Reproducibility Crisis, Open Science, and Computer Science

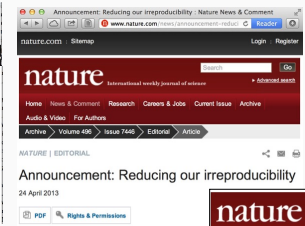
Arnaud Legrand

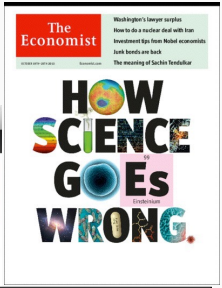
Université Grenoble Alpes

October 2020

Public Evidence for a Lack of Reproducibility

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





Courtesy V. Stodden, SC, 2015

Newsworthy Stories about Scientific Misconduct

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

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

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

Newsworthy Stories about Scientific Misconduct

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

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

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

Dieterik Stapel Professor, Social Psychology, Univ. Amsterdam, 2011

I failed as a scientist. I adapted research data and fabricated research. Not once, but several times, not for a short period, but over a longer period of time. [..] I am aware of the suffering and sorrow that I caused to my colleagues... I did not withstand the pressure to score, to publish, the pressure to get better in time. I wanted too much, too fast. In a system where there are few checks and balances, where people work alone, I took the wrong turn. 58 retracted publications

Newsworthy Stories about Scientific Misconduct

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

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

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

Dieterik Stapel Professor, Social Psychology, Univ. Amsterdam, 2011

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

58 retracted publications

Brian Wansink Professor, Psychological Nutrition, Cornell, 2016

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

17 retracted publications

A Credibility Crisis?

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

- Every domain has its black sheep
- The publish or perish pressure is a huge pain

Media attention inflates conspiracy opinions 😡

Scientific result are worthless. Stop the scientific dictatorship/lobby!

The Battle against Scientific Fraud

CNRS International Magazine

A Credibility Crisis?

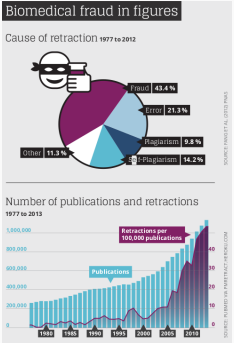
The Battle against Scientific Fraud
CNRS International Magazine

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

- Every domain has its black sheep
- The publish or perish pressure is a huge pain

Media attention inflates conspiracy opinions

Scientific result are worthless. Stop the scientific dictatorship/lobby!

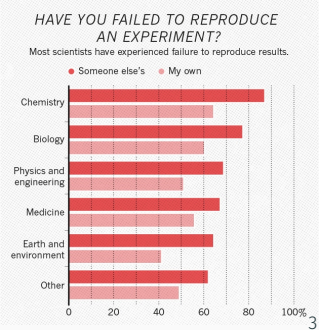


Fraud is the (uninteresting) visible part of the iceberg

- Failing to reproduce the results of others is common

  1,500 scientists lift the lid on reproducibility, Nature, May 2016

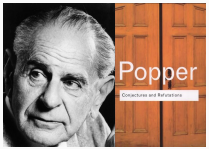
- How so? Why now? Why is this important? What can we do about it?



Reproducibility of experimental results is the *hallmark of science*

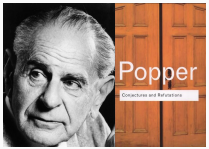
1934: Karl Popper puts the notions of falsifiability and crucial experiment as the hallmark of science

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



Reproducibility of experimental results is the *hallmark of science*

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



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

An ideal rather than the norm

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

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

Reproducibility: a *core value of science*

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

   Reproducibility acts as a Universality/Robustness control

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

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

   Reproducibility acts as a Quality control

Reproducibility: a *core value of science*

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

   Reproducibility acts as a Universality/Robustness control

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

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

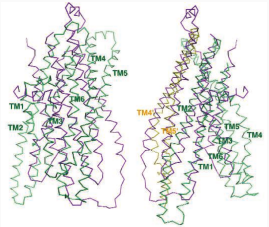
   Reproducibility acts as a Quality control

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

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

– Ben Marwick, The conversation, 2015

How computers broke science



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

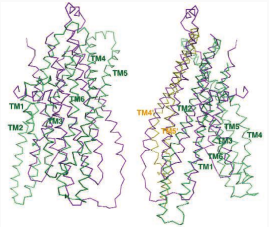
He specialized in structures of multidrug resistant transporter proteins in bacteria: MsbA de Escheria Choli (Science, 2001), Vibrio cholera (Mol. Biology, 2003), Salmonella typhimurium (Science, 2005)

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 retractations that motivate improved software engineering practices in comp. biology

How computers broke science



Geoffrey Chang (Scripps, UCSD) works on crystalography 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.*

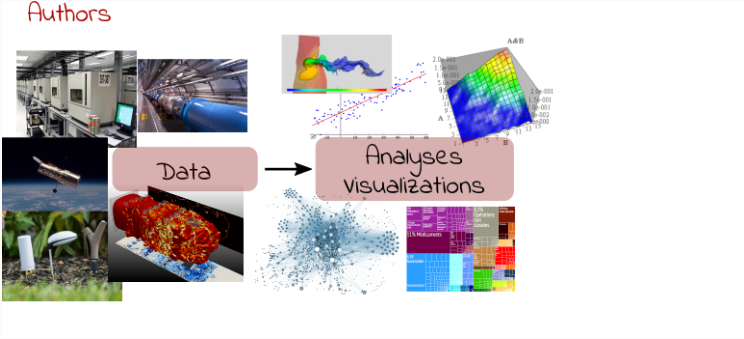
5 retractations that motivate improved software engineering practices in comp. biology

There is worse!

