

# Reproducible Research for Computer Scientists

Arnaud Legrand and Jean-Marc Vincent

Scientific Methodology and Performance Evaluation  
M2R MOSIG, Grenoble, September-December 2016

## ① The Reproducible Research Movement

How does it work in other sciences?

Is CS Concerned Really With This?

Reproducible Research/Open Science

Many Different Alternatives for Replicable Analysis

## ② Reporting Results

An IMRaD Report

Good Practice for Setting up a Laboratory Notebook

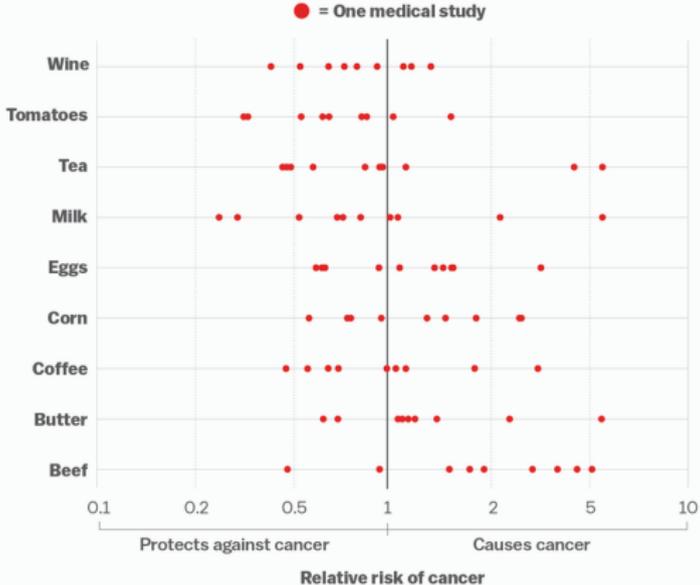
## ③ To do for the Next Time

## Inconsistencies

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

# Inconsistencies

## Everything we eat both causes and prevents cancer



SOURCE: Schoenfeld and Ioannidis, *American Journal of Clinical Nutrition*





# Public evidence for a Lack of Reproducibility

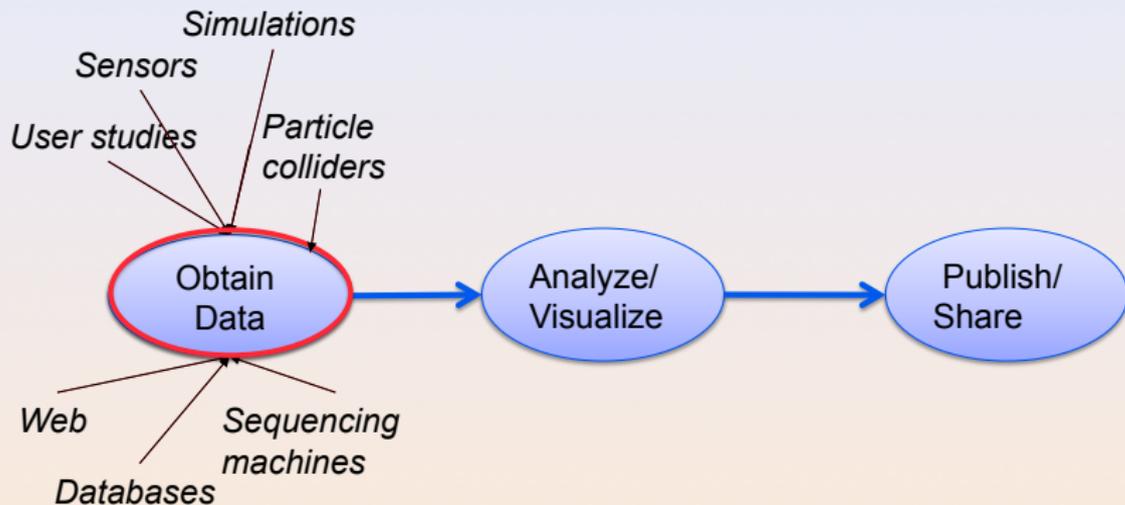
- J.P. Ioannidis. *Why Most Published Research Findings Are False* PLoS Med. 2005.
- *Lies, Damned Lies, and Medical Science*, The Atlantic, Nov, 2010



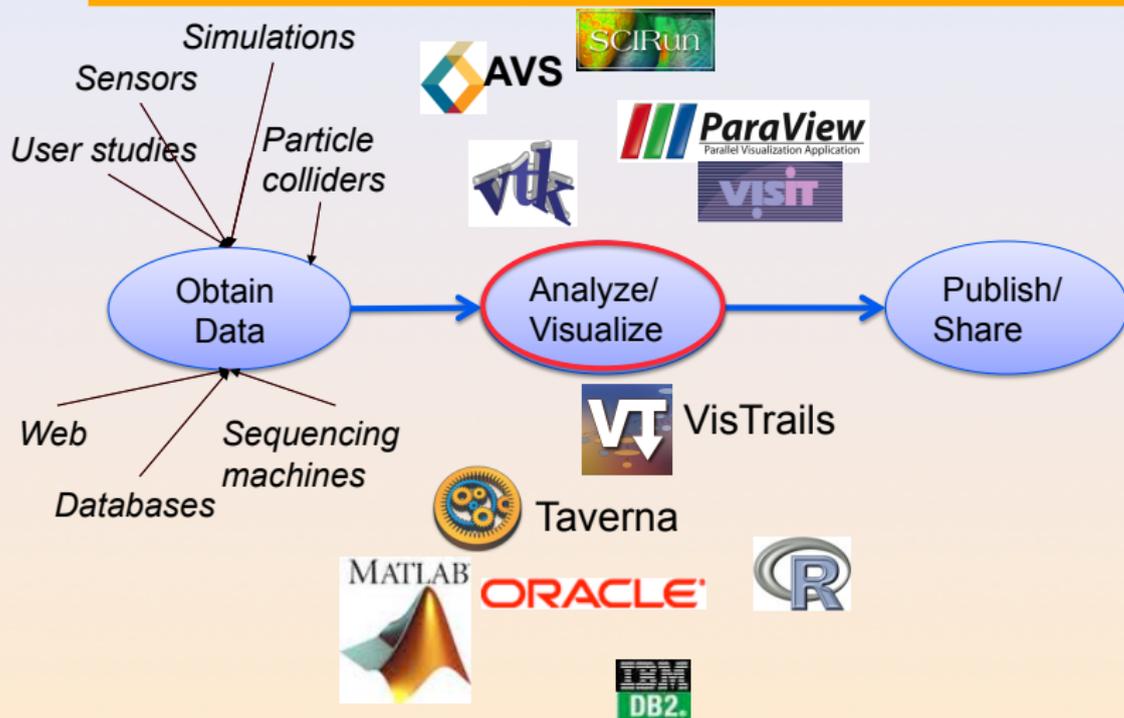
Courtesy V. Stodden, SC, 2015

Last Week Tonight with John Oliver:

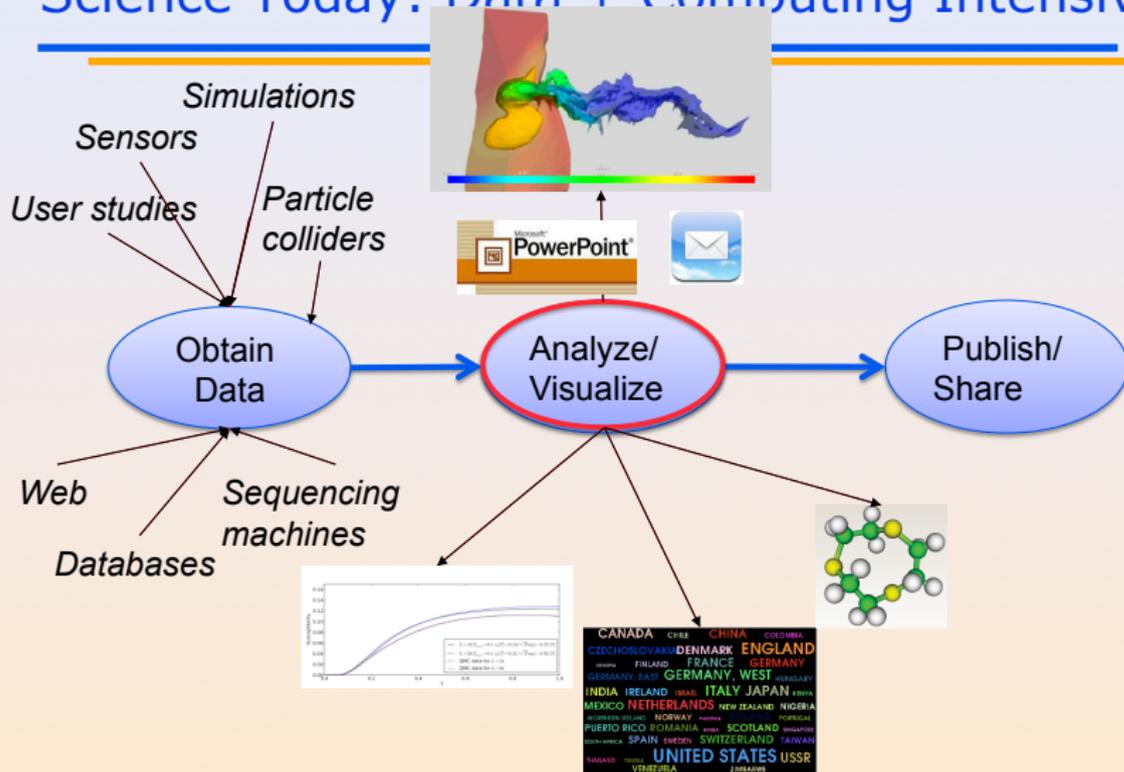
# Science Today: Data Intensive



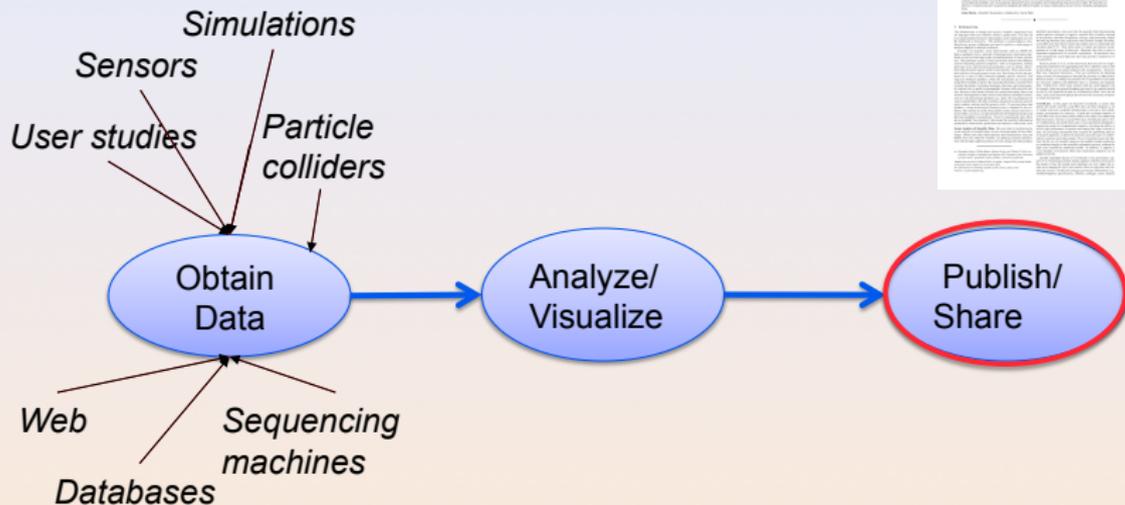
# Science Today: Data + Computing Intensive



# Science Today: Data + Computing Intensive



# Science Today: Data + Computing Inte



# 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



# 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 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]



## A few Words on Scientific Foundation

- **Falsifiability** or **refutability** of a statement, hypothesis, or theory is an inherent possibility to prove it to be false (not "*commit fraud*" but "*prove to be false*").
- Karl Popper makes falsifiability the demarcation criterion to **distinguish the scientific from the unscientific**  
*It is not only not right, it is not even wrong!*  
– Wolfgang Pauli
- Theories cannot be proved correct but they can be disproved. Only a few stand the test of batteries of **critical experiments**.
- It is not all black and white. There are many stories where scientists stick with their theories despite evidences and sometimes, they were even right to do so...

