Reproducible Research
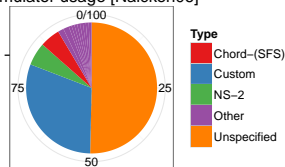
Arnaud Legrand

Journée Traces, Grenoble
March 24, 2015

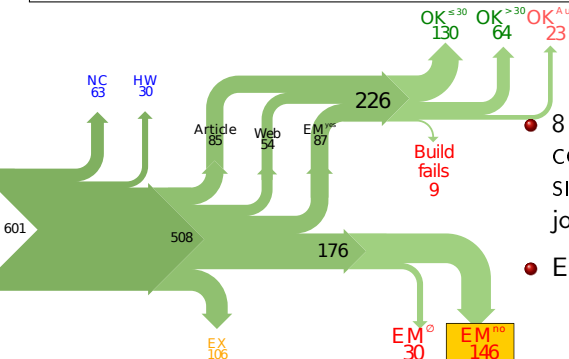Naicken, Stephen *et Al.*, *Towards Yet Another Peer-to-Peer Simulator*, HET-NETs'06.

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

- $EM^{no}=$ the code cannot be provided

The Dog Ate my Homework !!!

- Versionning 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.

- Versionning 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 !!!

- Versionning 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 !!!

- Versionning 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 !!!

- Versionning 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 !!!

- Versionning 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 !!!

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

- Commercial Code
- Proprietary Academic Code

Unfortunately, the ⟨SYSTEM⟩ sources are not meant to be opensource (the code is partially property of ⟨UNIVERSITY 1⟩, ⟨UNIVERSITY 2⟩ and ⟨UNIVERSITY 3⟩.)

If this will change I will let you know, albeit I do not think there is an intention to make the ⟨SYSTEM⟩ sources opensource in the near future.
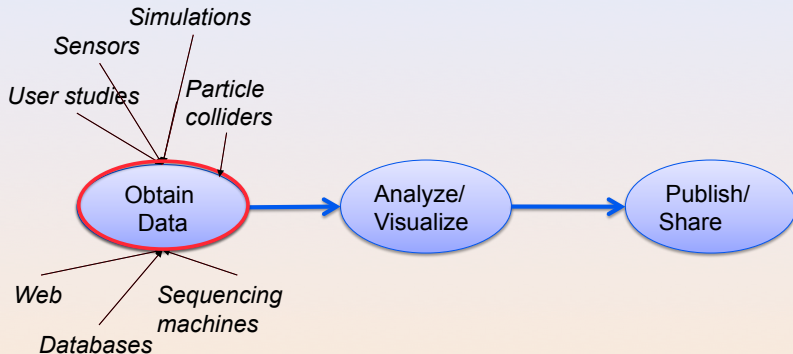
If you're interested in obtaining the code, we only ask for a description of the research project that the code will be used in (which may lead to some joint research), and we also have a software license agreement that the University would need to sign.
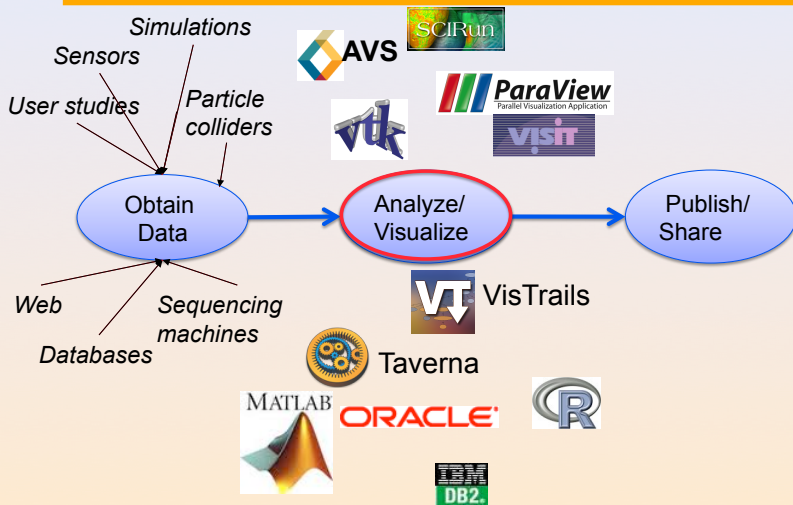
The Dog Ate my Homework !!!