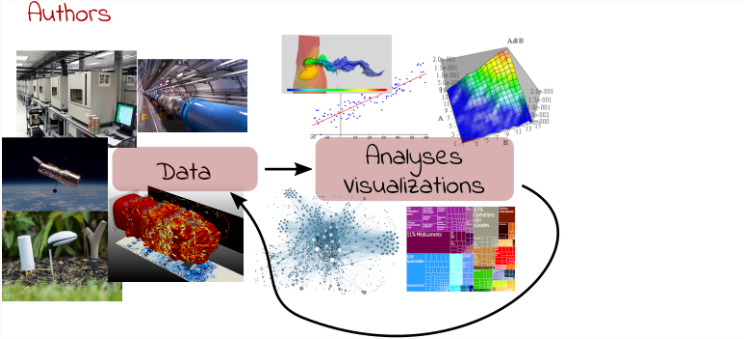
- The generalized and intensive use of spreadsheets (COVID tracing)
- Relying on black box statistical methods is infinitely easier than understanding them
  - Learning and Data Analytics frameworks are nuclear weapons
- Numerical errors and software environment unawareness

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

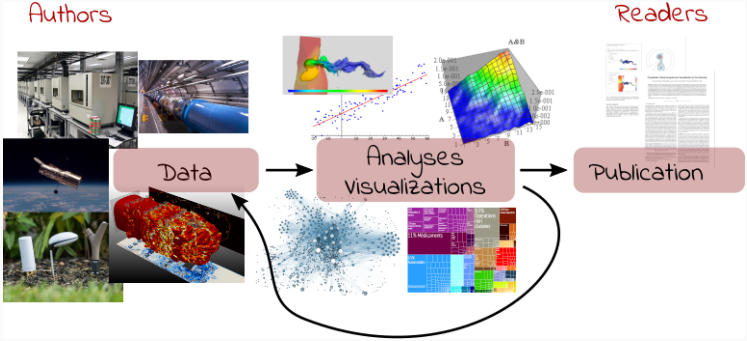
The processing steps between raw observations and findings have gotten increasingly numerous and complex.



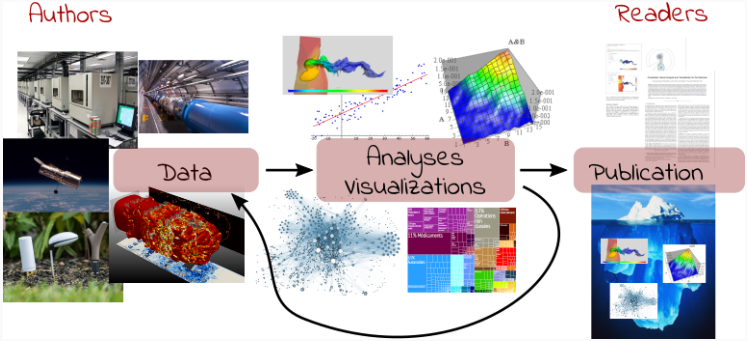
The processing steps between raw observations and findings have gotten increasingly numerous and complex.



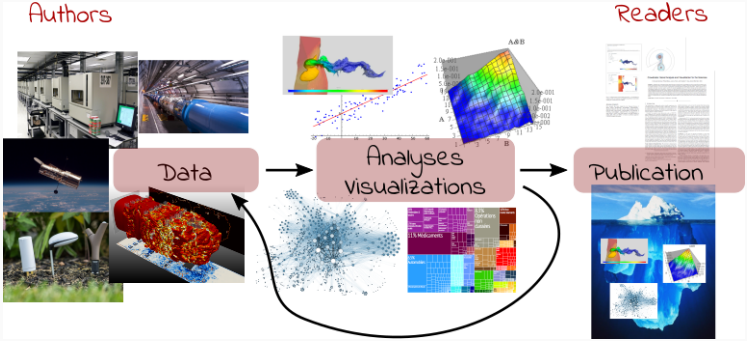
The processing steps between raw observations and findings have gotten increasingly numerous and complex.



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



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



Reproducible Research = Bridging the Gap by working Transparently

Different (but Converging) Reproducibility Concerns

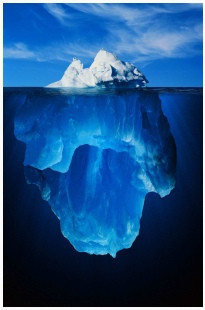
Reproducibility/robustness of the scientific fact, the statistical analysis, the computation, the observation, the process, ... ?

Different (but Converging) Reproducibility Concerns

Reproducibility/robustness of the scientific fact, the statistical analysis, the computation, the observation, the process, ... ?

Bad computer/statistic/publication practices "broke science" ☺

- Ensure articles and data are available
  - Open Access, Open Data, FAIR and DMP



Different (but Converging) Reproducibility Concerns

Reproducibility/robustness of the scientific fact, the statistical analysis, the computation, the observation, the process, ... ?