**Testing and checking is thus one of the basis of science**

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

# Why Are Scientific Studies so Difficult to Reproduce?

- Copyright/competition issue
- Publication **bias** (only the idea matters, not the gory details)
- Rewards for **positive results**
- Experimenter **bias**
- Programming **errors** or data manipulation **mistakes**
- Poorly selected statistical tests
- Multiple testing, multiple looks at the data, multiple statistical analyses
- ~~Lack of easy-to-use tools~~

# 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**

Courtesy of Adam J. Richards

# 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.

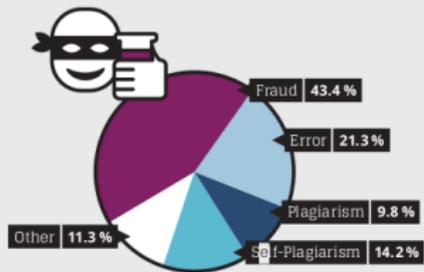
**Scientific fraud is bad but let's be careful** Have a look at the wikipedia *list of academic scandals*. On a totally different aspect, do not forget to also have a look at the *plagiarism* and *paper generation* entries at *having fun with h-index*

*The Battle against Scientific Fraud* in the CNRS International Magazine

# Is Fraud a new phenomenon?

## Biomedical fraud in figures

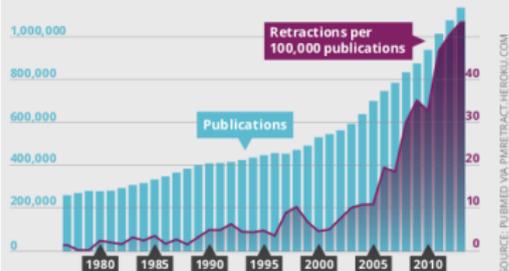
Cause of retraction 1977 to 2012



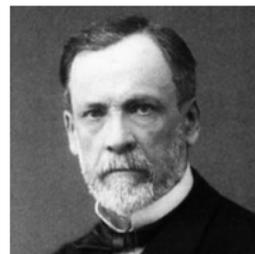
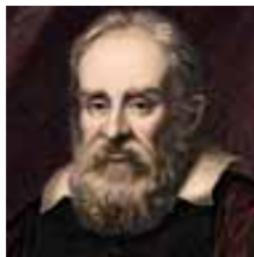
SOURCE: FANG ET AL. (2012) PNAS

Number of publications and retractions

1977 to 2013



SOURCE: PUBMED VIA PABRETRACT.HEROJU.COM



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

## ① The Reproducible Research Movement

How does it work in other sciences?

Is CS Concerned Really With This?

Reproducible Research/Open Science

Many Different Alternatives for Replicable Analysis

## ② Reporting Results

An IMRaD Report

Good Practice for Setting up a Laboratory Notebook

## ③ To do for the Next Time

# All this is about Natural Sciences. Should we care ?

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

**Model  $\neq$  Reality.**

# All this is about Natural Sciences. Should we care ?

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

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

- **Experiments:** Mytkowicz, Diwan, Hauswirth, Sweeney. **Producing wrong data without doing anything obviously wrong!**. SIGPLAN Not. 44(3), March 2009

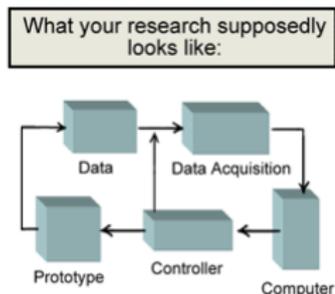


Figure 1. Experimental Diagram

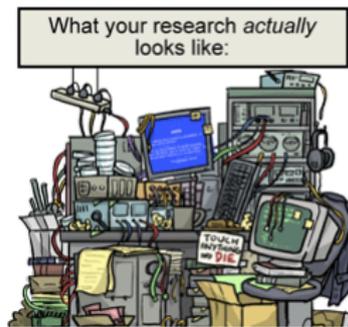


Figure 2. Experimental Mess

WWW.PHDCOMICS.COM JORGE CHAN © 2008

# All this is about Natural Sciences. Should we care ?

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

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

- **Experiments**: Mytkowicz, Diwan, Hauswirth, Sweeney. **Producing wrong data without doing anything obviously wrong!**. SIGPLAN Not. 44(3), March 2009
- **Statistics**: **Trouble at the lab**, The Economist 2013

*According to some estimates, three-quarters of published scientific papers in the field of machine learning are bunk because of this "overfitting".*

*Sandy Pentland, MIT*



# All this is about Natural Sciences. Should we care ?

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

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

- **Experiments:** Mytkowicz, Diwan, Hauswirth, Sweeney. **Producing wrong data without doing anything obviously wrong!**. SIGPLAN Not. 44(3), March 2009

- **Statistics:** **Trouble at the lab**, The Economist 2013

*According to some estimates, three-quarters of published scientific papers in the field of machine learning are bunk because of this "overfitting".*  
Sandy Pentland, MIT

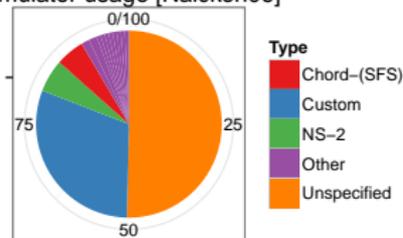
- **Numerical reproducibility:** change compiler, OS, machine and see what happens. **Ever tried to exploit a parallel architecture ?** 😊

