Scientific Integrity and Reproducibility: Data and Code Sharing

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

New England Statistics Symposium
Harvard University
April 17, 2010

Scientific Research is Changing

- Scientific computation becoming central to the scientific method,
  - Changing how research is conducted in many fields,
  - Changing the nature of how we learn about our world.

Relaxed practices regarding the communication of computational detail is creating a credibility crisis: Climategate 2009, Geoffrey Chang retractions 2006, fMRI correlation analysis 2005, Editorial Expression of Concern from *Science* in January 2010...

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)

Examples of Pervasiveness of Computational Methods

- Our own field:

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);
- Reaches into traditionally nonquantitative fields: Wordhoard project at Northwestern examining word distributions by Shakespearian play.

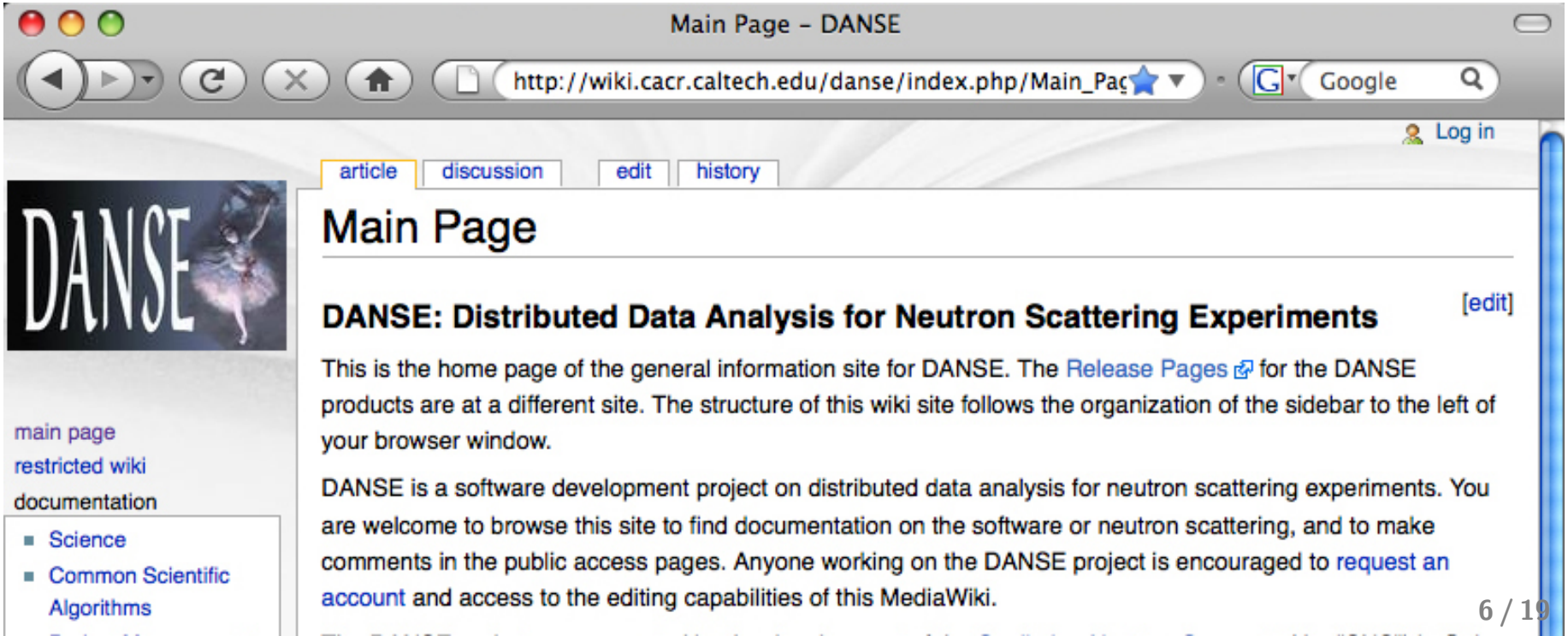
Examples of The Changing Nature of Scientific Discovery

- Climate Simulation: Community Climate Models (e.g. NCAR),
- High Energy Physics: LHC, Astrophysics Simulation Collaboratory (Washington U),
- Macromolecule Modeling: SaliLab UCSF, Doug Lauffenberger MIT
- Mathematical Proof by simulation and exhaustive grid search,
- more..