Bad computer/statistic/publication practices "broke science" ☺

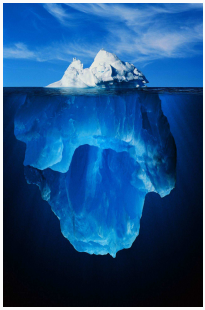
- Ensure articles and data are available
  - Open Access, Open Data, FAIR and DMP
- *Ensure* the experiment design and analysis is not broken
  - ~~data torture and hacking~~, pre-registration, education, ...

Different (but Converging) Reproducibility Concerns

Reproducibility/robustness of the scientific fact, the statistical analysis, the computation, the observation, the process, ... ?

Bad computer/statistic/publication practices "broke science" ☺

- Ensure articles and data are available
  - Open Access, Open Data, FAIR and DMP
- *Ensure* the experiment design and analysis is not broken
  - ~~data torture and hacking~~, pre-registration, education, ...
- Explain how the computations unfold and why
  - Notebooks (that can quickly get out of control)
  - Simple computational workflows



Different (but Converging) Reproducibility Concerns

Reproducibility/robustness of the scientific fact, the statistical analysis, the computation, the observation, the process, ... ?

Bad computer/statistic/publication practices "broke science" ☺

- Ensure articles and data are available
  - Open Access, Open Data, FAIR and DMP
- *Ensure* the experiment design and analysis is not broken
  - ~~data torture and hacking~~, pre-registration, education, ...
- Explain how the computations unfold and why
  - Notebooks (that can quickly get out of control)
  - Simple computational workflows
- Allow to inspect, rerun, make variations on computations
  - Access to code, data, ...

Different (but Converging) Reproducibility Concerns

Reproducibility/robustness of the scientific fact, the statistical analysis, the computation, the observation, the process, ... ?

Bad computer/statistic/publication practices "broke science" ☺

- Ensure articles and data are available
  - Open Access, Open Data, FAIR and DMP
- *Ensure* the experiment design and analysis is not broken
  - ~~data torture and hacking~~, pre-registration, education, ...
- Explain how the computations unfold and why
  - Notebooks (that can quickly get out of control)
  - Simple computational workflows
- Allow to inspect, rerun, make variations on computations
  - Access to code, data, ... options/parameters, environment, resources?

This requires first class software engineering practices instead of building on prototypes

Software factories, Archives, and Provenance Tracking tools

A few Computer Science Challenges

What is hiding behind a simple

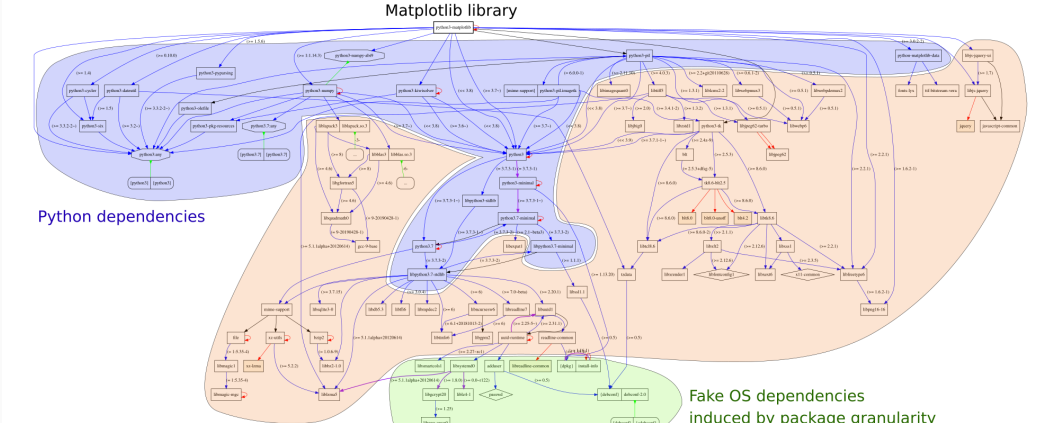
```
import matplotlib
```

```
Package: python3-matplotlib
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)
```

What is hiding behind a simple

```
import matplotlib
```

`Package: python3-matplotlib`



Software Environments Nightmare

Python and its rapidly evolving environment

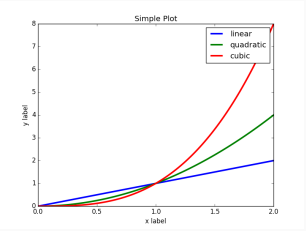
```
python2 -c "print(10/3)"
python3 -c "print(10/3)"
```

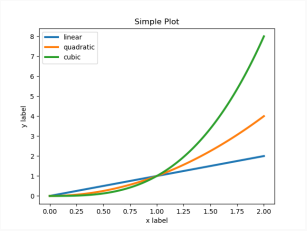
```
3
3.3333333333333335
```

Software Environments Nightmare

Python and its rapidly evolving environment

```
python2 -c "print(10/3)"
python3 -c "print(10/3)"
```

