Reproducible Research and Computer Science

Arnaud Legrand

Master 1 UGA
December 2024

What is Science about?

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

What is Science about?

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

Dictionary of science and technology

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

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

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

What is Science about?

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

Dictionary of science and technology

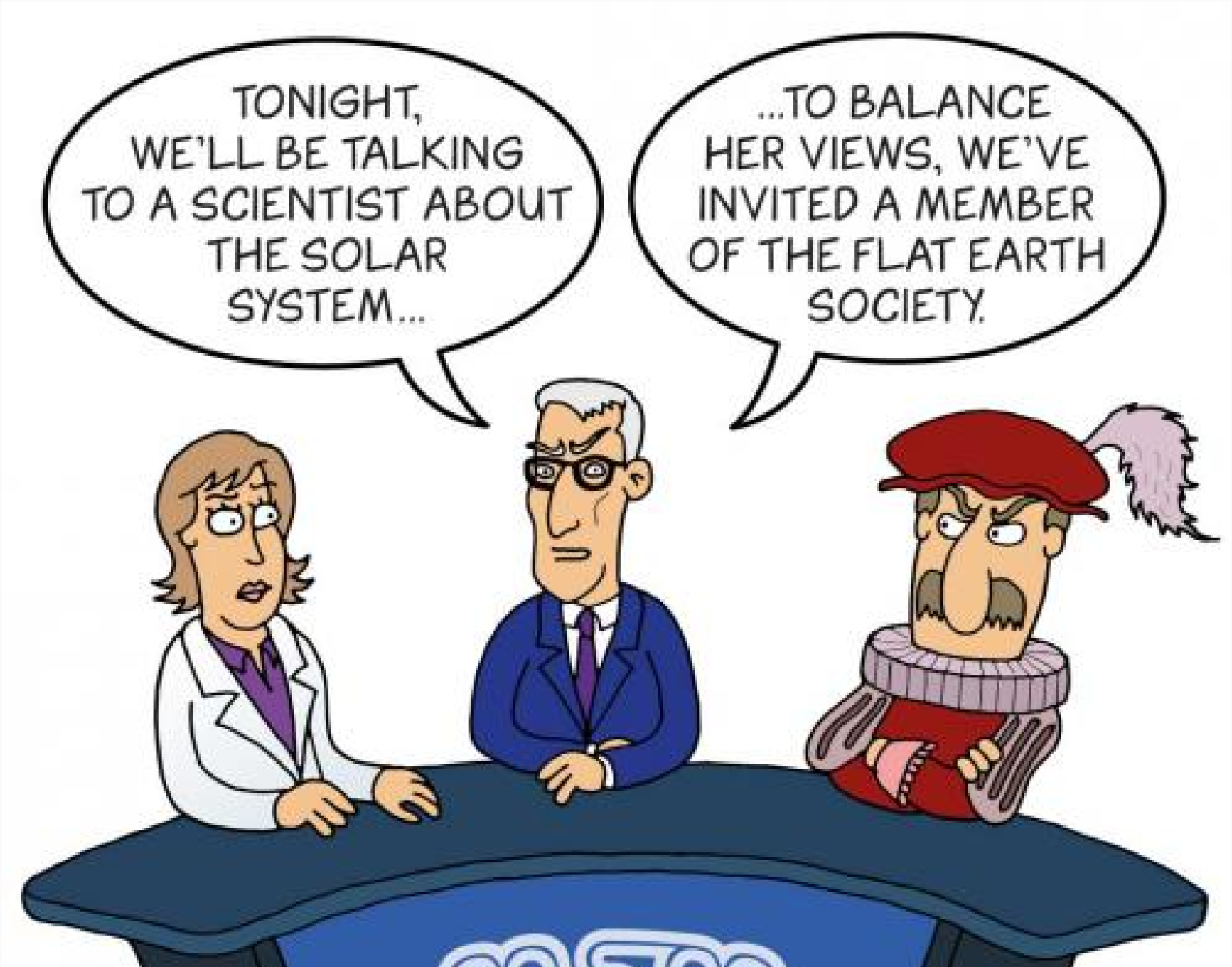
1. the study of the physical and natural world and phenomena, especially by using systematic observation and experiment
2. a particular area of study or knowledge of the physical world
3. a systematically organized body of knowledge about a particular subject

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

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

Building Reliable Knowledge

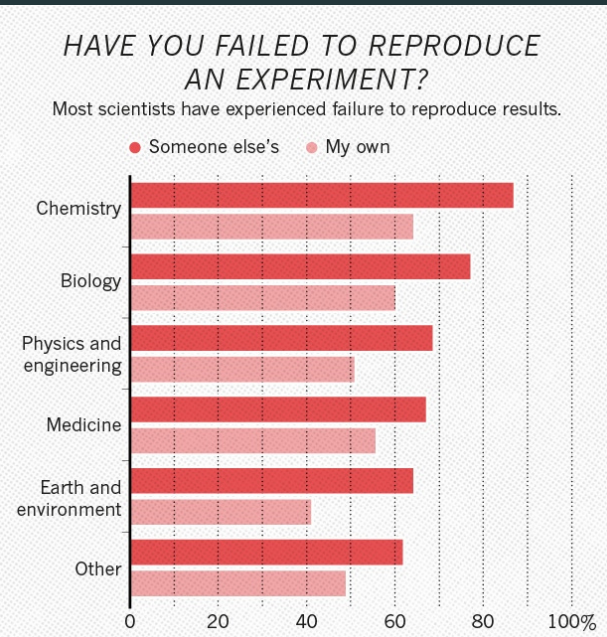
Scientific Consensus vs. Democracy and Freedom of Speech

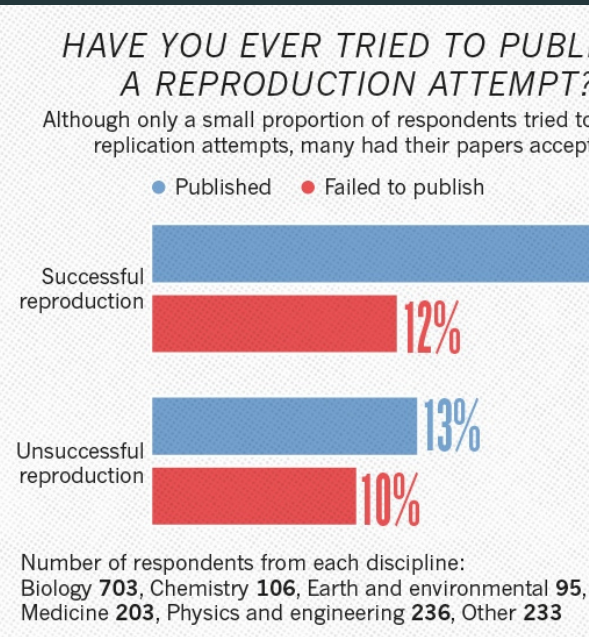


Reproducibility Crisis

Is there a reproducibility crisis

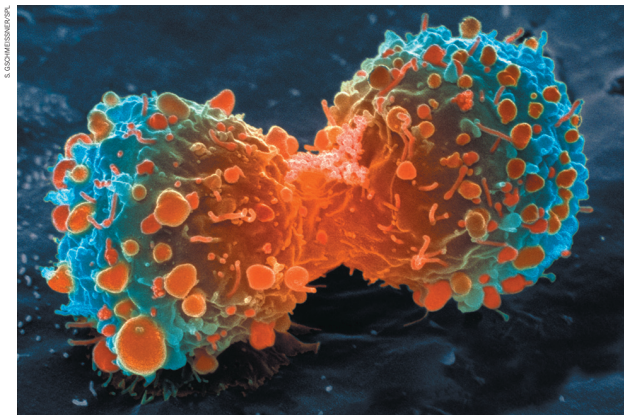
A Nature survey, 2016






COMMENT

AVIAN INFLUENZA Shift expertise to track mutations where they emerge p.534 | EARTH SYSTEMS Past climates give valuable clues to future warming p.537 | HISTORY OF SCIENCE Descartes' lost letter tracked using Google p.540 | OBITUARY Wylie Vale and an elusive stress hormone p.542


S. GSCHMEISSNER/SPL

Many landmark findings in preclinical oncology research are not reproducible, in part because of inadequate cell lines and animal models.

Raise standards for preclinical cancer research

C. Glenn Begley and **Lee M. Ellis** propose how methods, publications and incentives must change if patients are to benefit.

Efforts over the past decade to characterize the genetic alterations in human cancers have led to a better understanding of molecular drivers of this complex set of diseases. Although we in the cancer field hoped that this would lead to more effective drugs, historically, our ability to translate cancer research to clinical success has been remarkably low[1]. Sadly, clinical trials in oncology have the highest failure rate compared with other therapeutic areas. Given the high unmet need in oncology, it is understandable that barriers to clinical development may be lower than for other disease areas, and a larger number of drugs with suboptimal preclinical validation will enter oncology trials. However, this low success rate is not sustainable or acceptable, and investigators must reassess their approach to translating discovery research into greater clinical success and impact.

Many factors are responsible for the high failure rate, notwithstanding the inherently difficult nature of this disease. Certainly, the limitations of preclinical tools such as inadequate cancer-cell-line and mouse models[2] make it difficult for even ▸

Must try harder

Too many sloppy mistakes are creeping into scientific papers at the data — and at themselves.

Six red flags for suspect work

C. Glenn Begley explains how to recognize the precilinical papers in which the data won't stand up

Error prone

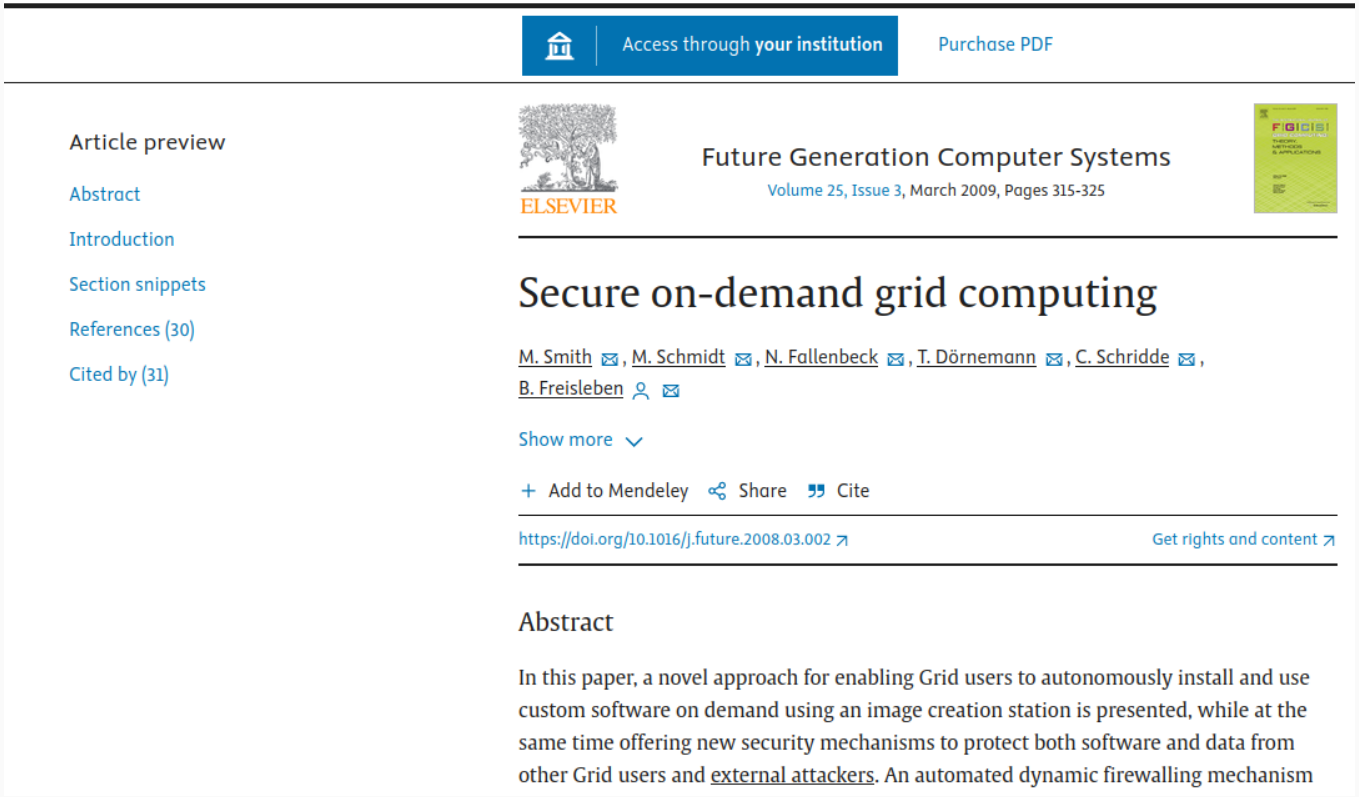
Biologists must realize the pitfalls of massive amount of data

Know when your numbers are significant

- Nekrutenko & Taylor, Nature Genetics (2012)
- Alsheikh-Ali et al. PLoS ONE (2011)
- Begley & Ellis Nature (2012)

Common Reproducibility Pitfalls

Go read the paper by Smith et. al. 2009



Access through your institution Purchase PDF

Article preview

Abstract

Introduction

Section snippets

References (30)

Cited by (31)

ELSEVIER

Future Generation Computer Systems

Volume 25, Issue 3, March 2009, Pages 315-325

Secure on-demand grid computing

M. Smith, M. Schmidt, N. Fallenbeck, T. Dörnemann, C. Schridde, B. Freisleben

Show more

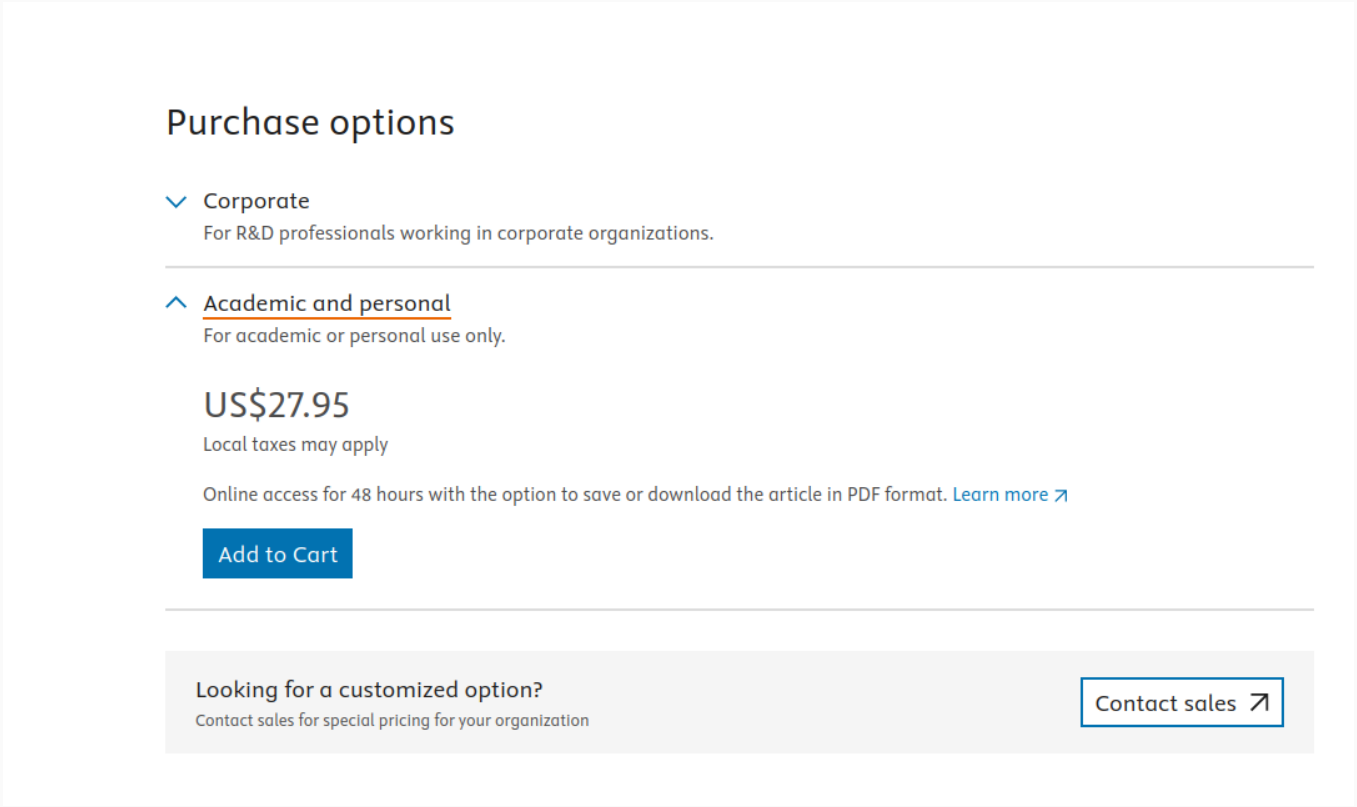
+ Add to Mendeley Share Cite

https://doi.org/10.1016/j.future.2008.03.002 Get rights and content

Abstract

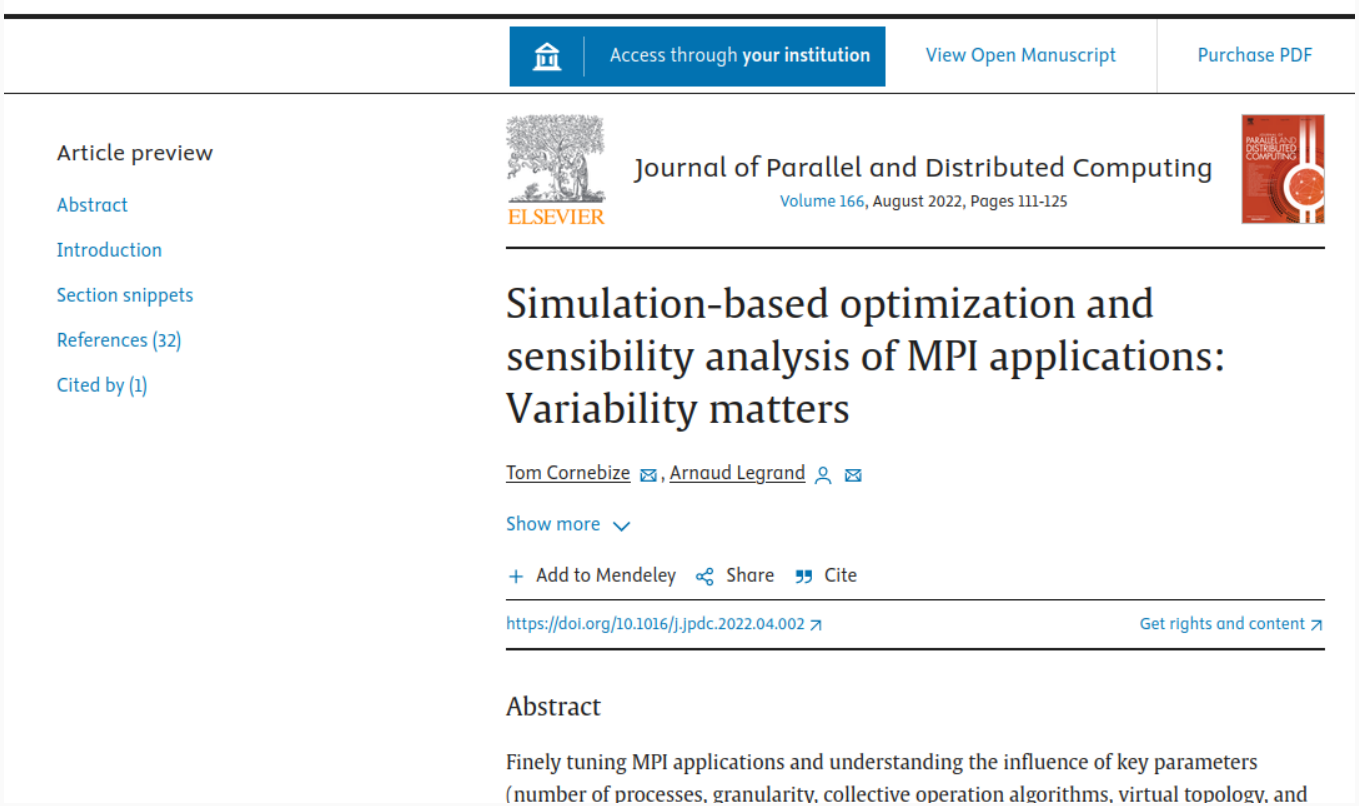
In this paper, a novel approach for enabling Grid users to autonomously install and use custom software on demand using an image creation station is presented, while at the same time offering new security mechanisms to protect both software and data from other Grid users and external attackers. An automated dynamic firewalling mechanism

Go read the paper by Smith et. al. 2009



- Use your institution subscription... or Sci-Hub 😉,

Go read the paper by Smith et. al. 2009



Access through your institution | View Open Manuscript | Purchase PDF

Article preview

Abstract

Introduction

Section snippets

References (32)

Cited by (1)

ELSEVIER

Journal of Parallel and Distributed Computing

Volume 166, August 2022, Pages 111-125

Simulation-based optimization and sensibility analysis of MPI applications: Variability matters

Tom Cornebize, Arnaud Legrand

Show more

+ Add to Mendeley | Share | Cite

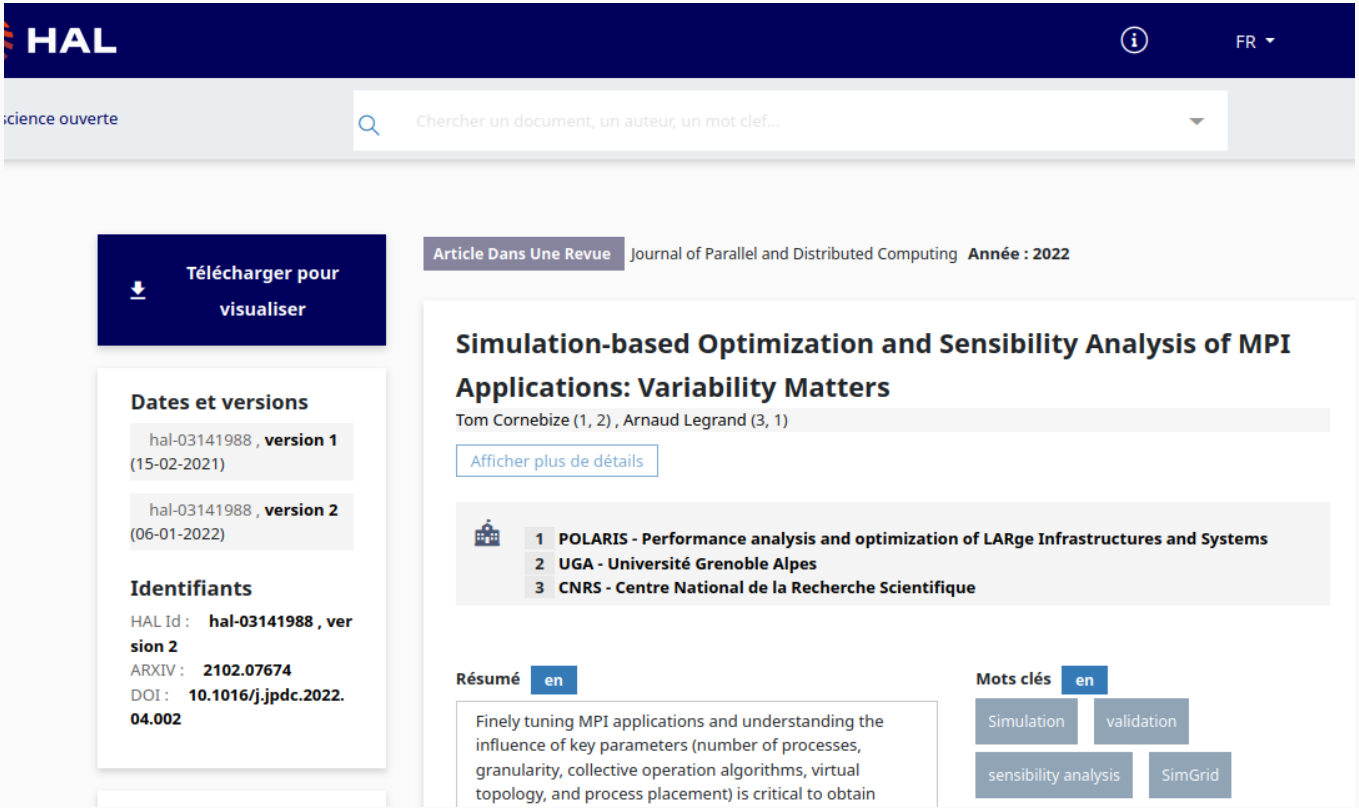
https://doi.org/10.1016/j.jpdc.2022.04.002 | Get rights and content

Abstract

Finely tuning MPI applications and understanding the influence of key parameters (number of processes, granularity, collective operation algorithms, virtual topology, and

- Use your institution subscription... or Sci-Hub 😉, or HAL/Arxiv

Go read the paper by Smith et. al. 2009



- Use your institution subscription... or Sci-Hub 😉, or HAL/Arxiv

Just compare to the Algorithm they proposed

Rodriguez et al., CONCUR'15

Unfolding-based Partial Order Reduction*

César Rodríguez[1], Marcelo Sousa[2], Subodh Sharma[3], and Daniel Kroening[4]

1 Université Paris 13, Sorbonne Paris Cité, LIPN, CNRS, France
2,4 Department of Computer Science, University of Oxford, UK
3 Indian Institute of Technology Delhi, India

Abstract

Partial order reduction (POR) and net unfoldings are two alternative methods to tackle state-space explosion caused by concurrency. In this paper, we propose the combination of both approaches in an effort to combine their strengths. We first define, for an abstract execution model, unfolding semantics parameterized over an arbitrary independence relation. Based on it, our main contribution is a novel stateless POR algorithm that explores at most one execution per Mazurkiewicz trace, and in general, can explore exponentially fewer, thus achieving a form of *super-optimality*. Furthermore, our unfolding-based POR copes with non-terminating executions and incorporates state-caching. Over benchmarks with busy-waits, among others, our experiments show a dramatic reduction in the number of executions when compared to a

Jul 2015

Just compare to the Algorithm they proposed

Rodriguez et al., CONCUR'15

Algorithm 1: An unfolding-based POR exploration algorithm.

```
1  Initially, set U := {⊥}, set G := ∅, and call Explore({⊥}, ∅, ∅).
2  Procedure Explore(C, D, A)
3      Extend(C)
4      if en(C) = ∅ return
5      if A = ∅
6          Choose e from en(C)
7      else
8          Choose e from A ∩ en(C)
9      Explore(C ∪ {e}, D, A ∖ {e})
10     if ∃J ∈ Alt(C, D ∪ {e})
11         Explore(C, D ∪ {e}, J ∖ C)
12     Remove(e, C, D)
13 Procedure Extend(C)
14     Add ex(C) to U
15 Procedure Remove(e, C, D)
16     Move {e} ∖ Q_{C,D,U} from U to G
17     foreach ê ∈ #^i_U(e)
18         Move [ê] ∖ Q_{C,D,U} from U to G
```

- Looks good!

Just compare to the Algorithm they proposed

Rodriguez et al., CONCUR'15

We give some new definitions. Let C be a configuration of $\mathcal{U}$. The *extensions* of C, written $ex(C)$, are all those events outside C whose causes are included in C. Formally, $ex(C) := \{e \in E : e \notin C \wedge \lceil e \rceil \subseteq C\}$. We let $en(C)$ denote the set of events *enabled* by C, i.e., those corresponding to the transitions enabled at $state(C)$, formally defined as $en(C) := \{e \in ex(C) : C \cup \{e\} \in conf(\mathcal{U})\}$. All those events in $ex(C)$ which are not in $en(C)$ are the *conflicting extensions*, $cex(C) := \{e \in ex(C) : \exists e' \in C,\ e \#^i e'\}$. Clearly, sets $en(C)$ and $cex(C)$ partition the set $ex(C)$. Lastly, we define $\#^i(e) := \{e' \in E : e \#^i e'\}$, and $\#^i_U(e) := \#^i(e) \cap U$. The difference between both is that $\#^i(e)$ contains events from *anywhere* in the unfolding structure, while $\#^i_U(e)$ can only *see* events in U.

The algorithm is given in Alg. 1. `Explore`(C, D, A), the main procedure, is given the configuration that is to be explored as the parameter C. The parameter D (for *disabled*) is the set of set of events that have already been explored and prevents that `Explore()` repeats work. It can be seen as a *sleep set* [7]. Set A (for *add*) is occasionally used to guide the direction of the exploration.

Additionally, a global set U stores all events presently known to the algorithm. Whenever some event can safely be discarded from memory, `Remove` will move it from U to G (for *garbage*). Once in G, it can be discarded at any time, or be preserved in G in order to save work when it is re-inserted in U. Set G is thus our *cache memory* of events.

The key intuition in Alg. 1 is as follows. A call to `Explore`(C, D, A) visits all maximal configurations of $\mathcal{U}$ which contain C and do not contain D; and the first one explored will

- Looks good! Err... not so simple. Depending on how you do this, you quickly move from polynomial to exponential.

Just compare to the Algorithm they proposed

Rodriguez et al., CONCUR'15

POSIX threads. The analyzer accepts deterministic programs, implements a variant of Alg. 1 where the computation of the alternatives is memoized, and supports cutoffs events with the criteria defined in § 5.

We ran Poet on a number of multi-threaded C programs. Most of them are adapted from benchmarks of the Software Verification Competition [17]; others are used in related works [8, 19, 2]. We investigate the characteristics of average program unfoldings (depth, width, etc.) as well as the frequency and impact of cutoffs on the exploration. We also compare Poet with Nidhugg [1], a state-of-the-art stateless model checking for multi-threaded C programs that implements Source-DPOR [2], an efficient but non-optimal DPOR. All experiments were run on an Intel Xeon CPU with 2.4 GHz and 4 GB memory. Tables 1 and 2 give our experimental data for programs with acyclic and non-acyclic state spaces, respectively.

For programs with acyclic state spaces (Table 1), Poet with and without cutoffs seems to perform the same exploration when the unfolding has no cutoffs, as expected. Furthermore, the number of explored executions also coincides with Nidhugg when the latter reports 0 sleep-set blocked executions (cf., § 4), providing experimental evidence of Poet's optimality.

The unfoldings of most programs in Table 1 do not contain cutoffs. All these programs are deterministic, and many of them highly sequential (Stf, Spin08, Fib), features known to make cutoffs unlikely. Ccnf(n) are concurrent programs composed of $n-1$ threads where thread i and $i+1$ race on writing one variable, and are independent of all remaining

[3] Source code and benchmarks available from: http://www.cs.ox.ac.uk/people/marcelo.sousa/poet/.