# A Few Edifying Examples

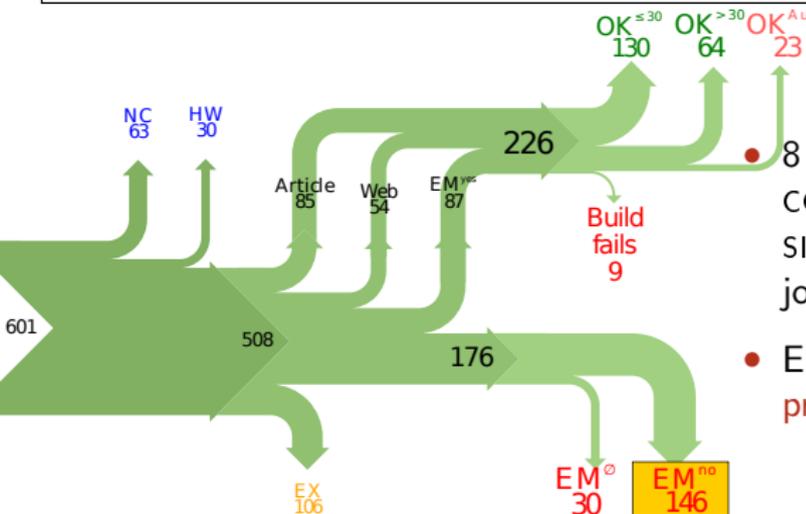
Naicken, Stephen *et Al.*, *Towards Yet Another Peer-to-Peer Simulator*, HET-NETs'06.

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

Simulator usage [Naicken06]



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



• 8 ACM conferences (ASPLOS'12, CCS'12, OOPSLA'12, OSDI'12, PLDI'12, SIGMOD'12, SOSP'11, VLDB'12) and 5 journals

• EM<sup>no</sup> = the code cannot be provided

# 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  $\langle$ system $\rangle$  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.*

## ① The Reproducible Research Movement

How does it work in other sciences?

Is CS Concerned Really With This?

Reproducible Research/Open Science

Many Different Alternatives for Replicable Analysis

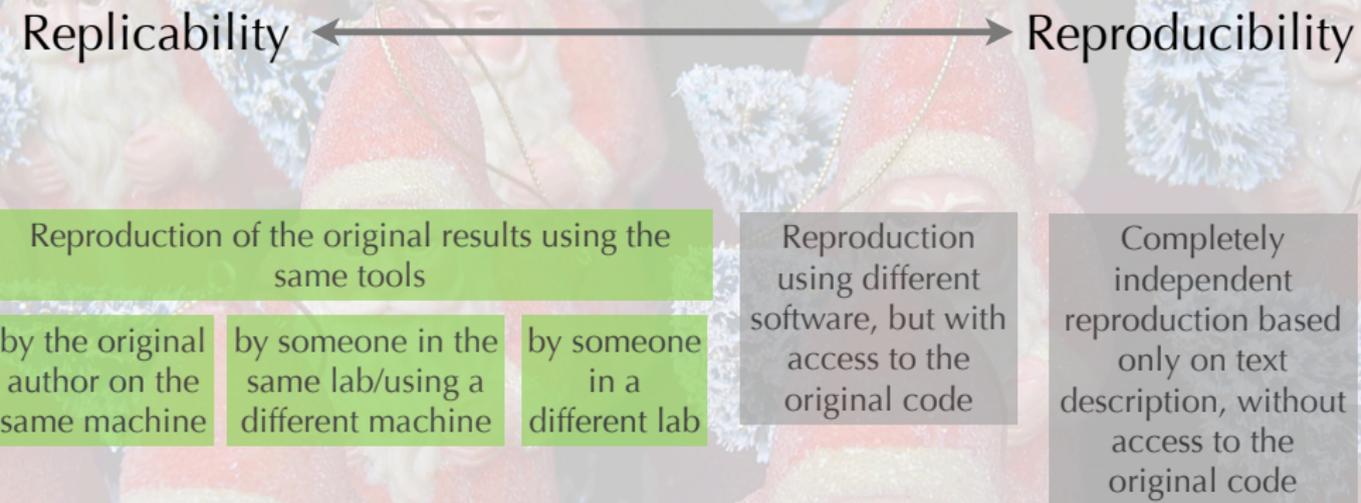
## ② Reporting Results

An IMRaD Report

Good Practice for Setting up a Laboratory Notebook

## ③ To do for the Next Time

# Reproducibility: What Are We Talking About?



# Reproducible Research: Trying to Bridge the Gap

Author



Published  
Article

Nature/System/...



Protocol

(Design of Experiments)