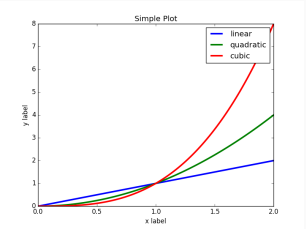


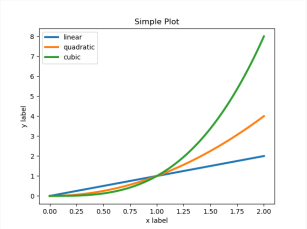


Software Environments Nightmare

Python and its rapidly evolving environment

```
python2 -c "print(10/3)"
python3 -c "print(10/3)"
```



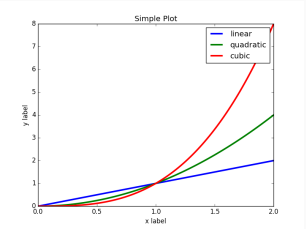


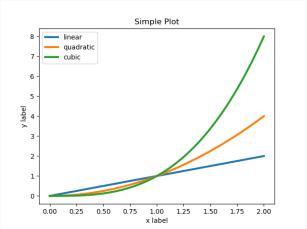
Cortical Thickness Measurements (PLOS ONE, June 2012): *FreeSurfer*: *differences were found between the Mac and HP workstations and between Mac OSX 10.5 and OSX 10.6.*

Software Environments Nightmare

Python and its rapidly evolving environment

```
python2 -c "print(10/3)"
python3 -c "print(10/3)"
```





Cortical Thickness Measurements (PLOS ONE, June 2012): *FreeSurfer*: *differences were found between the Mac and HP workstations and between Mac OSX 10.5 and OSX 10.6.*

Operating System Architectures

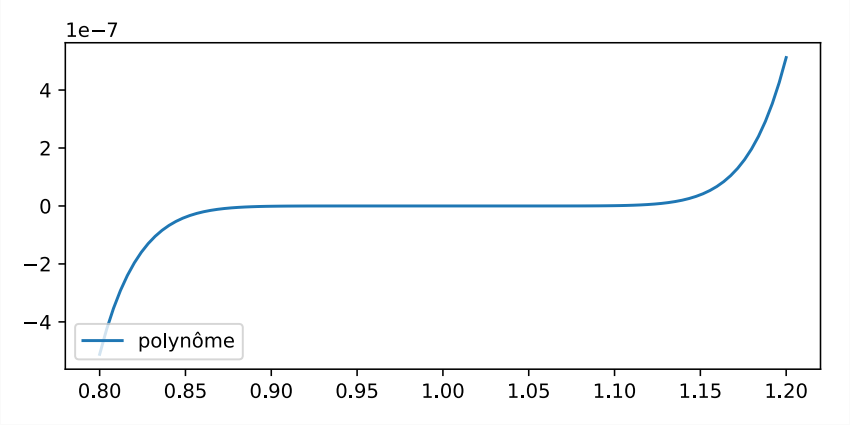
- Single-User/Single-Tasking operating system (e.g., DOS, Palm OS)
  - A single file system, a single code running at a time, no need for protection
- Single-User/Multi-Tasking operating system (Windows, "Android" 😉)
  - Requires isolation between processes (security, fairness)
- Multiple-User/Multi-Tasking operating system (UNIX)
  - Separate home directories with personal data
  - Shared program (single version)

Evolution is motivated by user needs but constrained by available technology

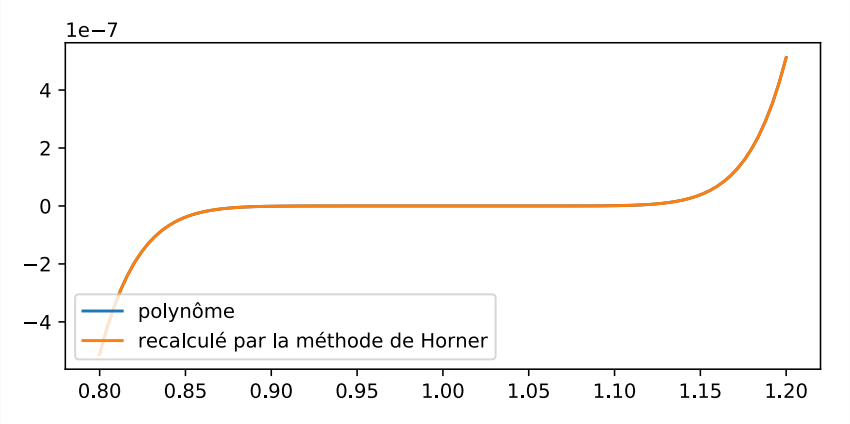
Example: Virtual machines

- 1970: VM/370 enables time-shared execution of DOS (*Full virtualization*)
- 1994: Java Virtual Machine (*Process virtualization*)
- 2000: FreeBSD jail/Linux Chroot and then containers like docker/LXC/... (*OS-level virtualization*)
- 2005: additional hardware to support full virtualization from Intel for KVM, XEN, VMWARE...

