

# The Digitization of Science and the Degradation of the Scientific Method

**Victoria Stodden**

Postdoctoral Associate in Law and  
Kauffman Fellow in Law and Innovation  
Yale Law School

Dean's Lecture  
School of Information  
UC Berkeley  
May 5, 2010

## What's the Problem?

Setting the Stage

Examples

The Credibility Crisis

## Survey of Machine Learning Community

## Legal Barriers to Sharing (and a solution)

Copyright

Responses in the Digital Realm

Reproducible Research Standard

## New Publication Modalities

Example: SparseLab

## Conclusions

# Scientific Research is Changing

Scientific computation is becoming central to the scientific method:

- ▶ Changing how research is conducted in many fields,
- ▶ Changing the nature of how we learn about our world.

Today's academic scientist probably has more in common with a large corporation's information technology manager than with a philosophy or English professor at the same university.

# I. Examples of Pervasiveness of Computational Methods

- ▶ For example, in statistics:

JASA June	Computational Articles	Code Publicly Available
1996	9 of 20	0%
2006	33 of 35	9%
2009	32 of 32	16%

- ▶ Social network data and the quantitative revolution in social science (Lazier et al. 2009);
- ▶ Computation reaches into traditionally nonquantitative fields: e.g. Wordhoard project at Northwestern examining word distributions by Shakespearian play.

## II. Examples of the Changing Nature of Scientific Discovery

### 1. Climate Simulation: Community Climate Models (e.g. NCAR),


The screenshot shows the homepage of the Community Climate System Model (CCSM) website. The browser is Firefox, and the URL is http://www.cesm.ucar.edu/. The page features a navigation menu with links for 'about', 'administration', 'working groups', 'research tools', 'events', 'news', 'publications', and 'support'. The main content area is titled 'Community Climate System Model' and includes a search bar. Below the title, there are three columns: 'RESEARCH TOOLS' with a photo of a beach and links to 'models', 'experiments', and 'support'; 'ADMINISTRATION' with a climate map and links to 'Scientific Steering Committee' and 'Advisory Board'; and 'WORKING GROUPS' with a satellite image and links to 'working groups' and 'lessons'. To the right, there is an 'ANNOUNCEMENTS' section for the '15th Annual CCSM Workshop Announcement' and a 'Welcome Dr. Jim Hurrell' message. At the bottom, there is an 'ABOUT CCSM' section and a 'CLIMATE NEWS' section.


Firefox File Edit View History Bookmarks Tools Window Help  
Community Climate System Model (CCSM): Home  
http://www.cesm.ucar.edu/ Google  
UCAR NCAR UOP Find People Contact/Visit  
CCSM about administration working groups research tools events news publications support


Community Climate System Model Search

WELCOME TO CCSM  
EXPAND COLLAPSE  
About CCSM  
CCSM Administration  
CCSM Working Groups  
CSEG  
CCSM Research Tools  
Events  
CCSM News  
Publications  
CCSM Support

CCSM PROJECT

Research Tools  
  
> models  
> experiments  
> support

Administration  
  
> Scientific Steering Committee  
> Advisory Board

Working Groups  
  
> working groups  
> lessons

ANNOUNCEMENTS

15th Annual CCSM Workshop Announcement  
We are pleased to announce that the 15th Annual CCSM Workshop will be held at the Great Divide Lodge in Breckenridge, Colorado the week of 6/28/10.

Welcome Dr. Jim Hurrell as chair of the CCSM Science Steering Committee (SSC), as part of his new position as Chief Scientist for Community Climate Projects in CGD.

NSF Climate Process and Modeling Teams (CPT) Call for Proposal  
The key aim of the Climate Process Modeling Team (CPT) concept is to speed development of global coupled climate models and reduce uncertainties in climate models. [more]

ABOUT CCSM  
CCSM belongs to an elite category of computer-based simulation models known as general-circulation models. Such models use mathematical formulas to recreate the chemical and physical processes that drive Earth's climate. What emerges from trillions of computer calculations is a picture of the world's climate in all its complexity. [More...]  
[CCSM Brochure]

CLIMATE NEWS

CCSM DISTINGUISHED ACHIEVEMENT

## II. Examples of the Changing Nature of Scientific Discovery

### 2. High Energy Physics: Large Hadron Collider

- ▶ 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

## II. Examples of the Changing Nature of Scientific Discovery

### 3. Astrophysics Simulation Collaboratory, University of Washington

The screenshot shows a Firefox browser window displaying the website for the Astrophysics Simulation Collaboratory. The browser's address bar shows the URL `http://wugrav.wustl.edu/ASC/project/progress.html`. The website header features the logo "ASC" with three red spheres and the text "Astrophysics Simulation Collaboratory". Below the header, the text reads "A Laboratory For Large Scale Simulations Of Relativistic Astrophysics".