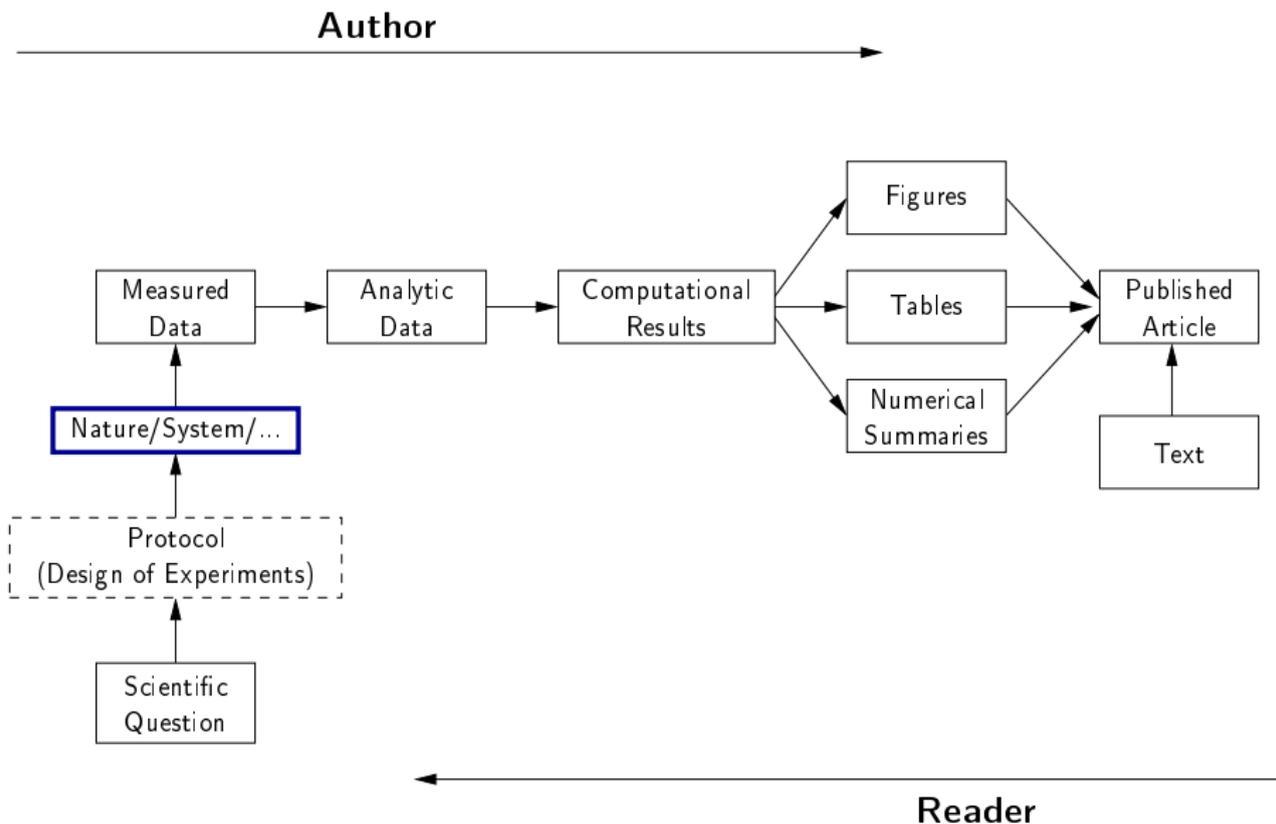
Scientific  
Question

Reader

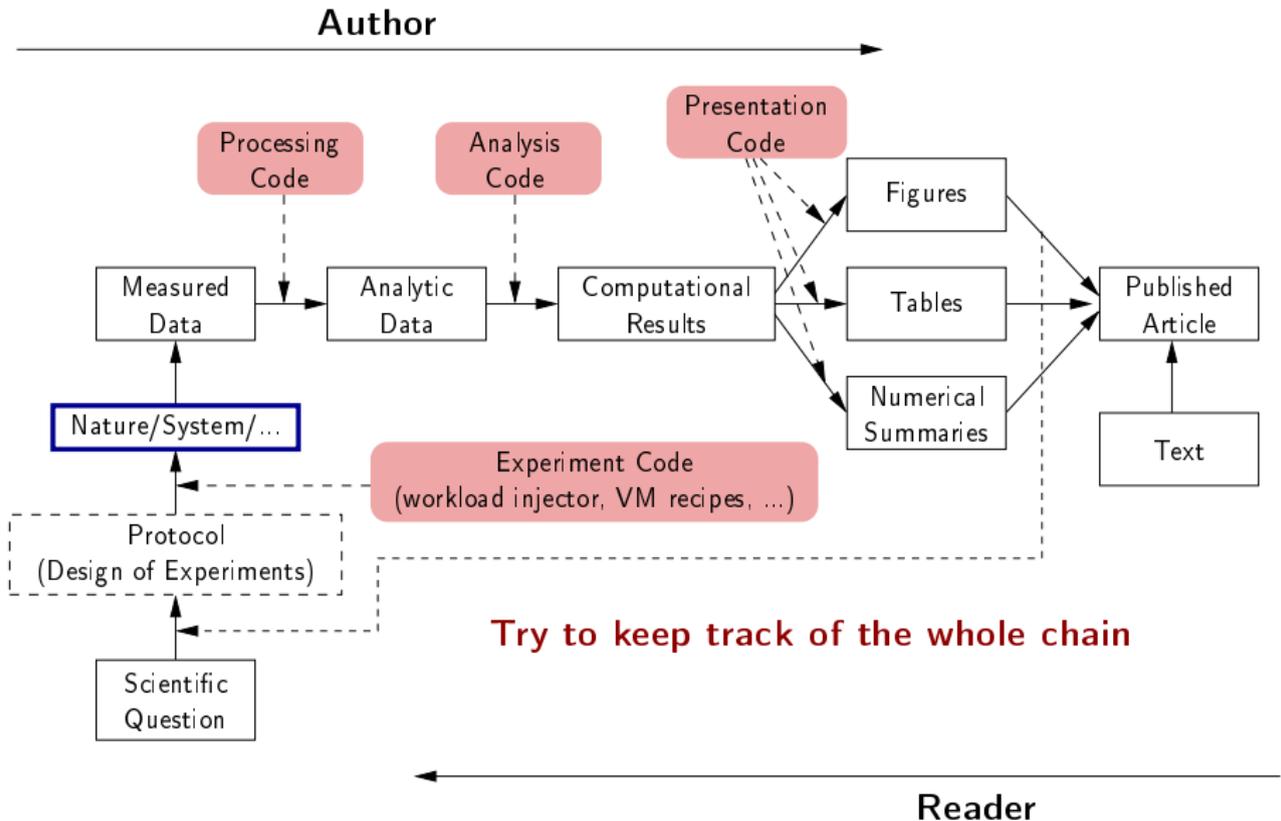


Inspired by Roger D. Peng's lecture on reproducible research, May 2014

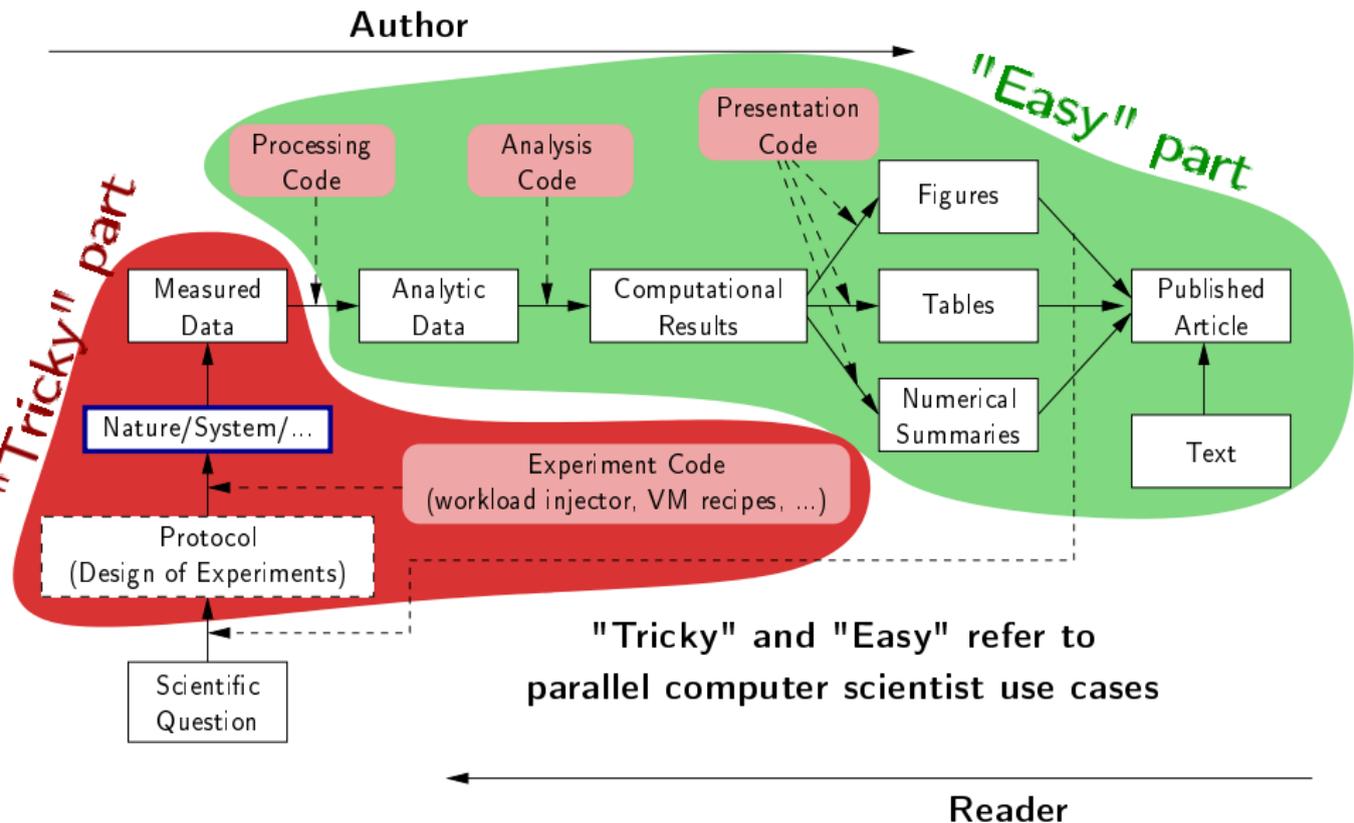
# Reproducible Research: Trying to Bridge the Gap



# Reproducible Research: Trying to Bridge the Gap



# Reproducible Research: Trying to Bridge the Gap



# Mythbusters: Science vs. Screwing Around



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

Handwritten notes on a clipboard:

COOL WATER TEST

Time	Temp	Temp	Temp
0	98.4	17	92.1
1	98.2	18	91.5
2	97.7	19	90.3
3	97.6	20	89.4
4	97.1	21	88.2
5	96.2	22	87.1
6	95.2	23	86.7
7	94.9	24	86.5
8	94.8	25	86.1
9		26	
10		27	
11		28	

Other notes include circled numbers (3), (4), (5) and a box containing "910" and "Sword".

## ① The Reproducible Research Movement

How does it work in other sciences?

Is CS Concerned Really With This?

Reproducible Research/Open Science

Many Different Alternatives for Replicable Analysis

## ② Reporting Results

An IMRaD Report

Good Practice for Setting up a Laboratory Notebook

## ③ To do for the Next Time

## 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 *Support different approaches*
  - 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?