- Looks good! Err... not so simple. Depending on how you do this, you quickly move from polynomial to exponential.
- Experiments!

Just compare to the Algorithm they proposed

Rodriguez et al., CONCUR'15

Table 1 Programs with acyclic state space. Columns are: $|P|$: nr. of threads; $|I|$: nr. of explored traces; $|B|$: nr. of sleep-set blocked executions; $t(s)$: running time; $|E|$: nr. of events in $\mathcal{U}$; $|E_{\text{cut}}|$: nr. of cutoff events; $|\Omega|$: nr. of maximal configurations; $\langle|U_\Omega|\rangle$: avg. nr. of events in U when exploring a maximal configuration. A $*$ marks programs containing bugs. <7K reads as "*fewer than 7000*".

Benchmark		Nidhugg			Poet (without cutoffs)				Poet (with cutoffs)				
Name	$\|P\|$	$\|I\|$	$\|B\|$	$t(s)$	$\|E\|$	$\|\Omega\|$	$\langle\|U_\Omega\|\rangle$	$t(s)$	$\|E\|$	$\|E_{\text{cut}}\|$	$\|\Omega\|$	$\langle\|U_\Omega\|\rangle$	$t(s)$
Stf	3	6	0	0.06	121	6	79	0.04	121	0	6	79	0.06
Stf*	3	-	-	0.05	-	-	-	0.02	-	-	-	-	0.03
Spin08	3	84	0	0.08	2974	84	1506	2.04	2974	0	84	1506	2.93
Fib	3	8953	0	3.36	<185K	8953	92878	305	<185K	0	8953	92878	704
Fib*	3	-	-	0.74	-	-	-	81.0	-	-	-	-	133
Ccnf(9)	9	16	0	0.05	49	16	46	0.07	49	0	16	46	0.06
Ccnf(17)	17	256	0	0.15	97	256	94	5.76	97	0	256	94	6.09
Ccnf(19)	19	512	0	0.28	109	512	106	22.5	109	0	512	106	22.0
Ssb	5	4	2	0.05	48	4	38	0.03	46	1	4	37	0.03
Ssb(1)	5	22	14	0.06	245	23	143	0.11	237	4	23	140	0.11
Ssb(3)	5	169	67	0.12	2798	172	1410	3.51	1179	48	90	618	0.90
Ssb(4)	5	336	103	0.15	<7K	340	3333	20.3	2179	74	142	1139	2.07
Ssb(8)	5	2014	327	0.85	<67K	2022	32782	4118	<12K	240	470	6267	32.1

Lastly, we note that this cutoff approach imposes no liability on what events shall be kept in the prefix, set G can be cleaned at discretion. Also, redefining (7) to use adequate orders [5] is straightforward, cf. App. C.1 (in our proofs we actually assume adequate orders).

- Looks good! Err... not so simple. Depending on how you do this, you quickly move from polynomial to exponential.
- Experiments! ⚠Possible *404, code not found!* ahead!!!

Just compare to the Algorithm they proposed

Rodriguez et al., CONCUR'15

- Looks good! Err... not so simple. Depending on how you do this, you quickly move from polynomial to exponential.
- Experiments! ⚠Possible *404, code not found!* ahead!!!
- Wait, what's this language? Did this ever run one day?

Just compare to the Algorithm they proposed

Rodriguez et al., CONCUR'15

- Looks good! Err... not so simple. Depending on how you do this, you quickly move from polynomial to exponential.
- Experiments! ⚠Possible *404, code not found!* ahead!!!
- Wait, what's this language? Did this ever run one day?
- Wow, I'll need `CPLEX` and `Gurobi` but all I have is `lpsolve` or `glpk`

Just compare to the Algorithm they proposed

Rodriguez et al., CONCUR'15

- Looks good! Err... not so simple. Depending on how you do this, you quickly move from polynomial to exponential.
- Experiments! ⚠Possible *404, code not found!* ahead!!!
- Wait, what's this language? Did this ever run one day?
- Wow, I'll need `CPLEX` and `Gurobi` but all I have is `lpsolve` or `glpk`
- Sweet, they provided a binary! Oh, wait, MacOSX in 2015 ?!?

Just compare to the Algorithm they proposed

Rodriguez et al., CONCUR'15

- Looks good! Err... not so simple. Depending on how you do this, you quickly move from polynomial to exponential.
- Experiments! ⚠Possible *404, code not found!* ahead!!!
- Wait, what's this language? Did this ever run one day?
- Wow, I'll need `CPLEX` and `Gurobi` but all I have is `lpsolve` or `glpk`
- Sweet, they provided a binary! Oh, wait, MacOSX in 2015 ?!?
- The GitHub webpage says it requires Foo, Bar, and Baz, but none of the versions I find appear to work.

Just compare to the Algorithm they proposed

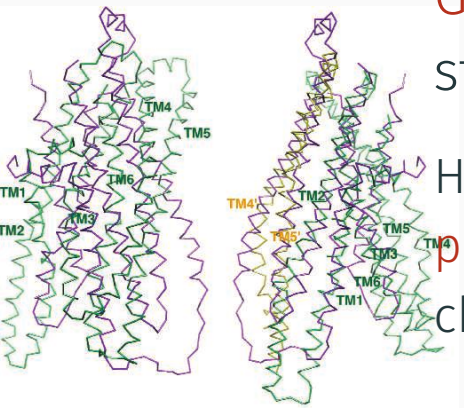
Rodriguez et al., CONCUR'15

- Looks good! Err... not so simple. Depending on how you do this, you quickly move from polynomial to exponential.
- Experiments! ⚠Possible *404, code not found!* ahead!!!
- Wait, what's this language? Did this ever run one day?
- Wow, I'll need `CPLEX` and `Gurobi` but all I have is `lpsolve` or `glpk`
- Sweet, they provided a binary! Oh, wait, MacOSX in 2015 ?!?
- The GitHub webpage says it requires Foo, Bar, and Baz, but none of the versions I find appear to work.
- With which parameters and data set do you run this code? And Why?

In the end, one new thesis to understand this paper and contribute.

Blaming "Computer Science"

How computers broke science



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

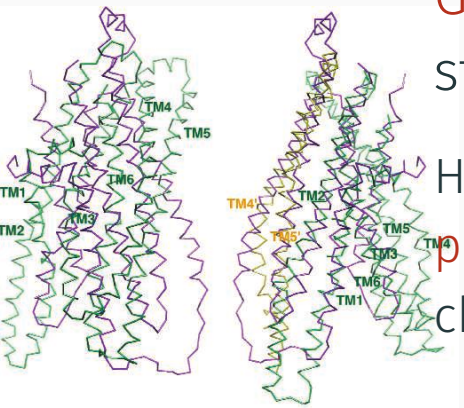
He specialized in structures of multidrug resistant transporter proteins in bacteria: MsbA de Escheria Choli (Science, 2001), Vibrio cholera (Mol. Biology, 2003), Salmonella typhimurium (Science, 2005)

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

Computers...

How computers broke science –

and what we can do about it

> *Most modern science is so complicated, and most journal articles so brief, it's impossible for the article to include details of many important methods and decisions made by the researcher as he analyzed his data on his computer. How, then, can another researcher judge the reliability of the results, or reproduce the analysis?*

– Ben Marwick,
The conversation, 2015

Point-and-click procedures are rampant but they hinder reproducibility.

Spreadsheets are generalized and intensively uses in biology:

- `Membrane-Associated Ring Finger (C3HC4) 1`,
  `E3 Ubiquitin Protein Ligase` → `MARCH1` → 2016-03-01 → 1456786800
- `2310009E13` → 2.31E+19

And more recently, we had the COVID tracing failure.

... and Statistics/Machine Learning

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".*
>
> *– Alex "Sandy" Pentland*

The Reproducibility Crisis in ML-based science (Princeton workshop 2022)

> *Reproducibility failures in ML-based science are systemic. We found 20 reviews across 17 scientific fields (medicine, neuroimaging, autism diagnosis, genomics, computer security, ...) that find errors in a total of 329 papers that use ML-based science and in some cases leading to wildly overoptimistic conclusion. [...] complex ML models don't perform substantively better than decades-old LR models.*
>
> *Data leakage: spurious relationship between the independent variables and the target variable that arises as an artifact of the data collection, sampling, or pre-processing strategy.*
>
> *– S. Kapoor and A. Narayanan*

This is about Computational Science. Should mathematicians care?

Computer Science is young and inherits from Mathematics, Engineering, Linguistic, Nat. Sciences, ...

Purely theoretical scientists whose practice is close to mathematics *may* not be concerned (can't publish a math article without releasing the proofs).

This is about Computational Science. Should mathematicians care?

Computer Science is young and inherits from Mathematics, Engineering, Linguistic, Nat. Sciences, ...

Purely theoretical scientists whose practice is close to mathematics *may* not be concerned (can't publish a math article without releasing the proofs).

Yet, incoherencies are common, especially in a fast moving field:

- E.g., definitions/concepts in book/article A and B are *slightly different* and the resulting theorems cannot be mixed
- Have a look at Vladimir Voevodsky's talk in 2014 at Princeton 😉
- ERC Nano bubbles: how, when and why does science fail to correct itself?

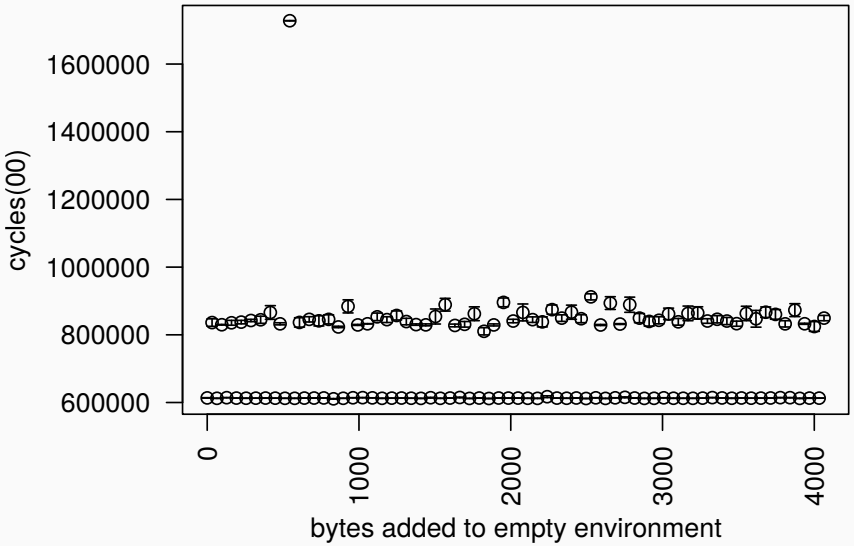
  Flagging incorrect nucleotide sequence reagents in biomedical papers: To what extent does the leading publication format impede automatic error detection? (Labbe et al., 2020)

Experimenting with computers

Machines are real!

Brendan Gregg: Shouting in the data center

Machines are complicated



Mytkowicz et al. *Producing wrong data without doing anything obviously wrong!* ACM SIGPLAN Not. 44(3), March 2009

Our reality evolves!!! The hardware keeps evolving so most results on old platforms quickly become obsolete (although, we keep building on such results 😉).

We need to regularly revisit and allow others to build on our work!

Computer performance ? Well, I design algorithms!

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

Algorithms are evaluated on particular workloads that impact both their running time and the quality of the solutions

Computer performance ? Well, I design algorithms!

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

Algorithms are evaluated on particular workloads that impact both their running time and the quality of the solutions

Image Processing: True horror stories, E. Meinhardt-Llopis, CANUM 2016

- *The proposed multigrid algorithm converges to the solution of the problem in O(N)* using biharmonic functions
- Surprisingly, our naive multi-scale Gauss-Seidel converges much faster

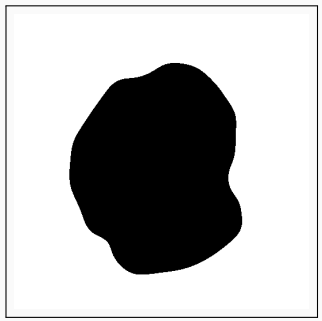
Computer performance ? Well, I design algorithms!

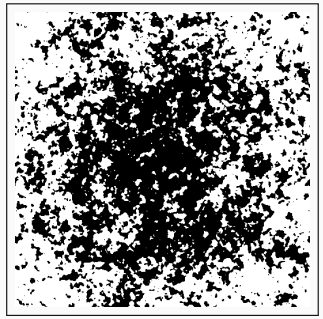
- ”Real” problems are all NP-hard, Log-APX, etc.
- Real workload = ~~NP-completeness proof widgets~~, regularities and properties (difficult to formally state but that should be exploited)

Algorithms are evaluated on particular workloads that impact both their running time and the quality of the solutions

Image Processing: True horror stories, E. Meinhardt-Llopis, CANUM 2016

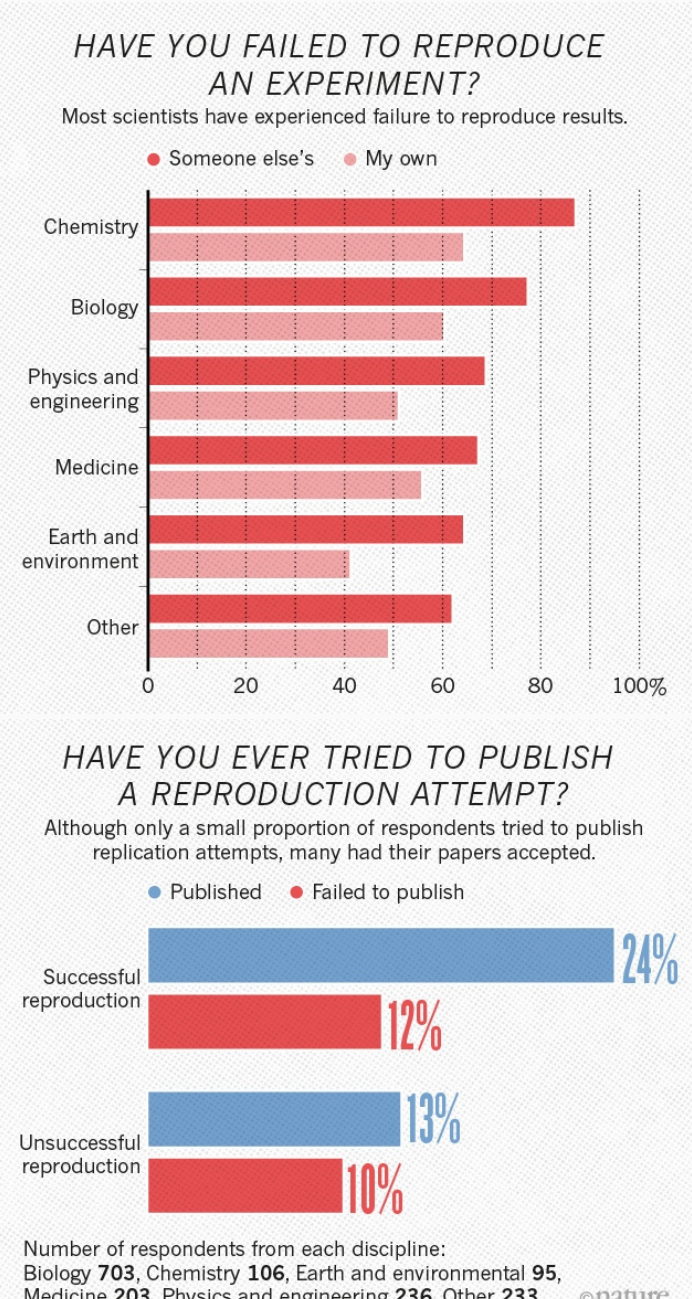
- *The proposed multigrid algorithm converges to the solution of the problem in O(N)* using biharmonic functions
- Surprisingly, our naive multi-scale Gauss-Seidel converges much faster





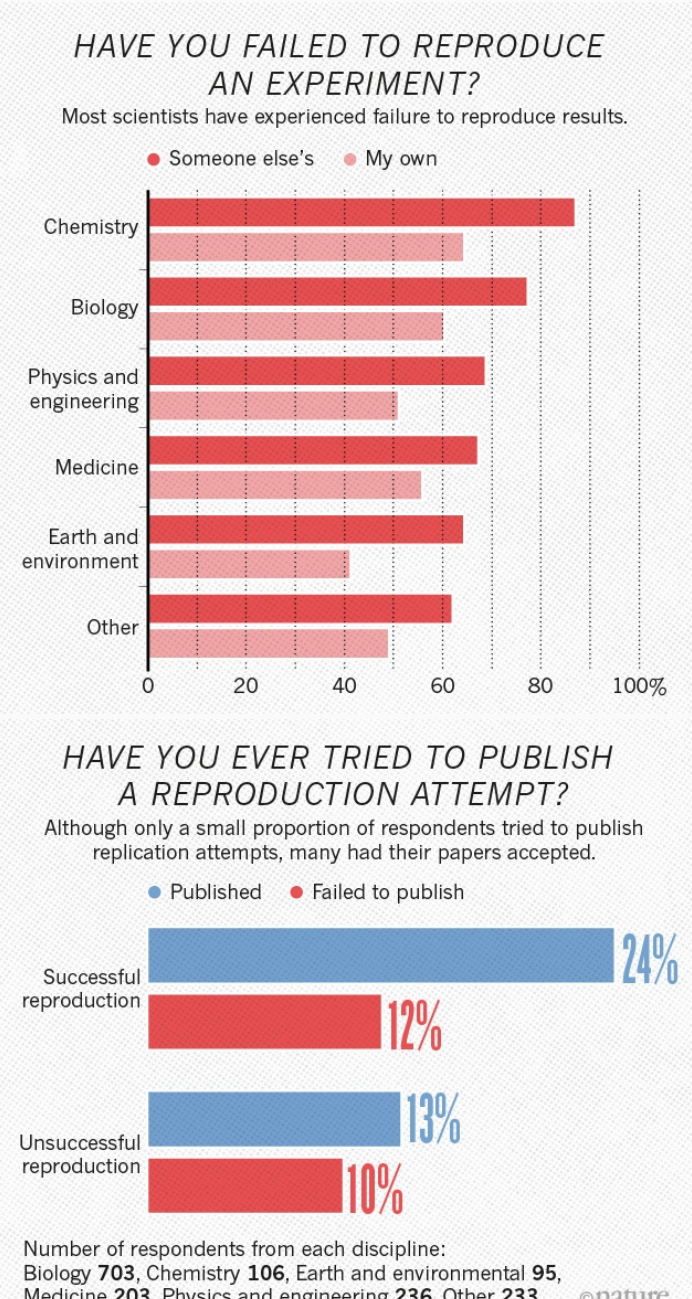
Different kinds of reproducibility

Socio-technical Challenges



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

Socio-technical Challenges

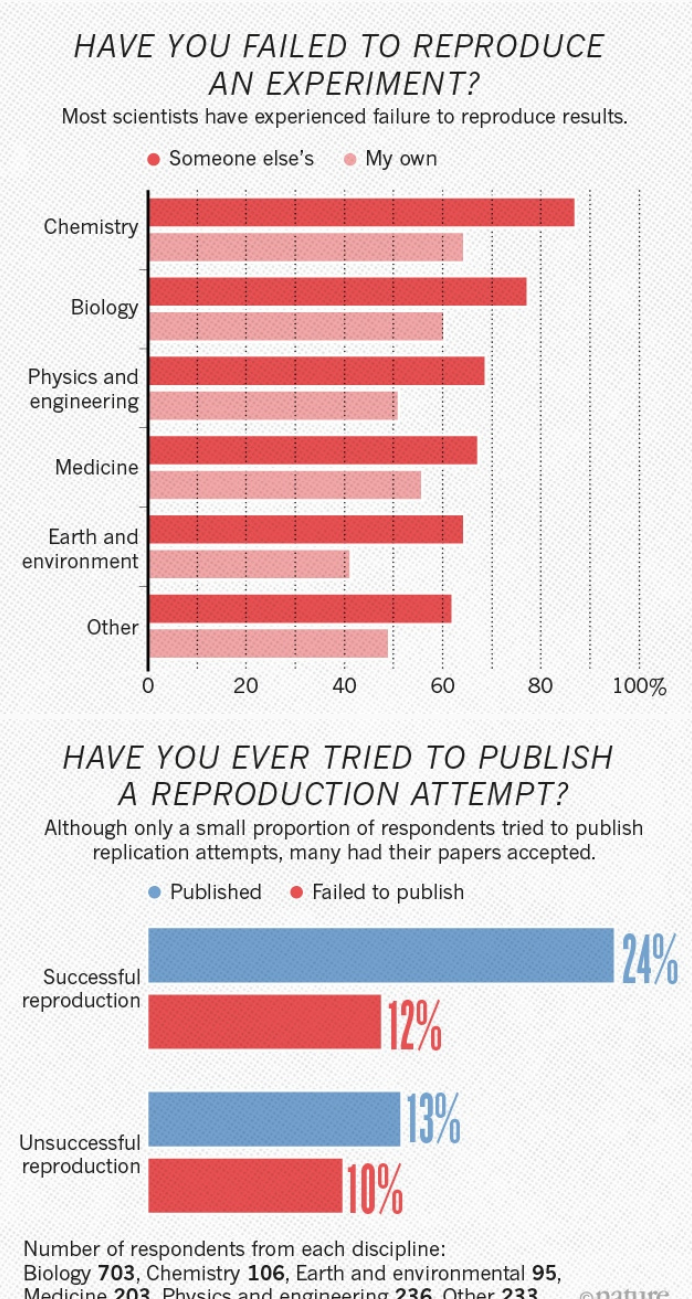


1,500 scientists lift the lid on reproducibility, Nature, May 2016

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: 1M+ articles per year!

Socio-technical Challenges

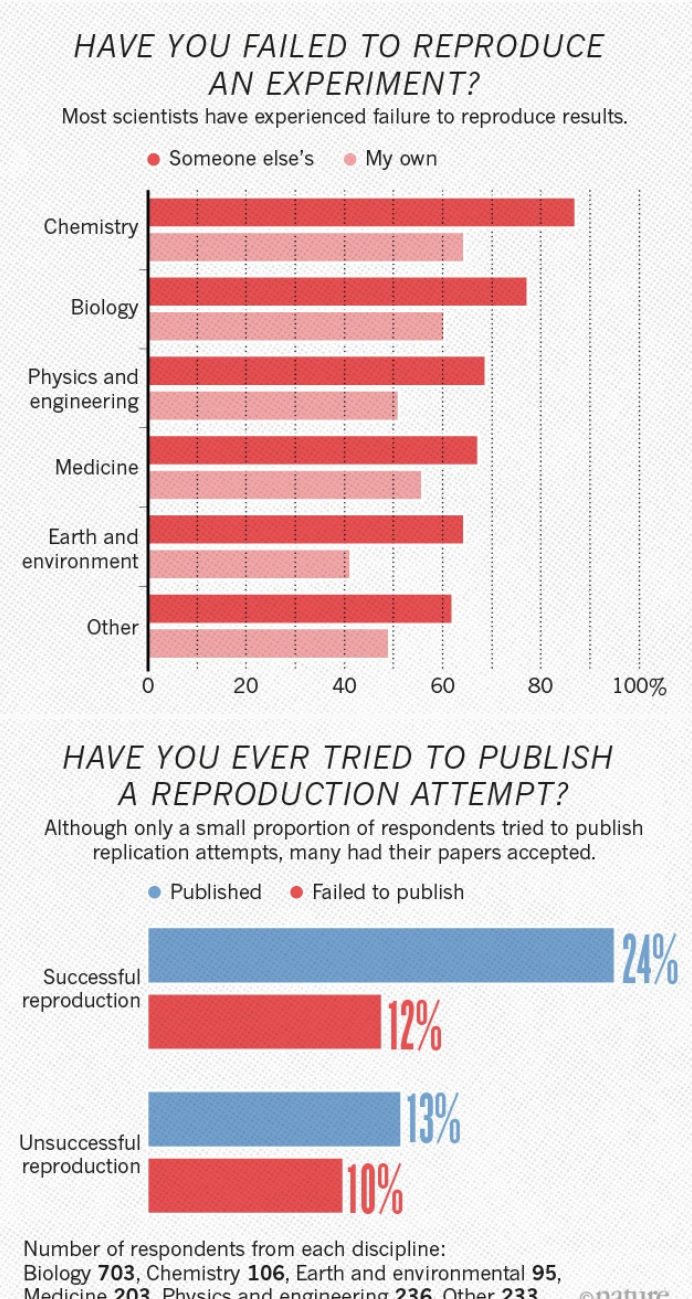


1,500 scientists lift the lid on reproducibility, Nature, May 2016

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: 1M+ articles per year!
- Emerging practices: DORA/Plan S/COARA, DMP and FAIR data, artefact evaluation, reproducibility badges, reproducibility challenges, open reviews, ...

Socio-technical Challenges



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

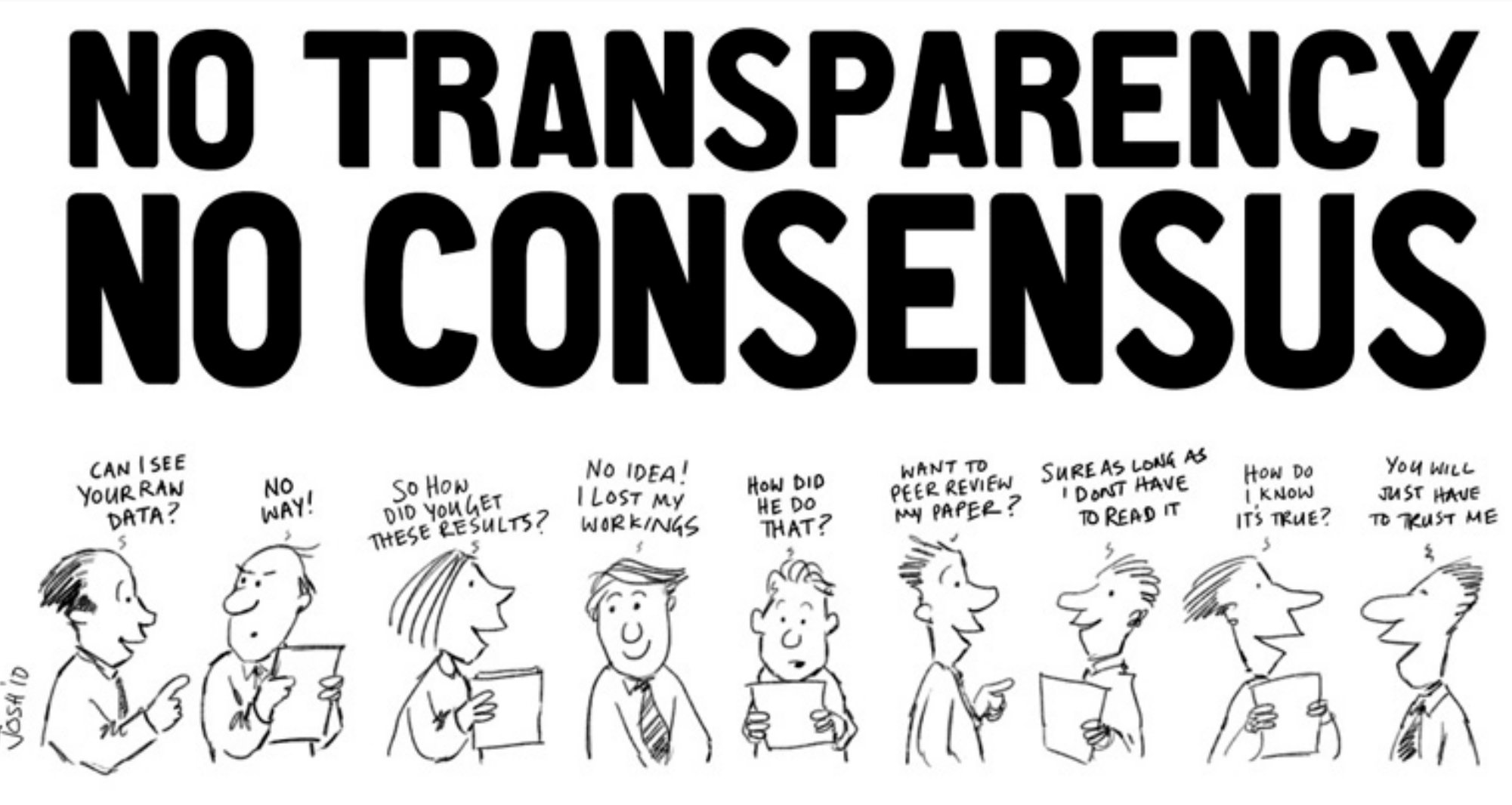
Social causes

- 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: 1M+ articles per year!
- Emerging practices: DORA/Plan S/COARA, DMP and FAIR data, artefact evaluation, reproducibility badges, reproducibility challenges, open reviews, ...

