

Reproducible Research, Open Science Motivation, Challenges, Approaches, . . .

Arnaud Legrand
CNRS, Inria, University of Grenoble

February 3, 2016 – Grenoble
Grid'5000 Winter School

Outline

① A Few Motivating Examples

② The Reproducible Research Movement

How does it work in "real" sciences?

Reproducible Research/Open Science

Illustrating Nice Ideas Through Different Tools

And In Practice ?

③ Where are we now?

Frustration



As an Author

- Advisor: "Did you take care of setting this?" Me: "Uh?"
- I thought I used the same parameters but I'm getting different results! I swear it **worked yesterday!**
- A new student wants to compare with the method I proposed last year
- 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 to generate this figure?
- 6 months later: **Why** did I do that?

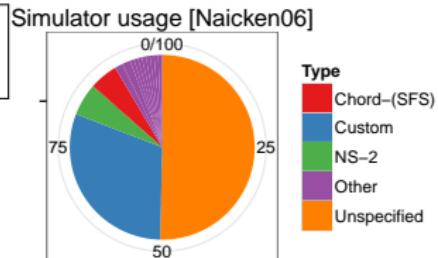
As a Reviewer This may be an interesting contribution but:

- There is no label/legend/... What is the **meaning of this graph**? If only I could access the generation script
- Why is this graph in **logscale**? How would it look like otherwise?
- This **average value** must hide something. As usual, no **confidence interval**... I wonder whether the difference is **significant** at all
- That can't be true, I'm sure they **removed some points** or decided to show only a **subset of the data**. I wonder what the rest looks like

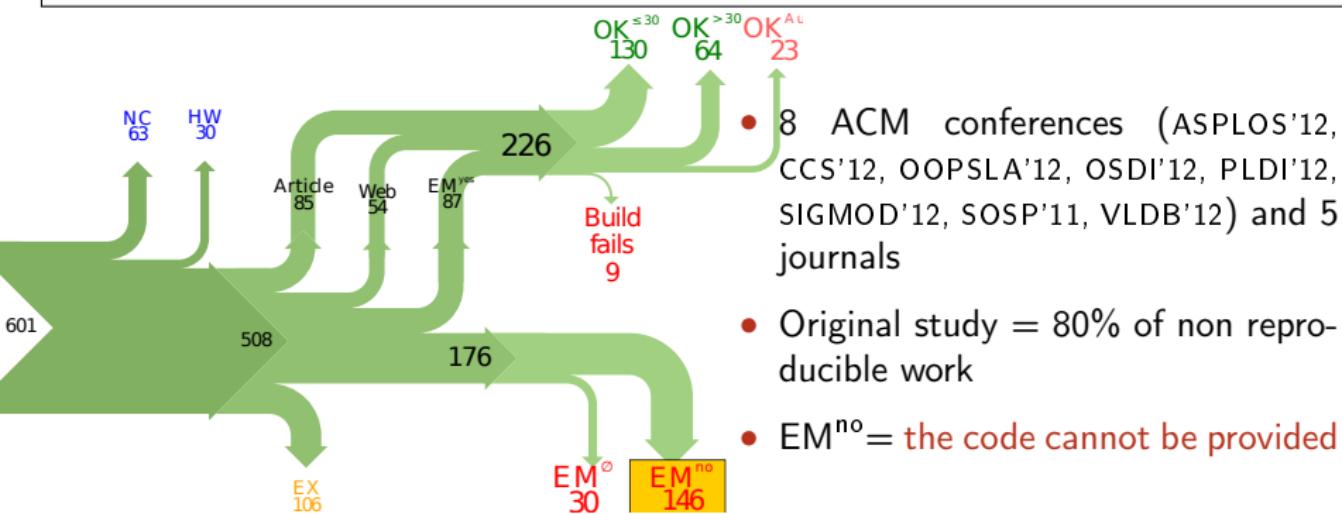
A Few Edifying Examples

Naicken, Stephen et Al., *Towards Yet Another Peer-to-Peer Simulator*, HET-NETs'06.

From 141 P2P sim.papers, 30% use a custom tool,
50% don't report used tool



Collberg, Christian et Al., *Measuring Reproducibility in Computer Systems Research*, <http://reproducibility.cs.arizona.edu/> 2014,2015



The Dog Ate my Homework !!!

- Versioning Problems

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

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

The Dog Ate my Homework !!!

- Versioning Problems
- Bad Backup Practices

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

The Dog Ate my Homework !!!

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

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

The Dog Ate my Homework !!!

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

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

The Dog Ate my Homework !!!

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

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

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

The Dog Ate my Homework !!!

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

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

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

The Dog Ate my Homework !!!

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

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

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

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

The Dog Ate my Homework !!!

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

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

My Feeling

Computer scientists have an incredibly poor training in probabilities, statistics, experiment management, Design of Experiments

Why should we? Computer are deterministic machines after all, right? 😊

Ten years ago, I've started realizing how lame the articles I reviewed (as well as those I wrote) were in term of experimental methodology.

- Yeah, I know, your method/algorithm is better than the others as demonstrated by the figures
- Not enough information to discriminate real effects from noise
- Little information about the workload. Would the “conclusion” still hold with a slightly different workload?
- I got tired of awful combination of tools (perl, gnuplot, sql, ...) and bad methodology

I got sick of struggling in vain when trying to build on the work of others

Outline

① A Few Motivating Examples

② The Reproducible Research Movement

How does it work in "real" sciences?

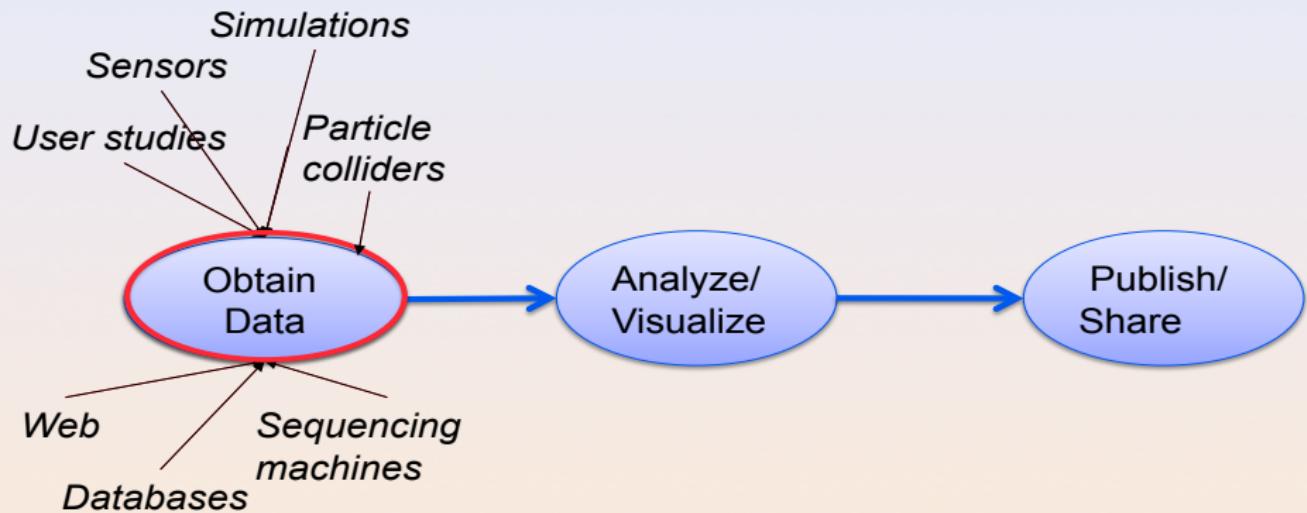
Reproducible Research/Open Science

Illustrating Nice Ideas Through Different Tools

And In Practice ?