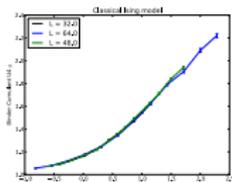
## An Provenance-Rich Paper: ALPS2.0

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

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

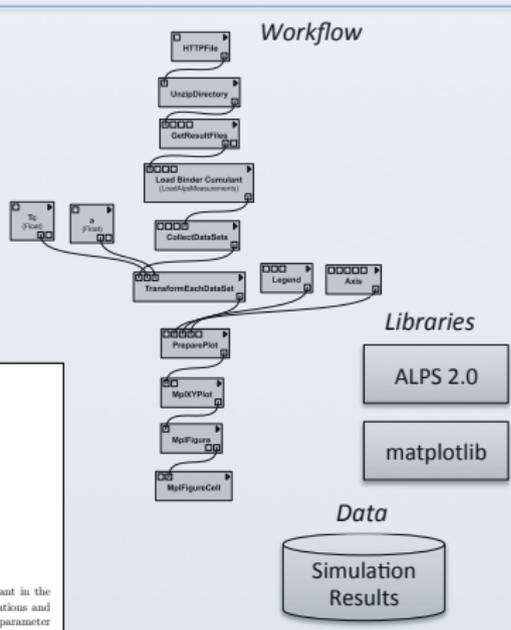
B. Bauer<sup>1</sup> L. D. Carr<sup>2</sup> H.G. Evertz<sup>3</sup> A. Feiguin<sup>4</sup> J. Freire<sup>5</sup>  
S. Fuchs<sup>6</sup> L. Gamper<sup>7</sup> J. Gukelberger<sup>8</sup> E. Gull<sup>9</sup> S. Guertler<sup>4</sup>  
A. Hehn<sup>10</sup> R. Igarashi<sup>11,10</sup> S.V. Isakov<sup>1</sup>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

<sup>1</sup>Theoretische Physik, ETH Zurich, 8005 Zurich, Switzerland  
<sup>2</sup>Department of Physics, Colorado School of Mines, Golden, CO 80401, USA  
<sup>3</sup>Institut für Theoretische Physik, Technische Universität Graz, A-8010 Graz, Austria  
<sup>4</sup>Department of Physics and Astronomy, University of Wyoming, Laramie, Wyoming 82071, USA  
<sup>5</sup>Scientific Computing and Imaging Institute, University of Utah, Salt Lake City, Utah 84112, USA  
<sup>6</sup>Institut für Theoretische Physik, Georg-August-Universität Göttingen, Göttingen, Germany  
<sup>7</sup>Columbia University, New York, NY 10027, USA  
<sup>8</sup>Bethe Center for Theoretical Physics, Universität Bonn, Nussallee 12, 53115 Bonn, Germany