Methodological/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

Scientific Consensus



Different Reproducibility Concerns in Modern Science

Biology, Oncology sample provenance, clinical trials $\rightsquigarrow$ standardized protocols

Psychology, Nutrition HARKING, p-hacking $\rightsquigarrow$ pre-registration

Different Reproducibility Concerns in Modern Science

Biology, Oncology sample provenance, clinical trials $\rightsquigarrow$ standardized protocols

Psychology, Nutrition HARKING, p-hacking $\rightsquigarrow$ pre-registration

Genomics software engineering, computational reproducibility, provenance

Computational fluid dynamics numerical chaos, parallel architectures

Different Reproducibility Concerns in Modern Science

Biology, Oncology sample provenance, clinical trials ⇝ standardized protocols

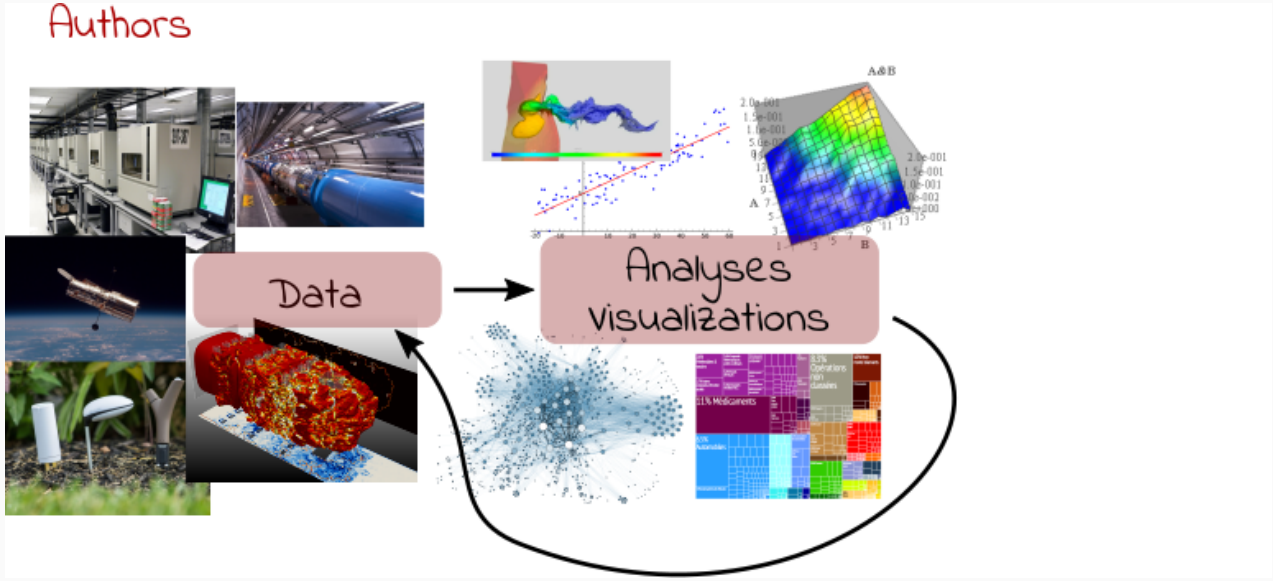
Psychology, Nutrition HARKING, p-hacking ⇝ pre-registration

Genomics software engineering, computational reproducibility, provenance

Computational fluid dynamics numerical chaos, parallel architectures

Artificial Intelligence most of the above 😉

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



Different Reproducibility Concerns in Modern Science

Biology, Oncology sample provenance, clinical trials ⇝ standardized protocols

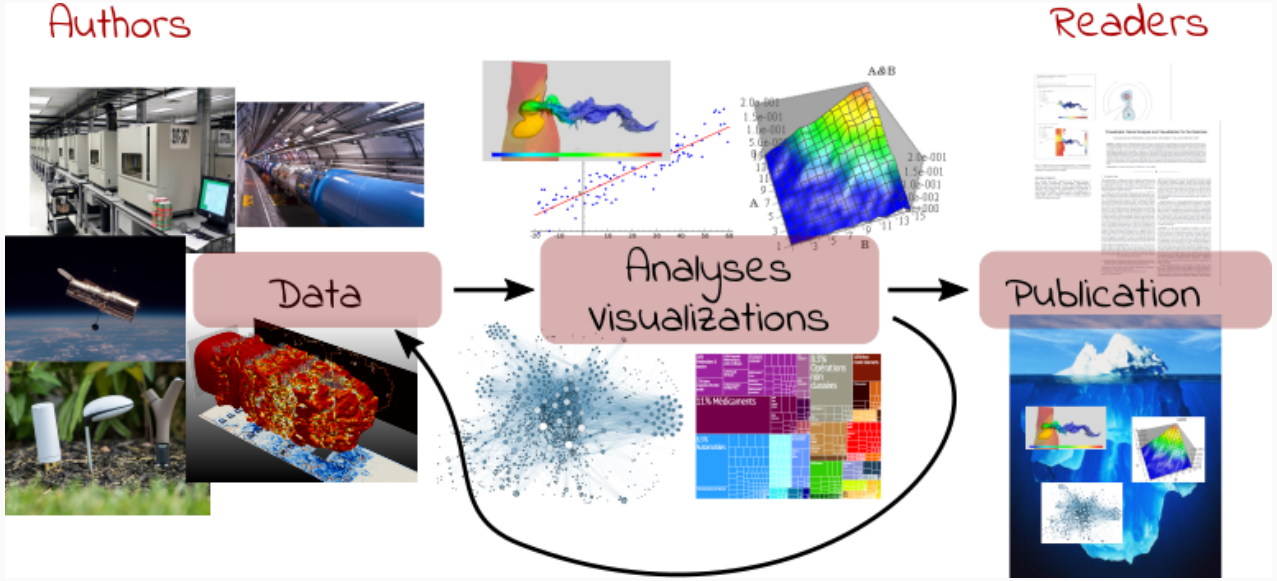
Psychology, Nutrition HARKING, p-hacking ⇝ pre-registration

Genomics software engineering, computational reproducibility, provenance

Computational fluid dynamics numerical chaos, parallel architectures

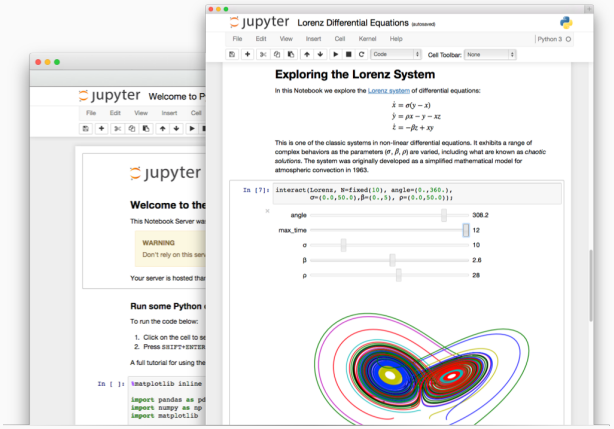
Artificial Intelligence most of the above 😉

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



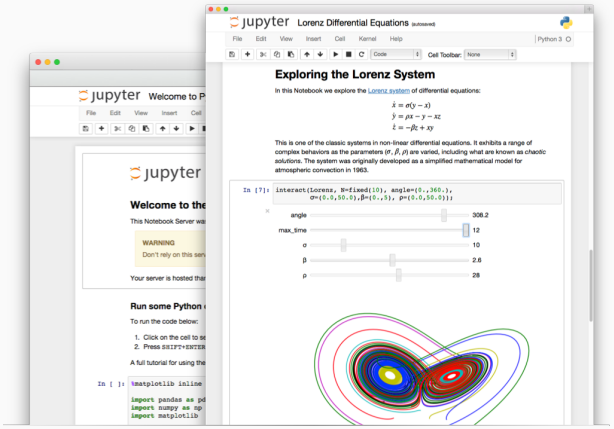
Reproducible Research = Bridging the Gap by working Transparently

Computation provenance: notebooks and workflows

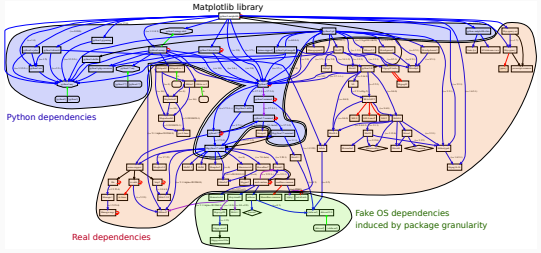


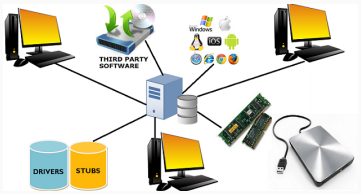
Reproducibility Issues Related to the use of Computers

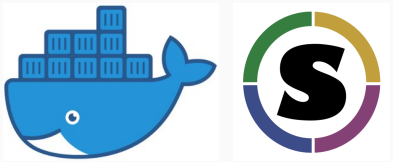
Computation provenance: notebooks and workflows



Software environments

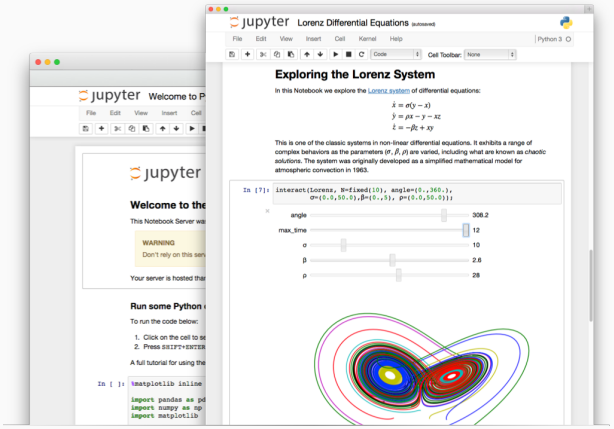




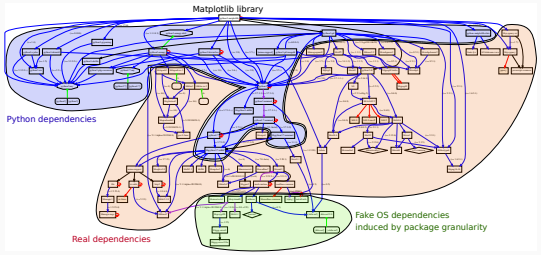


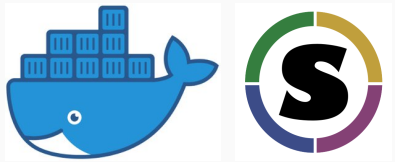
Reproducibility Issues Related to the use of Computers

Computation provenance: notebooks and workflows

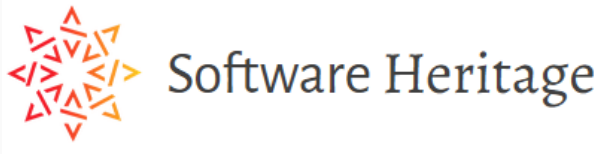


Software environments



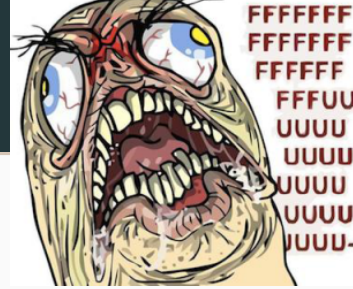


Sharing and Archiving



Good Practice #1
Taking Notes and Documenting

Frustration as an Author/Reviewer



Author

- I thought I used the same parameters but I'm getting different results!
- The new student wants to compare with the method I proposed last year
- My advisor asked me whether I took care of setting this or this but I can't remember
- The damned fourth reviewer asked for a major revision and wants me to change Figure 3. Which code and which data set did I use?
- It worked yesterday! 6 months later: Why did I do that?

Reviewer

- As usual, there is no confidence interval, I wonder about the variability and whether the difference is significant or not
- That can't be true, I'm sure they removed some points
- Why is this graph in logscale? How would it look like otherwise? I'm not even sure of what this value means. If only I could access the generation script

Tool 1: Computational Notebooks/Litterate Programming

Un document computationnel

Mon ordinateur m'indique que π vaut *approximativement*

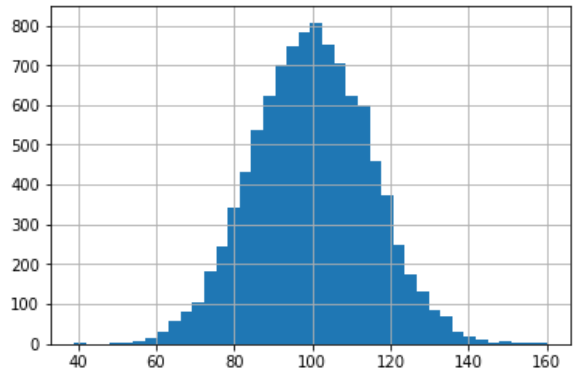
```
3.141592653589793
```

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

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

```
3.1437198694098765
```

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



Tool 1: Computational Notebooks/Litterate Programming

Document initial dans son environnement

jupyter example_pi (autosaved)

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

Code CellToolbar

Hide Prompts Hide Code Hide Outputs

```
# Un document computationnel
```

Hide Prompts Hide Code Hide Outputs

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

In [1]:

Hide Prompts Hide Code Hide Outputs

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

```
3.141592653589793
```

Hide Prompts Hide Code Hide Outputs

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

In [2]:

Hide Prompts Hide Code Hide Outputs

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

Out[2]: 3.1437198694098765

Hide Prompts Hide Code Hide Outputs

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

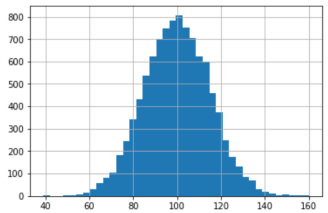
In [3]:

Hide Prompts Hide Code Hide Outputs

```
%matplotlib inline
import matplotlib.pyplot as plt

mu, sigma = 100, 15
x = mu + sigma*np.random.randn(10000)

plt.hist(x,40)
plt.grid(True)
plt.show()
```



Document final

Un document computationnel

Mon ordinateur m'indique que π vaut *approximativement*

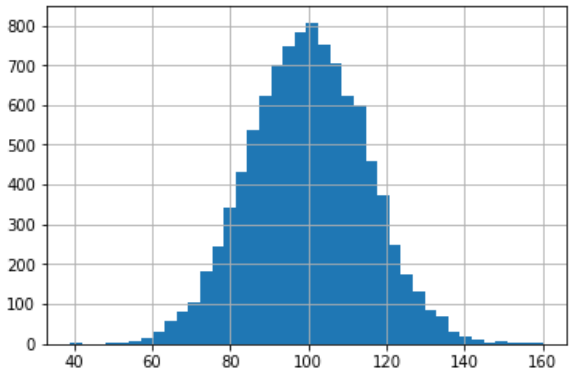
```
3.141592653589793
```

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

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

```
3.1437198694098765
```

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



Tool 1: Computational Notebooks/Litterate Programming

Document initial dans son environnement

Document final

jupyter example_pi (autosaved)

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

Code CellToolbar

Hide Prompts Hide Code Hide Outputs

```
# Un document computationnel
```

Hide Prompts Hide Code Hide Outputs

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

In [1]:

Hide Prompts Hide Code Hide Outputs

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

```
3.141592653589793
```

Hide Prompts Hide Code Hide Outputs

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

In [2]:

Hide Prompts Hide Code Hide Outputs

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

Out[2]: 3.1437198694098765

Hide Prompts Hide Code Hide Outputs

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

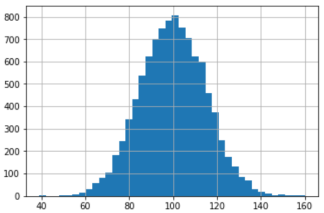
In [3]:

Hide Prompts Hide Code Hide Outputs

```
%matplotlib inline
import matplotlib.pyplot as plt

mu, sigma = 100, 15
x = mu + sigma*np.random.randn(10000)

plt.hist(x,40)
plt.grid(True)
plt.show()
```



Mark Down

Un document computationnel

Mon ordinateur m'indique que π vaut *approximativement*

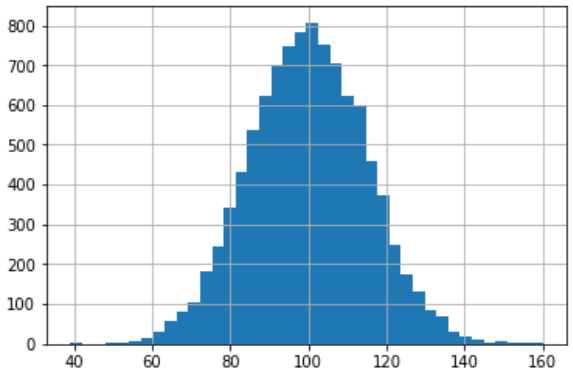
```
3.141592653589793
```

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

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

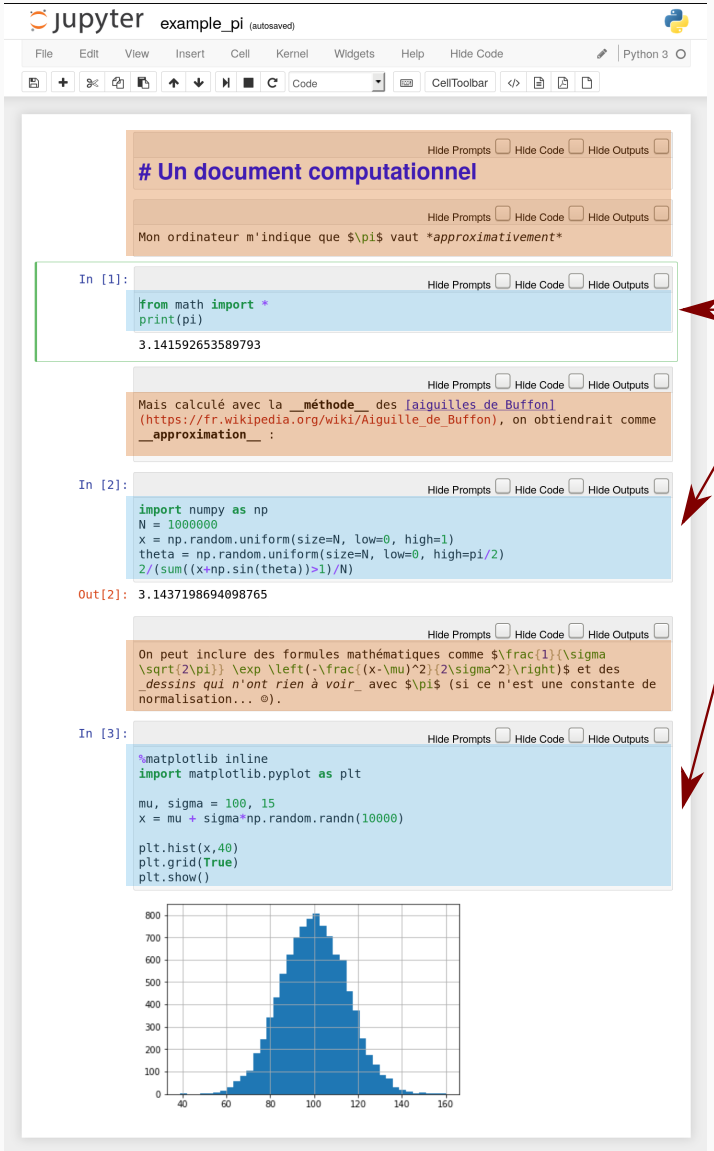
```
3.1437198694098765
```

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



Tool 1: Computational Notebooks/Litterate Programming

Document initial dans son environnement



Code

Document final

Un document computationnel

Mon ordinateur m'indique que π vaut *approximativement*

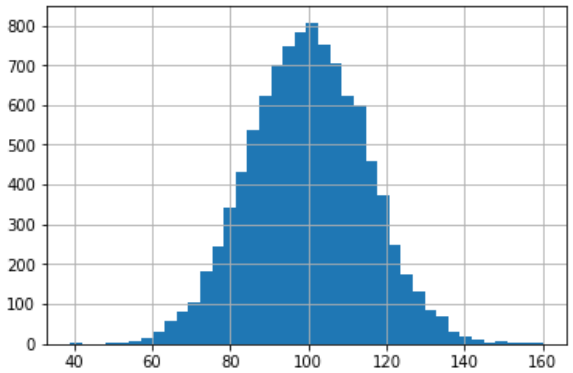
```
3.141592653589793
```

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... ☺).



Tool 1: Computational Notebooks/Litterate Programming

Document initial dans son environnement

jupyter example_pi (autosaved)

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

Code CellToolbar

Hide Prompts Hide Code Hide Outputs

```
# Un document computationnel
```

Hide Prompts Hide Code Hide Outputs

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

In [1]:

Hide Prompts Hide Code Hide Outputs

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

```
3.141592653589793
```

Hide Prompts Hide Code Hide Outputs

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

In [2]:

Hide Prompts Hide Code Hide Outputs

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

Out[2]: 3.1437198694098765

Hide Prompts Hide Code Hide Outputs

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

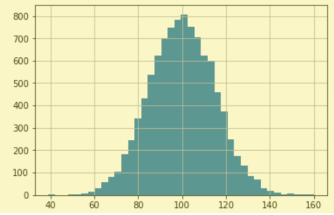
In [3]:

Hide Prompts Hide Code Hide Outputs

```
%matplotlib inline
import matplotlib.pyplot as plt

mu, sigma = 100, 15
x = mu + sigma*np.random.randn(10000)

plt.hist(x,40)
plt.grid(True)
plt.show()
```



Résultats

Document final

Un document computationnel

Mon ordinateur m'indique que π vaut *approximativement*

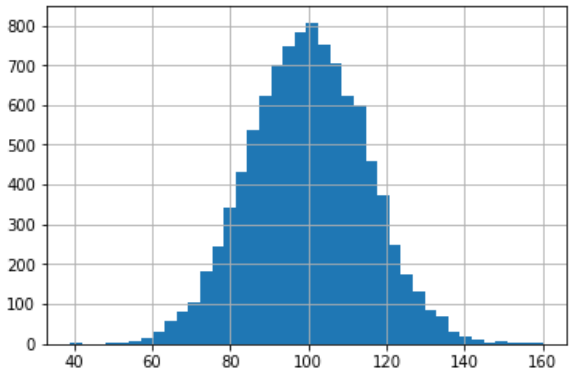
```
3.141592653589793
```

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)
```

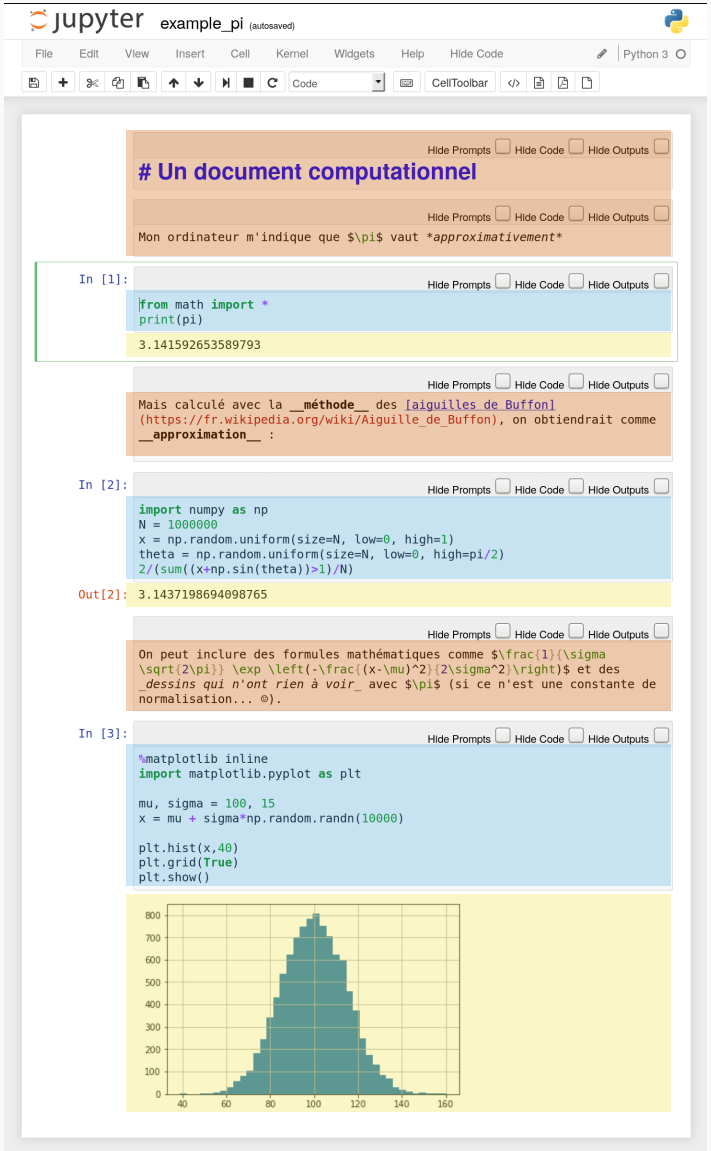
```
3.1437198694098765
```

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... ☺).



Tool 1: Computational Notebooks/Litterate Programming

Document initial dans son environnement



Export

Document final

Un document computationnel

Mon ordinateur m'indique que π vaut *approximativement*

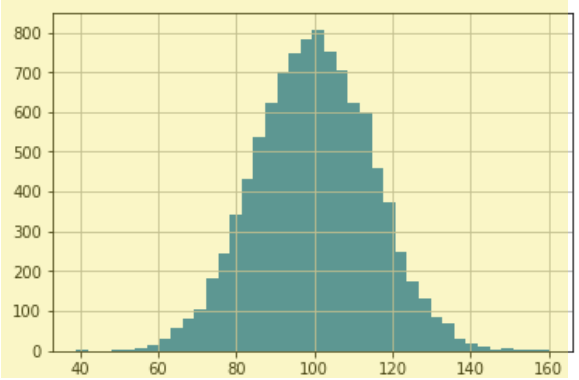
```
3.141592653589793
```

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

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

```
3.1437198694098765
```

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



Tool 1: Computational Notebooks/Litterate Programming

Document initial dans son environnement

jupyter example_pi (autosaved)

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

Code CellToolbar

Hide Prompts Hide Code Hide Outputs

```
# Un document computationnel
```

Hide Prompts Hide Code Hide Outputs

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

In [1]:

Hide Prompts Hide Code Hide Outputs

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

```
3.141592653589793
```

Hide Prompts Hide Code Hide Outputs

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

In [2]:

Hide Prompts Hide Code Hide Outputs

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

Out[2]: 3.1437198694098765

Hide Prompts Hide Code Hide Outputs

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

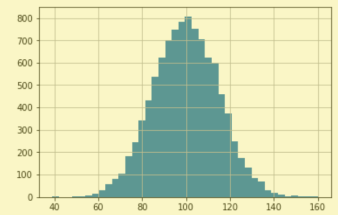
In [3]:

Hide Prompts Hide Code Hide Outputs

```
%matplotlib inline
import matplotlib.pyplot as plt

mu, sigma = 100, 15
x = mu + sigma*np.random.randn(10000)

plt.hist(x,40)
plt.grid(True)
plt.show()
```



Export

Document final

Un document computationnel

Mon ordinateur m'indique que π vaut *approximativement*

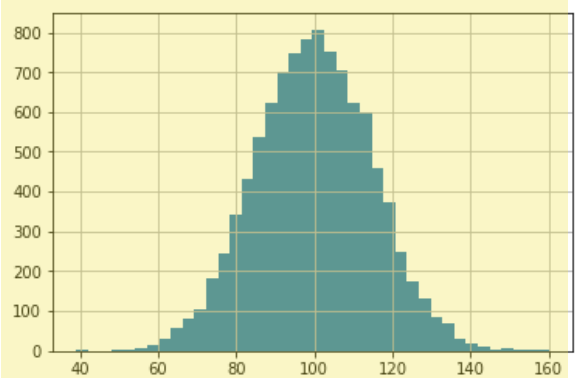
```
3.141592653589793
```

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

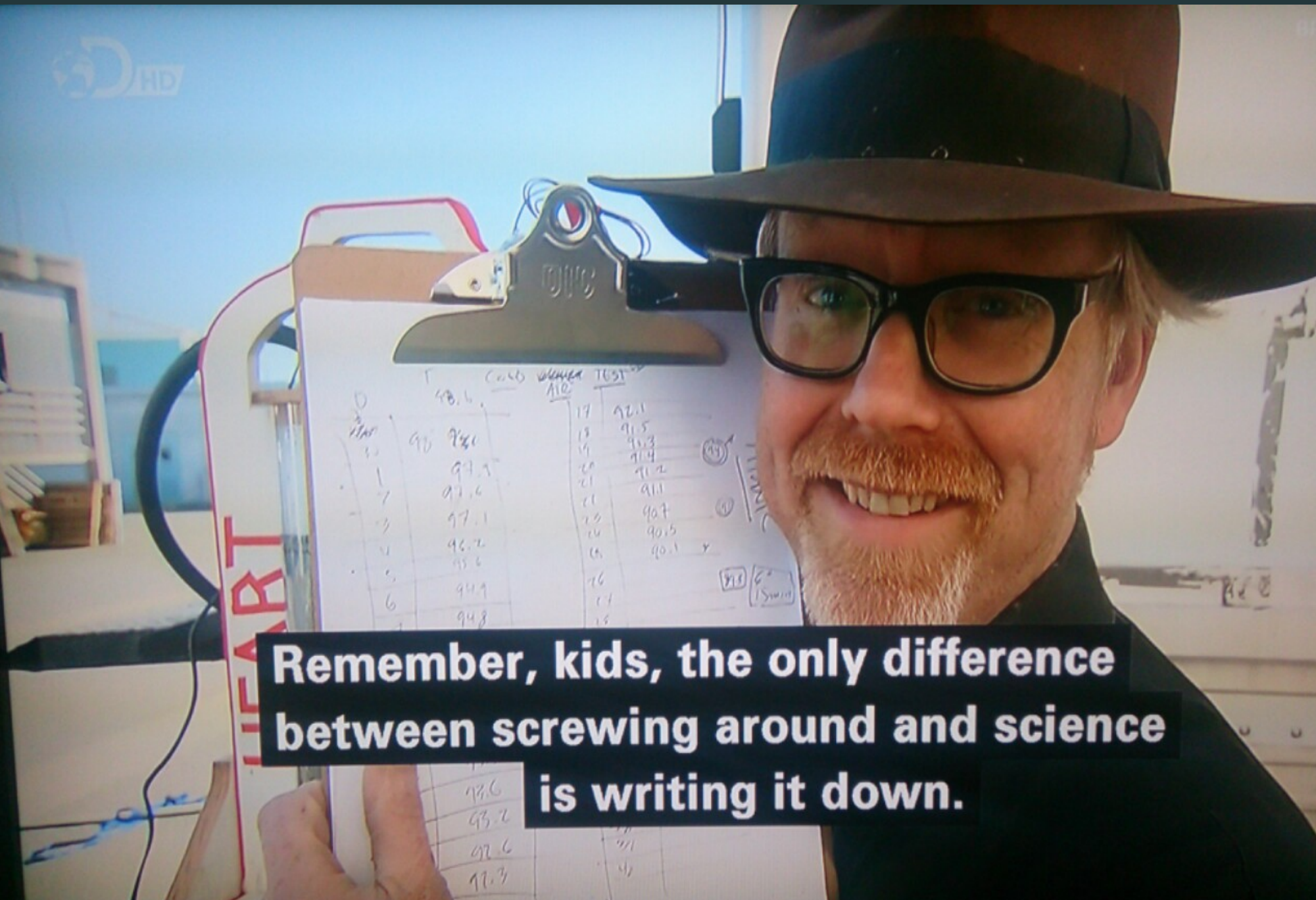
```
import numpy as np
N = 1000000
x = np.random.uniform(size=N, low=0, high=1)
theta = np.random.uniform(size=N, low=0, high=pi/2)
2/(sum((x+np.sin(theta))>1)/N)
```

```
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... ☺).