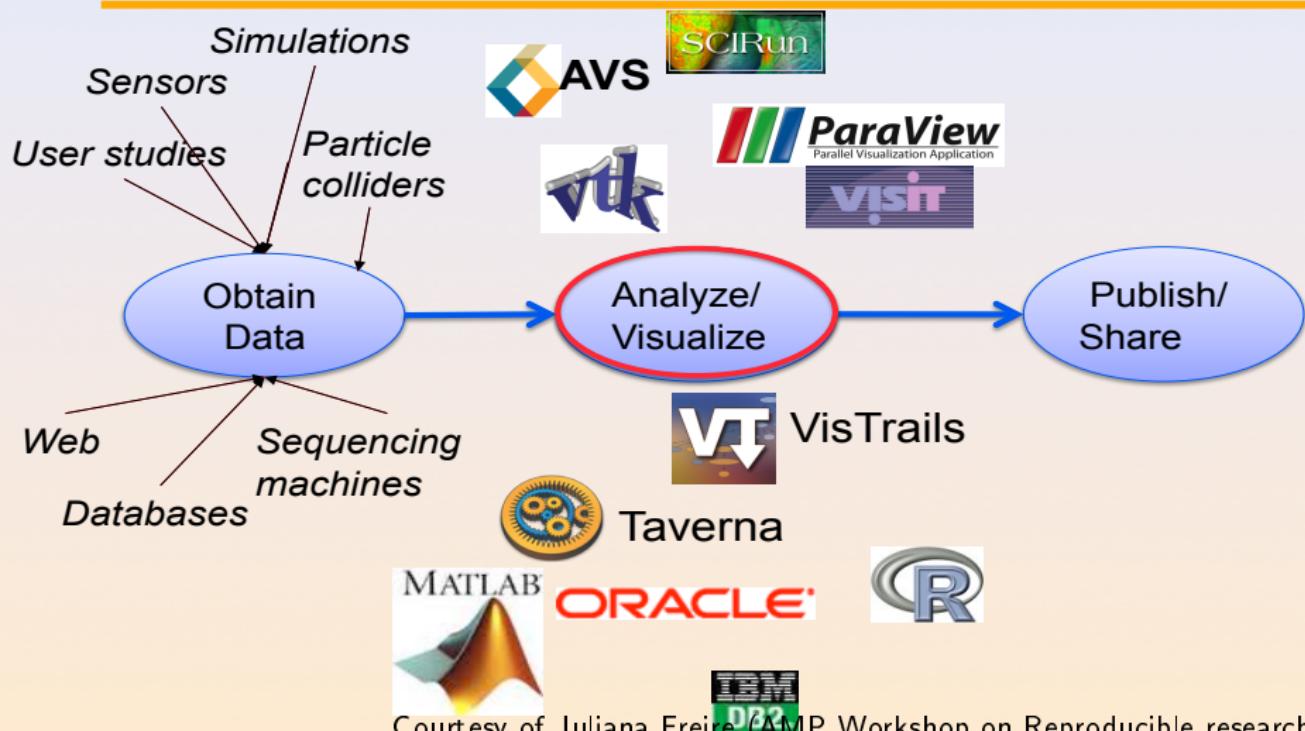
③ Where are we now?

Science Today: Data Intensive

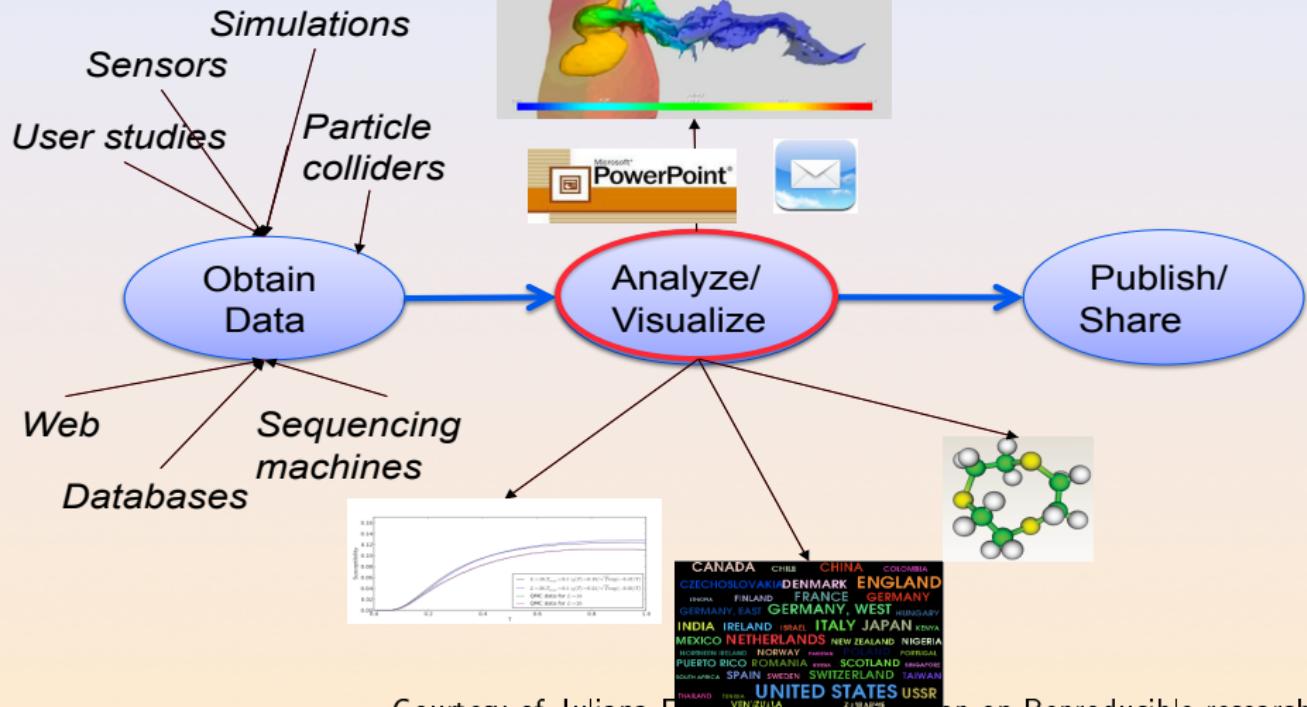


Courtesy of Juliana Freire (AMP Workshop on Reproducible research)

Science Today: Data + Computing Intensive

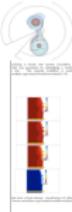
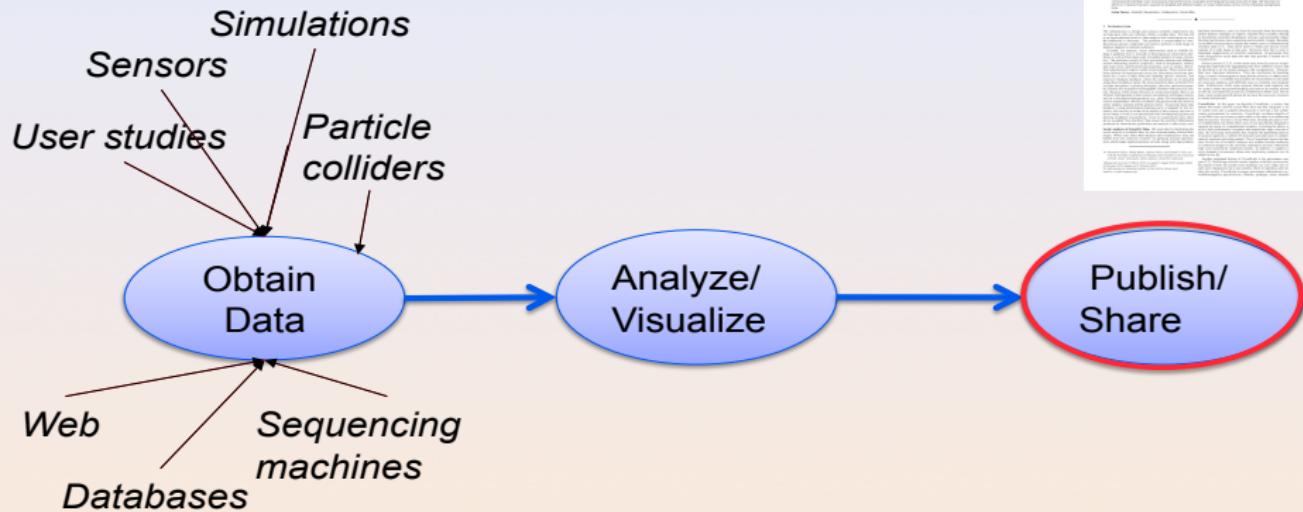


