Computational Research and the Scientific Method: A Third Branch?

Victoria Stodden
Yale Law School and Science Commons
vcs@stanford.edu

Statistics Seminar
Department of Statistics, Yale University
March 1, 2010

Agenda

1. Scientific research is being transformed by massive computation
2. Credibility crisis through lack of reproducibility
3. Survey results on barriers to sharing of scientific work
4. Solution to IP barriers: the *Reproducible Research Standard*
5. Data and Code Sharing Roundtable, Nov 2009

Transformation of the 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.

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

The Third Branch of the Scientific Method

- Branch 1: *Deductive/Theory*: e.g. mathematics
- Branch 2: *Empirical*: e.g. statistical data analysis of controlled experiments

- Simulation? Data-driven research?

Controlling Error is Central to Scientific Progress

"The scientific method's central motivation is the *ubiquity of error* - the awareness that mistakes and self-delusion can creep in absolutely anywhere and that the scientist's effort is primarily expended in recognizing and rooting out error." David Donoho et al. (2009)

Emerging Credibility Crisis in Computational Science

- Error control seems to be forgotten
- Typical scientific communication doesn't include code, data.
- Published computational science near impossible to replicate.
- Scientific method:
  - Replicability necessary for a discovery to be accepted as a contribution to the stock of knowledge.
  - Control over error a hallmark.

2nd Transformation: 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, journals…

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

- 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 Data and Code Sharing: Survey

Hypotheses:

1. Scientists are primarily motivated by personal gain or loss.

2. Scientists are primarily worried about being scooped.

Survey of Computational Scientists

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

Reported Sharing Habits

- 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.

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%

For example..

"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%

Preliminary Findings

- *Surprise*: Motivated to share by communitarian ideals.
- *Not surprising*: Reasons for not revealing reflect private incentives.
- *Surprise*: Scientists not that worried about being scooped.
- *Surprise*: Scientists quite worried about IP issues.

Legal Barriers to Reproducibility

- Original expression of ideas falls under copyright by default (written expression, code, figures, tables..)

- Copyright creates exclusive rights vested in the author to:
  - reproduce the work
  - prepare derivative works based upon the original
  - Exceptions and limitations: Fair Use, Academic purposes

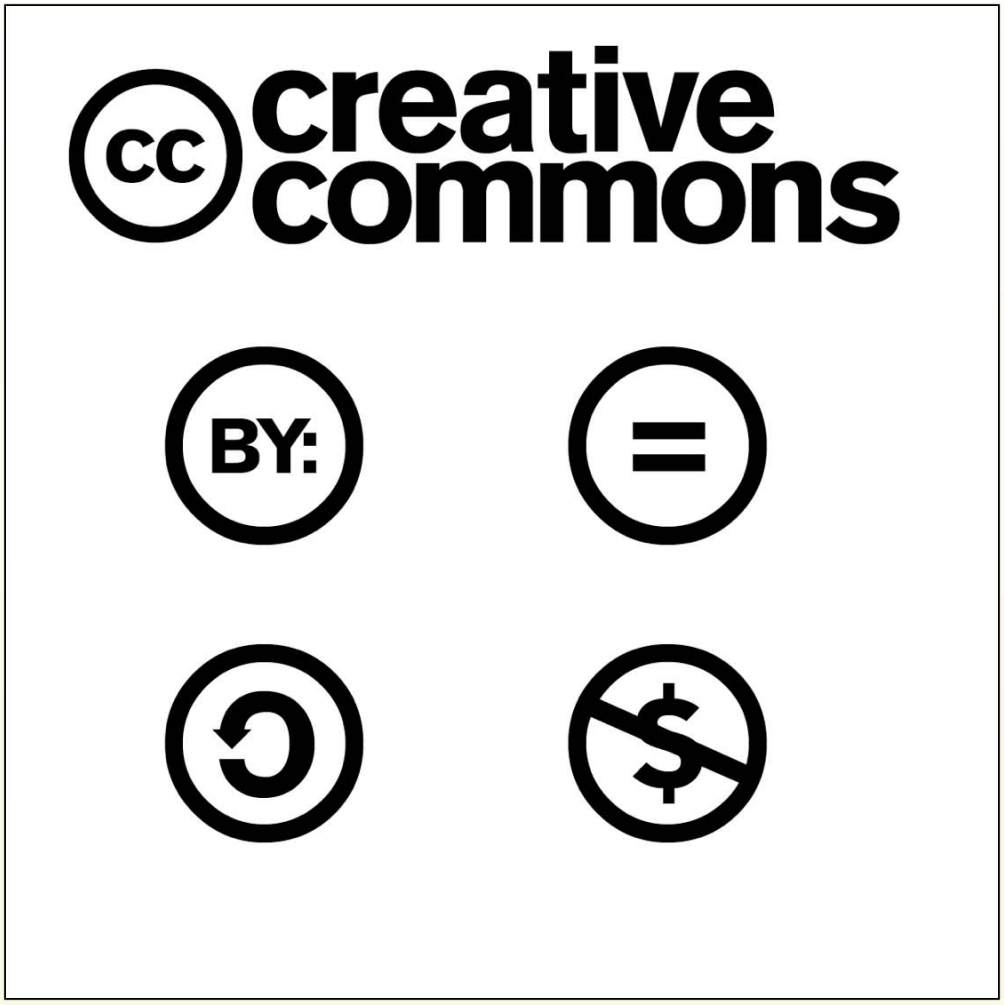
Creative Commons

- 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.