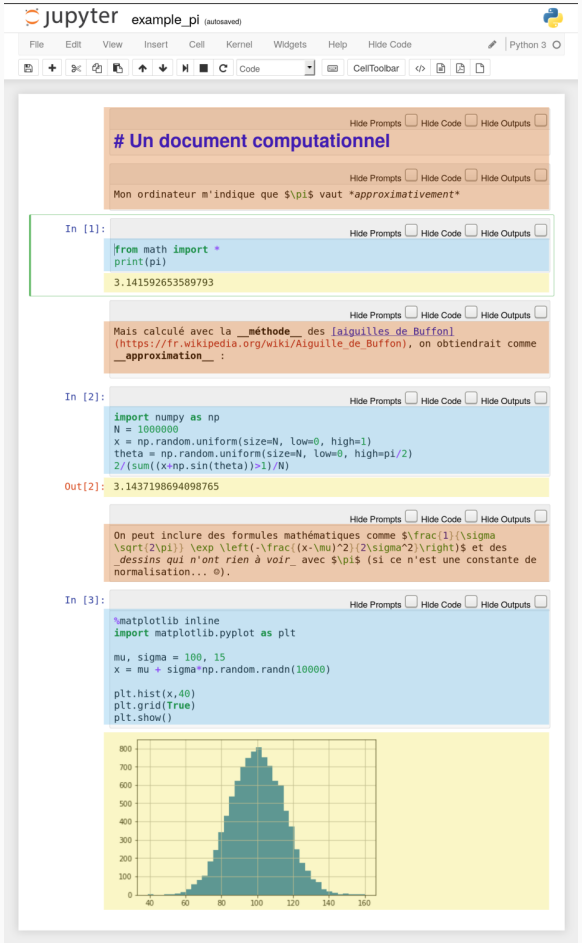


Tool 1 bis: Laboratory Notebooks, Computational Documents



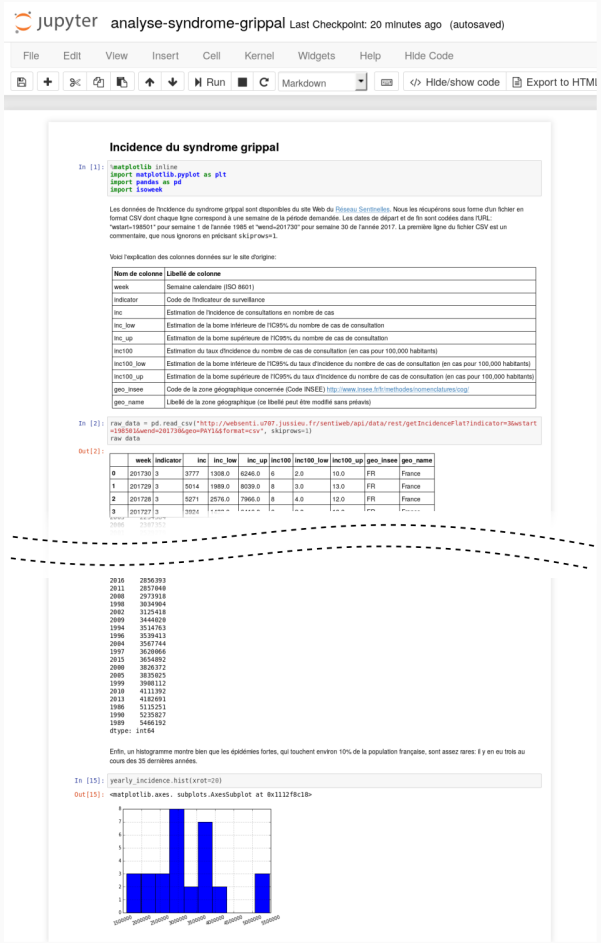
Tool 1 ter: Workflows

Notebooks are no panacea and do not help developing clean code



Tool 1 ter: Workflows

Notebooks are no panacea and do not help developing clean code



jupyter analyse-syndrome-grippal Last Checkpoint: 20 minutes ago (autosaved)

File Edit View Insert Cell Kernel Widgets Help Hide Code

Run Markdown Hide/show code Export to HTML

Incidence du syndrome grippal

```
In [1]: %matplotlib inline
        import matplotlib.pyplot as plt
        import pandas as pd
        import isoweek
```

Les données de l'incidence du syndrome grippal sont disponibles du site Web du Réseau Sentinelles. Nous les récupérons sous forme d'un fichier en format CSV dont chaque ligne correspond à une semaine de la période demandée. Les dates de départ et de fin sont codées dans l'URL : "sstart=198501" pour semaine 1 de l'année 1985 et "send=201730" pour semaine 30 de l'année 2017. La première ligne du fichier CSV est un commentaire, que nous ignorons en précisant `skiprows=1`.

Voici l'explication des colonnes données sur le site d'origine:

Nom de colonne	Libellé de colonne
week	Semaine calendaire (ISO 8601)
indicator	Code de l'indicateur de surveillance
inc	Estimation de l'incidence de consultations en nombre de cas
inc_low	Estimation de la borne inférieure de l'IC95% du nombre de cas de consultation
inc_up	Estimation de la borne supérieure de l'IC95% du nombre de cas de consultation
inc100	Estimation du taux d'incidence du nombre de cas de consultation (en cas pour 100,000 habitants)
inc100_low	Estimation de la borne inférieure de l'IC95% du taux d'incidence du nombre de cas de consultation (en cas pour 100,000 habitants)
inc100_up	Estimation de la borne supérieure de l'IC95% du taux d'incidence du nombre de cas de consultation (en cas pour 100,000 habitants)
geo_insee	Code de la zone géographique concernée (Code INSEE) http://www.insee.fr/fr/methodes/nomenclatures/cog/
geo_name	Libellé de la zone géographique (ce libellé peut être modifié sans préavis)

```
In [2]: raw_data = pd.read_csv("http://websenti.u707.jussieu.fr/sentiweb/api/data/rest/getIncidenceFlat?indicator=3&sstart
        =198501&sweend=201730&geo=PAYS1&format=csv", skiprows=1)
        raw_data
```

Out[2]:

	week	indicator	inc	inc_low	inc_up	inc100	inc100_low	inc100_up	geo_insee	geo_name
0	201730	3	3777	1308.0	6246.0	6	2.0	10.0	FR	France
1	201729	3	5014	1989.0	8039.0	8	3.0	13.0	FR	France
2	201728	3	5271	2576.0	7966.0	8	4.0	12.0	FR	France
3	201727	3	3924							

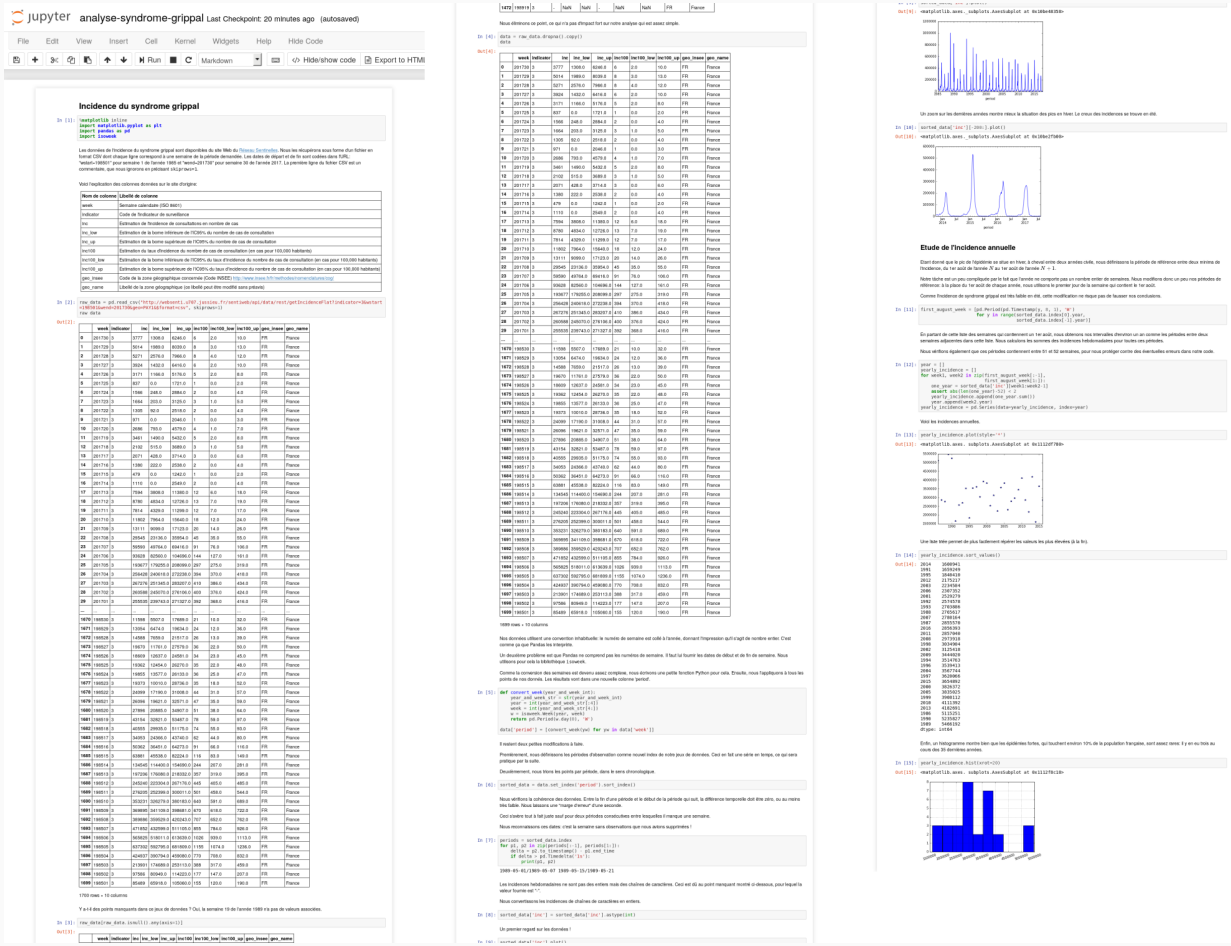
```
2016    2856393
2011    2857040
2008    2973918
1998    3034904
2002    3125418
2009    3444028
1994    3514763
1996    3539413
2004    3567744
1997    3620066
2015    3654892
2000    3826372
2005    3835025
1999    3908112
2010    4111392
2013    4182691
1986    5115251
1990    5235827
1989    5466192
dtype: int64
```

Enfin, un histogramme montre bien que les épidémies fortes, qui touchent environ 10% de la population française, sont assez rares: il y en eu trois au cours des 35 dernières années.

```
In [15]: yearly_incidence.hist(xrot=20)
Out[15]: <matplotlib.axes._subplots.AxesSubplot at 0x1112f8c18>
```

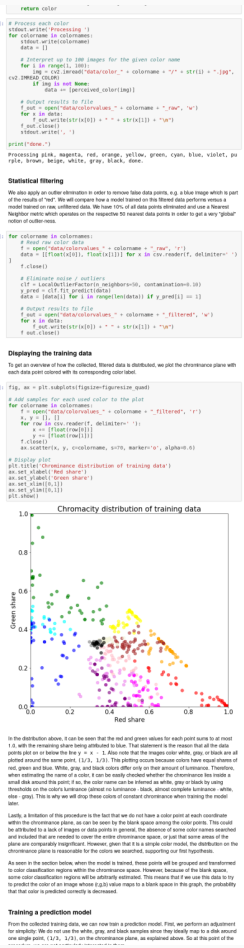
Tool 1 ter: Workflows

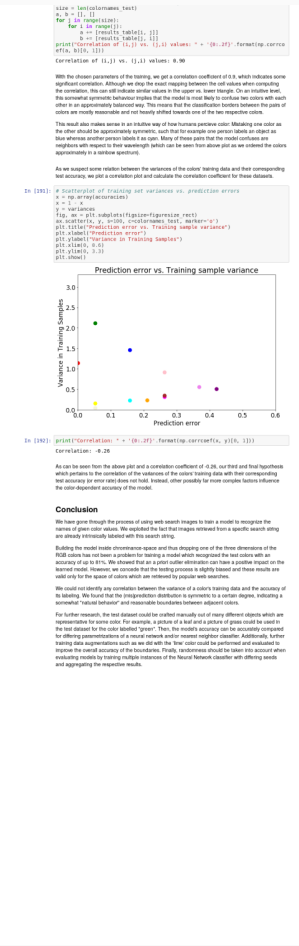
Notebooks are no panacea and do not help developing clean code



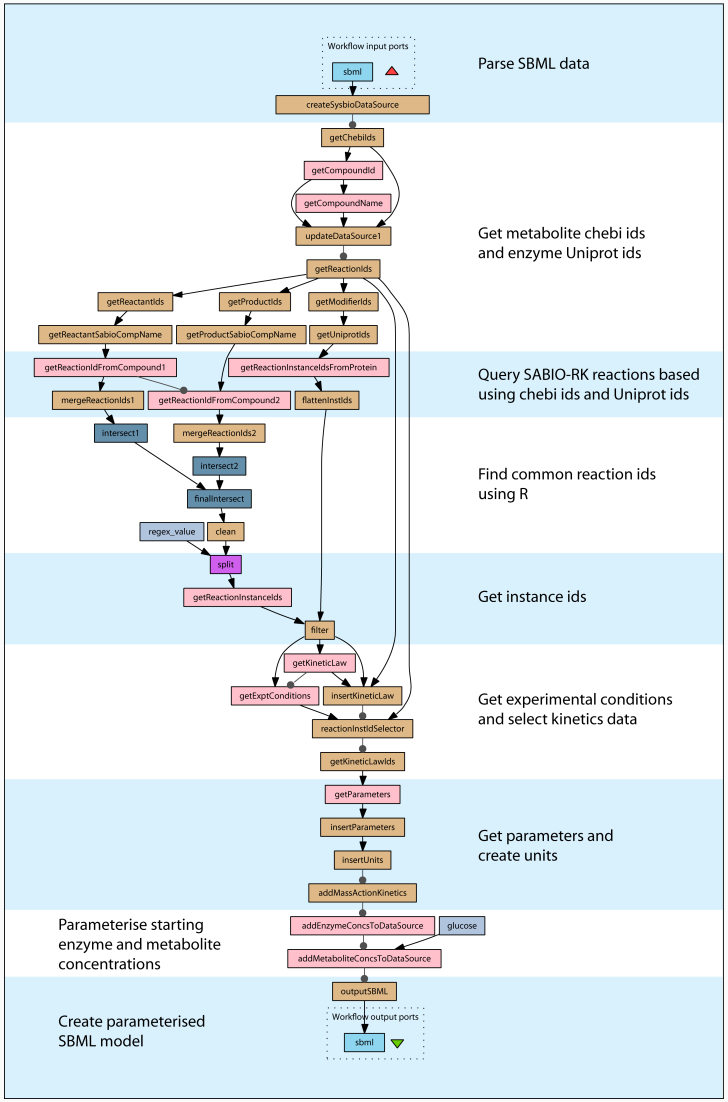
Tool 1 ter: Workflows

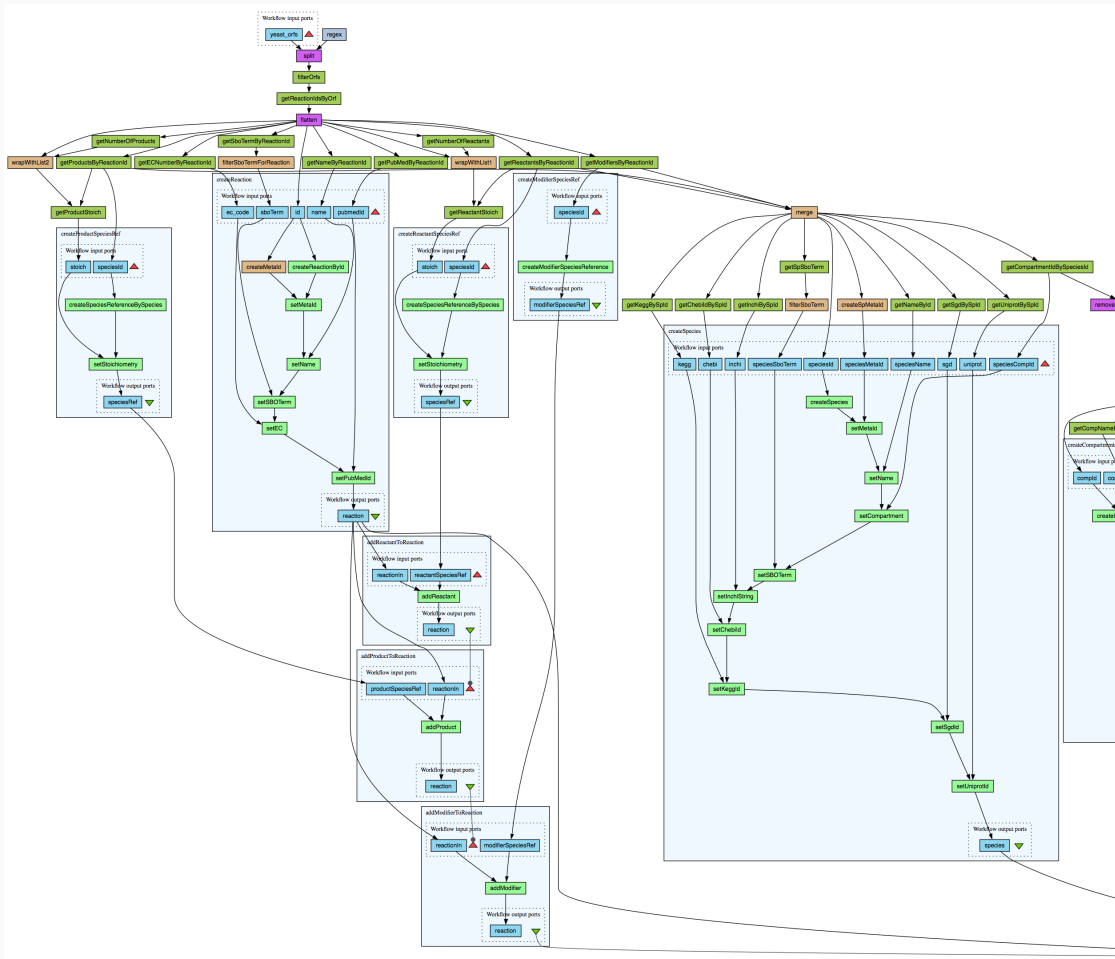
Notebooks are no panacea and do not help developing clean code





Tool 1 ter: Workflows





Tool 1 ter: Workflows

Workflows:

- Clearer high-level view
- Explicit composition of codes and data movement
- Safer sharing, reusing, and execution
- Notebooks are a variant that is both impoverished and richer
  - No simple/mature path from a notebook to a workflow

Examples:

- Galaxy, Kepler, Taverna, Pegasus, Collective Knowledge, VisTrails
- Light-weight: `make`, dask, drake, swift, `snakemake`, ...
- Hybrids: SOS-notebook, ...

Good Practice #2
Controling Software Environment

Argh... damned computers

- **Alice**: I got 3.123123 **Bob**: I got segfault
- Damned! It used to work!!! Whenever I upgrade my computer, things break so I try to stay away from this ☹

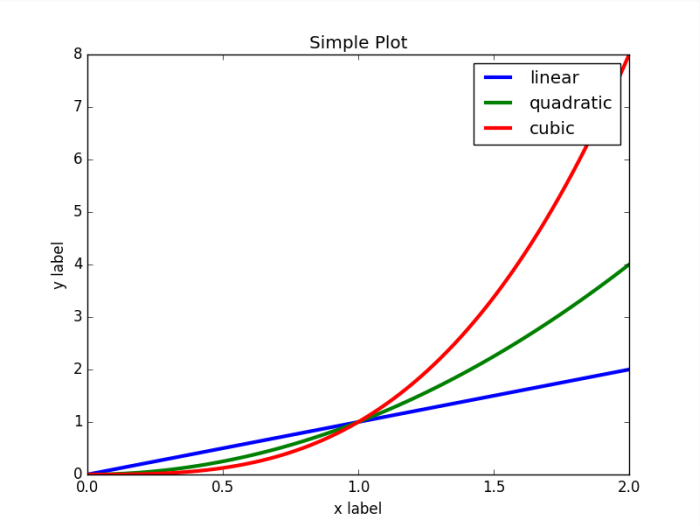
- Whenever trying the code of my colleague, I had to install `libFoo-1.5c` and `pip install blah` but I broke everything and now neither his code nor mine works! ☹

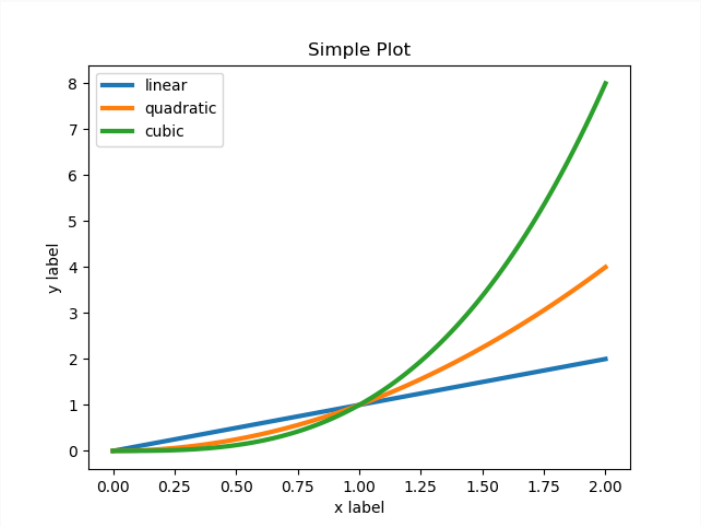
- But hey! Here is my code. It's on GitHub so feel free to play with it! I'm doing open science 😉
  1. No one will ever run/use your code if it isn't easy to install
  2. No one will ever manage to run your code if you don't document how to run it
  3. Others (even you) are unlikely to get the same results unless you control and share your software environment

Software dependencies: horror stories

Software dependencies: horror stories

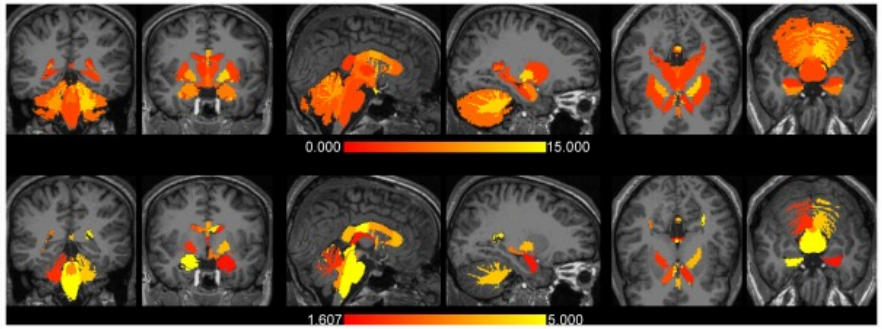
- Software environment evolution





Software dependencies: horror stories

- Software environment evolution
- OS heterogeneity

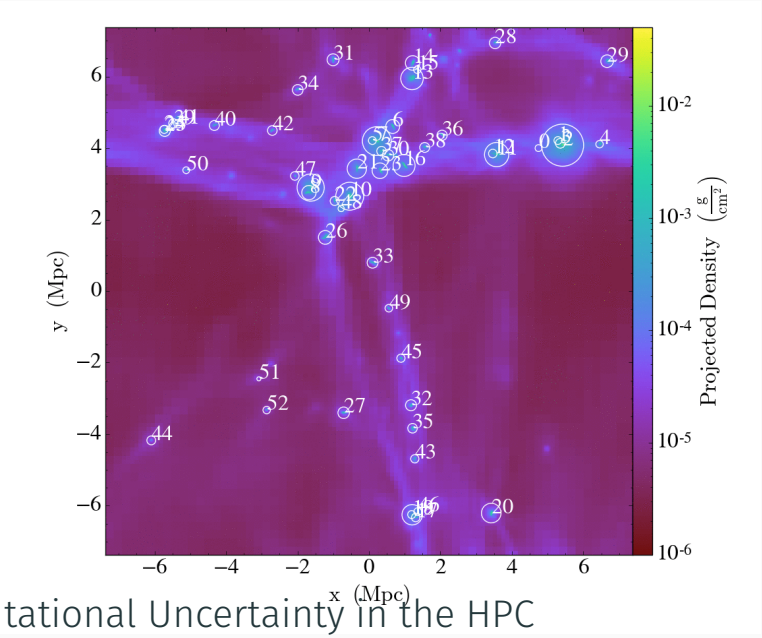


The Effects of FreeSurfer Version, Workstation Type, and Macintosh Operating System Version on Anatomical Volume and Cortical Thickness Measurements (PLOS ONE, 2012)

> *Significant differences in volume and cortical thickness were revealed across FreeSurfer versions:*
>
> - *volume:* *8.8 ± 6.6% (range 1.3-64.0%)*
> - *cortical thickness:* *2.8 ± 1.3% (range 1.1-7.7%)*
>
> *About a factor two smaller differences were found between the Mac and HP workstations and between Mac OSX 10.5 and OSX 10.6.*
>
> *In the context of an ongoing study, users are discouraged to update to a new major release of either FreeSurfer or operating system.*
>
> *Formal assessment of the accuracy of FreeSurfer is desirable.*

Software dependencies: horror stories

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

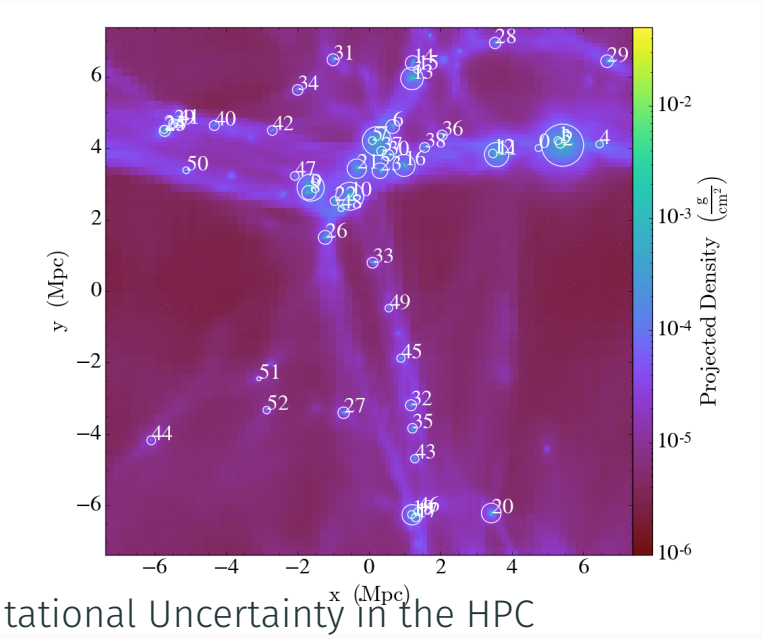


Assessing Reproducibility: An Astrophysical Example of Computational Uncertainty in the HPC Context (ResCuE-HPC, 2018)

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

Software dependencies: horror stories

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

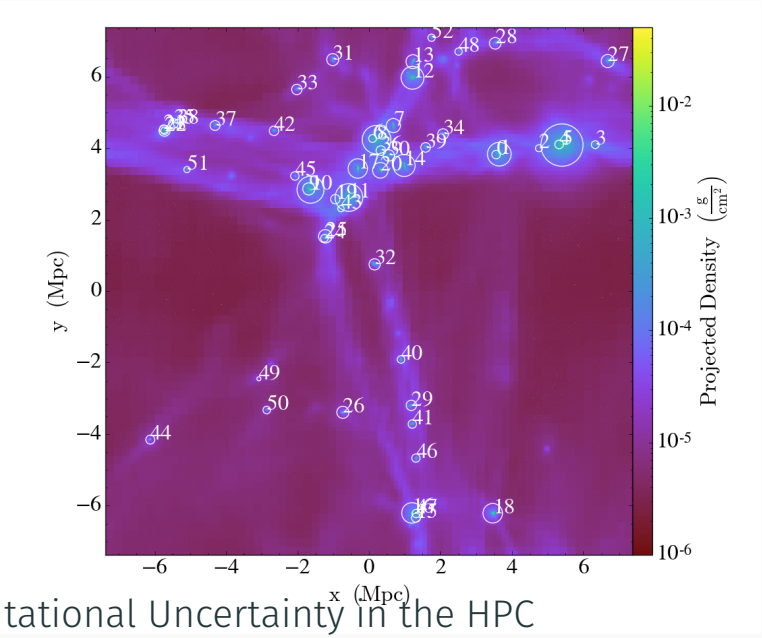


Assessing Reproducibility: An Astrophysical Example of Computational Uncertainty in the HPC Context (ResCuE-HPC, 2018)

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

Software dependencies: horror stories

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

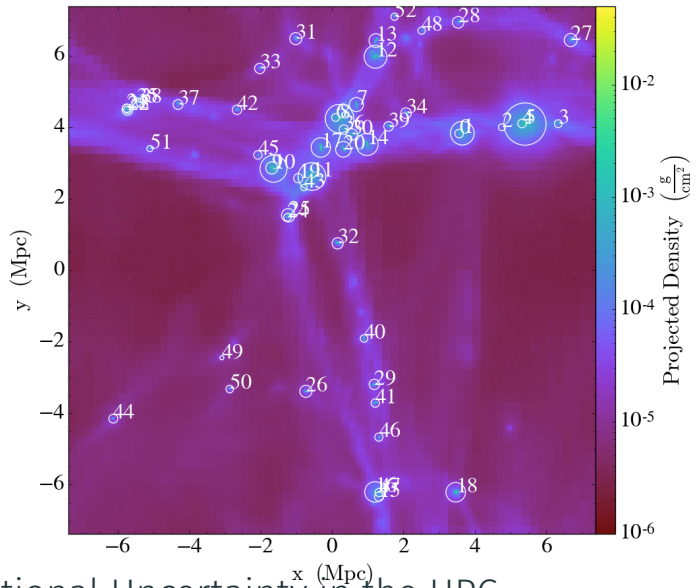


Assessing Reproducibility: An Astrophysical Example of Computational Uncertainty in the HPC Context (ResCuE-HPC, 2018)

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

Software dependencies: horror stories

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



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

Compiler	Optim.	Largest Halo Avg Mass.	Std. Err	Walltime
gcc@6.2.0	None	2.273E 46	1.069E 44	22h
gcc@6.2.0	Normal	2.266E 46	1.218E 44	10h
gcc@6.2.0	High	2.275E 46	1.199E 44	9h
intel@16.0.3	None	22.71 E 46	1.587E 44	39h
intel@16.0.3	Normal	43.30 E 46	1.248E 44	7h
intel@16.0.3	High	2.268E 46	1.414E 44	6h
cce@8.5.5	Low	43.11 E 46	1.353E 44	16h
cce@8.5.5	Normal	2.271E 46	1.261E 44	6h
cce@8.5.5	High	2.272E 46	1.341E 44	5h