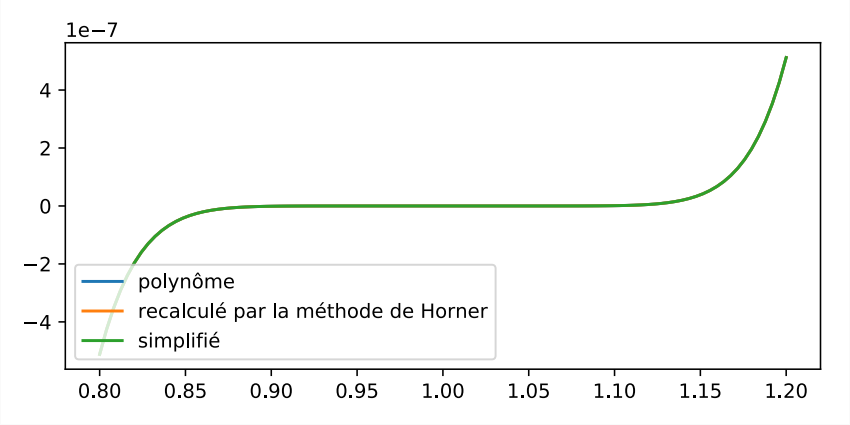
But should our problem be solved through OS architecture or through package management ?



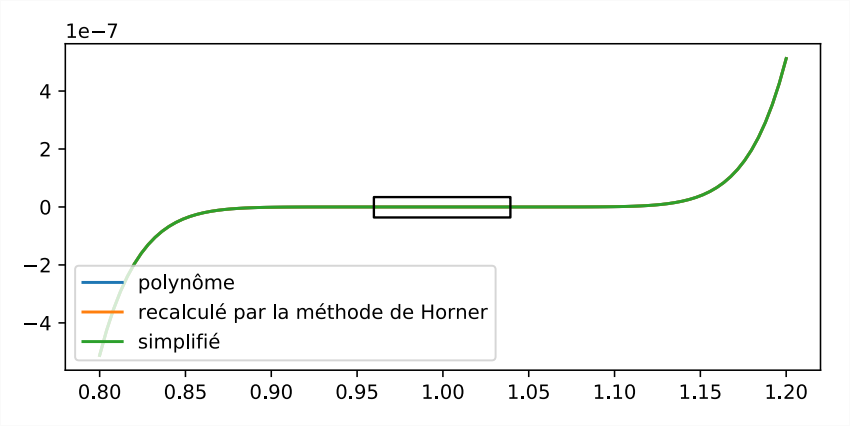

```
def polynome(x):
    return x**9 - 9.*x**8 + 36.*x**7 - 84.*x**6 + 126.*x**5 \
           - 126.*x**4 + 84.*x**3 - 36.*x**2 + 9.*x - 1.
```

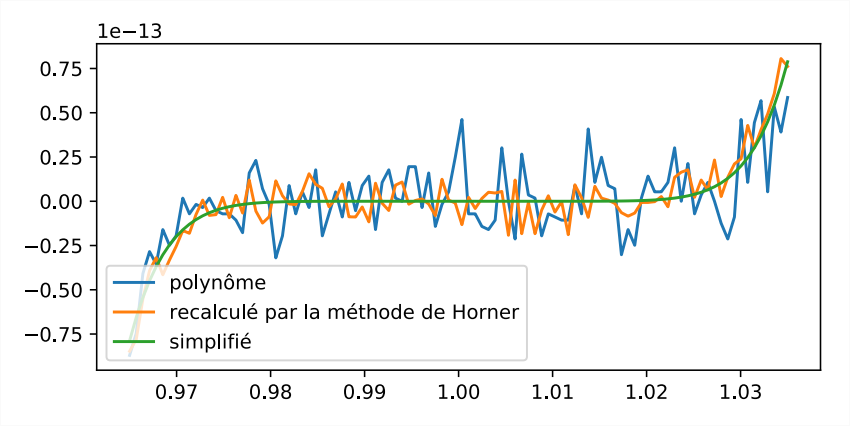



```
def horner(x):
    return x*(x*(x*(x*(x*(x*(x*(x*(x - 9.) + 36.) - 84.) + 126.) \
           - 126.) + 84.) - 36.) + 9.) - 1.
```




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


1e−7
4
2
0
−2
−4
0.80
0.85
0.90
0.95
1.00
1.05
1.10
1.15
1.20
polynôme
recalculé par la méthode de Horner
simplifié


1e−13
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
0.97
0.98
0.99
1.00
1.01
1.02
1.03
polynôme
recalculé par la méthode de Horner
simplifié

Rounding

- Every operation includes implicit rounding.
- a+b is actually $\texttt{round}(a+b)$.
- Unfortunately:

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

- Operation order therefore matters.

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

How to explain it to my compiler?

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

Two options for computing reproducibly:

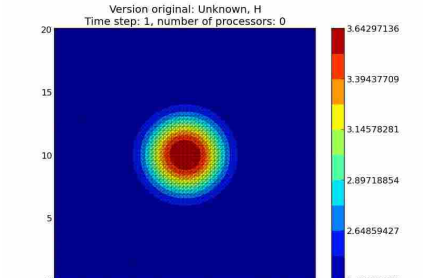
1. Insist on the preservation of operation order,
   - if the language permits it.
   - Example: Module ‘ieee_arithmetic‘ in Fortran 2003
2. Make compilation reproducible:
   - Record the precise compiler version
   - Record all compilation options

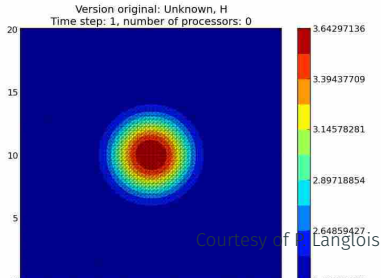
Telemac2D: the simplest gouttedo simulation

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, . . .)