Science Today: Data + Computing Intensive



Courtesy of Juliana Freire (AVIRI Workshop on Reproducible research)

Science Today: Data + Computing Inte



Courtesy of Juliana Freire (AMP Workshop on Reproducible research)

Science Today: Incomplete Publications

- ◆ Publications are just the tip of the iceberg
 - Scientific record is incomplete---to large to fit in a paper
 - Large volumes of data
 - Complex processes
- ◆ Can't (easily) reproduce results



Courtesy of Juliana Freire (AMP Workshop on Reproducible research)

Science Today: Incomplete Publications

- ◆ Publications are just the tip of the iceberg

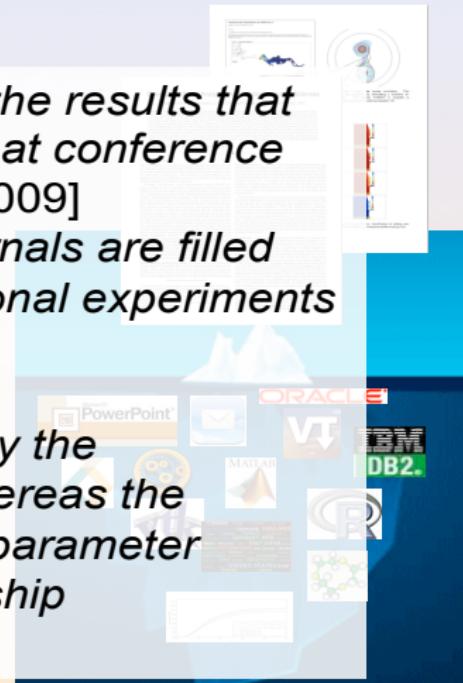
"It's impossible to verify most of the results that computational scientists present at conference and in papers." [Donoho et al., 2009]

-
-
-

"Scientific and mathematical journals are filled with pretty pictures of computational experiments"

- ◆ Can't that the reader has no hope of repeating." [LeVeque, 2009]

"Published documents are merely the advertisement of scholarship whereas the computer programs, input data, parameter values, etc. embody the scholarship itself." [Schwab et al., 2007]



Courtesy of Juliana Freire (AMP Workshop on Reproducible research)

Why Are Scientific Studies so Difficult to Reproduce?

Human error:

- Experimenter bias
- Programming errors or data manipulation mistakes
- Poorly selected statistical tests

There is just no real incentive in doing so:

- Legal barriers, copyright (*many ongoing thoughts on this in the US*)
- Competition issue (*researchware, bibliometry, . . .*)
- Publication bias (only the idea matters, not the gory details)
- Rewards for positive results, not for consolidating results

Technical difficulty:

- ~~Hardware and software evolve too quickly. It's not worth it~~
- ~~No resources for storing somuch data/information~~
- ~~Lack of easy-to-use tools~~

Evidence for a Lack of Reproducibility

- Studies showing that scientific papers commonly leave out experimental details essential for reproduction and showing difficulties with replicating published experimental results:
 - J.P. Ioannidis. *Why Most Published Research Findings Are False* PLoS Med. 2005 August; 2(8)
- High number of failing clinical trials.
 - *Do We Really Know What Makes Us Healthy?*, New-York Times — September 16, 2007
 - *Lies, Damned Lies, and Medical Science*, The Atlantic. Nov, 2010
- Increase in retracted papers:
 - Steen RG, Retractions in the scientific literature: is the incidence of research fraud increasing?
J Med Ethics 37: 249–253.

Courtesy V. Stodden, SC, 2015

The image displays three screenshots of scientific journals, each highlighting a different aspect of reproducibility issues:

- Los Angeles Times | BUSINESS**: A news article titled "Science has lost its way, at a big cost to humanity" discusses the decline in reproducibility in science.
- nature**: A screenshot of the journal's website with a banner at the top reading "Announcement: Reducing our irreproducibility". Below it, a news item states "Over the past year, Nature has published a string of articles that highlight the need to improve the reliability and reproducibility of published research. Colleagues in the field have responded by calling for changes to the way we do science."
- TheScientist**: A screenshot of the journal's website with a banner at the top reading "NIH Tackles Irreproducibility". Below it, a news item states "The federal agency speaks out about how to improve the quality of scientific research. The latest agency update speaks out about how to improve the quality of scientific research."

A Reproducibility Crisis?

The Duke University scandal with scientific misconduct on lung cancer

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

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

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

- Retractions: January 2011. Ten papers that Potti coauthored in prestigious journals were retracted for varying reasons
- Some people die and may be getting worthless information that is based on **bad science**

Definitely

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

- Falsified blood results to make it appear as though a vaccine he was working on had exhibited anti-HIV activity
- Han and his team received $\approx \$19$ million from NIH
- Retraction and resignation of university
- Han was sentenced in 2015 to 57 months imprisonment for fabricating and falsifying data in HIV vaccine trials. He was also fined US \$7.2 million!

We should avoid witch-hunt

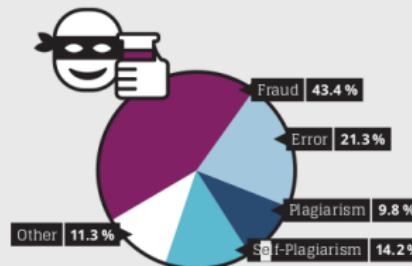
- August 5, 2014, Yoshiki Sasai (stem cell, considered for Nobel Prize) hanged in his laboratory at the RIKEN (Japan). Fraud suspicion...
- In 1986, a young postdoctoral fellow at MIT accused her director, Thereza Imanishi-Kari, of falsifying the results of a study published in Cell and co-signed by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

Scientific fraud is bad but let's be careful Have a look at the wikipedia *list of academic scandals*. On a totally different aspect, do not forget to also have a look at the *plagiarism* and *paper generation* wikipedia entries and at *having fun with h-index*

Is Fraud a new phenomenon?

Biomedical fraud in figures

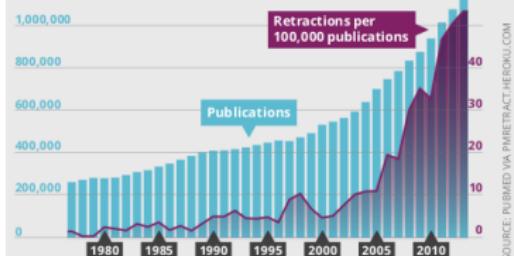
Cause of retraction 1977 to 2012



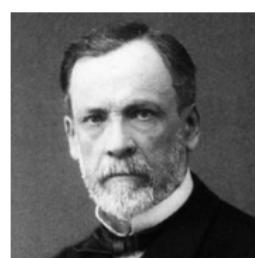
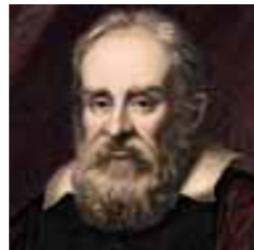
SOURCE: FANG ET AL. (2012) PNAS

Number of publications and retractions

1977 to 2013



SOURCE: PUBLISHED VIA PIRETRACT HEROKU.COM



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

Outline

① A Few Motivating Examples

② The Reproducible Research Movement

How does it work in "real" sciences?

Reproducible Research/Open Science

Illustrating Nice Ideas Through Different Tools

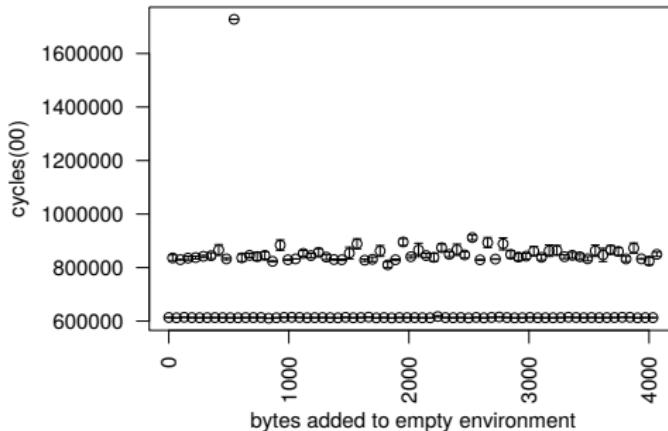
And In Practice ?