Complex ecosystems

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

```
3.5.1
```

Complex ecosystems

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

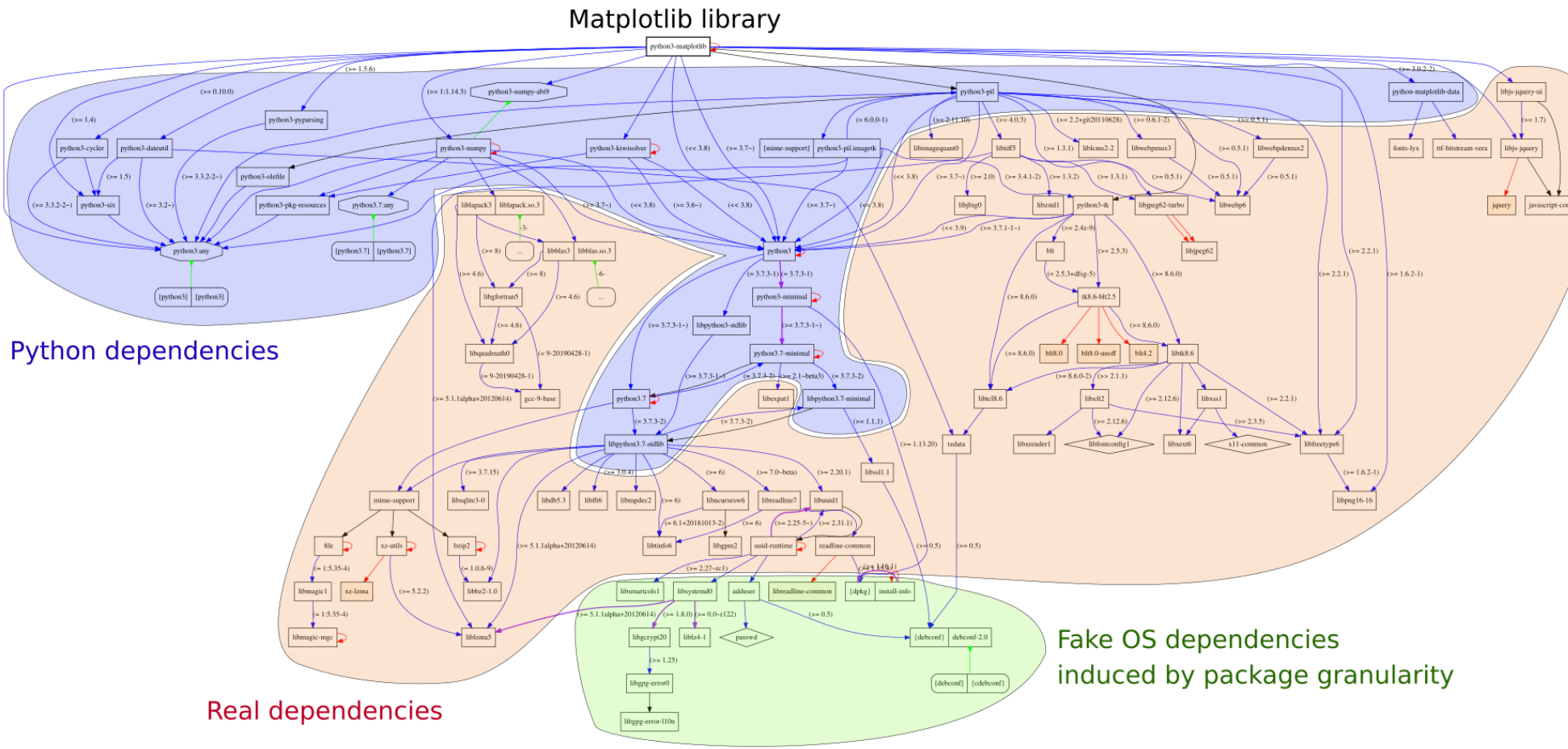
```
3.5.1
```

```
apt show python3-matplotlib
```

```
Package: python3-matplotlib
Version: 3.5.1-2+b1
Source: matplotlib (3.5.1-2)
Maintainer: Sandro Tosi <morph@debian.org>
Installed-Size: 27.6 MB
Depends: libjs-jquery, libjs-jquery-ui, python-matplotlib-data (>= 3.5.1),
         python3-dateutil, python3-pil.imagetk, python3-pyparsing (>= 1.5.6),
         python3-six (>= 1.4), python3-numpy (>= 1:1.20.0), python3-numpy-abi9,
         python3 (<< 3.11), python3 (>= 3.9~), python3-cycler (>= 0.10.0),
         python3-fonttools, python3-kiwisolver, python3-packaging, python3-pil,
         python3:any, libc6 (>= 2.29), libfreetype6 (>= 2.2.1),
         libgcc-s1 (>= 3.3.1), libqhull-r8.0 (>= 2020.1), libstdc++6 (>= 11)
Recommends: python3-tk
Suggests: dvipng, ffmpeg, fonts-staypuft, ghostscript, gir1.2-gtk-3.0, inkscape,
          ipython3, librsvg2-common, python-matplotlib-doc, python3-cairocffi,
          python3-gi, python3-gi-cairo, python3-gobject, python3-pyqt5,
          python3-scipy, python3-sip, python3-tornado, texlive-extra-utils,
          texlive-latex-extra
Enhances: ipython3
```

Complex ecosystems

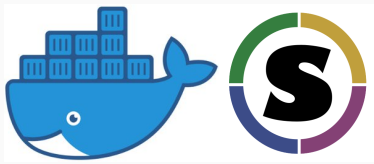
```
1 import matplotlib
2 print(matplotlib.__version__)
```




```
          python3-scipy, python3-sip, python3-tornado, texlive-extra-utils,
          texlive-latex-extra
Enhances: ipython3
```

Potential solutions: Containers and Package Managers

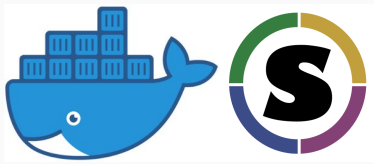
The good | The bad | The uggly



Automatic tracking

Potential solutions: Containers and Package Managers

The good The bad The uggly



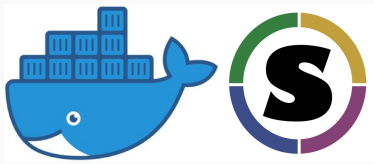

Automatic tracking

Containers

- Pros: Lightweight, Good isolation, Easy to use
  - Running as easy as `docker run <img> <cmd>`
  - Building images: `docker build -f <Dockerfile>`
  - Sharing through the Docker Hub: `docker pull/push <img>`

Potential solutions: Containers and Package Managers

The good The bad The uggly




Automatic tracking

Containers

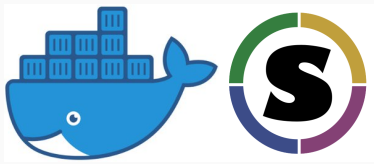
- Pros: Lightweight, Good isolation, Easy to use
- Cons: Opaque, Container build is generally not reproducible
  - Recipes rarely follow *reproducibility good practices*

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

  - Choose a stable image (and the smallest possible)
  - Include only the necessary libraries (e.g. no graphics libs)
  - Avoid system updates (instead freeze sources)

Potential solutions: Containers and Package Managers

The good The bad The uggly




Automatic tracking

Containers

- Pros: Lightweight, Good isolation, Easy to use
- Cons: Opaque, Container build is generally not reproducible

Package managers (the uggly and the good)

- Language specific: `pip/pipenv/virtualenv`, `conda`, `CRAN/Bioconductor`
  - Limits: version management, durability, permeable, language centric
- GUIX/NiX = Full-fledged functional package manager
  - Native support for environment (*à la git*)
  - Isolation through `--pure`
  - Recompile from source (cache recommended)

Debian dependencies

```
dpkg --status python3-matplotlib
```

```
Package: python3-matplotlib
Version: 3.6.3-1+b1
Source: matplotlib (3.6.3-1)
Depends: libjs-jquery, libjs-jquery-ui, python-matplotlib-data (>= 3.6.3),
         python3-dateutil, python3-pil.imagetk, python3-pyparsing (>= 1.5.6),
         python3-six (>= 1.4), python3-numpy (>= 1:1.22.0), python3-contourpy,
         python3 (<< 3.12), python3 (>= 3.11~), python3-numpy-abi9,
         python3-cycler (>= 0.10.0), python3-fonttools, python3-kiwisolver,
         python3-packaging, python3-pil, python3:any, libc6 (>= 2.34),
         libfreetype6 (>= 2.2.1), libgcc-s1 (>= 3.3.1),
         libqhull-r8.0 (>= 2020.1), libstdc++6 (>= 11)
```

Debian dependencies

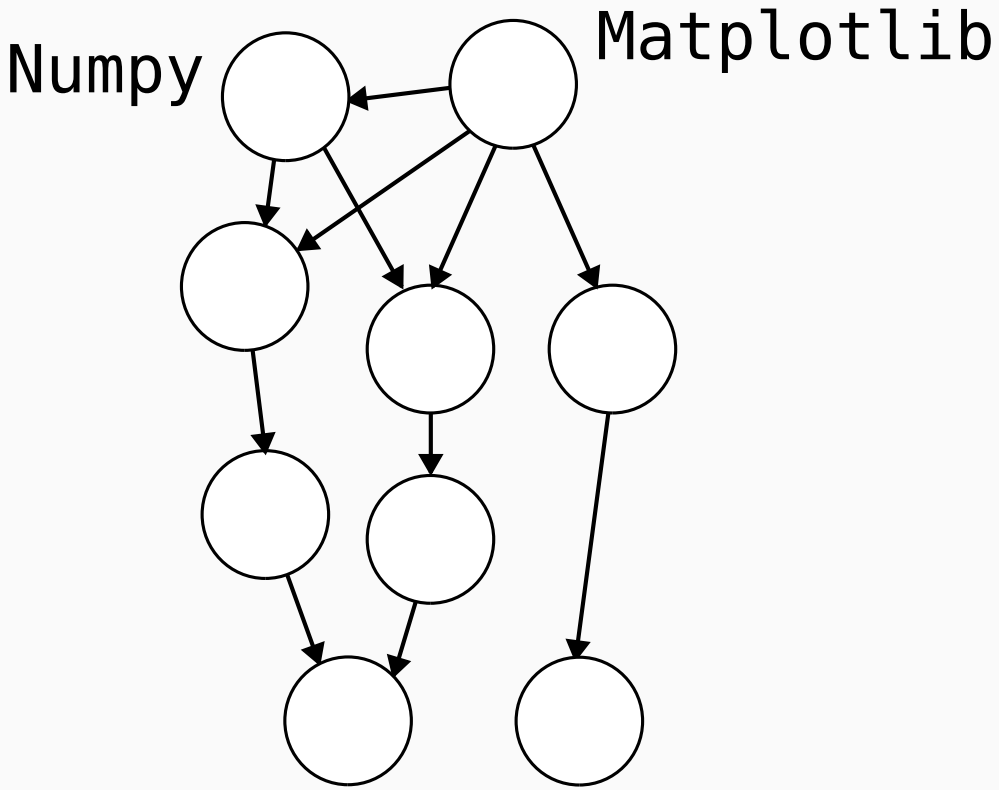
```
1 dpkg --status python3-matplotlib
```

```
Package: python3-matplotlib
Version: 3.6.3-1+b1
Source: matplotlib (3.6.3-1)
Depends: libjs-jquery, libjs-jquery-ui, python-matplotlib-data (>= 3.6.3),
         python3-dateutil, python3-pil.imagetk, python3-pyparsing (>= 1.5.6),
         python3-six (>= 1.4), python3-numpy (>= 1:1.22.0), python3-contourpy,
         python3 (<< 3.12), python3 (>= 3.11~), python3-numpy-abi9,
         python3-cycler (>= 0.10.0), python3-fonttools, python3-kiwisolver,
         python3-packaging, python3-pil, python3:any, libc6 (>= 2.34),
         libfreetype6 (>= 2.2.1), libgcc-s1 (>= 3.3.1),
         libqhull-r8.0 (>= 2020.1), libstdc++6 (>= 11)
```

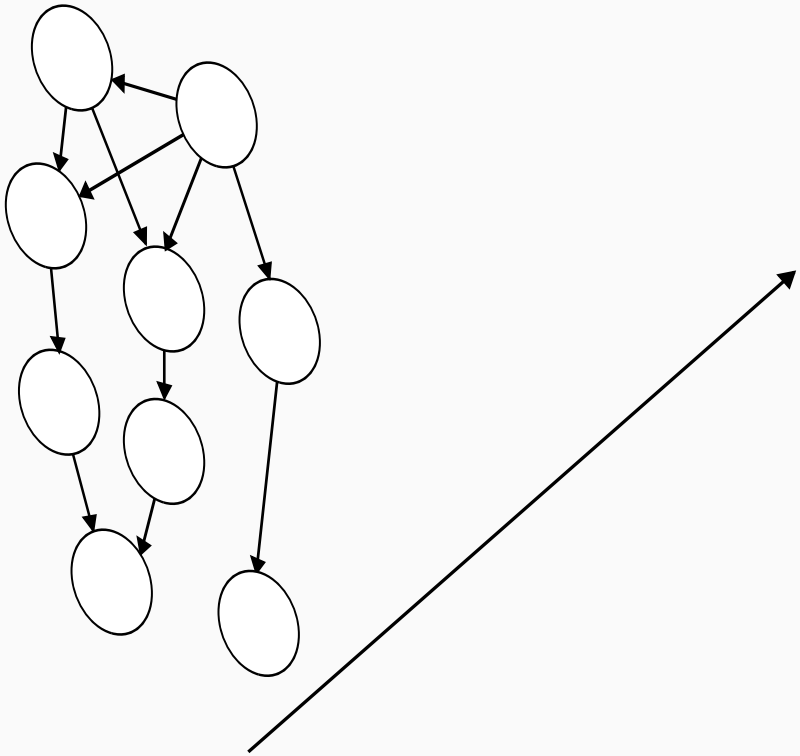
On a given day:

- Several versions of each package are available on the server
- Installing the latest version of a package may require upgrading some other packages

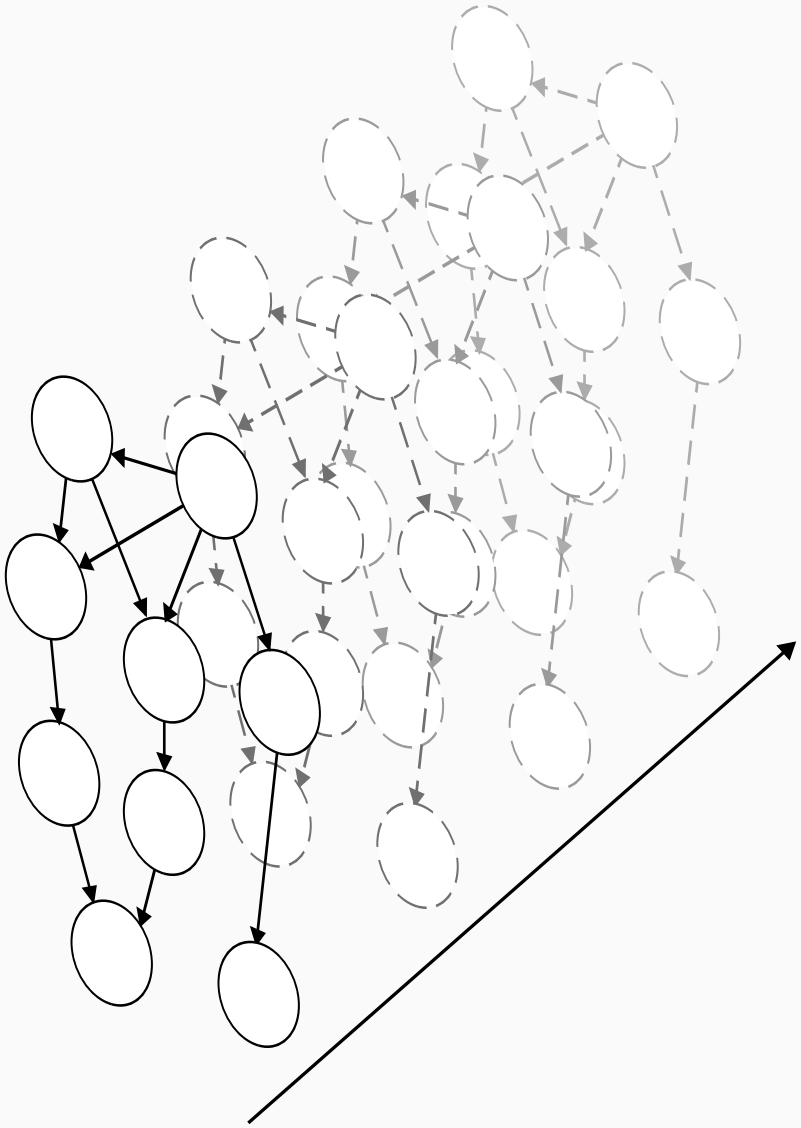
Loose vs. strict dependencies in picture



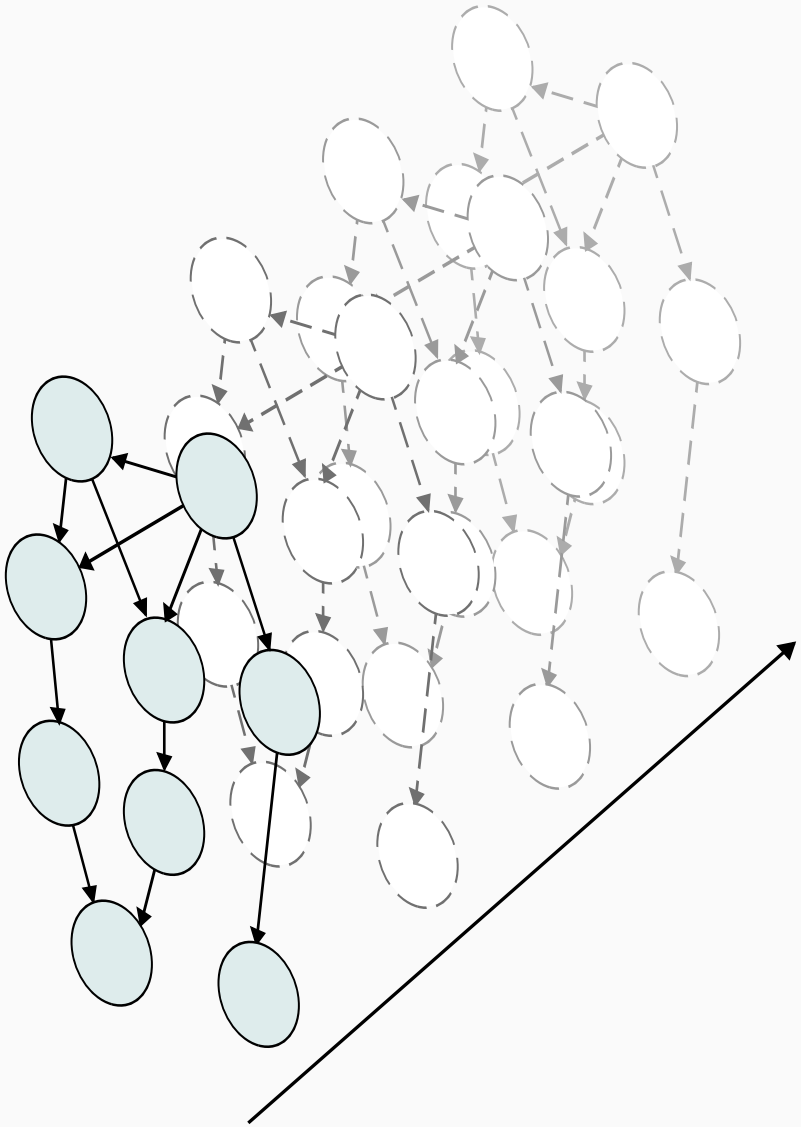
Loose vs. strict dependencies in picture



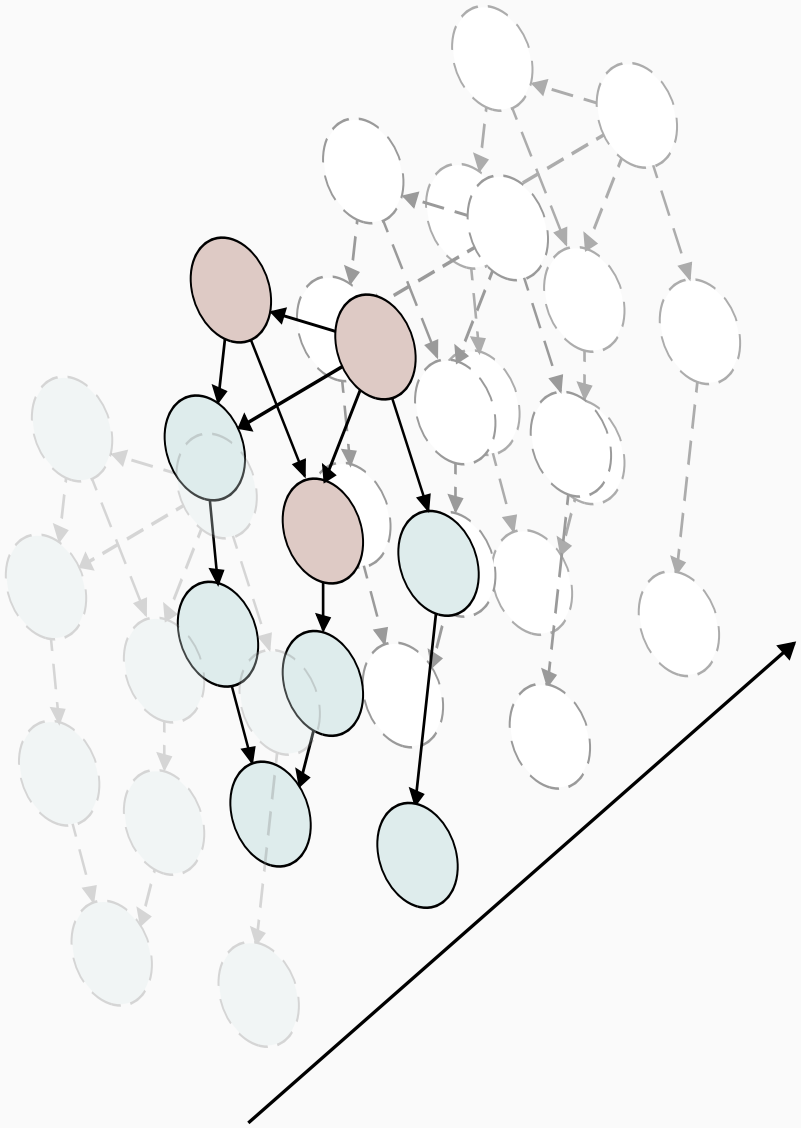
Loose vs. strict dependencies in picture



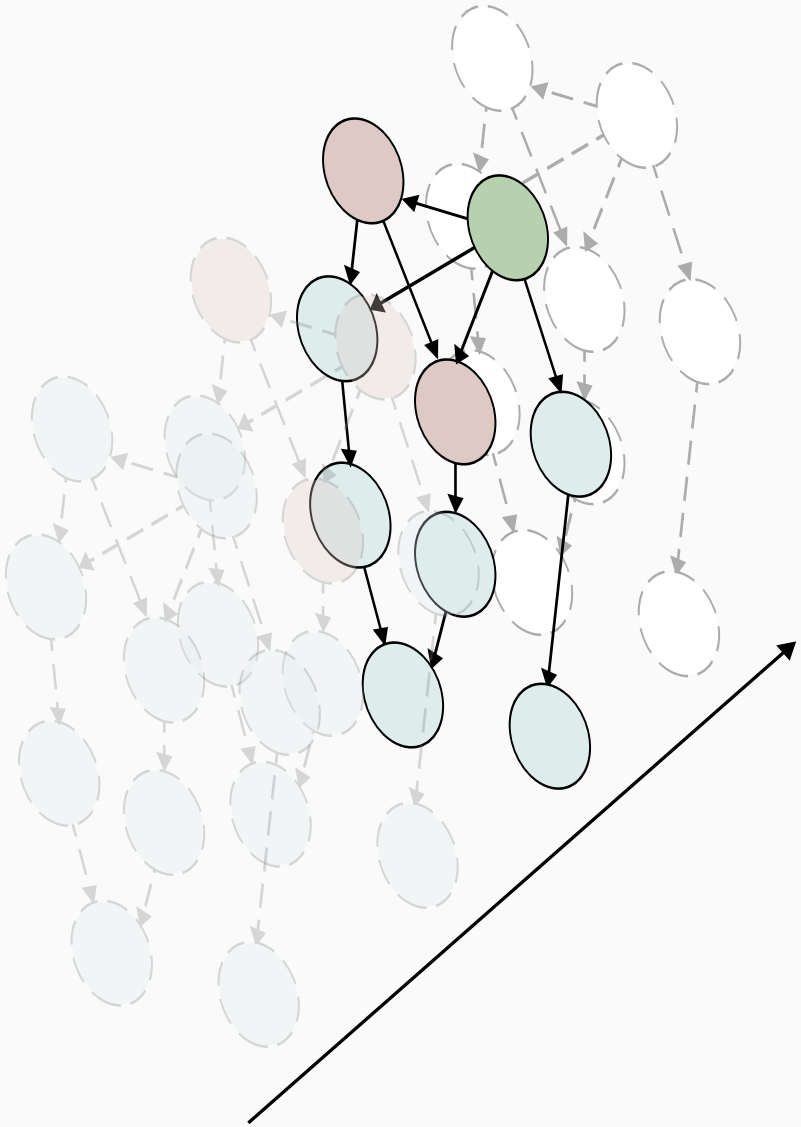
Loose vs. strict dependencies in picture



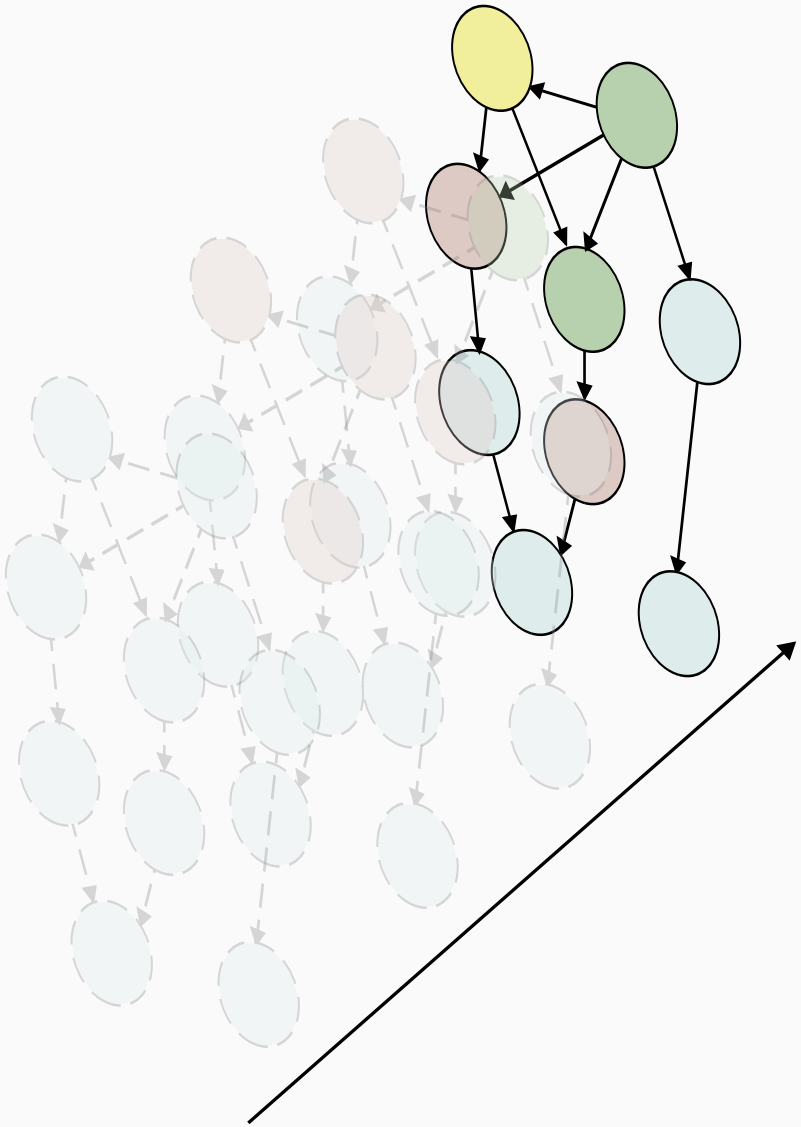
Loose vs. strict dependencies in picture



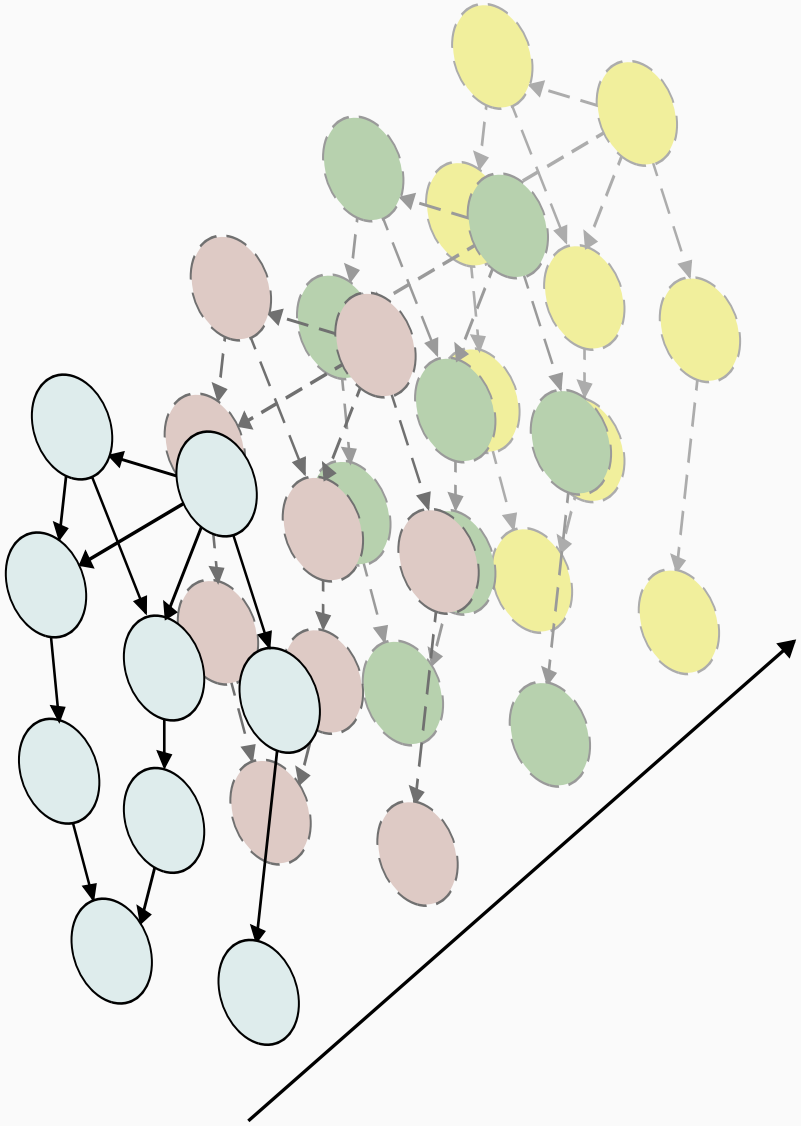
Loose vs. strict dependencies in picture



Loose vs. strict dependencies in picture



Loose vs. strict dependencies in picture



GUIX in a Nutshell (1/2)

GUIX is not a ~~container technology~~, it is meant for package management!

