

Reproducibility and Computational Science

The Real Story of Code and Data Sharing

Victoria Stodden

vcs@stanford.edu

SciFoo 2009

Agenda

1. The Scientific Method is being transformed by massive computation
 - New modes of knowledge discovery?
 - New standards for what we consider knowledge?
2. Facilitating reproducibility 1: the *Reproducible Research Standard*
3. Facilitating reproducibility 2: tools for attribution and research transmission
4. Why aren't researchers sharing? (new results)

Transformation of Scientific Enterprise

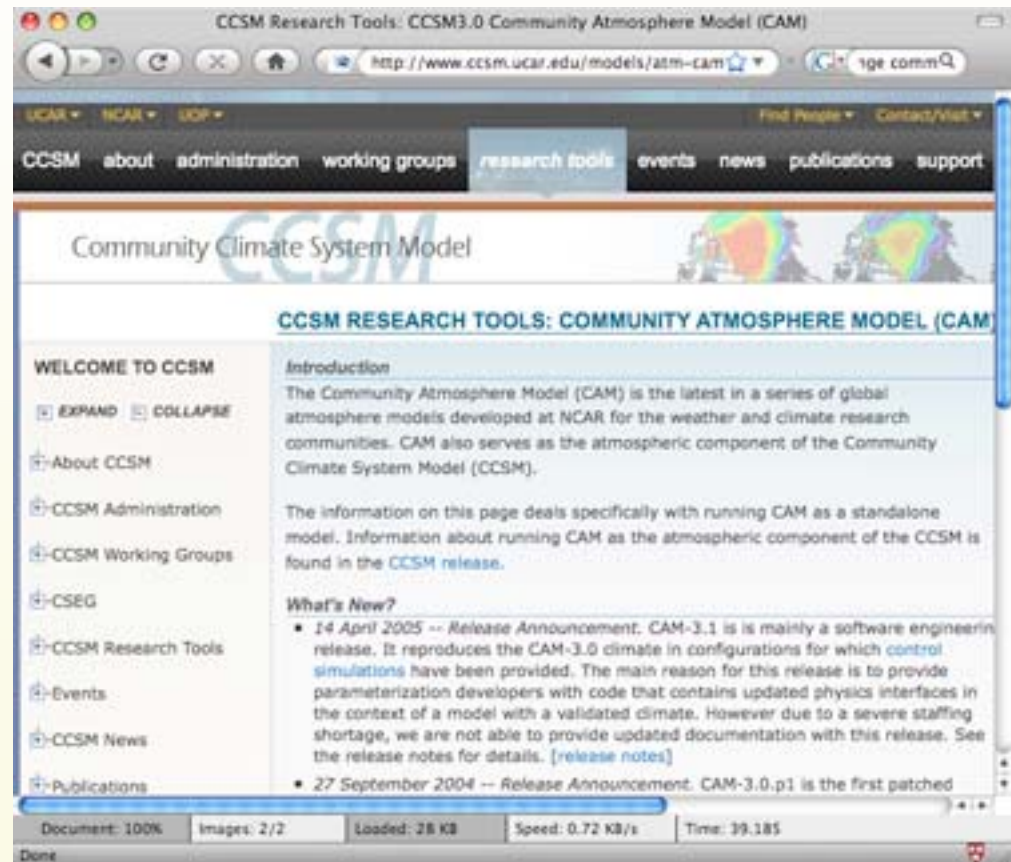
Massive Computation: emblems of our age include:

- data mining for subtle patterns in vast databases,
- massive simulations of a physical system's complete evolution repeated numerous times, as simulation parameters vary systematically.

Raises new questions about science..

Example: Community Climate Model (CCM)

- Collaborative system simulation
- Open code, data



Example: High Energy Physics

- 4 LHC experiments at CERN: 15 petabytes produced annually
- Data shared through grid to mobilize computing power
- Director of CERN (Heuer): “Ten or 20 years ago we might have been able to repeat an experiment. They were simpler, cheaper and on a smaller scale. Today that is not the case. So if we need to re-evaluate the data we collect to test a new theory, or adjust it to a new development, we are going to have to be able reuse it. That means we are going to need to save it as open data....” Computer Weekly, August 6, 2008