③ Where are we now?

But do we **really** have to care in CS?

Yes, although designed and built by human beings, computers are **so complex** that mistakes are easy to do...

- T. Mytkowicz, A. Diwan, M. Hauswirth, and P. F. Sweeney. **Producing wrong data without doing anything obviously wrong!**. SIGPLAN Not. 44(3), March 2009



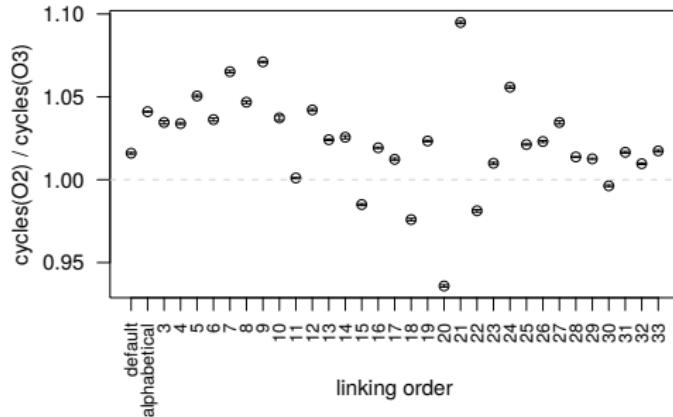
Key principles of experiment design

- Randomize to reduce bias
- Replicate (possibly in a smart way) to increase reliability
- Takes a few lectures on Design of Experiments to improve. Start by reading Jain's book on The Art of Computer Systems Performance Analysis

But do we **really** have to care in CS?

Yes, although designed and built by human beings, computers are **so complex** that mistakes are easy to do...

- T. Mytkowicz, A. Diwan, M. Hauswirth, and P. F. Sweeney. **Producing wrong data without doing anything obviously wrong!**. SIGPLAN Not. 44(3), March 2009



Key principles of experiment design

- Randomize to reduce bias
- Replicate (possibly in a smart way) to increase reliability
- Takes a few lectures on Design of Experiments to improve. Start by reading Jain's book on The Art of Computer Systems Performance Analysis

But do we **really** have to care in CS?

Yes, although designed and built by human beings, computers are **so complex** that mistakes are easy to do...

- T. Mytkowicz, A. Diwan, M. Hauswirth, and P. F. Sweeney. **Producing wrong data without doing anything obviously wrong!**. SIGPLAN Not. 44(3), March 2009

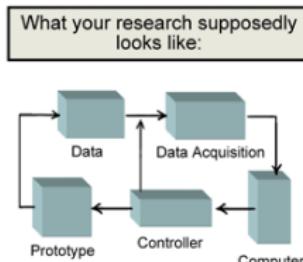


Figure 1. Experimental Diagram



Figure 2. Experimental Mess

- Rely on large, distributed, hybrid, prototype hardware/software
- Many parameters, very costly and hard to **reproduce**

Key principles of experiment design

- Randomize to **reduce bias**
- Replicate (possibly in a smart way) to **increase reliability**
- Takes a few lectures on **Design of Experiments** to improve. Start by reading Jain's book on **The Art of Computer Systems Performance Analysis**

Reproducible Research: the New Buzzword?

H2020-EINFRA-2014-2015

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

More and more workshops

- Workshop on Duplicating, Deconstructing and Debunking (WDDD) (2002-2014 edition)
- AMP Workshop. Reproducible Research: Tools and Strategies for Scientific Computing (2011)
- Working towards Sustainable Software for Science: Practice and Experiences (2013)
- REPPAR'14: 1st International Workshop on Reproducibility in Parallel Computing
- Reproducibility@XSEDE: An XSEDE14 Workshop
- Reproduce/HPCA 2014
- TRUST 2014, 2015
- Talk at SC by V. Stodden two weeks ago

Should be seen as opportunities to share experience

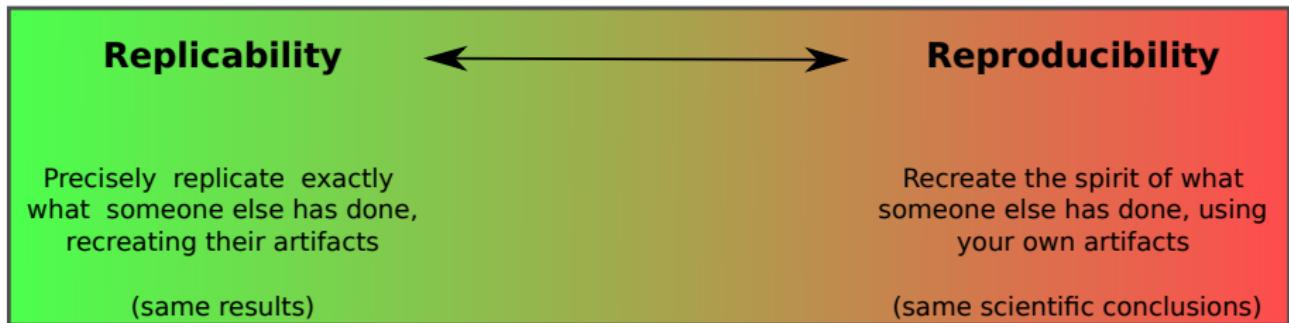
Reproducibility: What Are We Talking About?

1934: Karl Popper introduces the notion of **falsifiability** and **crucial experiment** and puts **reproducing the work of others** at the core of science

Reproducibility of experimental results is the hallmark of science

[Drummond, 2009]