- Versionning 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.
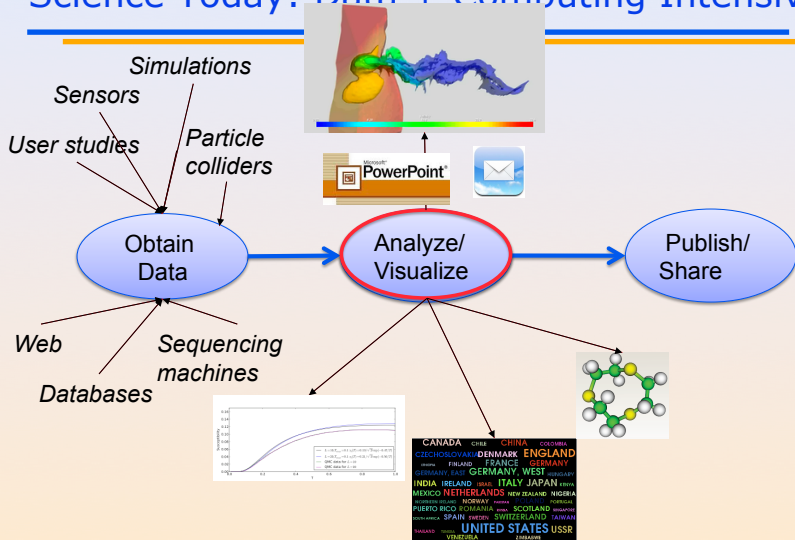
Science Today: Data Intensive

Simulations
Sensors
User studies
Particle colliders

Obtain Data → Analyze/Visualize → Publish/Share

Web
Sequencing machines
Databases

Courtesy of Juliana Freire (AMP Workshop on Reproducible research)

Science Today: Data + Computing Intensive

Simulations

Sensors

User studies

Particle colliders

AVS

SCIRun

ParaView
Parallel Visualization Application

vtk

VisIt

Obtain Data

Analyze/ Visualize

Publish/ Share

Web

Sequencing machines

Databases

VisTrails

Taverna

MATLAB

ORACLE

R

IBM DB2

Courtesy of Juliana Freire (AMP Workshop on Reproducible research)

Science Today: Data + Computing Intensive

Simulations

Sensors

User studies

Particle colliders

Obtain Data

Analyze/ Visualize

Publish/ Share

Web

Sequencing machines

Databases

CANADA

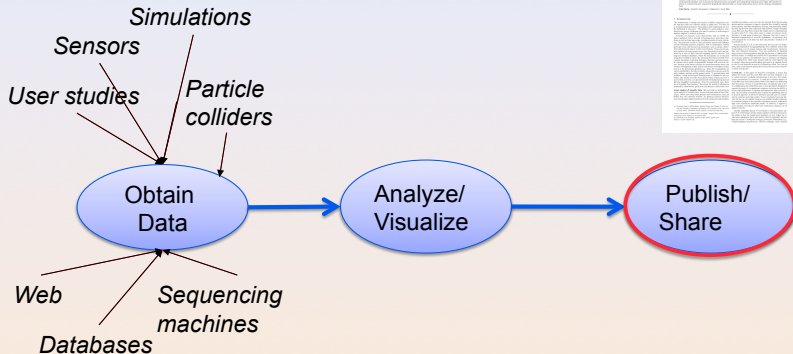Reproducible Research '11   UBC, Vancouver   Juliana Freire   4

Courtesy of Juliana Freire (AMP Workshop on Reproducible research)

Science Today: Data + Computing Inte

Simulations
Sensors
User studies
Particle colliders

Obtain Data → Analyze/ Visualize → Publish/ Share

Web
Databases
Sequencing machines

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.
- Scientific record is incomplete — too large to fit in a paper
- Large volumes of data
- Complex processes

◆ Can't reproduce or verify

> "*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]

Courtesy of Juliana Freire (AMP Workshop on Reproducible research)

- 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. 2010, Nov.
- 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 of Adam J. Richards

What your research supposedly looks like:
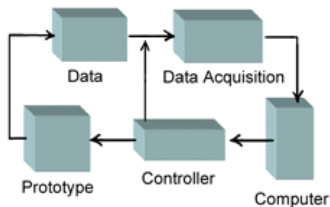
Figure 1. Experimental Diagram

What your research *actually* looks like:

Figure 2. Experimental Mess

JORGE CHAM © 2008   WWW.PHDCOMICS.COM
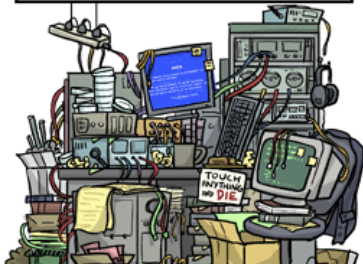
Experiments in HPC: Even Worse!