Example: Astrophysics Simulation Collaboratory

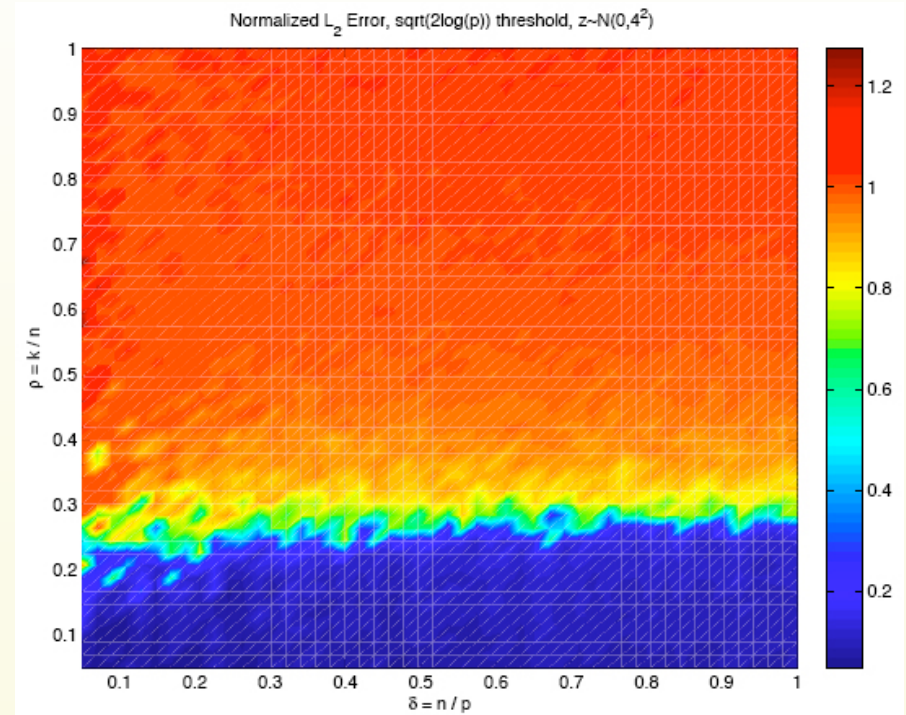
- Data and code sharing
- Interface for dynamic simulation
- mid 1930's: calculate the motion of cosmic rays in Earth's magnetic field..

The screenshot shows a web browser window titled "The Astrophysics Simulation Collaboratory". The address bar displays the URL "http://wugrav.wustl.edu/ASC/project/progress.html". The page features a navigation menu on the left with categories: Project (Progress, People, Goals, Developers), Portal (Login, Documentation, Credits), Grid/VMR (Machines, Resources, VMR Status), and Contact. The main content area has a header with the ASC logo and the text "Astrophysics Simulation Collaboratory". Below this is a sub-header: "A Laboratory For Large Scale Simulations Of Relativistic Astrophysics". A central diagram shows a red circle labeled "Astrophysics Simulation Collaboratory" connected to five surrounding grey ovals: "Collaboration ASC Portal", "Programming Framework Cactus, AMR", "Scientific Visualization, Vision, OpenDX, Ames", "Connections GridLab, EUNetwork, Cactus Development", and "Grid Computing". A text block at the bottom explains that the ASC provides a collaborative environment for geographically distributed projects through the ASC Portal, a specialized framework for the Cactus Computational Toolkit.

Document: 100% Images: 9/9 Loaded: 12 KB Speed: 6.83 KB/s Time: 1.759

Example: Proofs

- Mathematical proof via simulation, not deduction
- Breakdown point:
 $1/\sqrt{2\log(p)}$



- A valid proof?
- A contribution to the field of mathematics?

The Third Branch of the Scientific Method

- Branch 1: *Deductive/Theory*: e.g. mathematics; logic
- Branch 2: *Inductive/Empirical*: e.g. the machinery of hypothesis testing; statistical analysis of controlled experiments
- Branch 3: Large scale extrapolation and prediction: Knowledge from computation or tools for established branches?

Contention About 3rd Branch

- Anderson: The End of Theory. (Wired, June 2008)
- Hillis Rebuttal: We are looking for patterns first then create hypotheses as we always have.. (The Edge, June 2008)
- Idea (Weinstein): Simulation underlies branches
 1. Tools to build intuition (branch 1)
 2. Hypotheses to test (branch 2)
- Manipulation of systems you can't fit in a lab
- ~New: differential analyzers of 50's and 60's, chaos research in 70's

Controlling Error is Central to the Scientific Method



In stochastic modeling “the possibility of erroneous decisions cannot be eliminated, and the best one can do is to seek methods of making decisions that, in a sense, minimize the risk of mistakes.” Jerzy Neyman, “Statistics - Servant of All Sciences,” *Science*, 1955, p. 401

Computation is Increasingly Pervasive

- JASA June 1996: 9 of 20 articles computational
- JASA June 2006: 33 of 35 articles computational

Emerging Credibility Crisis in Computational Science

- Error control forgotten? Typical scientific communication doesn't include code, data.
- Published computational science near impossible to replicate.
- JASA June 1996: none of the 9 made code or data available
- JASA June 2006: 3 of those 33 articles had code publicly available.
- A second change to the scientific method due to computation?