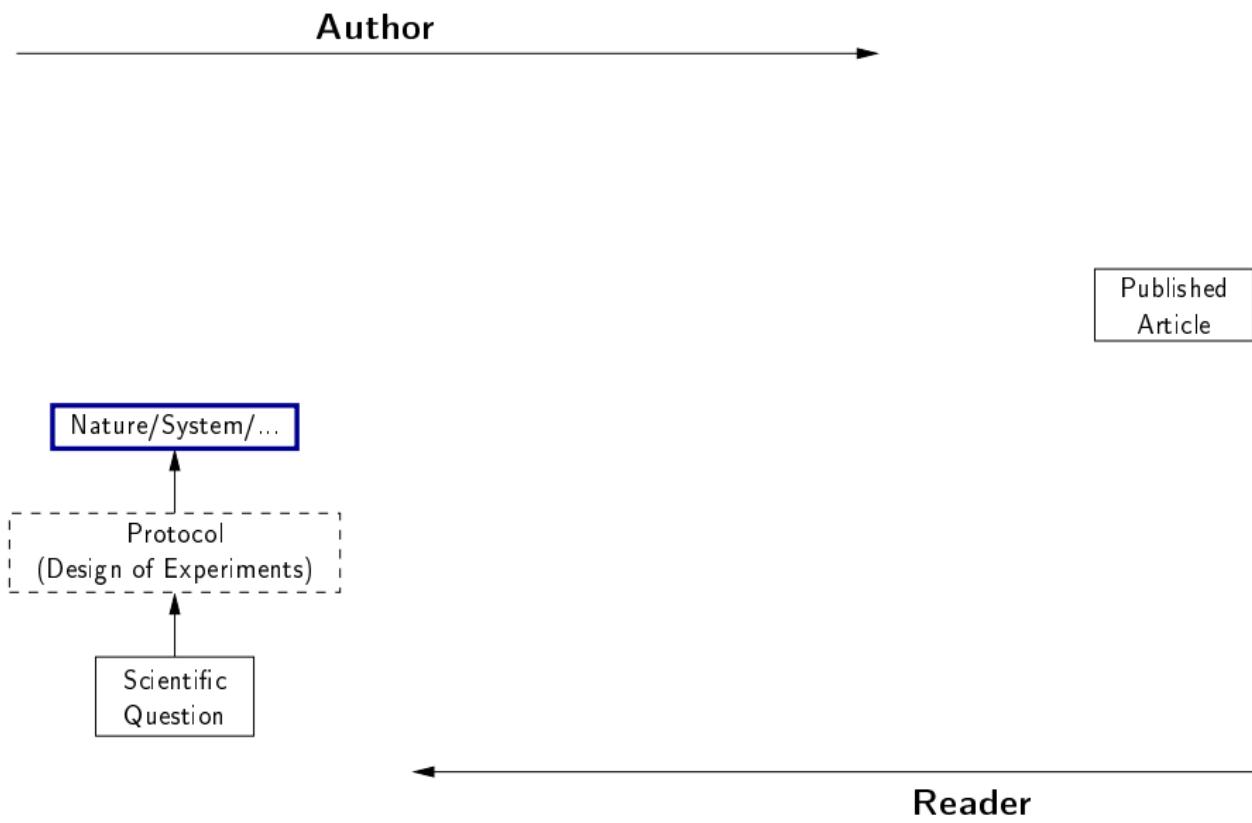
⚠ Terminology varies ⚠



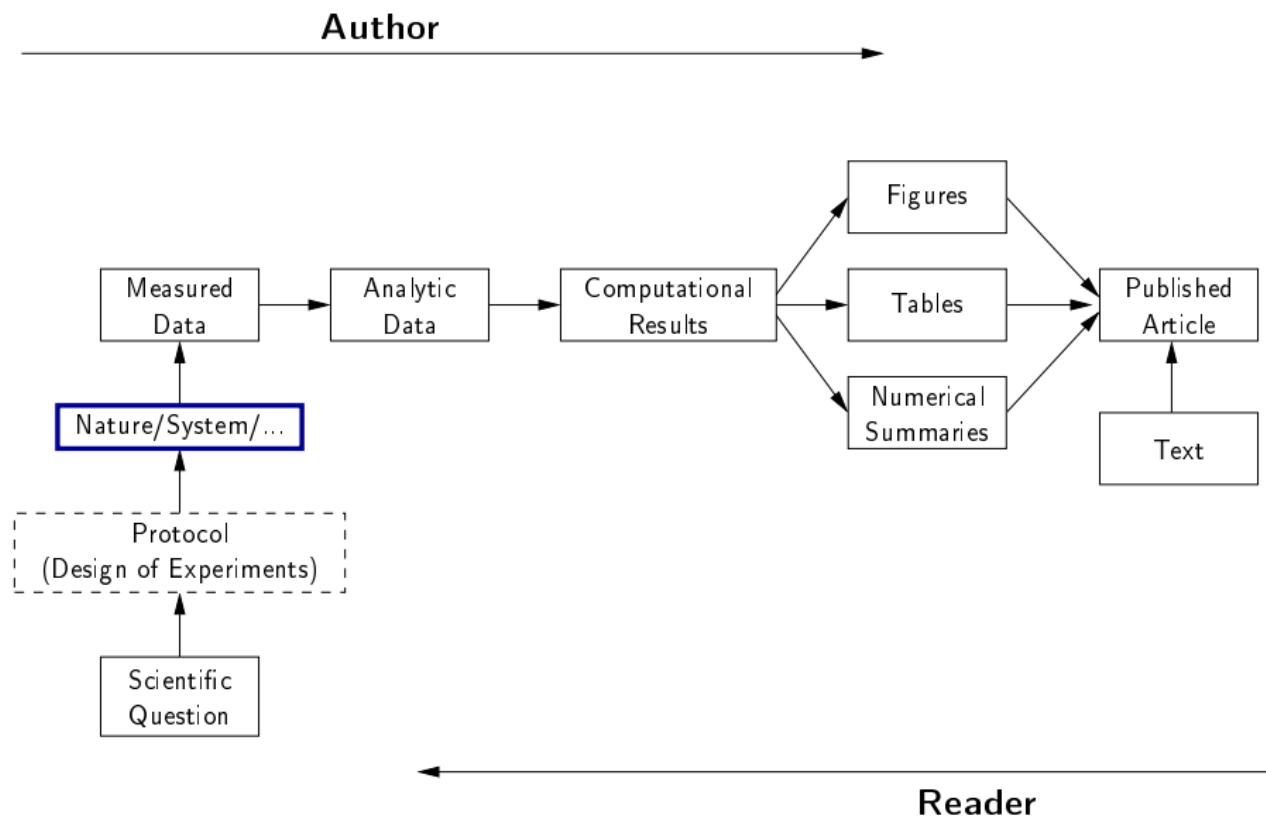
Inspired by Andrew Davison (AMP Workshop on Reproducible research) and [Feitelson, 2015]

Further readings: **A Summary of Scientific Method**, Peter Kosso, Springer

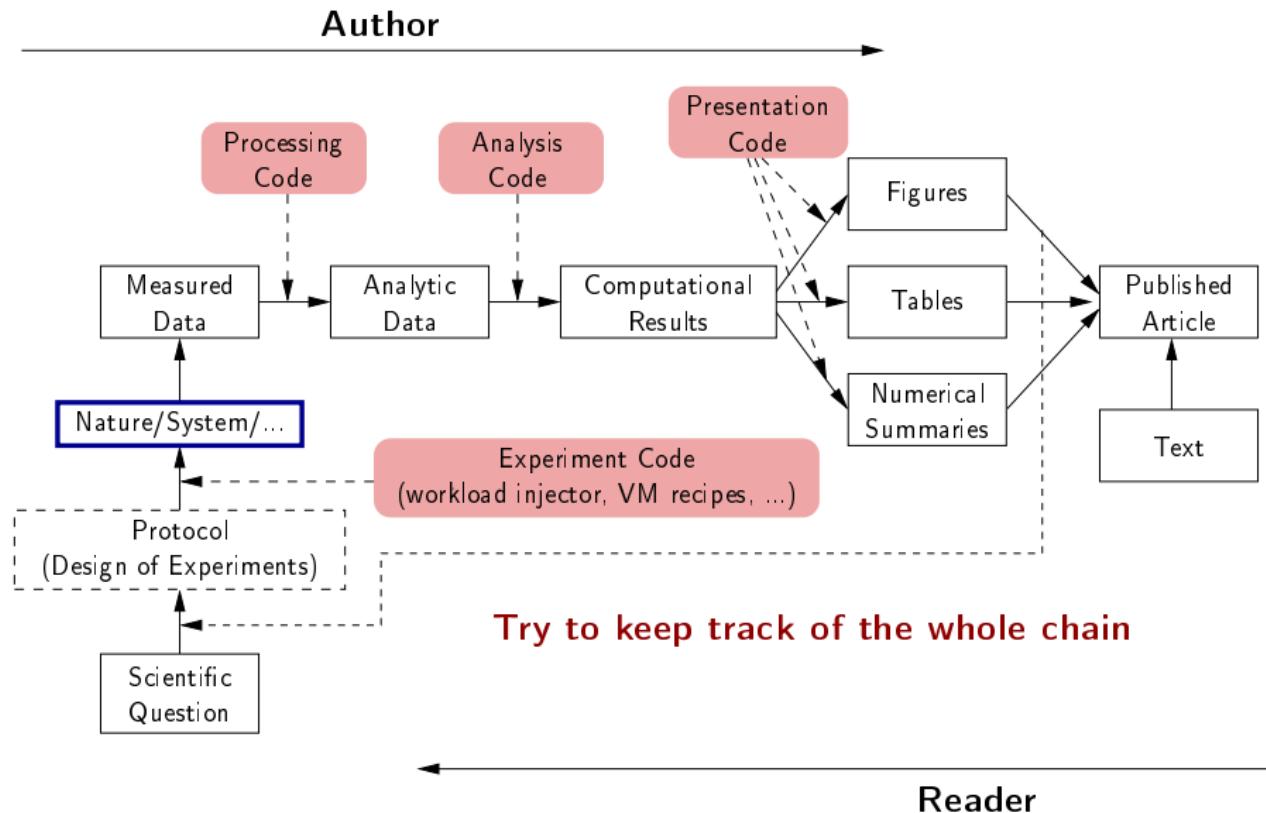
Reproducible Research: Trying to Bridge the Gap