GUIX in a Nutshell (1/2)

GUIX is not a ~~container technology~~, it is meant for package management!

- A GUIX recipe (aka `manifest.scm`)

```
(specifications->manifest
   (list "hello" "coreutils"))
```

```
guix shell -C -m manifest.scm -- ls /   # -C = --container
```

GUIX in a Nutshell (1/2)

GUIX is not a ~~container technology~~, it is meant for package management!

- A GUIX recipe (aka `manifest.scm`)

```
(specifications->manifest
   (list "hello" "coreutils"))
```

```
guix shell -C -m manifest.scm -- ls /   # -C = --container
```

- The description of the versions is in the `manifest.scm`

```
guix describe -f channels > channels.scm
```

```
(list (channel
       (name 'guix)
       (url "https://git.savannah.gnu.org/git/guix.git")
       (branch "master")
       (commit
         "d09a4cc7c739f4201821623f62c69bcef9c20f52")
       (introduction
         (make-channel-introduction
           "9edb3f66fd807b096b48283debdcddccfea34bad"
           (openpgp-fingerprint
             "BBB0 2DDF 2CEA F6A8 0D1D  E643 A2A0 6DF2 A33A 54FA")))))
```

GUIX in a Nutshell (1/2)

GUIX is not a ~~container technology~~, it is meant for package management!

- A GUIX recipe (aka `manifest.scm`)

```
(specifications->manifest
   (list "hello" "coreutils"))
```

```
guix shell -C -m manifest.scm -- ls /   # -C = --container
```

- The description of the versions is in the `manifest.scm`

```
guix describe -f channels > channels.scm
```

```
(list (channel
       (name 'guix)
       (url "https://git.savannah.gnu.org/git/guix.git")
       (branch "master")
       (commit
         "d09a4cc7c739f4201821623f62c69bcef9c20f52")
       (introduction
         (make-channel-introduction
           "9edb3f66fd807b096b48283debdcddccfea34bad"
           (openpgp-fingerprint
             "BBB0 2DDF 2CEA F6A8 0D1D  E643 A2A0 6DF2 A33A 54FA")))))
```

- A time-machine

```
guix time-machine -C channels.scm -- shell -m manifest.scm -- ls /
```

Guix in a Nutshell (2/2)

- Under the hood:
  - A deamon compiles everything in a collection of directories `/gnu/store/8fpk2cja3f07xls48jfnpgrzrljpqivr-coreutils-8.32/`
  - All the directories are assembled (with symlinks) in a `/gnu/store/j5964hh821p2h5mcadpvj16l1m9330gv-profile/` dir
  - Environment variables (`PATH`, `LD_LIBRARY_PATH`, ...) are updated accordingly

Guix in a Nutshell (2/2)

- Under the hood:
  - A deamon compiles everything in a collection of directories `/gnu/store/8fpk2cja3f07xls48jfnpgrzrljpqivr-coreutils-8.32/`
  - All the directories are assembled (with symlinks) in a `/gnu/store/j5964hh821p2h5mcadpvj16l1m9330gv-profile/` dir
  - Environment variables (`PATH`, `LD_LIBRARY_PATH`, ...) are updated accordingly
- Several containerization options

```
1  guix shell --container coreutils -- ls
2  guix shell coreutils -- ls   # Fully permeable: expends the env, ...
```

Guix in a Nutshell (2/2)

- Under the hood:
  - A deamon compiles everything in a collection of directories `/gnu/store/8fpk2cja3f07xls48jfnpgrzrljpqivr-coreutils-8.32/`
  - All the directories are assembled (with symlinks) in a `/gnu/store/j5964hh821p2h5mcadpvj16l1m9330gv-profile/` dir
  - Environment variables (`PATH`, `LD_LIBRARY_PATH`, ...) are updated accordingly
- Several containerization options

```
1  guix shell --container coreutils -- ls
2  guix shell coreutils -- ls   # Fully permeable: expends the env, ...
```

- Various export formats (`docker`, `squashfs`, `debian`, `tarball`, `module`, `relocatable`...)

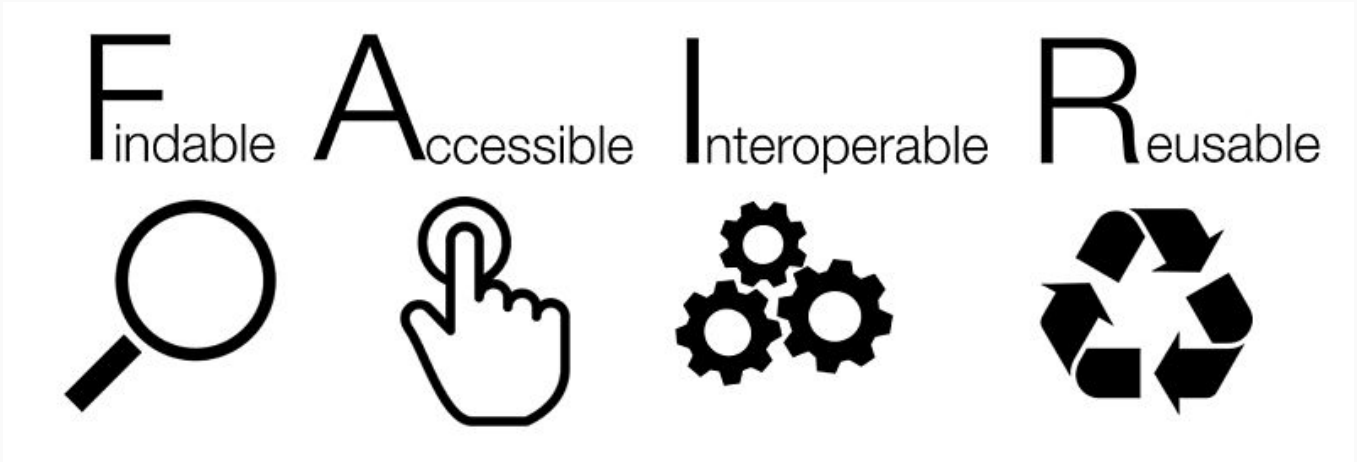
```
1  guix pack --format=docker --save-provenance -m manifest.scm
```

Allows to carefully control/nest environments and adapt to your context

- For more information: https://hpc.guix.info/
  - Lastest Guix Workshop on HPC was held in Bordeaux on Nov. 7, 2024, after the JCAD
  - Check out the Café Guix

Good Practice #3
Version Control and Archiving

FAIR principles



https://www.go-fair.org/fair-principles/

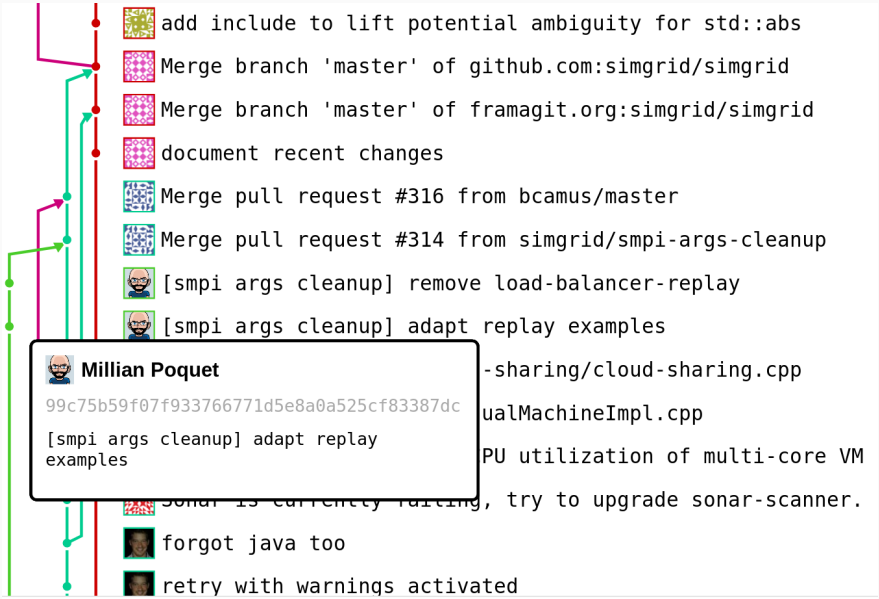
- *"Open as much as possible and close as much as necessary"*
- Management, publication, annotation (metadata), archiving
- Source code = specific data with specific consideration

Let's go beyond general principles!

Tool 3: Version Control and Forge

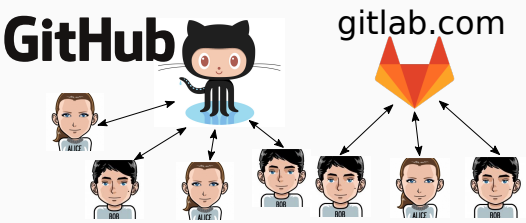
Git = version control

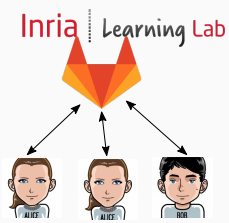
- Developed in 2005 by Linus Torvalds for the kernel development
- Local and efficient rollbacks
- Distributed: everyone has a full copy of the history

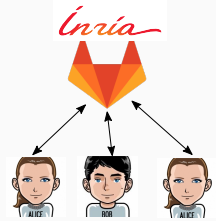


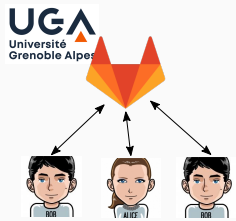
GitHub, GitLab, and Co

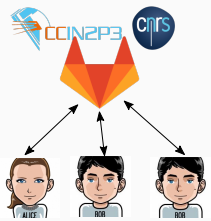
- Free hosting of public projects, social network











Limitation

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

Tool 3bis: Fighting Information Loss with Archives

 or  = awesome collaborations (≠ archive)

- D. Spinellis. *The Decay and Failures of URL References*. CACM, 46(1), 2003
  The half-life of a referenced URL is approximately 4 years from its publication date.
- P. Habibzadeh. *Decay of References to Web sites in Articles Published in General Medical Journals: Mainstream vs Small Journals*. Applied Clinical Informatics. 4 (4), 2013
  half life ranged from 2.2 years in EMHJ to 5.3 years in BMJ
- Discontinuated forges: Code Space, Gitorious, Google code, Inria Gforge

Tool 3bis: Fighting Information Loss with Archives

or

= awesome collaborations (≠ archive)

- D. Spinellis. *The Decay and Failures of URL References*. CACM, 46(1), 2003
  The half-life of a referenced URL is approximately 4 years from its publication date.
- P. Habibzadeh. *Decay of References to Web sites in Articles Published in General Medical Journals: Mainstream vs Small Journals*. Applied Clinical Informatics. 4 (4), 2013
  half life ranged from 2.2 years in EMHJ to 5.3 years in BMJ
- Discontinuated forges: Code Space, Gitorious, Google code, Inria Gforge

Article archives arXiv.org HAL archives-ouvertes.fr

Data archives figshare zenodo

Software Archive Software Heritage

INTERNET ARCHIVE

Collect/Preserve/Share

Controling the whole software/compiling stack is not sufficient

Floating Points Rounding: the other root of all evil ?

- Every operation includes implicit rounding.

```
1  print(2.1-2.0 == 0.1)
```

```
False
```

Floating Points Rounding: the other root of all evil ?

- Every operation includes implicit rounding.

```
1  print(2.1-2.0 == 0.1)
```

```
False
```

- Unfortunately: round(round(a+b)+c) $\neq$ round(a+round(b+c))
  Hence, operation order matters. For a reproducible computation, operation order should be preserved!!! Which order is more relevant is an other debate 😉

Floating Points Rounding: the other root of all evil ?

- Every operation includes implicit rounding.

```
1  print(2.1-2.0 == 0.1)
```

```
False
```

- Unfortunately: round(round(a+b)+c) $\neq$ round(a+round(b+c))
  Hence, operation order matters. For a reproducible computation, operation order should be preserved!!! Which order is more relevant is an other debate 😉
- Numerical instability may be closer than you think [Rump, 1988]

$$f(x,y) = 333.75y^6 + x^2(11x^2y^2 - y^6 - 121y^4 - 2)2 + 5.5y^8 + \frac{x}{2y}$$

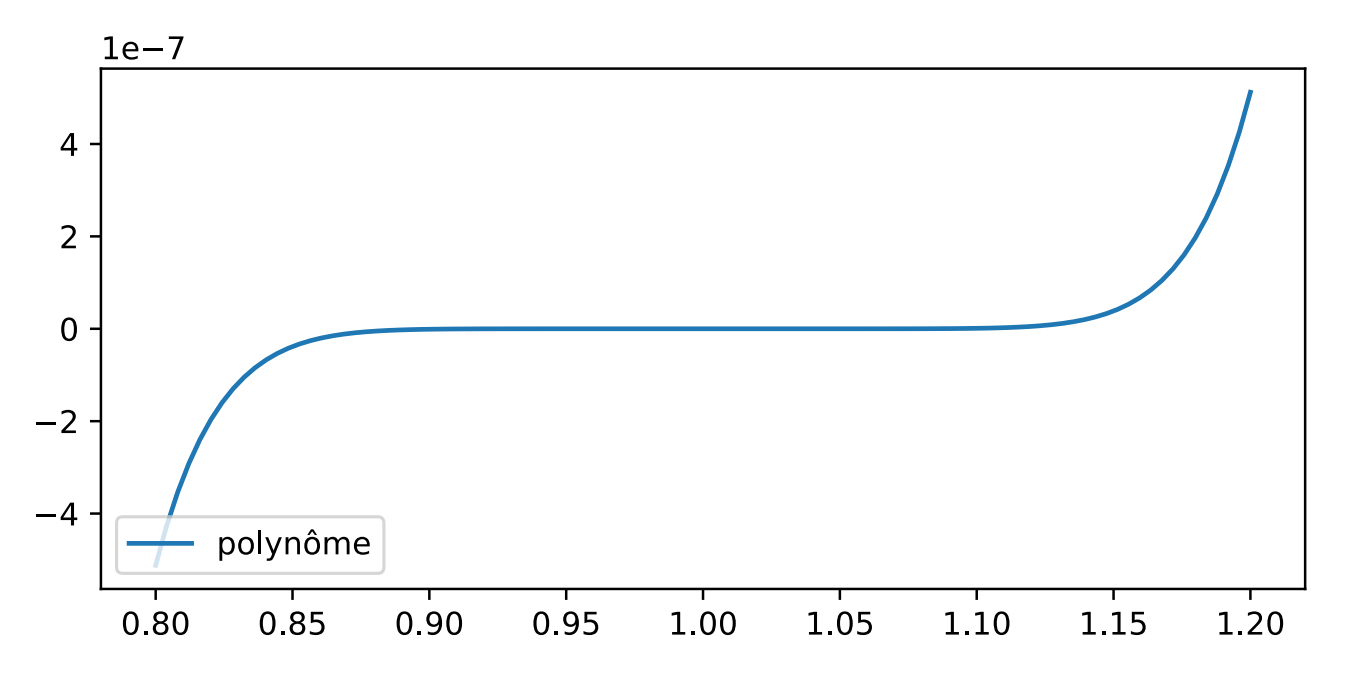
Evaluation of $f(77617.0, 33096.0)$

Single precision	1.172603
Double precision	1.1726039400531
Extended precision	1.172603940053178
MPFI (multiple precision interval arithmetic)	[-0.827396059946821368141165..., -0.827396059946821368141165...]

Courtesy of Christophe Denis

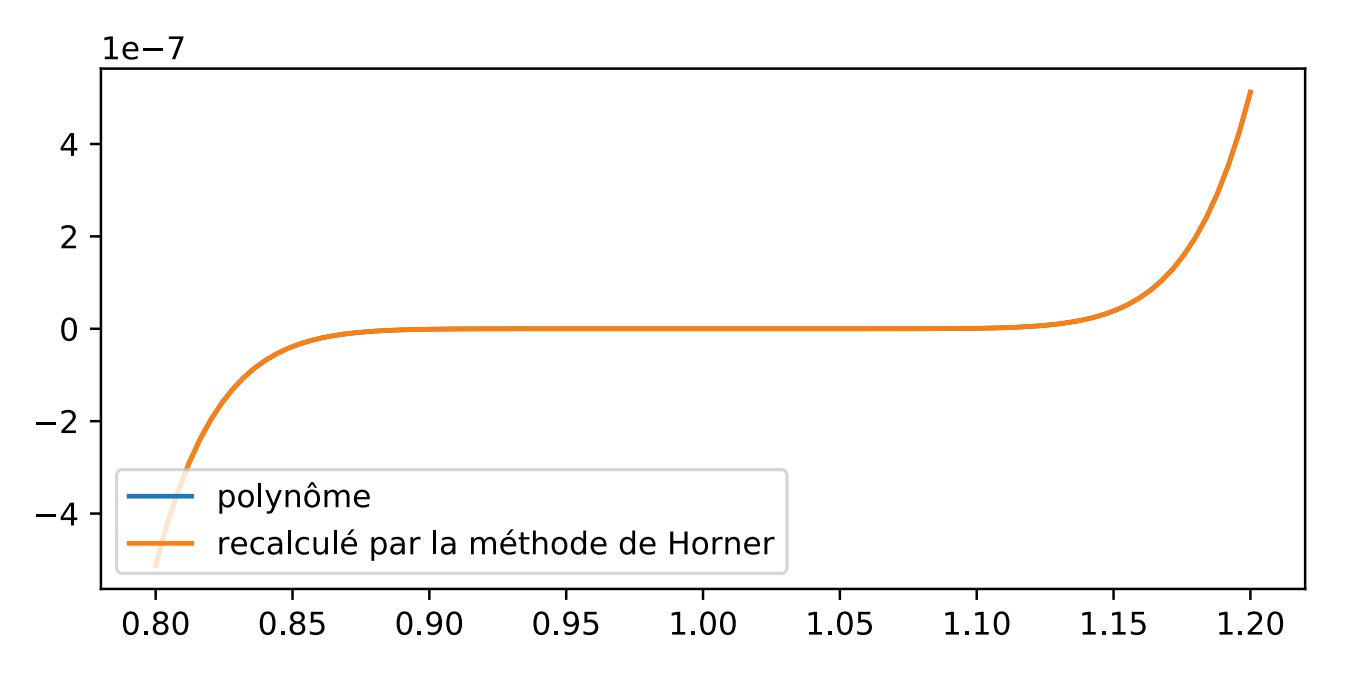
At scale (complex code + non-determinism), all this can become particularly harmful and painful.

All I care about is the algorithm output (FP)



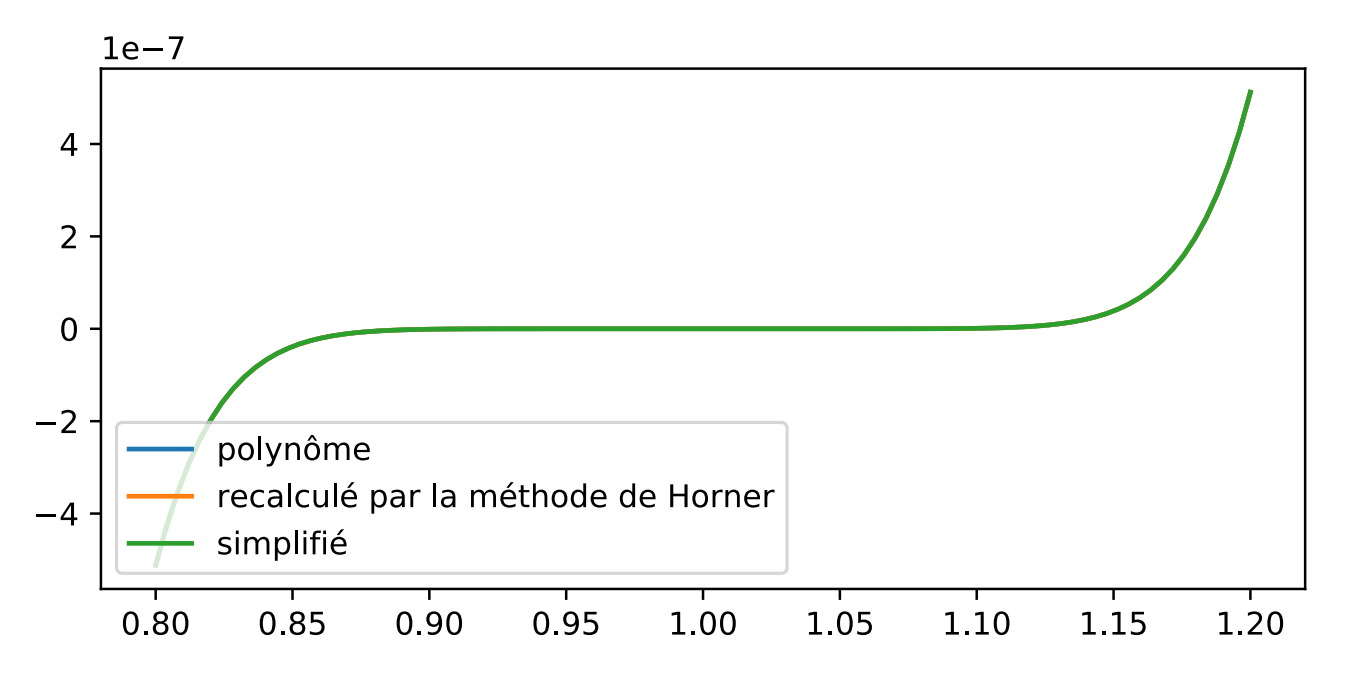

```
1  def polynome(x):
2      return x**9 - 9.*x**8 + 36.*x**7 - 84.*x**6 + 126.*x**5 \
3             - 126.*x**4 + 84.*x**3 - 36.*x**2 + 9.*x - 1.
```

Floating-point arithmetic



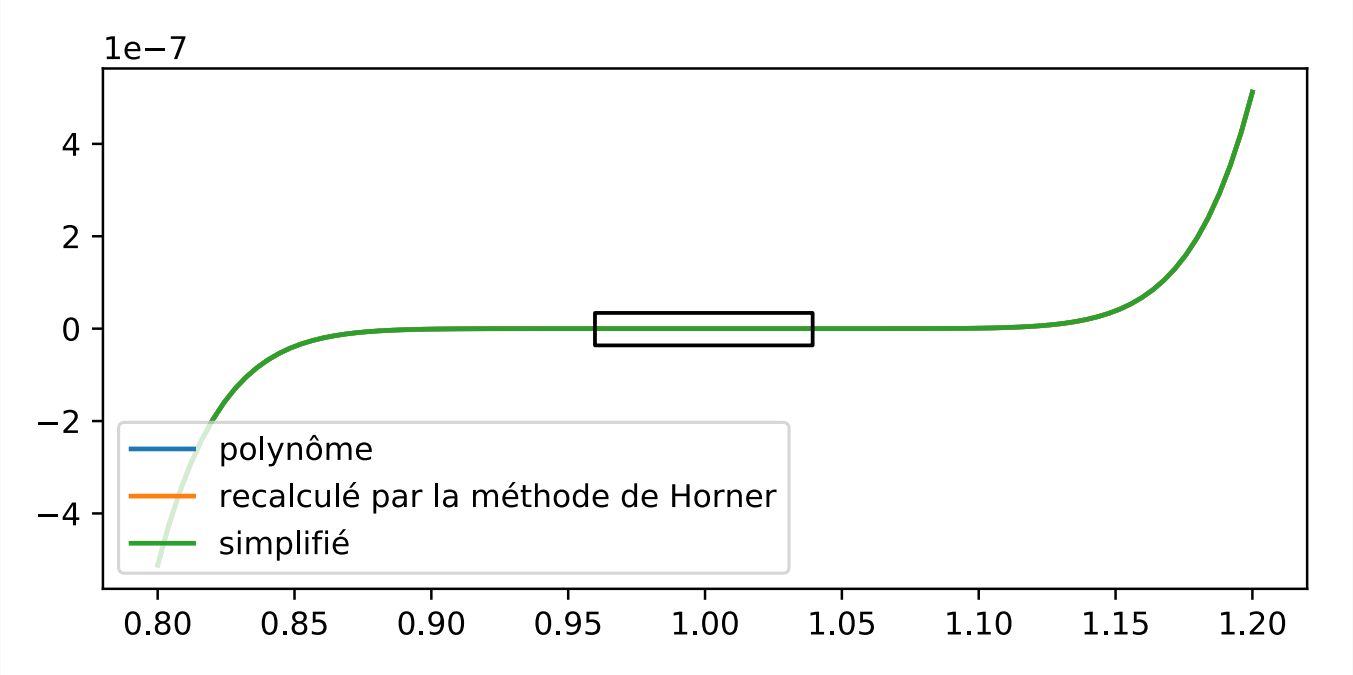

```
1  def horner(x):
2      return x*(x*(x*(x*(x*(x*(x*(x*(x - 9.) + 36.) - 84.) + 126.) \
3             - 126.) + 84.) - 36.) + 9.) - 1.
```

Floating-point arithmetic

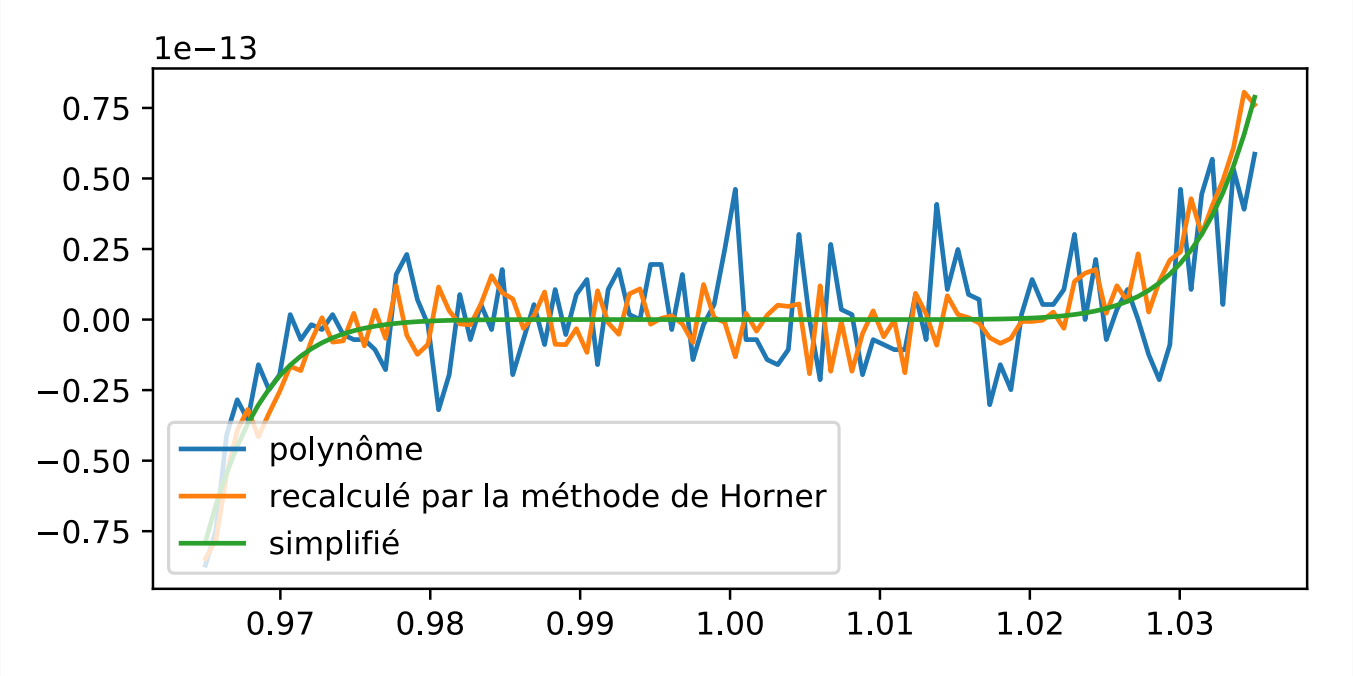



```
1  def simple(x):
2      return (x-1.)**9
3  # Easy! ;)
```

Floating-point arithmetic



Floating-point arithmetic

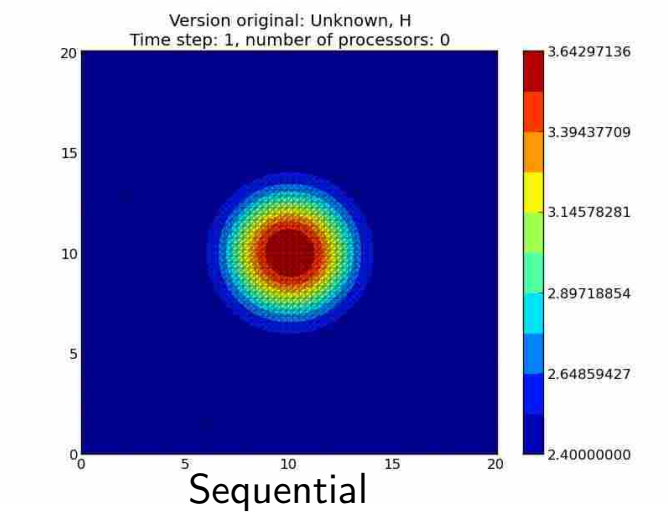


Telemac2D: the simplest gouttedo simulation

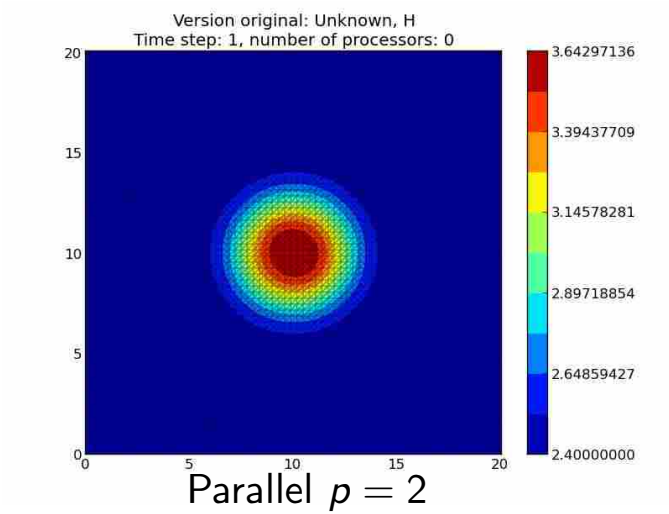
The gouttedo test case

- 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 $p = 2$

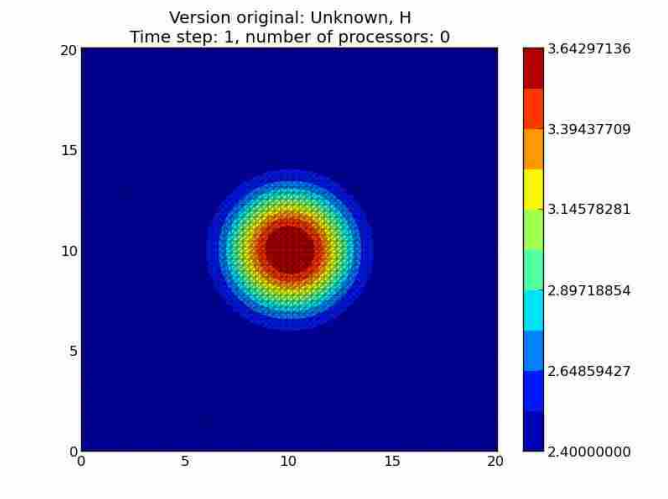
Courtesy of P. Langlois and R. Nheili

Did I mention we have parallel machines nowadays?