1 Correspond

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



## 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
>
```

## 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
```

## Word-processor plugin App

### LaTeX source

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

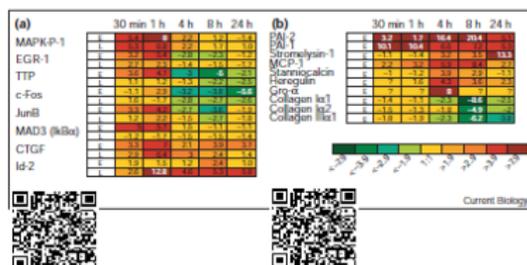
### LaTeX source with VCR package

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

Permanently bind printed graphics to underlying result content

**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.



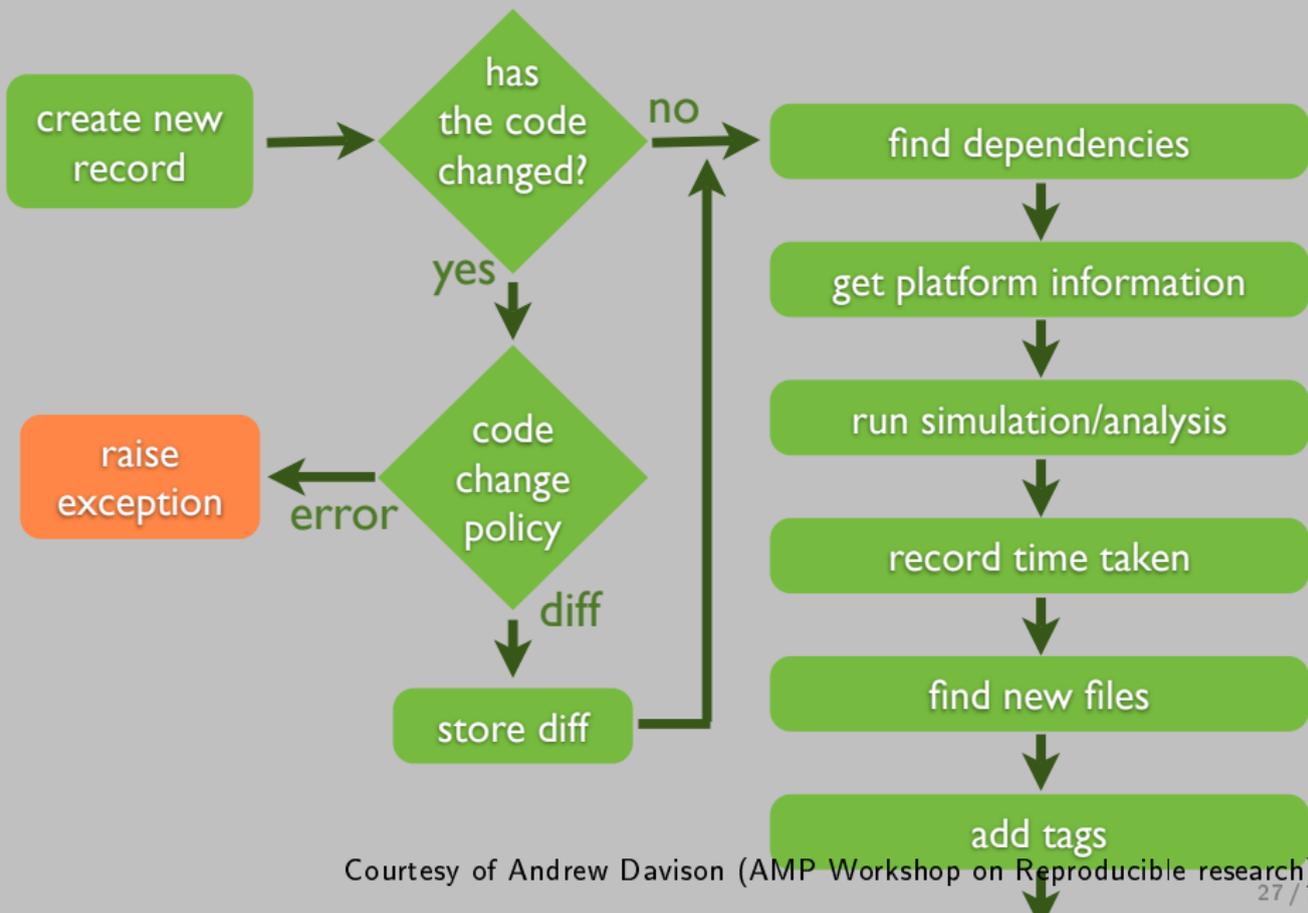
senescence response appears to overlap substantially with gene expression patterns observed in activated fibroblasts during wound healing [24–26]. MCP-1, Gro- $\alpha$ , IL-1 $\beta$  and IL-15 are strong effectors of macrophage and neutrophil recruitment and activation [27,28]. The upregulation of Toll (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- $\kappa$ 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- $\kappa$ B-dependent transcription [30] and also participates in inflammatory disease processes [28].

Deficiencies in the response of senescent cells to serum stimulation have been reported, and include an inability to induce the expression of *c-fos* mRNA [31] and markers of late G1 and S phase [32]. In response to serum, expression of inflammatory chemokines, matrix-degrading proteases and their modulators is induced in early-passage dermal fibroblasts, and expression of matrix collagens is reduced. This transient burst of activity may represent the natural response of the cells to wound repair [24]. Id-2 transcripts were hyper-induced in serum-stimulated senescent

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

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

# 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

http://127.0.0.1:8002/ Google

### 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"/>	<a href="#">20100709-154255</a>		'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"/>	<a href="#">20100709-154309</a>			0.59 s		Python	2.5.2	/Users/andrew/tmp/SumatraTest	main.py	396c2020ca50	09/07/2010	15:43:09	
<input type="checkbox"/>	<a href="#">hagging</a>	'determine whether the gourd is worth 3 or 4 shekels'	'apparently, it is worth NaN shekels.'	0.59 s		Python	2.5.2	/Users/andrew/tmp/SumatraTest	main.py	396c2020ca50	09/07/2010	15:43:20	fooba
<input type="checkbox"/>	<a href="#">20100709-154338</a>	'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"/>	<a href="#">hagging_repeat</a>	Repeat experiment hagging	The new record exactly matches the original.	0.58 s		Python	2.5.2	/Users/andrew/tmp/SumatraTest	main.py	396c2020ca50	09/07/2010	15:43:47	