The main content area contains a central diagram with a red oval labeled "Astrophysics Simulation Collaboratory" at its center. Six surrounding grey ovals are connected to the center by lines, each containing a project or service name and its associated technologies:

- Collaboration ASC Portal**
- Programming Framework**: Cactus, AMR
- Scientific Visualization, Vision, OpenDX, Amira**
- Connections**: GridLab, EUNetwork, Cactus Development
- Grid Computing**
- Astrophysics**: BH, NS, collapse, etc; Zeus, MACH, CactusEinstein, EOS

On the left side of the page, there is a navigation menu with the following sections and links:

- Project**
  - [Progress](#)
  - [People](#)
  - [Goals](#)
  - [Developers](#)
- Portal**
  - [Login](#)
  - [Documentation](#)
  - [Credits](#)
- Grid/VMR**
  - [Machines](#)

## II. Examples of the Changing Nature of Scientific Discovery

### 4. Dynamic modeling of macromolecules: SaliLab UCSF

COCEBI-649; NO OF PAGES 12

ARTICLE IN PRESS



ELSEVIER

Available online at [www.sciencedirect.com](http://www.sciencedirect.com)

ScienceDirect

Current Opinion in

Cell Biology

#### The structural dynamics of macromolecular processes

Daniel Russel<sup>1</sup>, Keren Lasker<sup>1,2</sup>, Jeremy Phillips<sup>1,3</sup>,Dina Schneidman-Duhovny<sup>1</sup>, Javier A Velázquez-Muriel<sup>1</sup> and Andrej Sali<sup>1</sup>

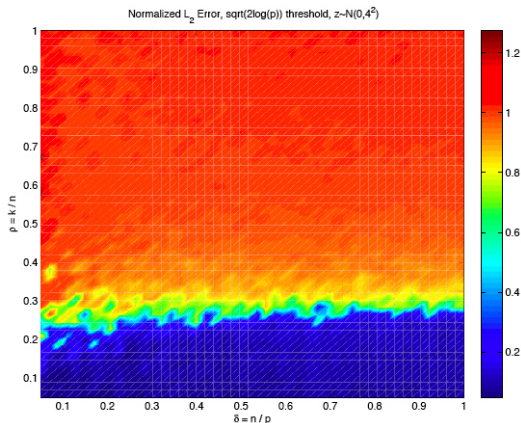
Dynamic processes involving macromolecular complexes are essential to cell function. These processes take place over a wide variety of length scales from nanometers to micrometers, and over time scales from nanoseconds to minutes. As a result, information from a variety of different experimental and computational approaches is required. We review the relevant sources of information and introduce a framework for integrating the data to produce representations of dynamic processes.

No single technique, computational or experimental, is able to span all relevant spatial and temporal scales (Figure 3). For static complexes, for example, X-ray crystallography can generate atomic structures of the components, while single particle cryo-electron microscopy (cryo-EM) can provide average mass density maps of the whole assembly at nanometer resolution for the whole assembly. For processes, computer simulations are beginning to reach the microsecond time scale, while



## II. Examples of the Changing Nature of Scientific Discovery

### 5. Mathematical proof by simulation and exhaustive grid search



(Stodden 2006)

## Evidence of a problem..

Relaxed practices regarding the communication of computational details is creating a credibility crisis in computational science, not only among scientists, but as a basis for policy decisions and in the public mind.

Recent prominent examples,

- ▶ Climategate 2009,
- ▶ Microarray-based clinical trials underway at Duke University.

# Climategate

- ▶ 19 Nov: Emails and documents from CRU appear illegally on the internet; Climate skeptics say the e-mails show that data is being manipulated; HARRY\_README.txt
- ▶ 22 Nov: Professor Mike Mann under (continuing) internal investigation at Penn
- ▶ 1 Dec: Man at centre of controversy, Professor Phil Jones, stands down while inquiry is conducted
- ▶ 3 Dec: Saudi chief negotiator says row proves climate change is not caused by humans
- ▶ 3 Dec: UEA commissions Sir Muir Russell to chair an independent inquiry
- ▶ 4 Dec: Head of UN climate science body says matter cannot be swept “under the carpet”
- ▶ 4 May: Virginia AG demands UVA documents related to Mann

## Clinical trials based on flawed genomic studies

### Timeline:

- ▶ Potti et al (2006), Nature Medicine: Main conclusion is that microarray data from cell lines can be used to define drug response “signatures,” that predict whether patients will respond,
- ▶ Coombes, Wang, Baggerly at M.D. Anderson Cancer Center cannot replicate, and find simple flaws: genes misaligned by one row, column labels flipped, genes repeated and missing from analysis..
- ▶ Clinical trials initiated in 2007 (Duke), 2008 (Moffitt).
- ▶ Baggerly & Coombes (2009) conducts “forensic bioinformatics” to replicate studies on a particular studies for drugs in clinical trials,