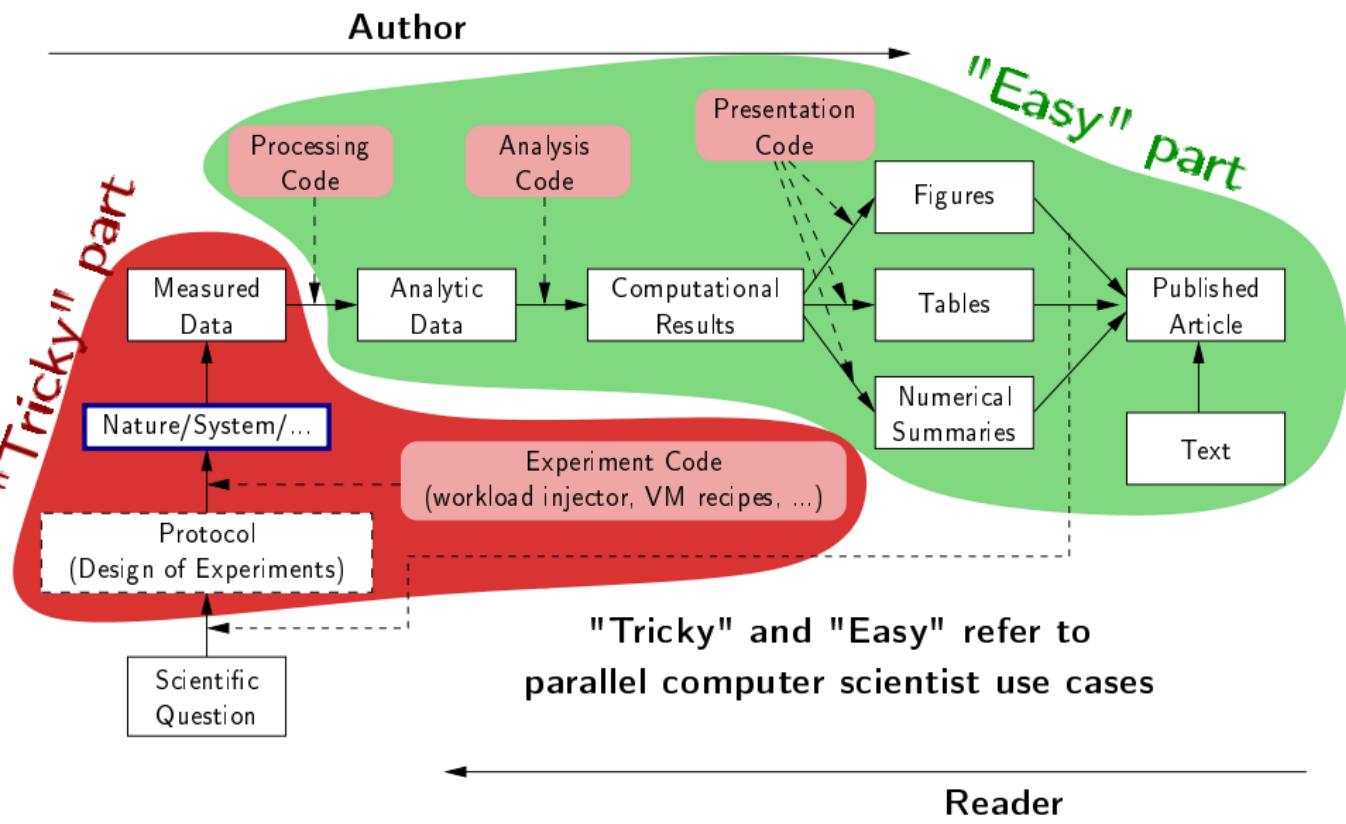
Reproducible Research: Trying to Bridge the Gap



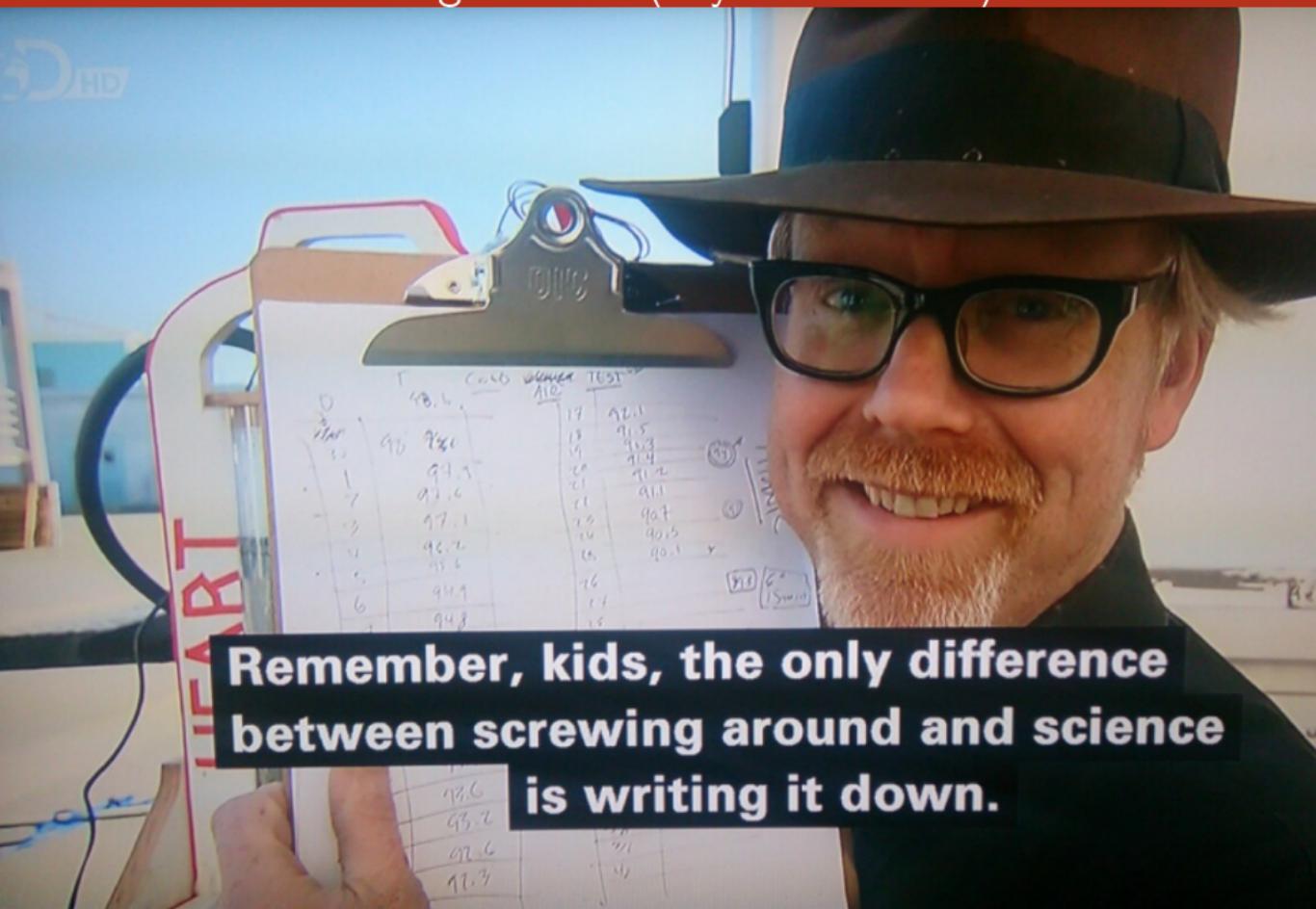
Reproducible Research: Trying to Bridge the Gap



Reproducible Research: Trying to Bridge the Gap



Science vs. Screwing Around (Mythbusters 😊)



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

Outline

① A Few Motivating Examples

② The Reproducible Research Movement

How does it work in "real" sciences?

Reproducible Research/Open Science

Illustrating Nice Ideas Through Different Tools

And In Practice ?

③ Where are we now?