Courtesy of Andrew Davison (AMP Workshop on Reproducible research)

# New Tools for Computational Reproducibility

- Dissemination Platforms:

[ResearchCompendia.org](http://ResearchCompendia.org)

[IPOL](http://IPOL)

[Madagascar](http://Madagascar)

[MLOSS.org](http://MLOSS.org)

[thedatahub.org](http://thedatahub.org)

[nanoHUB.org](http://nanoHUB.org)

[Open Science Framework](http://Open Science Framework)

[The DataVerse Network](http://The DataVerse Network)

[RunMyCode.org](http://RunMyCode.org)

- Workflow Tracking and Research Environments:

[VisTrails](http://VisTrails)

[Kepler](http://Kepler)

[CDE](http://CDE)

[Galaxy](http://Galaxy)

[GenePattern](http://GenePattern)

[Synapse](http://Synapse)

[Sumatra](http://Sumatra)

[Taverna](http://Taverna)

[Pegasus](http://Pegasus)

- Embedded Publishing:

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

[Verifiable Computational Research](http://Verifiable Computational Research)

[Sweave](http://Sweave)

[knitr](http://knitr)

[Collage Authoring Environment](http://Collage Authoring Environment)

[SHARE](http://SHARE)

And also: **Figshare**, **ActivePapers**, **Elsevier executable paper**, ...

## ① The Reproducible Research Movement

How does it work in other sciences?

Is CS Concerned Really With This?

Reproducible Research/Open Science

Many Different Alternatives for Replicable Analysis

## ② Reporting Results

An IMRaD Report

Good Practice for Setting up a Laboratory Notebook

## ③ To do for the Next Time

# Structure

Research articles are often structured in this basic order:

**Introduction** Why was the study undertaken? What was the research question, the tested hypothesis or the purpose of the research?

**Methods** When, where, and how was the study done? What materials/hardware were used? How was it configured?

**Results** What answer was found to the research question; what did the study find? Was the tested hypothesis true? **Present useful results in a synthetic way with a logical order.**

**Discussion** What might the answer imply and why does it matter? How does it fit in with what other researchers have found? What are the possible bias and points to improve? What are the perspectives for future research?

Such structure **facilitates literature review** and is a very effective way to convey information.

If the report is a few pages long then **an abstract is required.**

## ① The Reproducible Research Movement

How does it work in other sciences?

Is CS Concerned Really With This?

Reproducible Research/Open Science

Many Different Alternatives for Replicable Analysis

## ② Reporting Results

An IMRaD Report

Good Practice for Setting up a Laboratory Notebook

## ③ To do for the Next Time

# Step 0: Taking Notes

## Document your:

- **Hypotheses**: keep track of your ideas/line of thoughts
- **Experiments**: details on how and why an experiment was run, including failed or ambiguous attempts.
- **Initial analysis or interpretation** of these experiments: was the outcome conform to the expectation or not? does it (in)validate the hypothesis?
- **Organization**: keep track of things to do/fix/test/improve

## Structure:

- ① General information about the document and organization **conventions** (e.g., directory structure, notebook structure, experimental result storing mechanism, ...)
- ② Documentation of **commonly used commands** and of how to set up experiments (e.g., git cloning, environment deployment, connection to machines, compiling scripts)
- ③ Experiment results can be either structured **by dates** (↪ add tags) or **by experiment campaigns** (↪ add date/time)

# Which format should I use ?

- **Wikis** are encouraged to favor collaboration but I do not find them really effective
- **Blogging** systems are also a way of managing such notebook but they should rather be considered as an effective way to share information with others
- I recommend to use basic **plain-text** format and to **structure it hierarchically**

Here is a **link** to an excerpt of the journal of one of my PhD student, managed with git/org-mode. More detailed links are given in slide ??.

Last but not least:

Provide links to **Raw Data!!!**

# When/How Often Should I Use it?

I have a very intense usage (demo to **general journal** and specific **BOINC journal**) and I tend to capture a lot of information but you do not have to be as extreme as I am. Here are a few advices:

- 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, write what the next steps are
  - **Summarize in a 2-4 lines** (for your advisor) what you did, what was difficult, what you learnt.

## Step 1: Sharing Code and Data

### What kinds of systems are available?

- "Good" - The cloud (Dropbox, Google Drive, Figshare)
- Better - Version control systems (SVN, Git and Mercurial)
- "Best" - Version control systems on the cloud (GitHub, Bitbucket)

Depends on the level of privacy you expect but you probably already know these tools.

**Few handle GB files...**

### Is this enough?

- 1 Use a workflow that documents both data and process
- 2 Use the machine readable CSV format
- 3 Provide raw data and meta data, not just statistical outputs
- 4 Never do data manipulation and statistical tests by hand
- 5 Use R, Python or another free software to read and process raw data (ideally to produce complete reports with code, results and prose)

## Step 2: Literate Programming

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

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

### Org-mode (requires emacs)

My favorite tool.

- plain text, very smooth, works both for html, pdf, ...
- allows to combine all my favorite languages even with sessions

### Ipython notebook

If you are a python user, go for it! Web app, easy to use/setup...

### KnitR (a.k.a. Sweave)

For non-emacs users and as a first step toward *reproducible papers*:

- Click and play with a modern IDE (e.g., Rstudio)

If you use literate programming, scripts, and version control on a daily basis, writing a reproducible article will be painless.

Reviewers (and advisors) will greatly appreciate the effort

Next lecture: Causality, correlation and covariance

- 1 **Install** R and Rstudio to learn literate programming and check how to **create small reproducible analysis** and publish them on! **rpubs** with Rstudio.
- 2 **Setup** your own laboratory notebook and start using it to collect information. Emacs **org-mode** is a great tool for this.
- 3 **Select** a topic on which you would like to apply the techniques we will present you.