License Logos



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 framework 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.

“ShareAlike” Inappropriate

“ShareAlike”: licensing provision that requires identical licensing of downstream libraries,

Issue 1: Control of independent scientists’ work,

Issue 2: Incompatibility of differing licenses with this provisions.

=> GPL not suitable for scientific code.

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…

Reproducibility is an Open Problem

- 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
  - Provenance and workflow tracking tools (Mesirov)

- Tools for attribution:
  - Generalized contribution tracking
  - Legal attribution/license tracking tracking and search (RDFa)

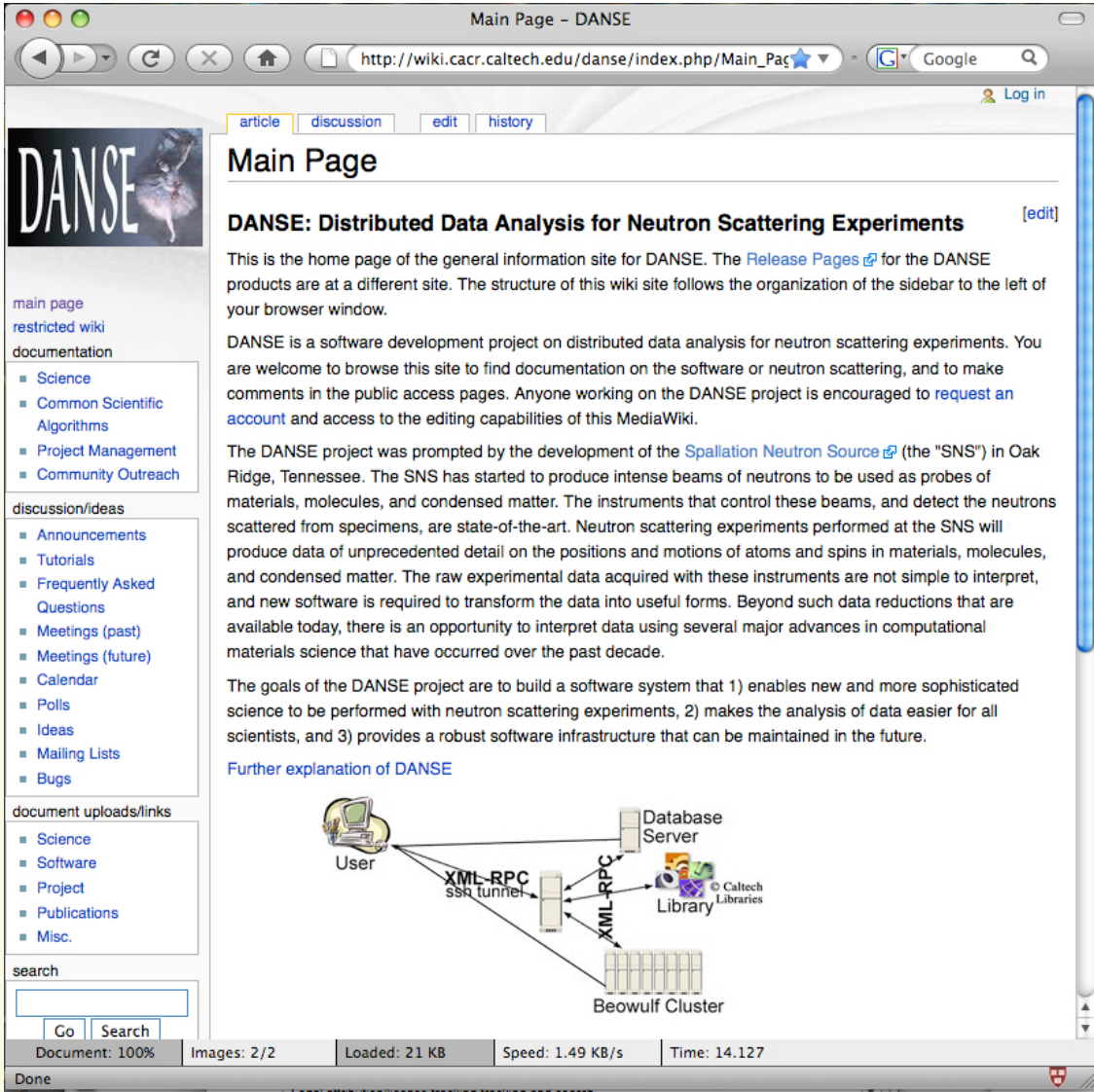
Case Study: mloss.org

- Machine Learning Open Source Software
- Active code repository

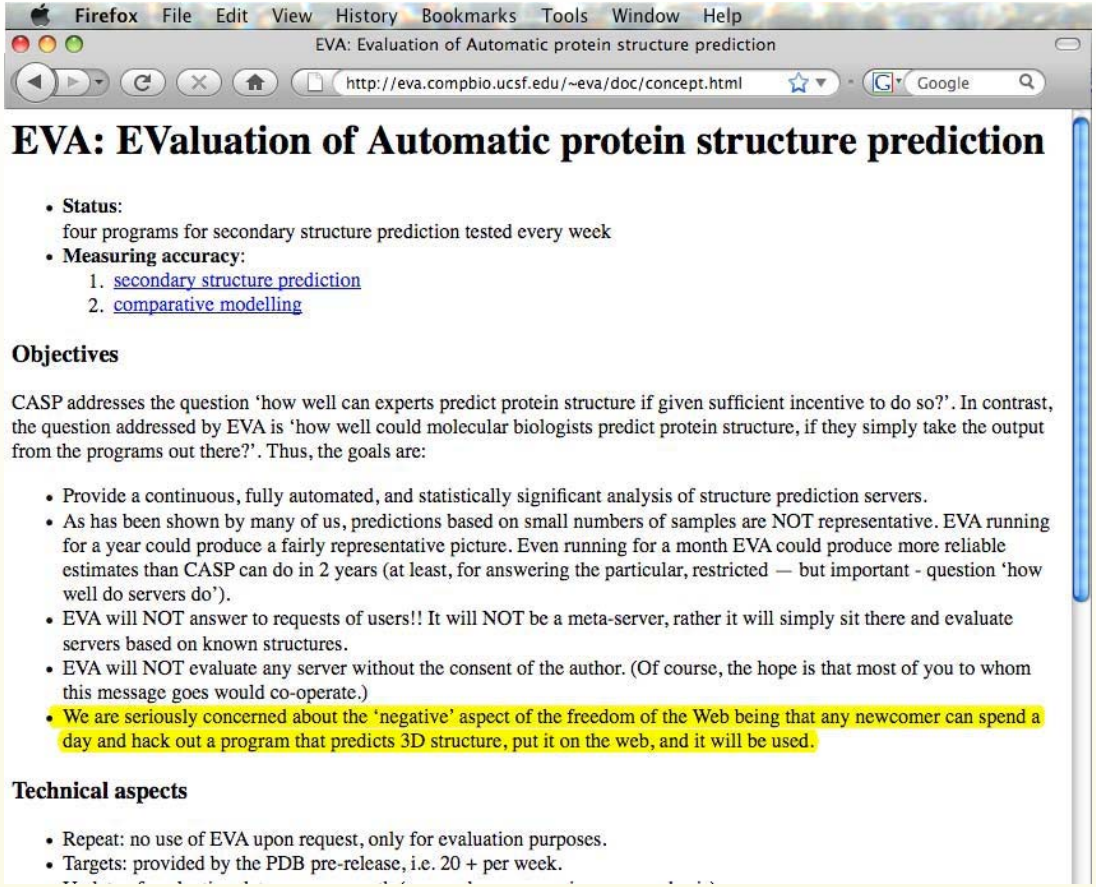
- Code release at least as important as data release
- Open question: software support

Case Study: DANSE

- Neutron scattering
- Make new data available
- Unify software for analysis



Openness and Taleb's Criticism



- Open Access movement removes the notion of a scientific community

Real and Potential Wrinkles

- Reproducibility neither necessary nor sufficient for correctness, but essential for dispute resolution,
- Software "lock-in" and the evolution of scientific ideas (standards lock-in),
- 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

Data and Code Sharing Roundtable, Nov 21 2009

Followed spirit of International strategy meetings on Human DNA Sequencing:

- Bermuda 1996
- Fort Lauderdale 2003
- Amsterdam 2008
- Toronto 2009

Goal: Discussion of reproducibility issues across the computational sciences broadly; Declaration of Best Practices.

Data and Code Sharing Roundtable

- 29 researchers, faculty, funders, journal representatives
- 4 sessions:
- Problem Framing
- Legal Barriers
- Computational Solutions
- Norms and Incentives
- Submission of thought pieces on reproducibility,
- Jointly composed list of recommended practices and “dream goals.”

Papers and Links

- "The Scientific Method in Practice: Reproducibility in the Computational Sciences"
- "15 Years of Reproducible Research in Computational Harmonic Analysis"
- "Enabling Reproducible Research: Open Licensing for Scientific Innovation"

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

http://www.stanford.edu/~vcs/Conferences/RoundtableNov212009/