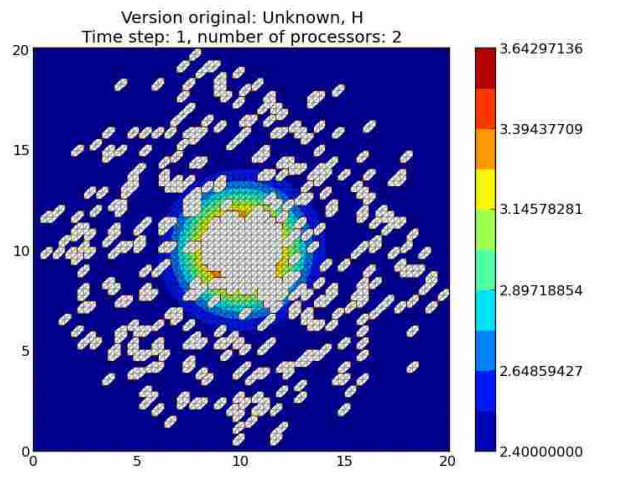
A white plot displays a non-reproducible value

Numerical reproducibility?

time step = 1



Sequential



Parallel $p = 2$

14 / 54

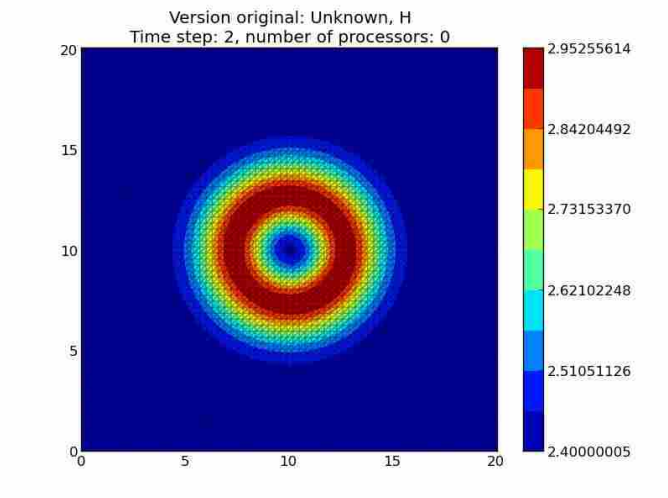
Courtesy of P. Langlois and R. Nheili

Did I mention we have parallel machines nowadays?

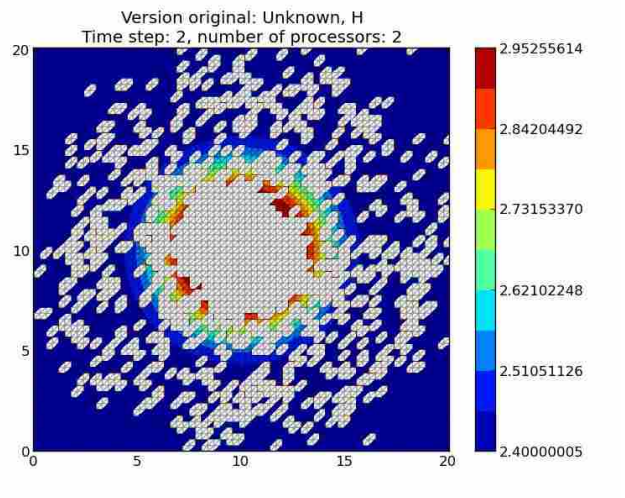
A white plot displays a non-reproducible value

Numerical reproducibility?

time step = 2



Sequential



Parallel $p = 2$

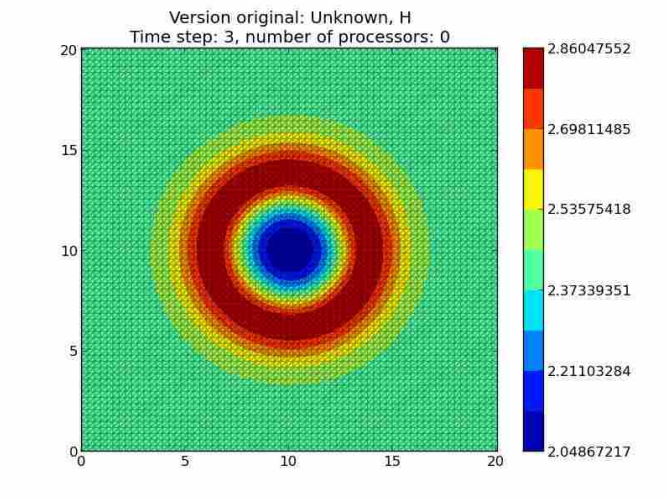
Courtesy of P. Langlois and R. Nheili

Did I mention we have parallel machines nowadays?

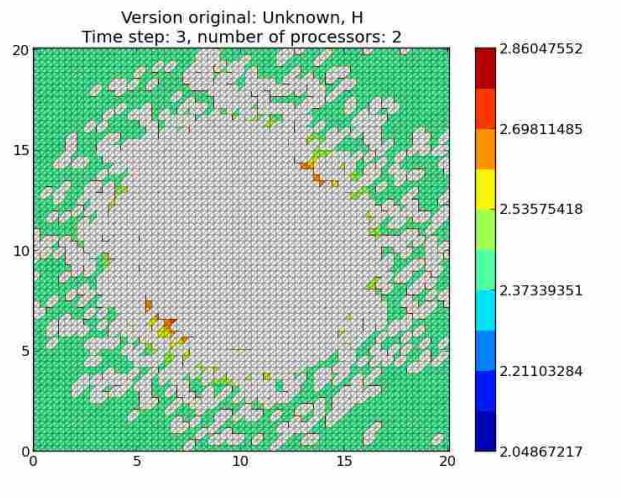
A white plot displays a non-reproducible value

Numerical reproducibility?

time step = 3



Sequential



Parallel $p = 2$

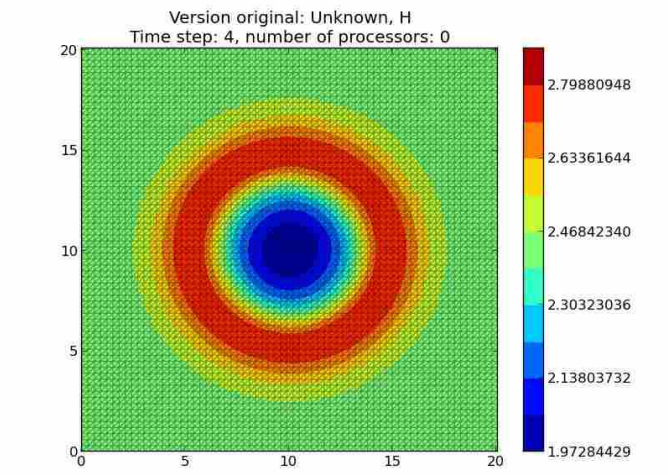
Courtesy of P. Langlois and R. Nheili

Did I mention we have parallel machines nowadays?

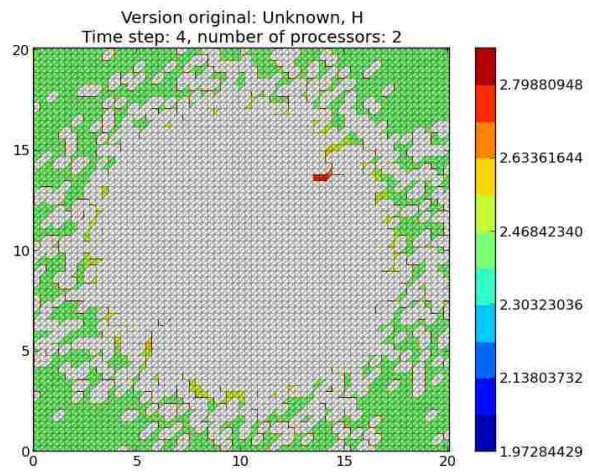
A white plot displays a non-reproducible value

Numerical reproducibility?

time step = 4



Sequential



Parallel $p = 2$

14 / 54

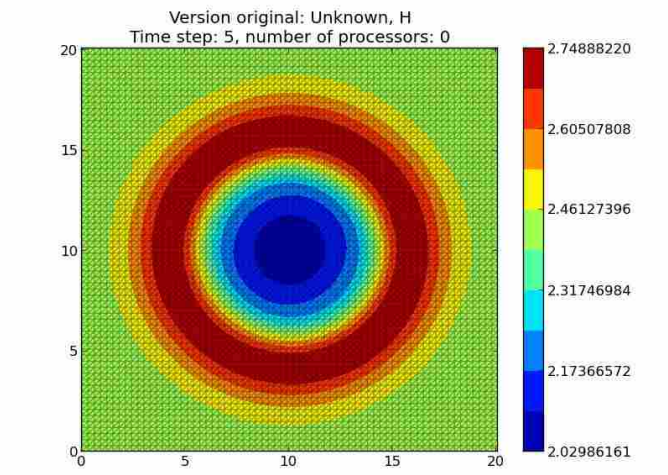
Courtesy of P. Langlois and R. Nheili

Did I mention we have parallel machines nowadays?

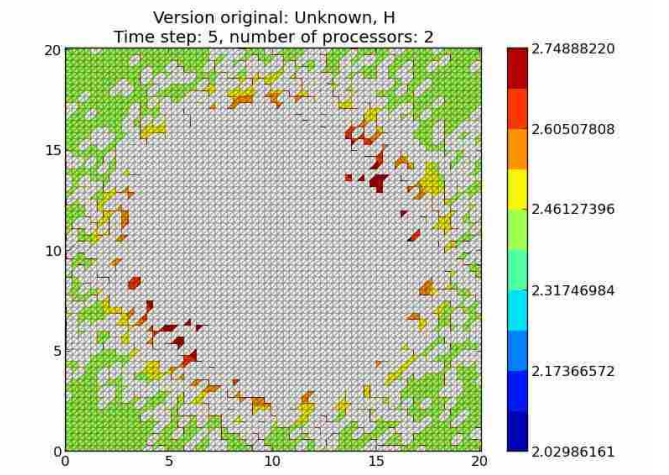
A white plot displays a non-reproducible value

Numerical reproducibility?

time step = 5



Sequential



Parallel $p = 2$

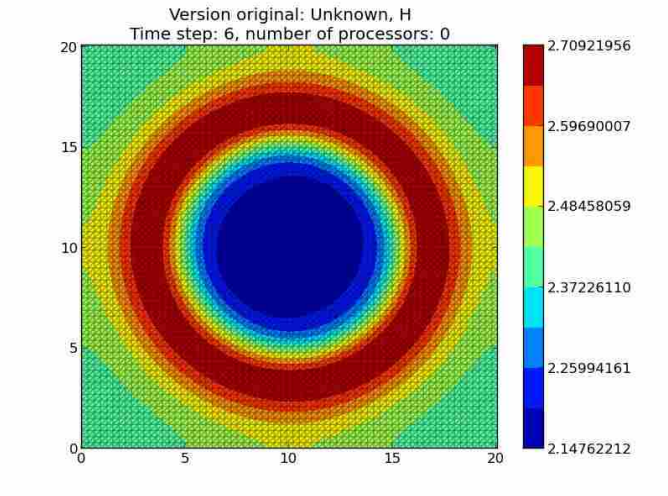
Courtesy of P. Langlois and R. Nheili

Did I mention we have parallel machines nowadays?

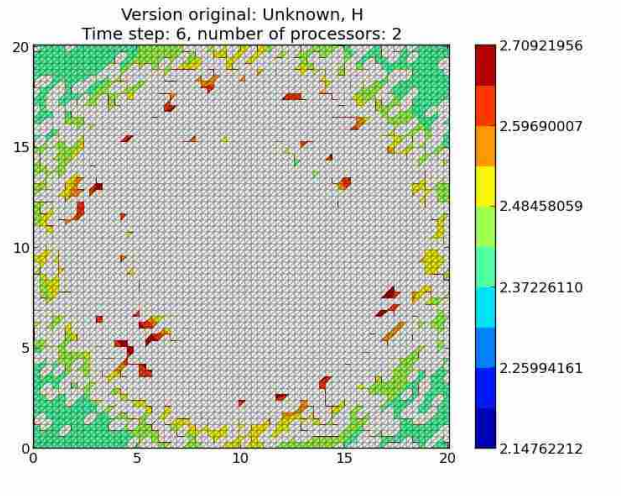
A white plot displays a non-reproducible value

Numerical reproducibility?

time step = 6



Sequential



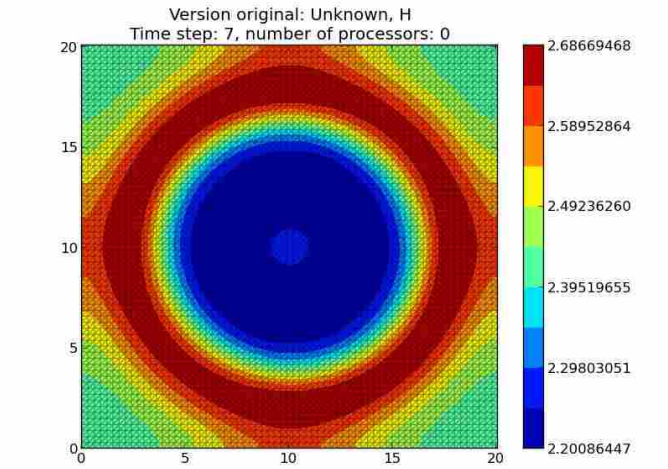
Parallel $p = 2$

Courtesy of P. Langlois and R. Nheili

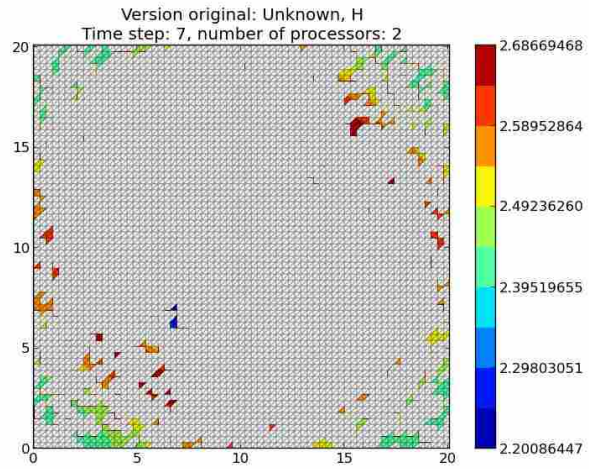
A white plot displays a non-reproducible value

Numerical reproducibility?

time step = 7



Sequential



Parallel $p = 2$

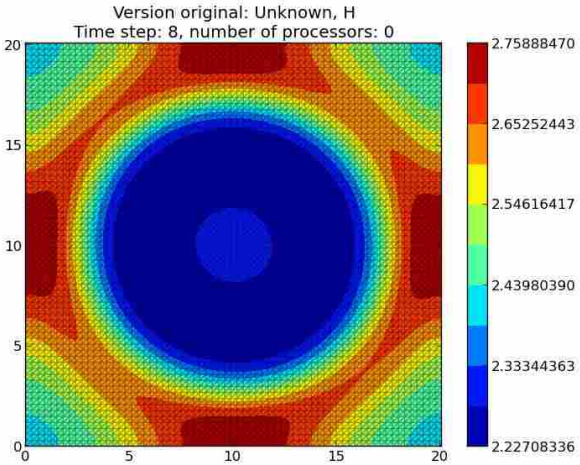
14 / 54

Courtesy of P. Langlois and R. Nheili

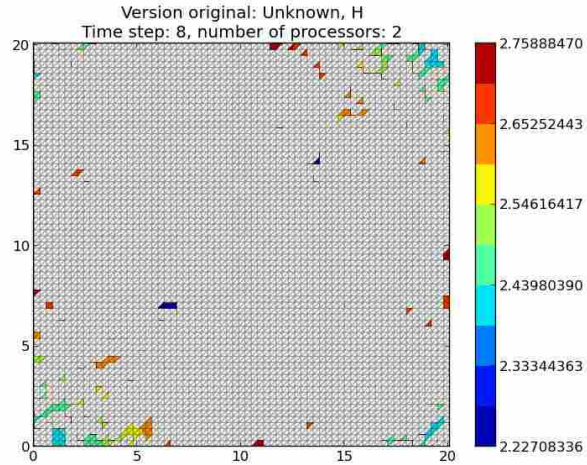
A white plot displays a non-reproducible value

Numerical reproducibility?

time step = 8



Sequential



Parallel $p = 2$

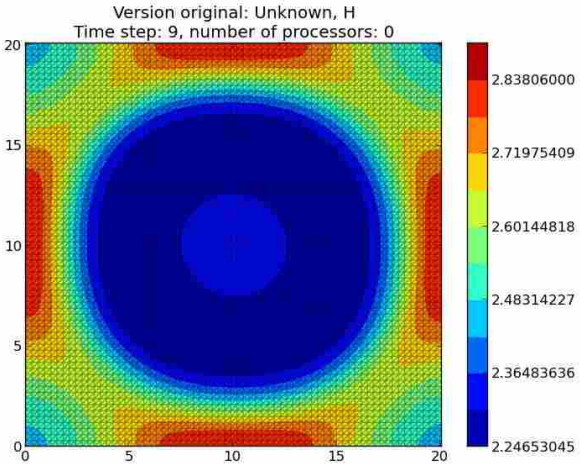
14 / 54

Courtesy of P. Langlois and R. Nheili

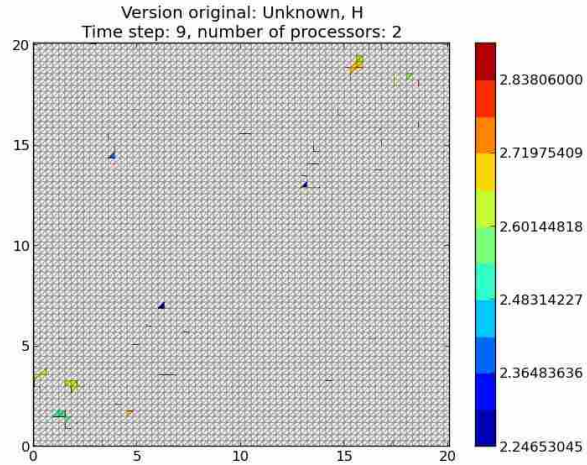
A white plot displays a non-reproducible value

Numerical reproducibility?

time step = 9



Sequential



Parallel $p = 2$

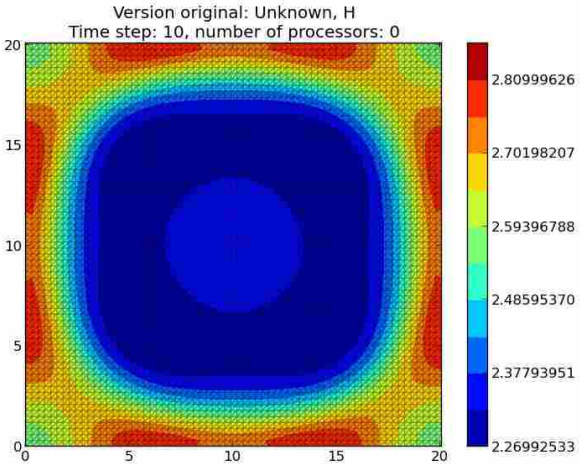
Courtesy of P. Langlois and R. Nheili

Did I mention we have parallel machines nowadays?

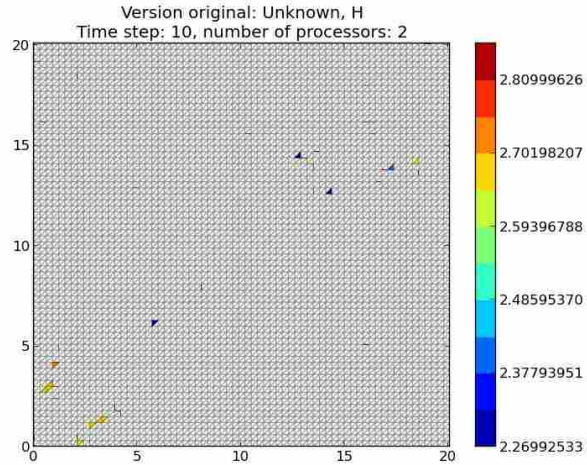
A white plot displays a non-reproducible value

Numerical reproducibility?

time step = 10



Sequential



Parallel $p = 2$

14 / 54

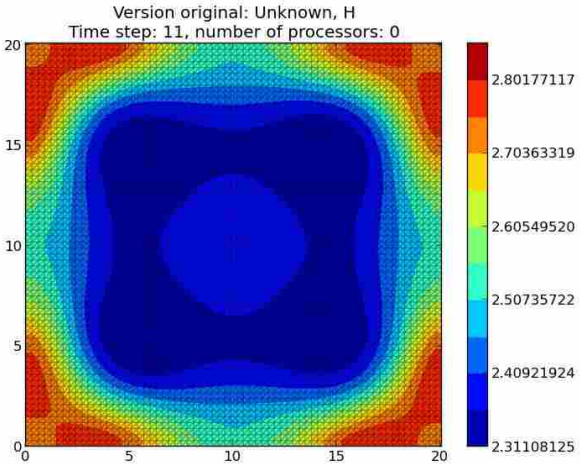
Courtesy of P. Langlois and R. Nheili

Did I mention we have parallel machines nowadays?

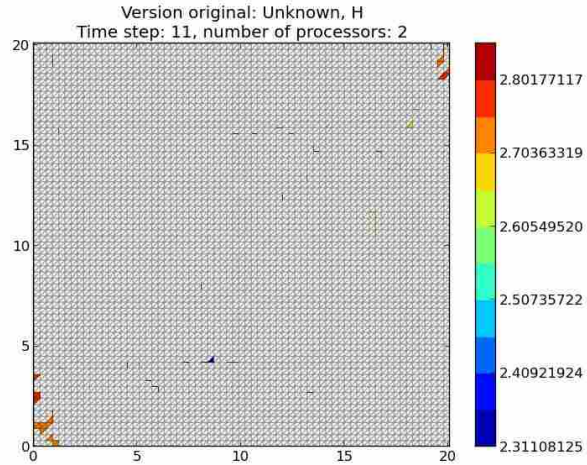
A white plot displays a non-reproducible value

Numerical reproducibility?

time step = 11



Sequential



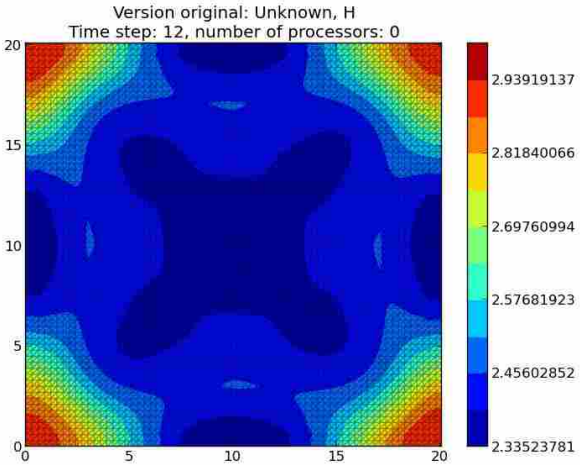
Parallel $p = 2$

Courtesy of P. Langlois and R. Nheili

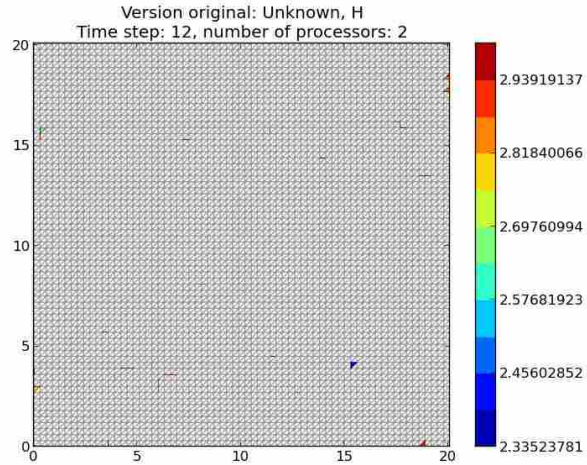
A white plot displays a non-reproducible value

Numerical reproducibility?

time step = 12



Sequential



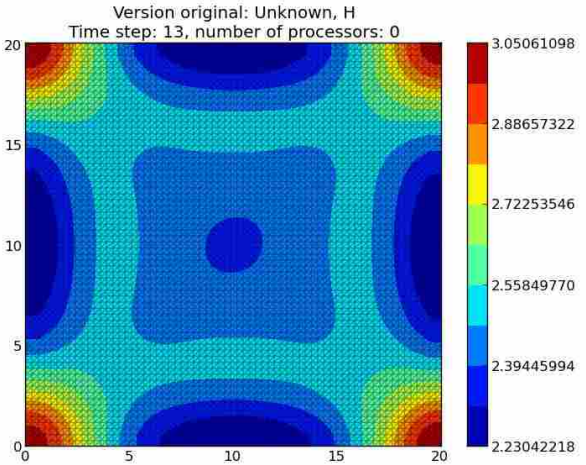
Parallel $p = 2$

Courtesy of P. Langlois and R. Nheili

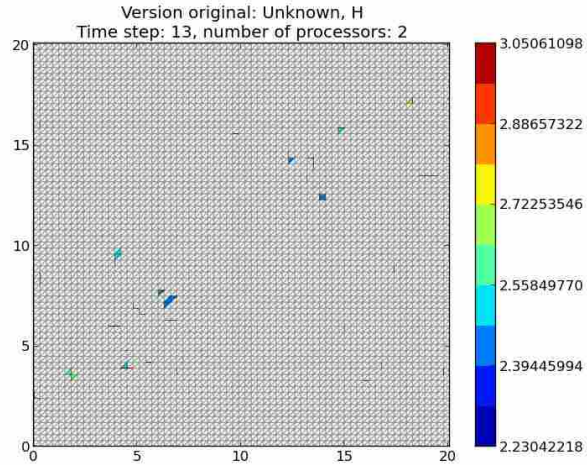
A white plot displays a non-reproducible value

Numerical reproducibility?

time step = 13



Sequential



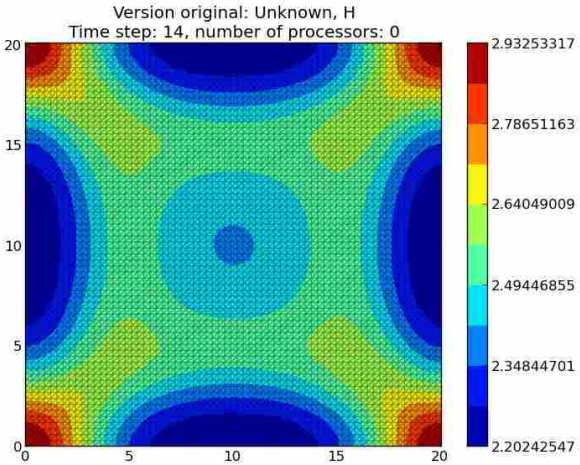
Parallel $p = 2$

Courtesy of P. Langlois and R. Nheili

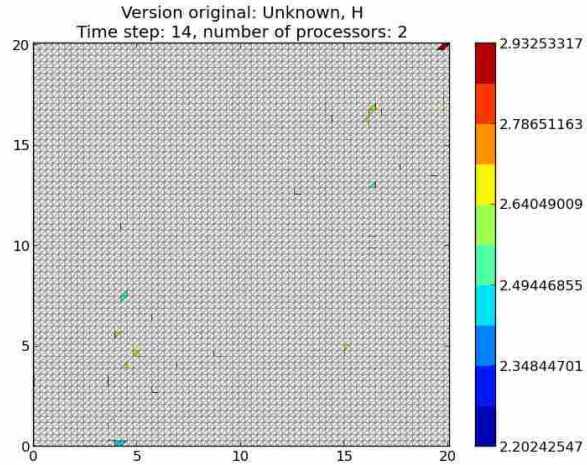
A white plot displays a non-reproducible value

Numerical reproducibility?

time step = 14



Sequential



Parallel $p = 2$

14 / 54

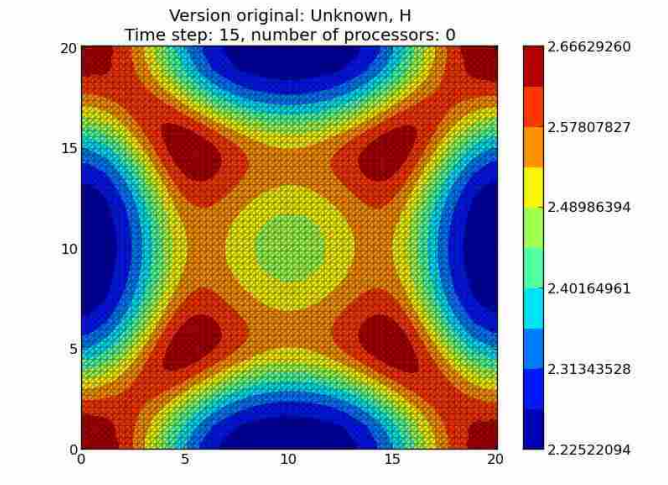
Courtesy of P. Langlois and R. Nheili

Did I mention we have parallel machines nowadays?

