

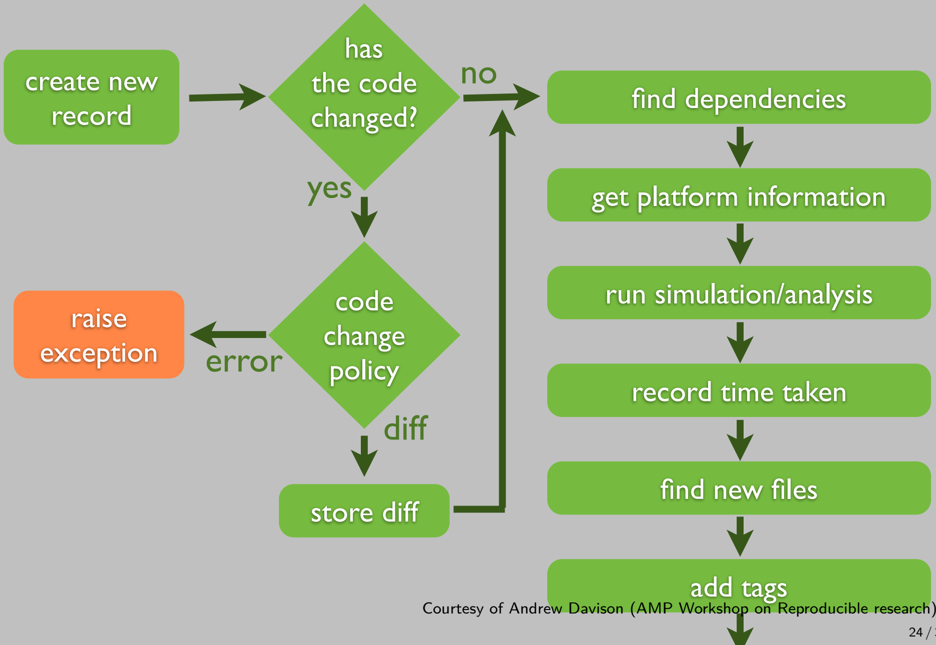
‡ Correspond

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)

[Bauer et al., JSTAT 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."
```

Courtesy of Andrew Davison (AMP Workshop on Reproducible research)

Sumatra: an "experiment engine" that helps taking notes

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

Courtesy of Andrew Davison (AMP Workshop on Reproducible research)

Sumatra: an "experiment engine" that helps taking notes

Sumatra: TestProject: List of records

http://127.0.0.1:8002/

Google

TestProject: List of records

Delete include data	Label	Reason	Outcome	Duration	Processes	Simulator Name	Simulator Version	Script Repository	Script Main file	Script Version	Date	Time	Tags
	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	
	20100709-154309			0.59 s		Python	2.5.2	/Users/andrew/tmp/SumatraTest	main.py	396c2020ca50	09/07/2010	15:43:09	
	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	fooba
	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	
	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)

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
- $\approx$ 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

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

```
3.10.1+dfsg1
```

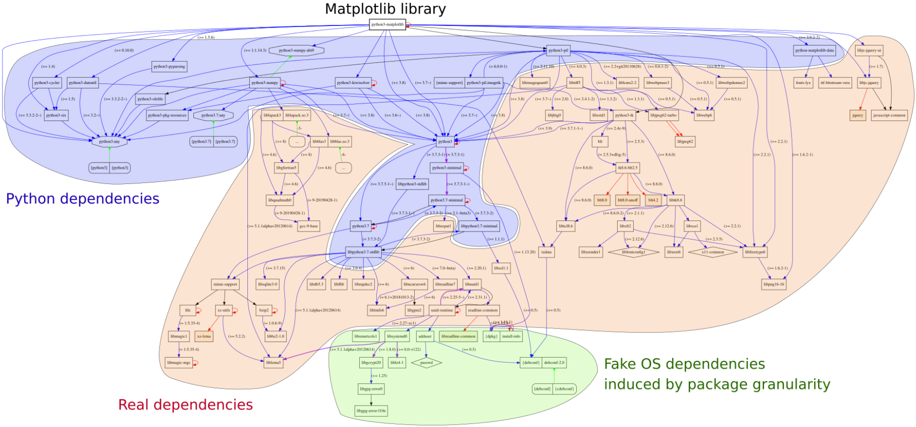
```python
import matplotlib
print(matplotlib.__version__)
```

```
3.10.1+dfsg1
```

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

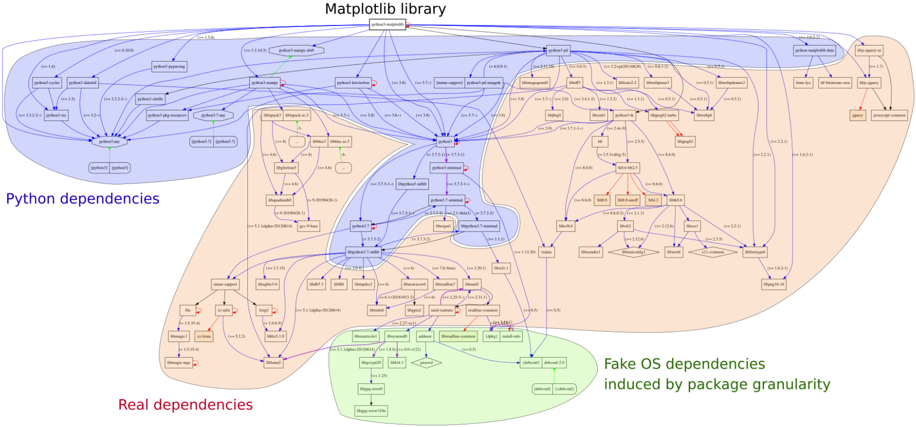
1

2


Matplotlib library
Python dependencies
Real dependencies
Fake OS dependencies
induced by package granularity

1

2



Virtual images or *Docker containers* allow to freeze environements

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)

  although nothing specific to G5K

  - Plush (2006-, Planetlab)
  - OMF (2009-, Wireless)
  - Splay (2008), ...

  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

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

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

ReScience
Reproducible science is good. Replicated science is better.

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 publically funded research projects – *Science with and for Society, 27 October 2017*

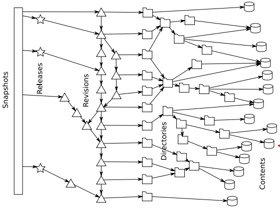
Outline

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