Changes in Scientific Communication

- Internet: communication of all computational research details/data possible
- Scientists often post papers but not their complete body of research
- Changes coming: Madagascar, Sweave, individual efforts, journal requirements...

Potential Solution: Really Reproducible Research



Pioneered by Jon Claerbout

“An article about computational science in a scientific publication is not the scholarship itself, it is merely advertising of the scholarship. The actual scholarship is the complete software development environment and the complete set of instructions which generated the figures.”

(quote from David Donoho, “Wavelab and Reproducible Research,” 1995)

Reproducibility

- (Simple) definition: A result is reproducible if a member of the field can independently verify the result.
- Typically this means providing the original code and data, but does not imply access to proprietary software such as Matlab, or specialized equipment or computing power.

Barriers to Sharing 1: Legal

- Original expression of ideas falls under copyright by default
- Copyright creates exclusive right of the author to:
 - reproduce the work
 - prepare derivative works based upon the original

Creative Commons



- Founded by Larry Lessig to make it easier for artists to share and use creative works
- A suite of licenses that allows the author to determine terms of use attached to works

Creative Commons Licenses

- A notice posted by the author removing the default rights conferred by copyright and adding a selection of:
- BY: if you use the work attribution must be provided,
- NC: work cannot be used for commercial purposes,
- ND: derivative works not permitted,
- SA: derivative works must carry the same license as the original work.

Open Source Software Licensing

- Creative Commons follows the licensing approach used for open source software, but adapted for creative works
- Code licenses:
 - BSD license: attribution
 - GNU GPL: attribution and share alike
 - Hundreds of software licenses..

Apply to Scientific Work?

- Remove copyright's block to fully reproducible research
- Attach a license with an attribution component to *all* elements of the research compendium (including code, data), encouraging full release.

Solution: *Reproducible Research Standard*

Reproducible Research Standard

Realignment of legal rights with scientific norms:

- Release media components (text, figures) under CC BY.
- Release code components under Modified BSD or similar.
- Both licenses free the scientific work of copying and reuse restrictions and have an attribution component.

Releasing Data?

- Raw facts not copyrightable.
- Original “selection and arrangement” of these facts is copyrightable. (Feist Publ’ns Inc. v. Rural Tel. Serv. Co., 499 U.S. 340 (1991))

Benefits of RRS

- Focus becomes release of the entire research compendium
- Hook for funders, journals, universities
- Standardization avoids license incompatibilities
- Clarity of rights (beyond Fair Use)
- IP framework supports scientific norms
- Facilitation of research, thus citation, discovery...

Barriers to Sharing 2

Hypotheses:

1. Scientists are primarily motivated by personal gain or loss.
2. Scientists are worried about being scooped.

Survey of Computational Scientists

- *Subfield*: Machine Learning
- *Sample*: American academics registered at top Machine Learning conference (NIPS).
- *Respondents*: 134 responses from 638 requests.

Sharing Habits

- Reported an average of 32% of their code available on the web, 48% of their data,
- 81% claim to reveal some code and 84% claim to reveal some data.
- Visual inspection of their websites: 30% had some code posted, 20% had some data posted.

Sharing Elements of Research

Proportion Comfortable Sharing on the Web...

Final Paper	99%
Draft Paper	26%
Pre-publication Data	13%
Post-publication Data	67%
Pre-publication Code	21%
Post-publication Code	74%
Unpublished research ideas, ie. through a blog	21%

Top Reasons Not to Share

<i>Code</i>		<i>Data</i>
77%	Time to document and clean up	54%
44%	Not receiving attribution	42%
40%	Possibility of patents	-
34%	Legal barriers (ie. copyright)	41%
-	Time to verify release with admin	38%
30%	Potential loss of future publications	35%
52%	Dealing with questions from users	34%
30%	Competitors may get an advantage	33%
20%	Web/Disk space limitations	29%

For example..



"Behind one door is tenure - behind the other is flipping burgers at McDonald's."