- Rely on large, distributed, hybrid, prototype hardware/software
- Measure execution times (makespans, traces, . . .)
- Many parameters, very costly and hard to *reproduce*

Reproducing in the Wild

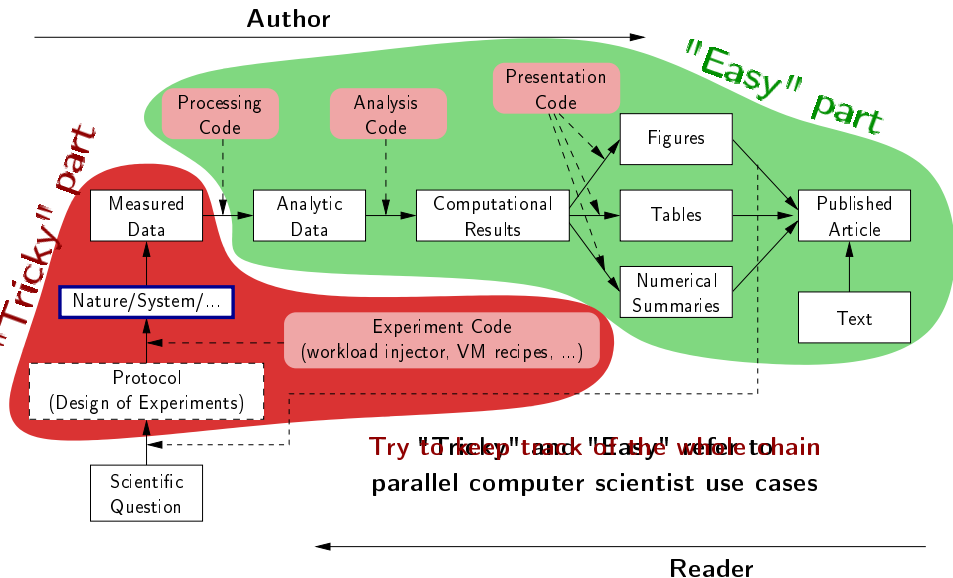## In my Community

Reproducing results of others is not in the culture

- Rewards for positive results and novelty rather than consolidating existing results
- Excuse #1: Hardware and software evolve very quickly
- Excuse #2: Reproducible research is not mandatory

Why is it so Hard in General?

- Copyright/competition issue (would give advantage to "others")
- Publication bias (only the idea matters, not the gory details)
- Experimenter bias
- Programming errors or data manipulation/analysis mistakes
- ~~Lack of easy-to-use tools~~

Why bother?

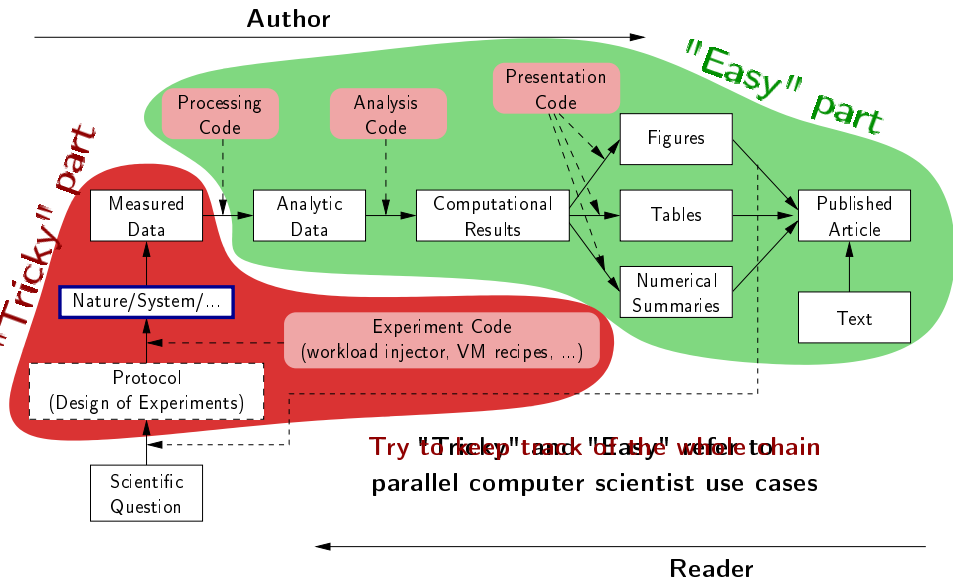Inspired by Roger D. Peng's lecture on reproducible research, May 2014