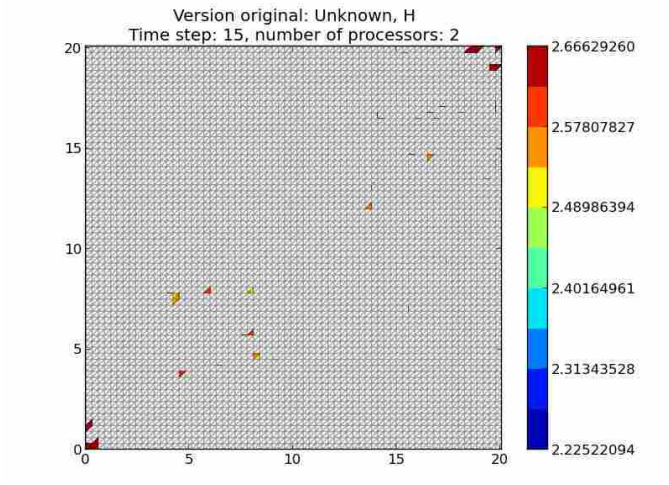
A white plot displays a non-reproducible value

NO numerical reproducibility!

time step = 15



Sequential

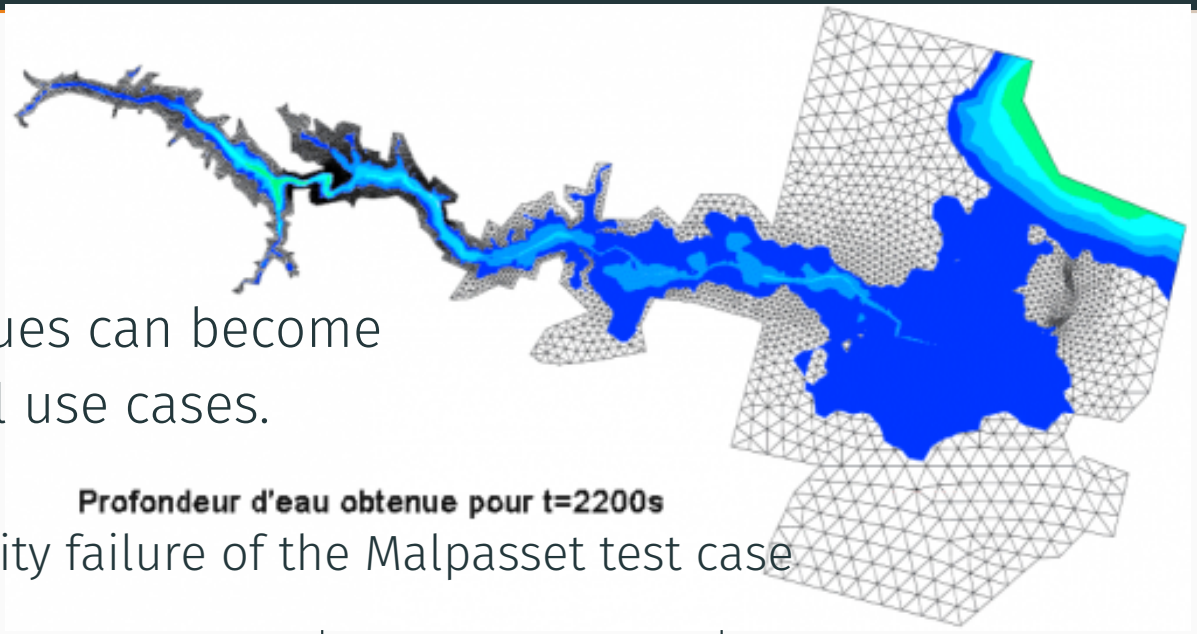


Parallel $p = 2$

14 / 54

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 the algorithm, in its implementation, in its execution.

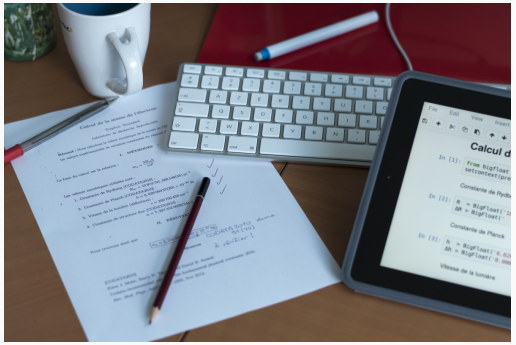
The whole chain needs to be revisited.

Courtesy of P. Langlois and R. Nheili

Advertisement

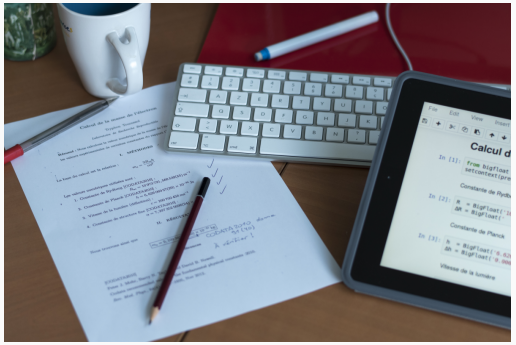
MOOCs on Reproducible Research:

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



- Konrad Hinsen, Christophe Pouzat
- 3rd Edition: March 2020 – … (16,800+)
- Notebooks, version control, simple data formats

MOOCs on Reproducible Research:



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

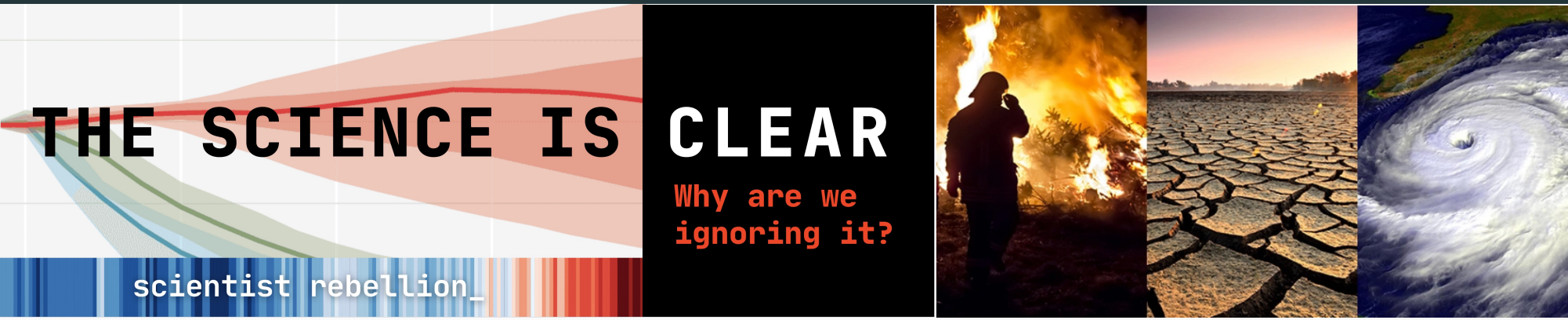
- Konrad Hinsen, Christophe Pouzat
- 3rd Edition: March 2020 – … (16,800+)
- Notebooks, version control, simple data formats

MOOC Reproducible Research II: Practices and tools for managing computations and data (May-Sep 2024, ≈ 2,000)

- Managing data (`FITS/HDF5`, `git annex`)
- Software environment control (`docker`, `singularity`, `guix`)
- Scientific workflow (`make`, `snakemake`)
- ~~Statistics~~, ~~Numerical Chaos~~

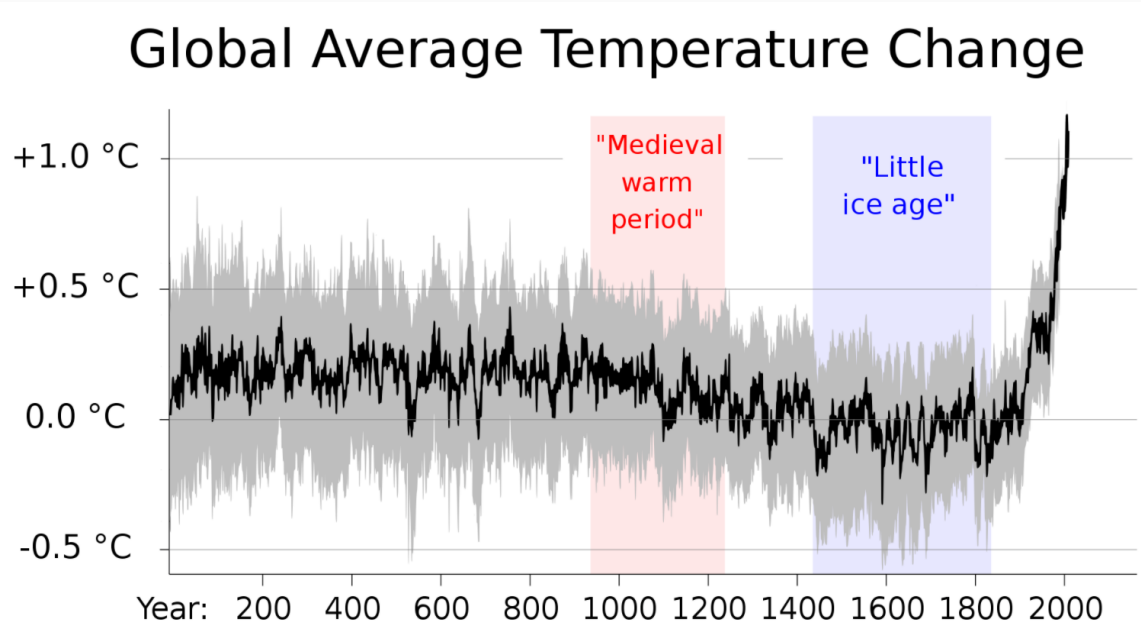
What's the point ?

The Elephant in the Room: Climate Change

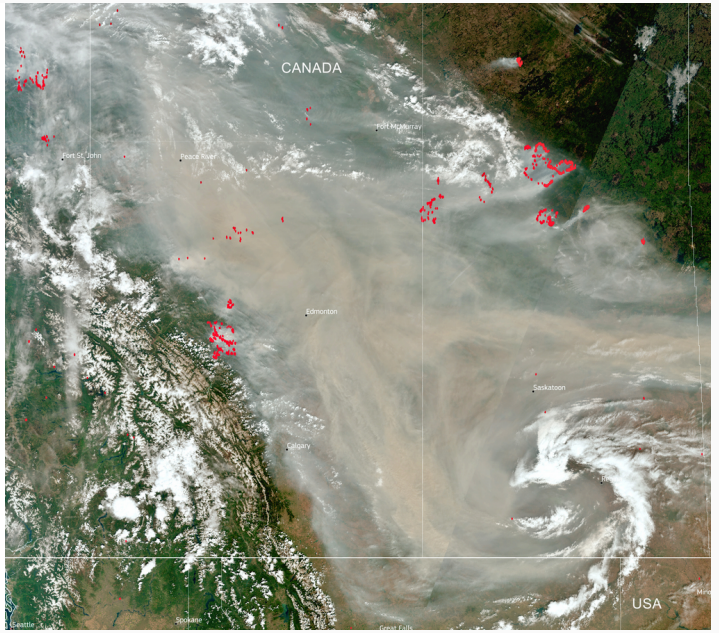


IPCC, IPBES, https://climate.nasa.gov/

1. Global climate change is **not** a future problem

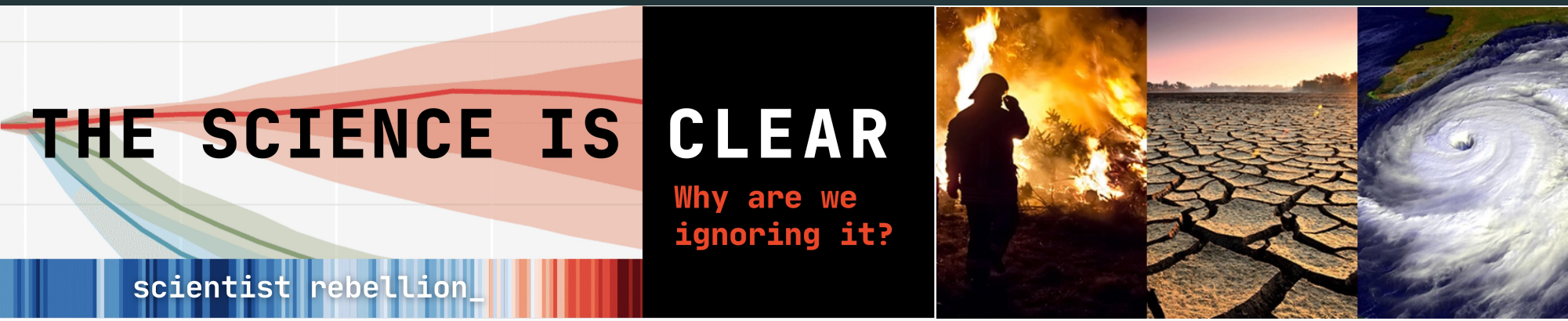


https://en.wikipedia.org/wiki/Global_temperature_record



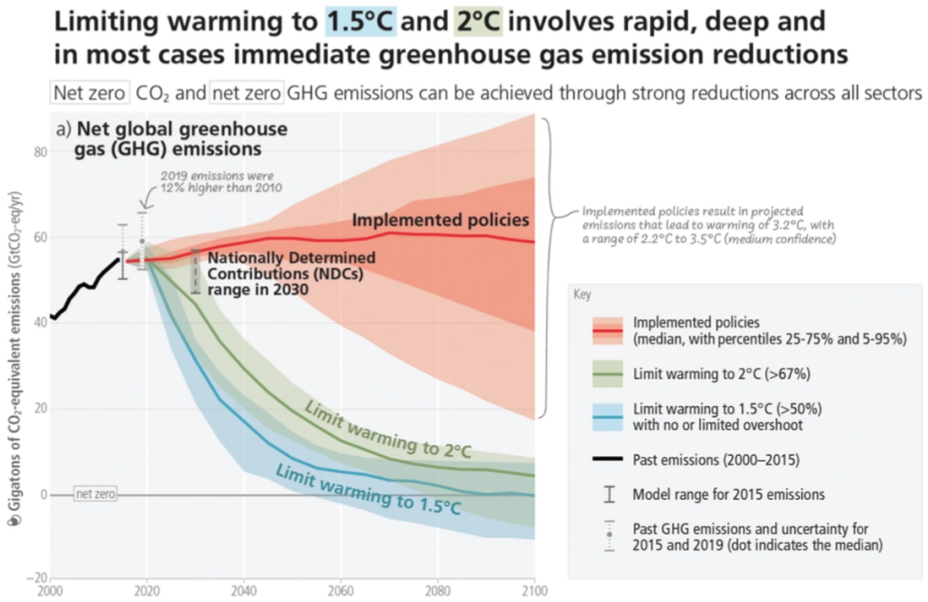
2023 Alberta wildfires (> 1 Mha)

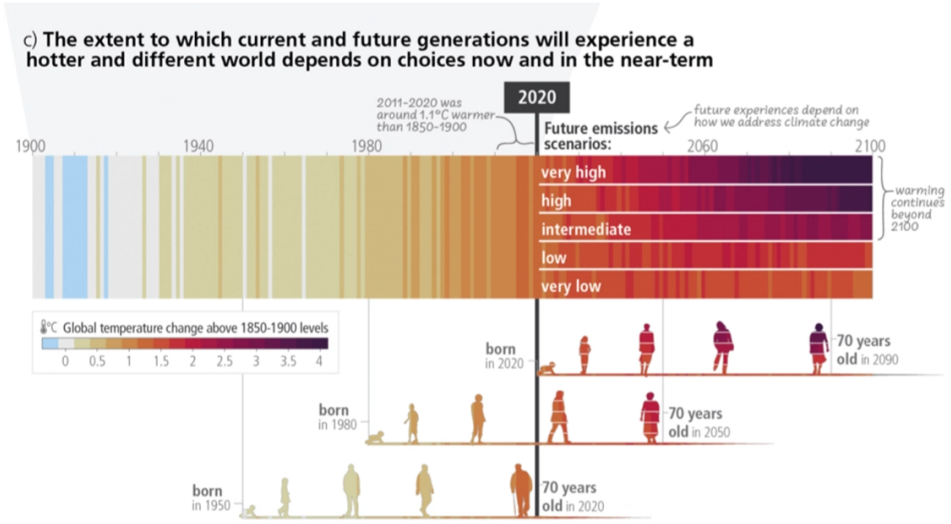
The Elephant in the Room: Climate Change



IPCC, IPBES, https://climate.nasa.gov/

1. Global climate change is **not** a future problem
2. It is **entirely** due to human activity

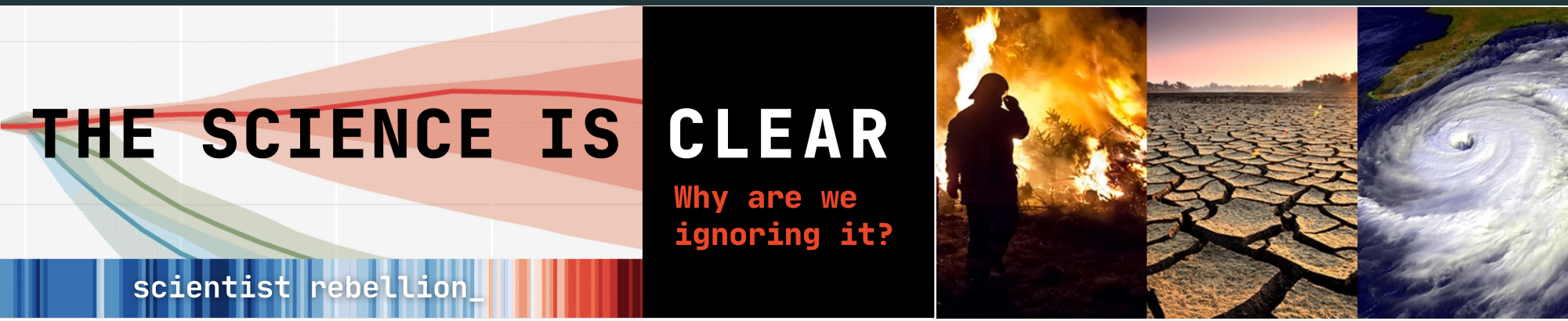




Paris Agreement'15 ∼ Net Zero by 2050

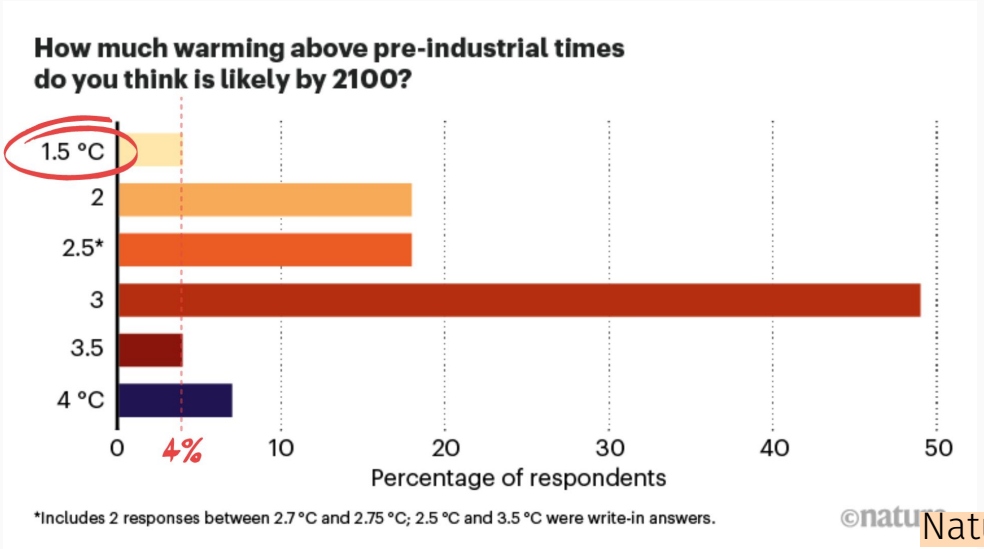
Latest IPCC report

The Elephant in the Room: Climate Change



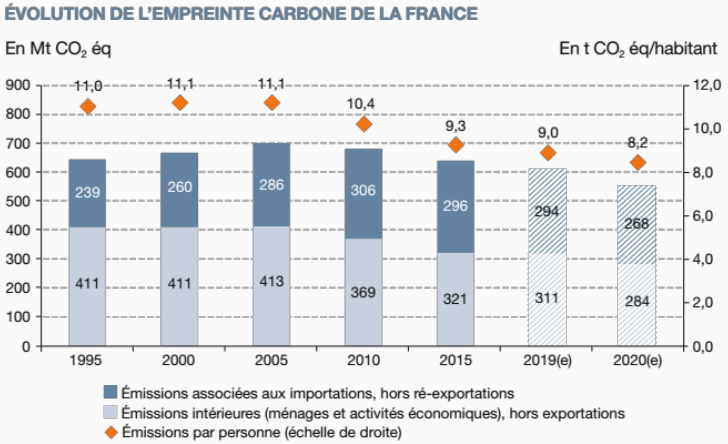
IPCC, IPBES, https://climate.nasa.gov/

1. Global climate change is **not** a future problem
2. It is **entirely** due to human activity
3. *9 out of 10 IPCC scientists believe overshoot is likely*

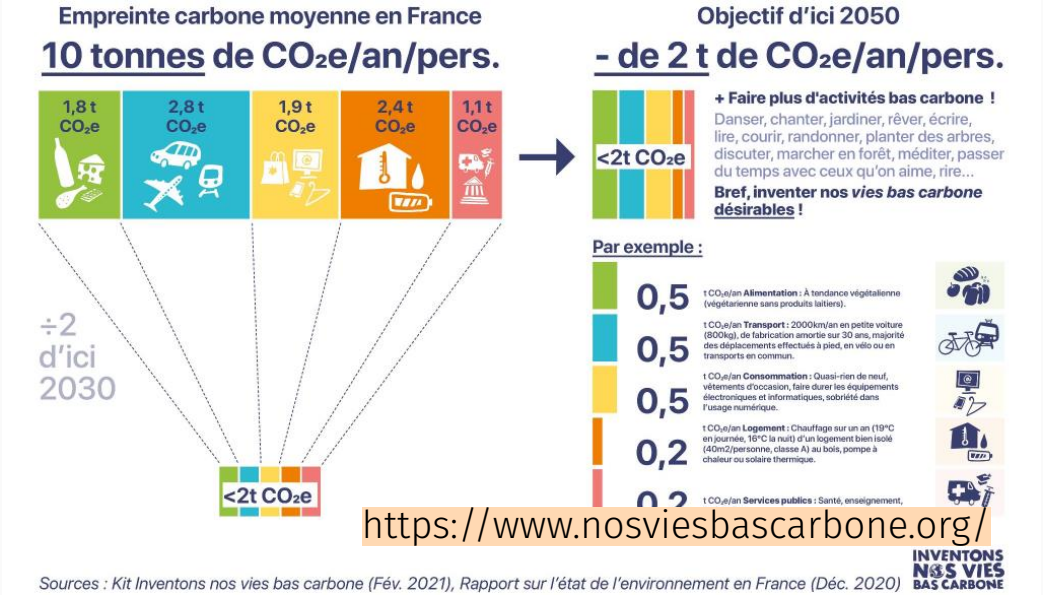


Nature survey, Nov. 2021

Put aside biodiversity loss, pollution, freshwater, land system change...

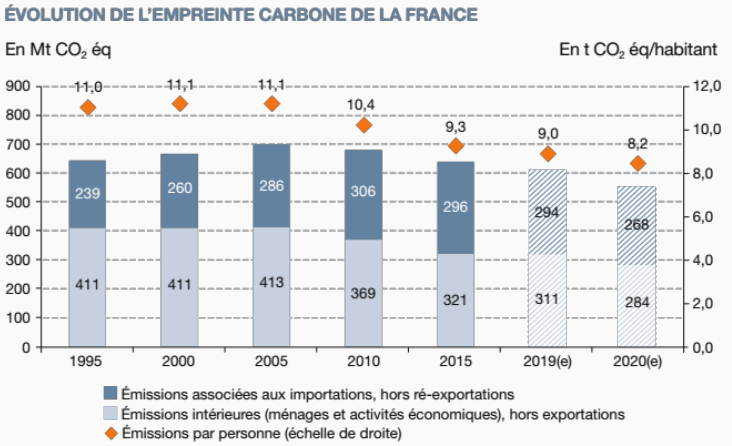


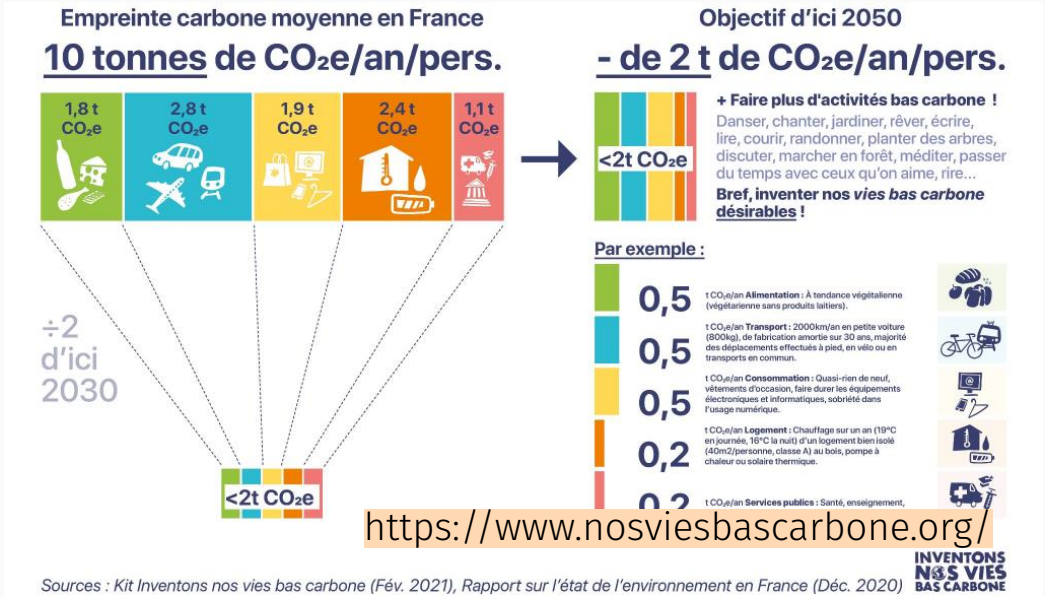
(e) = estimations.
Note : l'empreinte carbone porte sur les trois principaux gaz à effet de serre (CO_2, CH_4, N_2O). En 2021, la méthodologie a été ajustée afin de mieux tenir compte de l'évolution des coûts du pétrole brut, du gaz et du charbon ; l'ensemble de la série a ainsi été révisée, l'essentiel des ajustements portant sur les émissions importées de CH_4.
Champ : périmètre Kyoto (métropole et outre-mer appartenant à l'UE).
***Sources :** Citepa ; AIE ; FAO ; Douanes ; Eurostat ; Insee. Traitement : SDES, 2021*



Sources : Kit Inventons nos vies bas carbone (Fév. 2021), Rapport sur l'état de l'environnement en France (Déc. 2020)

Put aside biodiversity loss, pollution, freshwater, land system change...

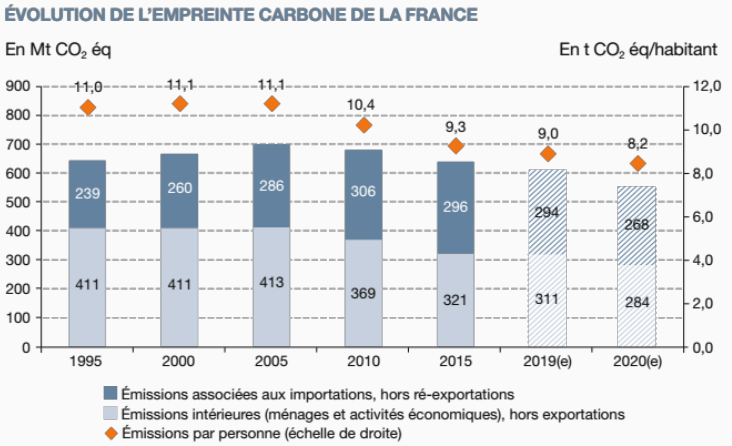


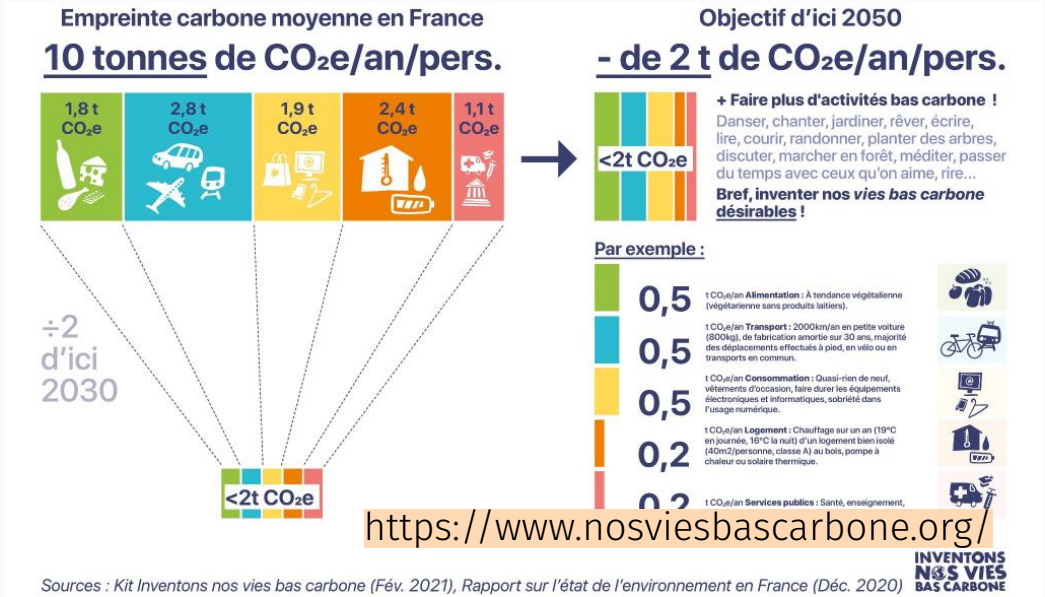


French government response

- *Verdissement de l'industrie: « pause » sur les normes environnementales*
- *Loi de programmation militaire (+41%)*
- *Nous devons préparer la France à une élévation de la température de 4 °C*
- Academia ? PEPR 5G, Cloud, NUMPEX, Quantique, IA, Agroécologie et numérique

Put aside biodiversity loss, pollution, freshwater, land system change...





French government response

- *Verdissement de l'industrie: « pause » sur les normes environnementales*
- *Loi de programmation militaire (+41%)*
- *Nous devons préparer la France à une élévation de la température de 4 °C*
- Academia ? PEPR 5G, Cloud, NUMPEX, Quantique, IA, Agroécologie et numérique

~~Several~~ scenarios on the table

- What will research/CS look like/be used for in such a world?
- Energy optimization/saving ≠ sobriety and frugality