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

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

Support different approaches

Vistrails: a Workflow Engine for Provenance Tracking

An Provenance-Rich Paper: ALPS2.0

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

B. Bauer¹ L. D. Carr² H.G. Evertz³ A. Feiguin⁴ J. Freire⁵
S. Fuchs⁶ L. Gamper¹ J. Gukelberger⁶ E. Gulf⁷ S. Guertler⁸
A. Hehn⁹ R. Igashiri¹⁰ S.V. Isakov¹ D. Koop² P.N. Ma¹¹
P. Mates^{1,2} H. Matsuo¹¹ O. Parcollet¹² G. Pawłowski¹³
J.D. Picon¹⁴ L. Pollet¹⁵ E. Santos¹⁶ V.W. Scarola¹⁶
U. Schollwöck¹⁷ C. Silva¹⁸ B. Surer¹⁹ S. Todo^{11,20} S. Trebst¹⁶
M. Troyer¹¹ M. L. Wall²¹ P. Werner¹ S. Wessel^{1,20}

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

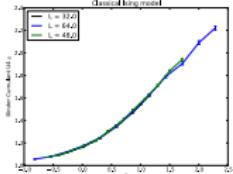
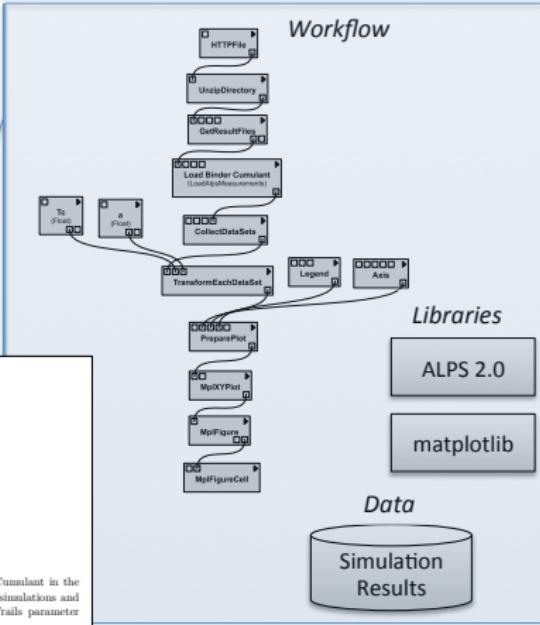


Figure 3 shows a plot of the relative cumulant versus temperature T for the classical Ising model. The x-axis ranges from -0.5 to 0.5, and the y-axis ranges from 1.0 to 2.0. Three data series are plotted for system sizes L = 32, 64, and 128. The L = 32 curve (red) shows significant deviation from the others at low temperatures. As the system size increases, the curves converge towards a single blue line, which represents the critical behavior of the model.



The diagram illustrates the workflow for generating a plot. It starts with 'Simulation Results' (represented by a cylinder) which feeds into 'ALPS 2.0' (represented by a rectangle). 'ALPS 2.0' then feeds into 'matplotlib' (also represented by a rectangle). Finally, the data is plotted using 'matplotlib' (represented by a scatter plot icon).

```
graph TD; SR((Simulation Results)) --> ALPS[ALPS 2.0]; ALPS --> Matplotlib[matplotlib]; Matplotlib --> Plot[Plot]
```

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

¹ Correspondence

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

VCR: A Universal Identifier for Computational Results

Chronicling computations in real-time

VCR computation platform Plugin = Computation recorder

Regular program code

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

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

VCR: A Universal Identifier for Computational Results

Chronicling computations in real-time

VCR computation platform Plugin = Computation recorder

Program code with VCR plugin

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

```
> vcr.nature.com approved:
```

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

VCR: A Universal Identifier for Computational Results

Word-processor plugin App

LaTeX source

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

LaTeX source with VCR package

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

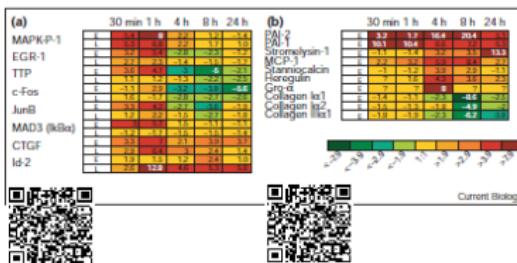
Permanently bind printed graphics to underlying result content

VCR: A Universal Identifier for Computational Results

Research Paper Analysis of replicative senescence Shelton et al. 943

Figure 3

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



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

Deficiencies in the response of senescent cells to serum stimulation have been reported, and include an inability to induce the expression of *c-fos* mRNA [31] and markers of late G1 and S phase [32]. In response to serum, expression of inflammatory chemokines, matrix-degrading proteases and their modulators is induced in early-passage dermal fibroblasts, and expression of matrix collagens is reduced. This transient burst of activity may represent a natural progression of events in early-passage cells. Transcripts were hyper-induced in serum-stimulated senescent cells, and included markers of the immune system, such as

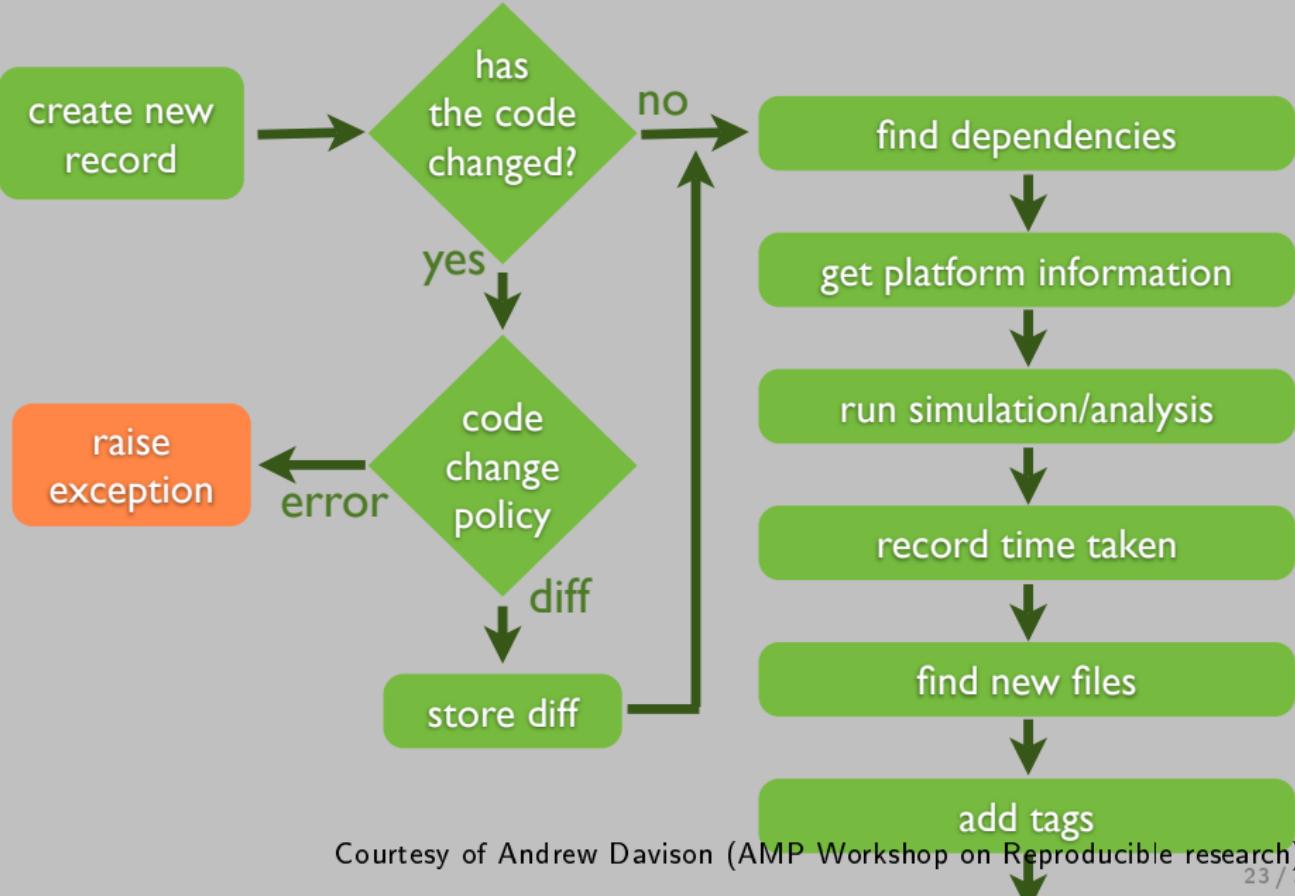
states overlap substantially with those in telomere-induced senescence (W.F., D.N.S., R. Allsopp, S. Lowe, and G. Ferbeyre, unpublished observations) and thus are likely to use many of the same activation processes.