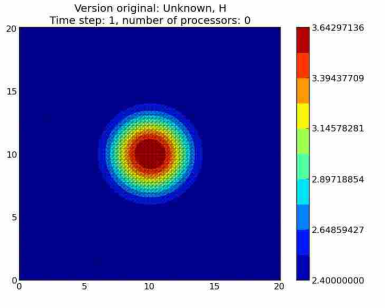




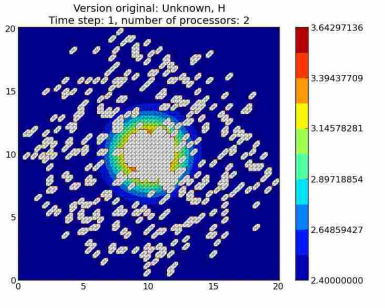
A white plot displays a non-reproducible value

Numerical reproducibility?

time step = 1



Sequential

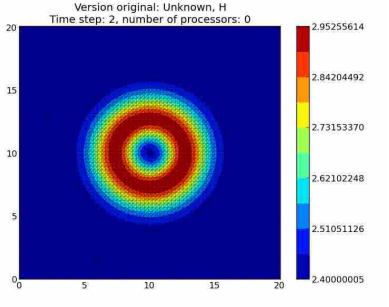


Parallel $p = 2$

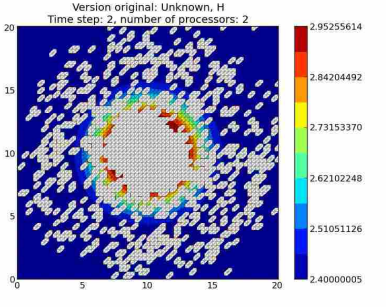
A white plot displays a non-reproducible value

Numerical reproducibility?

time step = 2



Sequential



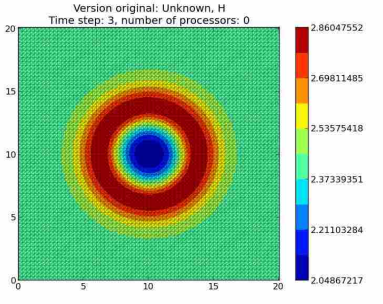
Parallel $p = 2$

Courtesy of P. Langlois and R. Nheili

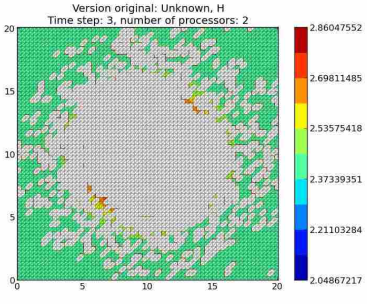
A white plot displays a non-reproducible value

Numerical reproducibility?

time step = 3



Sequential



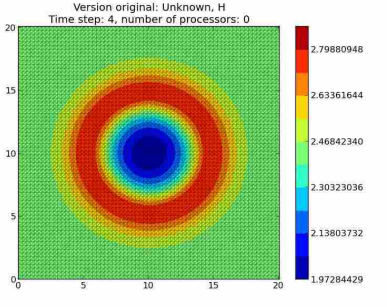
Parallel $p = 2$

Courtesy of P. Langlois and R. Nheili

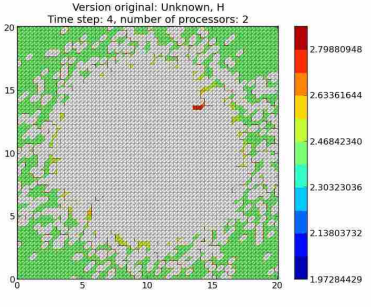
A white plot displays a non-reproducible value

Numerical reproducibility?

time step = 4



Sequential



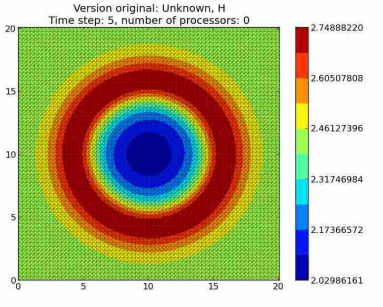
Parallel $p = 2$

Courtesy of P. Langlois and R. Nheili

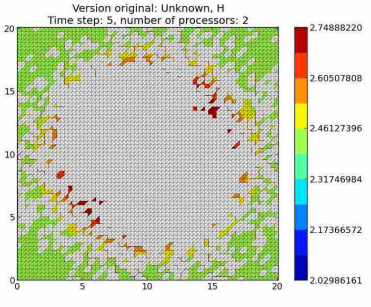
A white plot displays a non-reproducible value

Numerical reproducibility?