Top Reasons to Share

<i>Code</i>		<i>Data</i>
91%	Encourage scientific advancement	81%
90%	Encourage sharing in others	79%
86%	Be a good community member	79%
82%	Set a standard for the field	76%
85%	Improve the caliber of research	74%
81%	Get others to work on the problem	79%
85%	Increase in publicity	73%
78%	Opportunity for feedback	71%
71%	Finding collaborators	71%

Preliminary Findings

- *Surprise*: Scientists are generally motivated to share by communitarian ideals.
- *Not surprising*: Reasons for not revealing reflect private incentives.
- *Surprise*: Scientists not that worried about being scooped.

Perception of Theft...

Idea Theft	Count	Proportion
At least one publication scooped	53	0.51
2 or more scooped	31	0.30
No ideas stolen	50	0.49

Real and Potential Wrinkles

- Reproducibility neither necessary nor sufficient for correctness
- Attribution in digital communication:
 - Legal attribution and academic citation not isomorphic
 - Contribution tracking (RDFa)
- RRS: Need for individual scientist to act
- “progress depends on artificial aids becoming so familiar they are regarded as natural” I.J. Good (“How Much Science Can You Have at Your Fingertips” , 1958)

Papers

- “Enabling Reproducible Research: Open Licensing for Scientific Innovation”
- “15 Years of Reproducible Research in Computational Harmonic Analysis”
- “The Legal Framework for Reproducible Research in the Sciences: Licensing and Copyright”

<http://www.stanford.edu/~vcs>

Reproducibility is Subtle

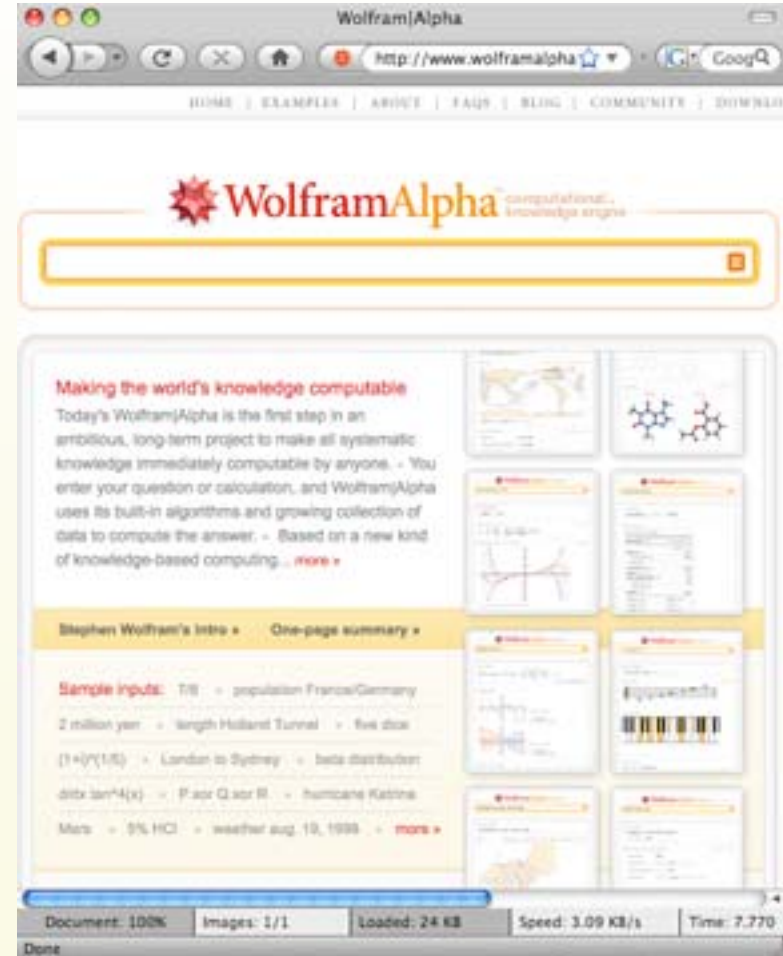
- Simple case: open data and small scripts. Suits simple definition.
- Hard case: Inscrutable code, organic programming.
- Harder case: massive computing platforms, streaming data.
- Can we have reproducibility in the hard cases?

Solutions for Harder Cases

- Tools for reproducibility:
 - Standardized testbeds
 - Open code for continuous data processing, flags for “continuous verifiability”
 - Standards and platforms for data sharing
- Tools for attribution:
 - Generalized contribution tracking
 - Legal attribution/license tracking tracking and search (RDFa)

Case Study: Wolfram | Alpha

- Obscure code - testbeds for verifiability
- Dataset construction methods opaque



Appendix: Attribution

- Legal attribution and academic citation not isomorphic.
- Minimize administrative burden
- Evolving norms / field specific norms / technology
- “keep intact all copyright notices for the Work and provide, reasonable to the medium or means You are utilizing... .”