## Clinical trials based on flawed genomic studies

### Timeline continued:

- ▶ Duke launches internal investigation Sept 2009; all three trials suspended in Oct 2009,
- ▶ Oct 2009: results reported validated, regardless of errors, because data blinded,
- ▶ Baggerly finds data is not blinded as submitted to EORTC investigators, published in *Cancer Letter*, 2009,
- ▶ Jan 2010: Duke clinical trials resume, patients allocated to treatment and control groups. *"Neither the review nor the raw data are being made available at this time."* A future paper will explain their methods.

## A Credibility Crisis on Computational Science..

Other examples come to light..

- ▶ Geoffrey Chang retractions 2006,
- ▶ fMRI correlation analysis 2005,
- ▶ Editorial Expression of Concern from *Science* in January 2010,
- ▶ more...

## Controlling Error is Central to Scientific Progress



“The scientific methods central motivation is the ubiquity of error - the awareness that mistakes and self-delusion can creep in absolutely anywhere and that the scientists effort is primarily expended in recognizing and rooting out error.”  
David Donoho et al. (2009)

## 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, using simulation and other data-intensive methods.



## Toward a Resolution of the Credibility Crisis

- ▶ Typical scientific communication doesn't include code, data, test suites.
- ▶ Most published computational science is near impossible to replicate.

**Thesis:** Computational science cannot be elevated to a third branch of the scientific method until it generates *routinely verifiable knowledge*. (Donoho, Stodden, et al. 2009)

Sharing of underlying code and data is a necessary part of this solution, enabling *Reproducible Research*.

## Question: How do we share computational work?

Goal: encourage reproducibility and verifiability, and permit others to build on the work.

Prototypical example, the Caltech-based DANSE project seeks to share neutron scattering data and code among researchers:

The screenshot shows a web browser window titled "Main Page - DANSE". The address bar contains the URL "http://wiki.cacr.caltech.edu/danse/index.php/Main\_Page". The page content includes a sidebar with a "DANSE" logo and navigation links for "main page", "restricted wiki", and "documentation". The main content area features a "Main Page" heading, a sub-heading "DANSE: Distributed Data Analysis for Neutron Scattering Experiments", and introductory text about the project's purpose and access policies. A "Log in" link is visible in the top right corner of the page.

article | discussion | edit | history

## Main Page

**DANSE: Distributed Data Analysis for Neutron Scattering Experiments** [edit]

This is the home page of the general information site for DANSE. The [Release Pages](#) for the DANSE products are at a different site. The structure of this wiki site follows the organization of the sidebar to the left of your browser window.

DANSE is a software development project on distributed data analysis for neutron scattering experiments. You are welcome to browse this site to find documentation on the software or neutron scattering, and to make comments in the public access pages. Anyone working on the DANSE project is encouraged to [request an account](#) and access to the editing capabilities of this MediaWiki.

main page  
restricted wiki  
documentation

- Science
- Common Scientific Algorithms

Log in

18 / 31

# Surveying the Machine Learning Community (Stodden 2010)

**Question:** Why isn't reproducibility practiced more widely?  
Answer builds on literature of free revealing and open innovation in industry, and the sociology of science.

**Hypothesis 1:** Scientists are motivated to share or not share work by perceptions of personal gain or loss.

**Hypothesis 2:** The willingness to reveal work reflects a scientists desire to belong to a community and gain feedback on work.

- ▶ Sample: American academics registered at the Machine Learning conference NIPS.
- ▶ Respondents: 134 responses from 593 requests (~23%).

## Top Reasons Not to Share

Code		Data
77%	Time to document and clean up	54%
52%	Dealing with questions from users	34%
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%
30%	Competitors may get an advantage	33%
20%	Web/Disk space limitations	29%



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

## Top Reasons to Share

Code		Data
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%

## Findings

Not surprising:

- ▶ Reasons for not revealing reflect private incentives.
- ▶ Reasons for revealing include community membership and opportunities for feedback.

Several surprises:

- ▶ Computational scientists motivated to share by communitarian ideals.
- ▶ Computational scientists not that worried about being scooped.
- ▶ Computational scientists quite worried about Intellectual Property issues when sharing data and code.
- ▶ Attribution matters for those who share vs those who do not share.