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

surprisingly, early-passage cells also overexpress many of the markers associated with senescence in dermal fibroblasts, such as *PAI-2*, *Cathepsin-1*, *MCP-1*, *Gro- α* , *IL-15*, *Collagen I α 1*, *Collagen III α 1*, and *Collagen IV α 1*.

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

Sumatra: an "experiment engine" that helps taking notes



Courtesy of Andrew Davison (AMP Workshop on Reproducible research)

Sumatra: an "experiment engine" that helps taking notes

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

Sumatra: an "experiment engine" that helps taking notes

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

Sumatra: an "experiment engine" that helps taking notes

Sumatra: TestProject: List of records

TestProject: List of records

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

So many new tools

New Tools for Computational Reproducibility

- Dissemination Platforms:

[ResearchCompendia.org](#)

[IPOL](#)

[Madagascar](#)

[MLOSS.org](#)

[thedatahub.org](#)

[nanoHUB.org](#)

[Open Science Framework](#)

[The DataVerse Network](#)

[RunMyCode.org](#)

- Workflow Tracking and Research Environments:

[VisTrails](#)

[Kepler](#)

[CDE](#)

[Galaxy](#)

[GenePattern](#)

[Synapse](#)

[Sumatra](#)

[Taverna](#)

[Pegasus](#)

- Embedded Publishing: Courtesy of Victoria Stodden (UC Davis, Feb 13, 2014)

[Verifiable Computational Research](#) [Sweave](#) [knitR](#)

[Collage Authoring Environment](#) [SHARE](#)

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

Outline

① A Few Motivating Examples

② The Reproducible Research Movement

How does it work in "real" sciences?

Reproducible Research/Open Science

Illustrating Nice Ideas Through Different Tools

And In Practice ?

③ Where are we now?

A Difficult Trade-off

Many different tools/approaches developed in various communities

But mainly two approaches:

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

A Difficult Trade-off

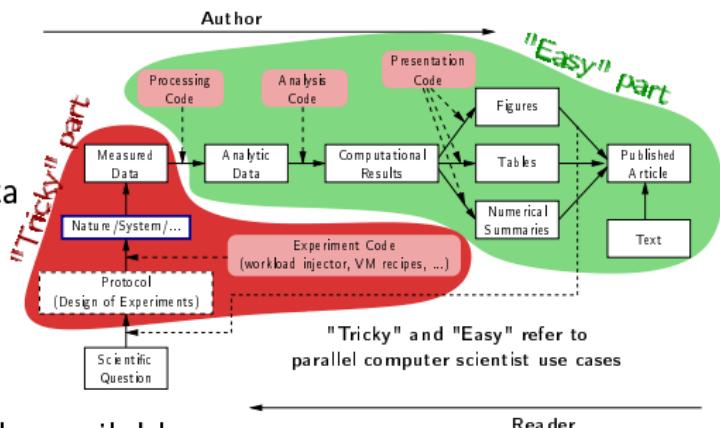
Many different tools/approaches developed in various communities

But mainly two approaches:

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

And the following key points:

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



1. Replicable article (Literate programming)

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

I'm way too 3133t to program this way 😊 but that's exactly what we need for writing a reproducible article/analysis!

Org-mode (my favorite! requires emacs though)

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

Ipython/Jupyter notebook

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

Knitr (a.k.a. Sweave)

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

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

A replicable article with Org-Mode

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

Here are the following important features to exploit:

Structure highly hierarchical

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

Export in several output formats

- Fine control with #+BEGIN_LaTeX
- Unfortunate need for verbose headers (because of LATEX \frown) and black magic in the end of the file (for emacs portability \frown)

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

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

2. Logging your activity (Laboratory Notebook)

Pioneered by Mathematica (TBOMK)

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

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

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

3. Logging and backup your data

What are the options?

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

4. Organizing your data

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

5. Mastering your environment

What are the options?

- Nothing 😊
- Restrict your tools/dependencies to the bare minimum (e.g., python)
 - List them all manually in a README
 - Use custom shell scripts or sosreport that log all the dependencies you are aware. Ask your friends to check whether this is sufficient...
 - Combine everything in activepapers, i.e., an HDFS5 file combining datasets and programs working on these datasets in a single package, along with meta data, history, ...
- Create and distribute your own virtual image (VM, docker, Singularity)
- Have tools that automatically keep track of dependencies/files and packages up the Code, Data, and Environment
 - CDE (Guo et al., 2011) ReproZip (Freire et al., 2013), CARE (Janin et al., 2014),
 - See Preserve the Mess or Encourage Cleanliness? (Thain et al., 2015)
- Use a specific tool to generate customized appliances (kvm, LXC, Virtualbox, iso, ...): recipes with steps and aliases, execution in contexts, checkpoints, ... (*Kameleon*)

6. Controlling your experiments

- Naive way: sh + ssh + ...

Parallel/distributed experiments differ from computational science and come with their own difficulties

- **Expo** (2007-, G5K)
 - **XPflow** (2012-, G5K)
 - **Execo** (2013-, G5K)
- } although nothing specific to G5K

- Plush (2006-, PlanetLab)
- OMF (2009-, Wireless testbeds and Planetlab)
- Splay (2008, distributed algorithm comparison), ...

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

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

7. Making your data/code/article available

- Your webpage 😞
- Figshare, Zenodo 😊, ...
- Companion websites ([elsevier executable paper](#) 😞, [runmycode](#), [exec&share](#) 😊, ...)
- Inria Forge/Gitlab, Github (damn, they're good! 😊), ...

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

- Arbitrary limits can make your life painful
- Perennity ([Roberto Di Cosmo's talk at R⁴](#))
 - CodeSpaces murdered on Amazon, Google Code termination, Gitorious shutdown, ...
 - Disruption of the web of reference: URLs decay (half-life of 4 years), DOIs have little guarantee, ...

Outline

① A Few Motivating Examples

② The Reproducible Research Movement

How does it work in "real" sciences?

Reproducible Research/Open Science

Illustrating Nice Ideas Through Different Tools

And In Practice ?

③ Where are we now?

What is needed?

- Many **legal aspects** about data/code/idea sharing
 - I do not really care as I am a civil servant and I strongly believe research is a team sport
 - I am naive. This is an important topic we do not want to leave to bureaucrats and lawyers...
- Changes in **funding agency** requirements
 - Starting ? But I hardly see how they could really enforce things
- Changes in journal/conferences **publication requirements**
 - Several attempts (reproducibility labels)
 - V. Stodden seems confident (progressive policies rapidly adopted, journals with high impact factors)
- **Cultural changes** in our **relation to publication**

What is needed?

- Many legal aspects about data/code/idea sharing
 - I do not really care as I am a civil servant and I strongly believe research is a team sport
 - I am naive. This is an important topic we do not want to leave to bureaucrats and lawyers...
- Changes in funding agency requirements
 - Starting ? But I hardly see how they could really enforce things
- Changes in journal/conferences publication requirements
 - Several attempts (reproducibility labels)
 - V. Stodden seems confident (progressive policies rapidly adopted, journals with high impact factors)
- Cultural changes in our relation to publication

I think the change has to be profound and cannot be top-down

- Train our researchers and students to use better tools, better research methodology, Statistics/Design of Experiments, performance evaluation, ...
- Many French researchers (not only computer scientists) have started acting. Inria asked me to animate/coordinate such activity and open it way beyond Inria so that it is effective at national scale

Reproducible research actions

Webinars (1/month ?) with interactions, hands on keyboards when relevant.

- ① Reproducible research, challenges, ethic
- ② Numerical reproducibility
- ③ Provenance tracking of experimental data
- ④ Large scale experimental platforms
- ⑤ Code and Data archiving
- ⑥ Workflows
- ⑦ Online journals, companion websites
- ⑧ Evaluation campaign/challenges/benchmarks
- ⑨ Environment archiving (docker, VM, ...)
- ⑩ ...

We intend to start on March 7, 2016