time step = 5



Sequential



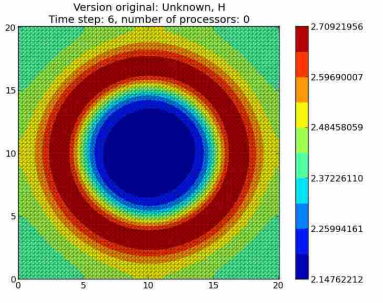
Parallel $p = 2$

Courtesy of P. Langlois and R. Nheili

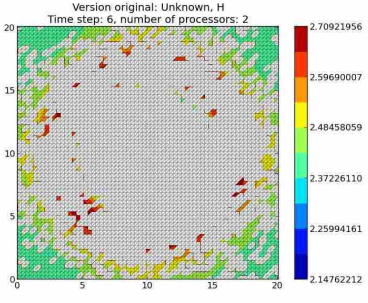
A white plot displays a non-reproducible value

Numerical reproducibility?

time step = 6



Sequential



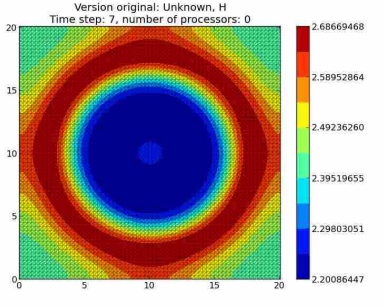
Parallel $p = 2$

Courtesy of P. Langlois and R. Nheili

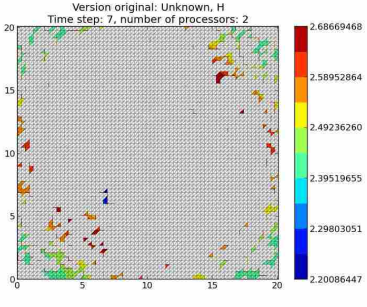
A white plot displays a non-reproducible value

Numerical reproducibility?

time step = 7



Sequential



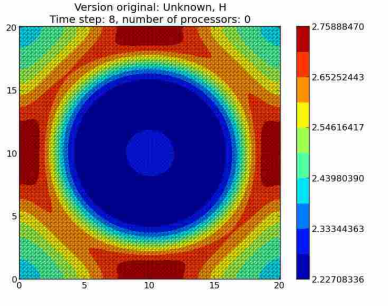
Parallel $p = 2$

Courtesy of P. Langlois and R. Nheili

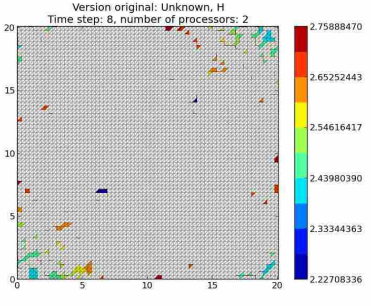
A white plot displays a non-reproducible value

Numerical reproducibility?

time step = 8



Sequential

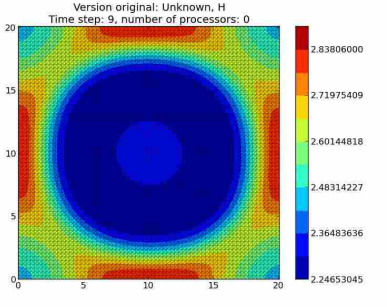


Parallel $p = 2$

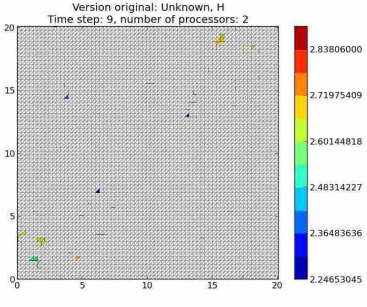
A white plot displays a non-reproducible value

Numerical reproducibility?

time step = 9



Sequential

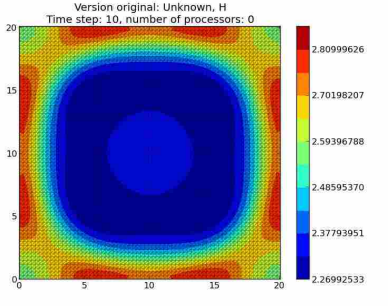


Parallel $p = 2$

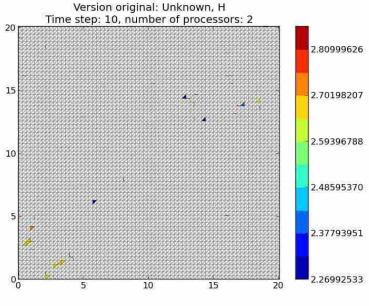
A white plot displays a non-reproducible value

Numerical reproducibility?

time step = 10



Sequential

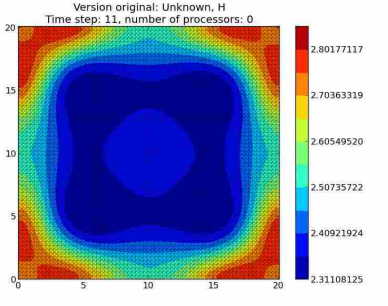


Parallel $p = 2$

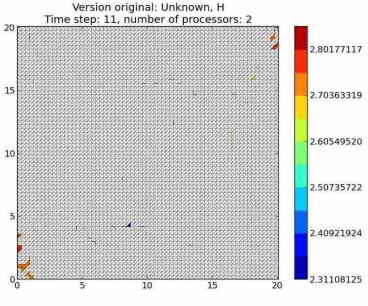
A white plot displays a non-reproducible value