## Legal Barriers to Reproducibility: Copyright

To promote the Progress of Science and useful Arts, by securing for limited Times to Authors and Inventors the exclusive Right to their respective Writings and Discoveries. (U.S. Const., art. I, §8, cl. 8)

- ▶ Original expression of ideas falls under copyright *by default* (papers, code, figures, tables..)
- ▶ Copyright secures exclusive rights vested in the author to:
  - ▶ reproduce the work
  - ▶ prepare derivative works based upon the original
  - ▶ limited time: generally life of the author + 70 years

Exceptions and limitations: Fair Use: "the fair use of a copyrighted work . . . for purposes such as criticism, comment, news reporting, teaching (including multiple copies for classroom use), scholarship, or research, is not an infringement of copyright." 17 U.S.C. §107.



## Responses Outside the Sciences 1: Open Source Software

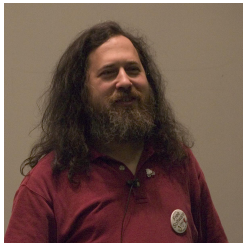
Software with licenses that communicate alternative terms of use to code developers, rather than the default assigned by copyright law.

Richard Stallman created the GNU Public License (GPL) in 1989 to ensure distribution of source code, with compiled programs. Majority of open source code under GPL.

Since then hundreds of software licenses have been created with varying terms:

- ▶ (Modified) BSD license
- ▶ MIT license
- ▶ Apache 2.0
- ▶ "Lesser" GPL v3
- ▶ ... (see <http://www.opensource.org/licenses/alphabetical>)

# Open Source Software: The Movement



## Free Software Foundation

- ▶ Richard Stallman, Founder, 1985
- ▶ "the leading civil liberties group defending your rights in the digital world."

## Responses Outside the Sciences 2: Creative Commons



Larry Lessig, Founder, 2001

- ▶ Adapts Open Source Software approach to artistic and creative works
- ▶ Provides a suite of licenses:
  - ▶ 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.

## Response from Within the Sciences: The Reproducible Research Standard (Stodden 2009)

- ▶ Remove copyright's barrier to reproducible research,
- ▶ Realign the IP framework with longstanding scientific norms.  
A suite of license recommendations for computational science:
  1. Release media components (text, figures) under CC BY,
  2. Release code components under Modified BSD or similar,
  3. Release data to public domain (CC0) or attach an attribution license.

Winner of the Access to Knowledge Kaltura Award in 2008.

## Releasing Data?

- ▶ Raw facts not copyrightable.
- ▶ Original “selection and arrangement” of these facts is copyrightable. (Feist Publns Inc. v. Rural Tel. Serv. Co., 499 U.S. 340 (1991)).
- ▶  $\implies$  the possibility of a residual copyright in data (attribution licensing or public domain certification).
- ▶ Law doesn't match reality on the ground: What constitutes a “raw” fact?

## Benefits and Difficulties of the 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

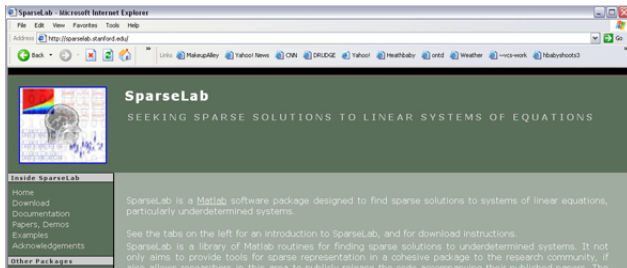
### Difficulties:

- ▶ Massive codes, software support, streaming data,...
- ▶ Tools for ease of implementation (ie. data provenance and workflow),
- ▶ “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.

## Publishing, SparseLab, and Reproducible Research

*SparseLab*: a MATLAB toolbox that makes software solutions for sparse systems available.

- ▶ A platform for code/data sharing: 13 papers and 12 authors.
- ▶ Standardized tools could advance the research community;
- ▶ Demos, exercises, documentation, download and install script, acknowledgments, guidance for contributors included;
- ▶ Over 7000 downloads in 2008.



## Conclusions

1. Massive computation revolutionizing scientific research, including quantitative social science.
2. New paradigm(s) for publication and verification of results: legal standard and open platforms.
3. Questions emerging regarding adherence to the scientific method, and replicability of our published computational results.
4. Barriers to reproducibility, including Copyright.
5. New directions for improving reproducibility: e.g. software development for provenance and workflow tracking; citation standards; funder and journal requirements.



## References:

- ▶ “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,”
- ▶ “The Scientific Method in Practice: Reproducibility in the Computational Sciences”

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

Data and Code Sharing Roundtable, Nov 2009:  
[http://www.stanford.edu/~vcs/Conferences/  
RoundtableNov212009/](http://www.stanford.edu/~vcs/Conferences/RoundtableNov212009/)