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

Reproducible Research: Trying to Bridge the Gap

Author

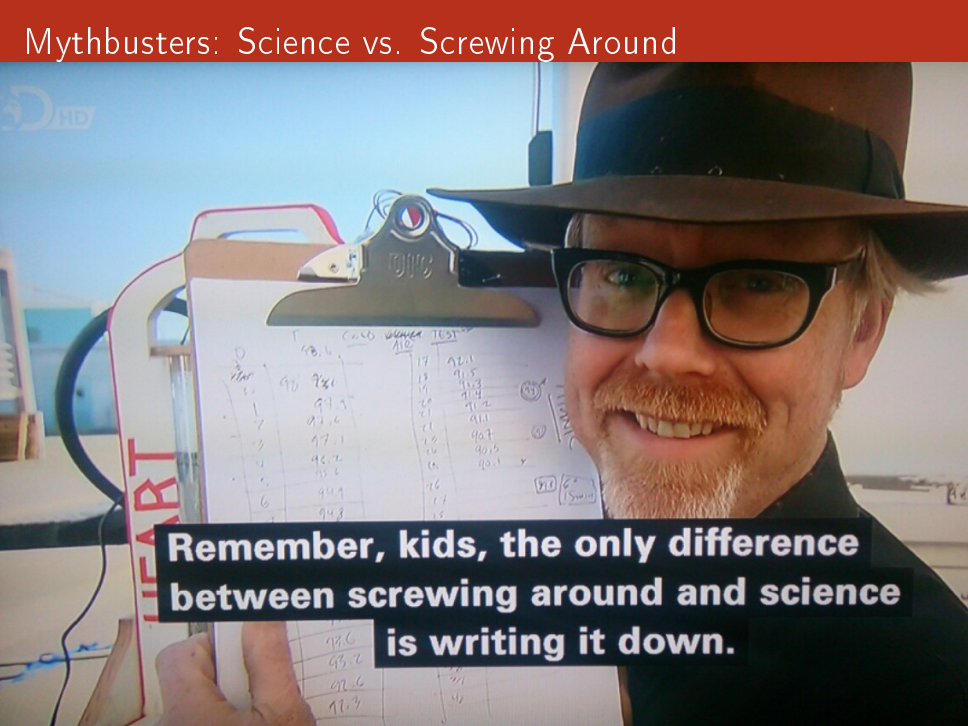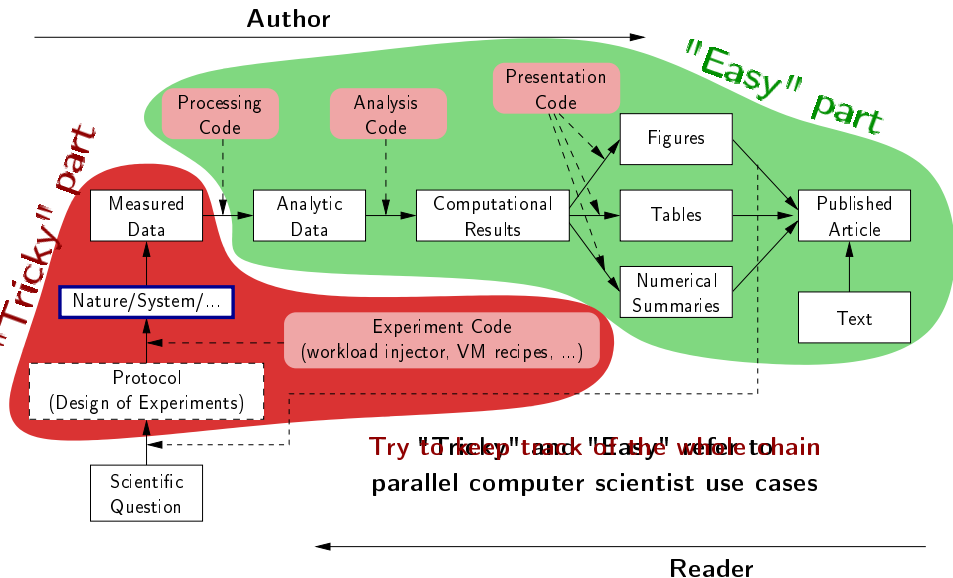"Easy" part

"Tricky" part

Processing Code

Analysis Code

Presentation Code

Figures

Measured Data

Analytic Data

Computational Results

Tables

Nature/System/...

Numerical Summaries

Experiment Code (workload injector, VM recipes, ...)

Published Article

Text

Protocol (Design of Experiments)

Scientific Question

Try to keep track of the whole chain
Try "Tricky" and "Easy" when to
parallel computer scientist use cases

Reader

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

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

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

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: **Figshare/Zenodo**, **ActivePapers**, **Elsevier executable paper**,...

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) (2014 edition)
- Reproducible Research: Tools and Strategies for Scientific Computing (2011)
- Working towards Sustainable Software for Science: Practice and Experiences (2013)
- Reproducibility@XSEDE: An XSEDE14 Workshop
- TRUST 2014
- REPPAR'15: 2nd International Workshop on Reproducibility in Parallel Computing
- Reproducibility of Computation Based Research/NTMS'15

Should be seen as opportunities to share experience.

recherche-reproductible@listes.univ-orleans.fr