Numerical reproducibility?

time step = 11



Sequential

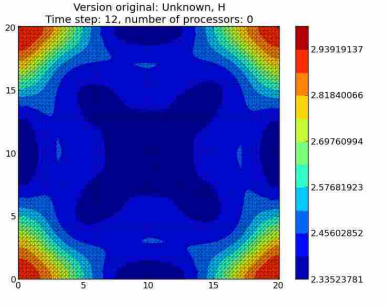


Parallel $p = 2$

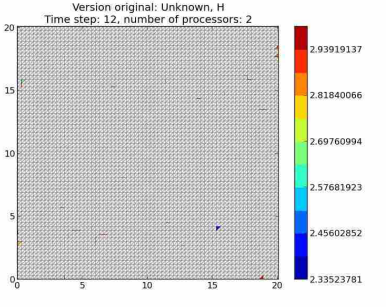
A white plot displays a non-reproducible value

Numerical reproducibility?

time step = 12



Sequential

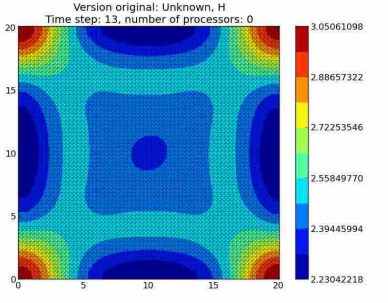


Parallel $p = 2$

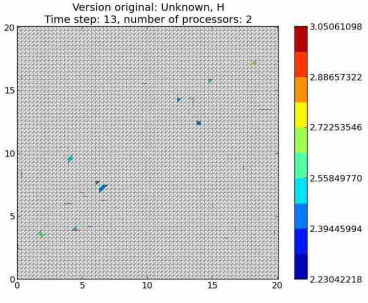
A white plot displays a non-reproducible value

Numerical reproducibility?

time step = 13



Sequential

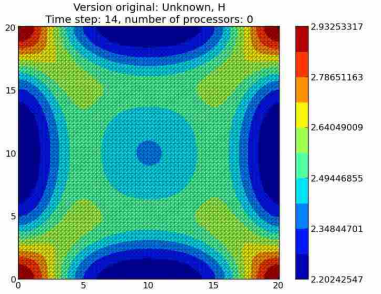


Parallel $p = 2$

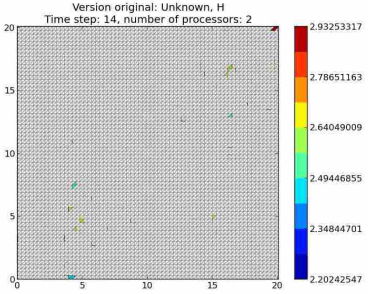
A white plot displays a non-reproducible value

Numerical reproducibility?

time step = 14



Sequential



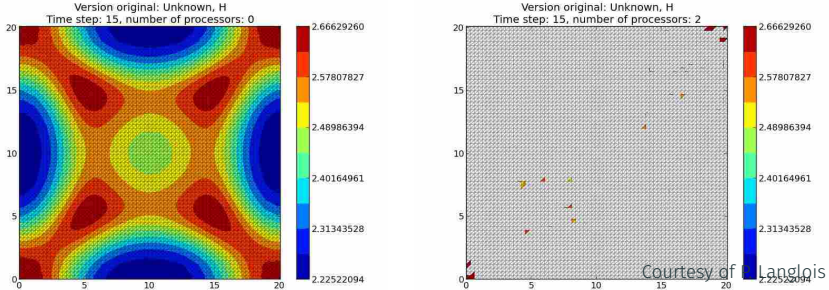
Parallel $p = 2$

Courtesy of P. Langlois and R. Nheili

A white plot displays a non-reproducible value

NO numerical reproducibility!

time step = 15



Courtesy of M. Langlois and R. Nheili

Did I mention we have parallel machines nowadays?

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

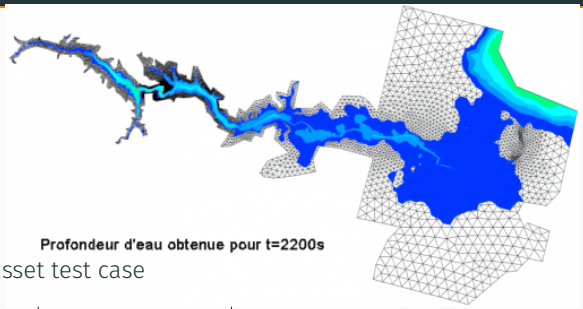


TABLE 1.1: Reproducibility failure of the Malpasset test case

	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 the algorithm, in its implementation, in its execution.

The whole chain needs to be revisited.

Courtesy of P. Langlois and R. Nheili

Resources and Acknowledgments

Resources and Acknowledgments

https://github.com/alegrand/SMPE/raw/master/lectures/talk_20_10_08_DUISN.pdf

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

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

Resources and Acknowledgments

https://github.com/alegrand/SMPE/raw/master/lectures/talk_20_10_08_DUISN.pdf

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

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

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

- Konrad Hinsen, Christophe Pouzat, Alexandre Hocquet
- 3rd Edition: March 2020 – March 2021
- MOOC RR "Advanced" planned for 2021