Question: How do we share this work?

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

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



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 ($\sim$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

Copyright is a near-ubiquitous barrier to reproducibility

- 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 in 2001 by to make it easier for artists to share and use creative works
  - Provides a suite of licenses that allows the author to determine terms of use attached to works
  - Apply to scientific works?

Reproducible Research Standard

Reproducible Research: underlying code and data available such that published results can be replicated by a researcher in the field.

The Reproducible Research Standard (Stodden 2009) addresses transparency and reproducibility issues in computational science:

- to realign IP rights with scientific norms,
- reinstate reproducibility in computational science,
- notion of research compendium (Gentleman and Lang 2004)
- release of scientific research, including code and data, for verification, wide reuse, development...

To satisfy the Reproducible Research Standard, use attribution-only licensing on text, figures, code, and certify data in the public domain.

Legal Nuts and Bolts for Open Reproducible Science

1. Creative Commons Attribution license (CC BY) on media such as text, figures,
2. Attribution license on code: such as Apache 2.0, MIT, LGPL,
3. Data under CC0 or Science Commons Open Access Data Protocol,
4. "original selection and arrangement" of the data, under CC BY or attribution open source license.

Open Problem: ownership details to be sorted out in various cases (e.g. confidentiality issues, proprietary data/code..).

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)).
- $\Longrightarrow$ 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 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

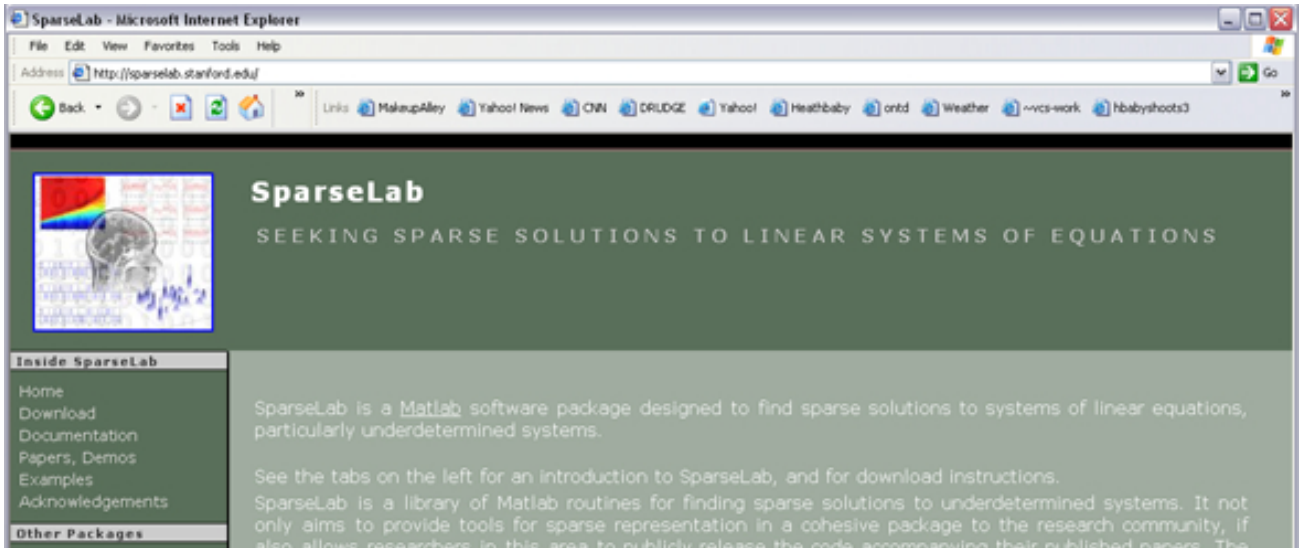
Difficulties of the RRS

- Massive codes, software support, streaming data,...
- Tools for ease of implementation (ie. data provenance and workflow: "Accessible Reproducible Research" (Mesirov, Science, 2010)),
- citation and micro-citation for code and data contributions,
- "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. Survey results to understand barriers to wide sharing of data and code underlying published results.
4. New directions for improving reproducibility: e.g. software development for provenance and workflow tracking; citation standards; funder and journal requirements.