Open Licensing and Scientific Reproducibility

Victoria Stodden
Information Society Project @ Yale Law School
<vcs@stanford.edu>

Commons-based Peer Production
UC Berkeley School of Information
April 28, 2010

Agenda

1. The Scientific Backdrop: The Onslaught of Massive Computation
2. Reproducibility Needed to Comply with the Scientific Method
3. Survey: Barriers to Open Code/Data
4. Untangling Intellectual Property Issues
5. Historical Legacies in Open Science

Science is Changing

- Scientific computing emerging as central to the scientific enterprise
  - 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 details is creating a credibility crisis:
  - Climategate 2009, Geoffrey Chang retractions 2006, fMRI correlation analysis 2005, Editorial Expression of Concern from Science in January 2010…

Scientific Publication is Changing

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

Example: Community Climate System Model (CCSM)

- Collaborative system simulation
- Code available by permission
- Data output files by permission

Example: High Energy Physics

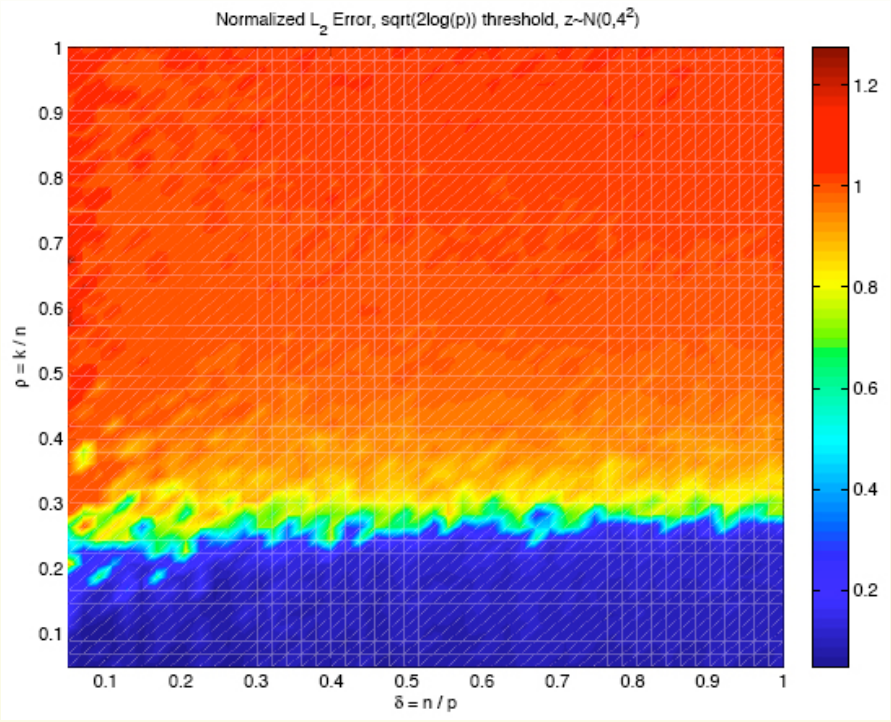
- 4 LHC experiments at CERN: 15 petabytes produced annually
- Data shared through grid to mobilize computing power
- Director of CERN (Heuer): "Ten or 20 years ago we might have been able to repeat an experiment.They were simpler, cheaper and on a smaller scale. Today that is not the case. So if we need 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

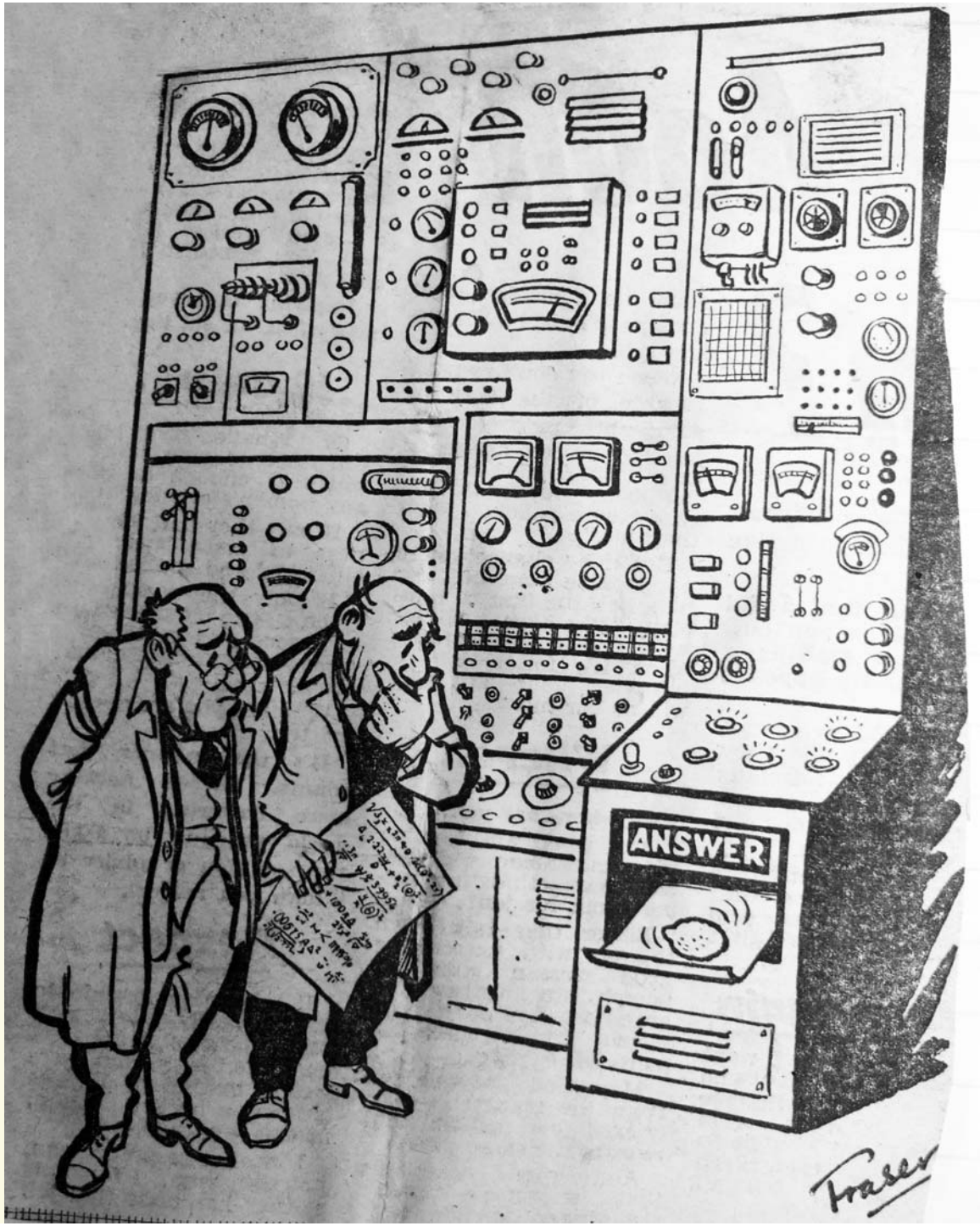
- Data and code sharing within community
- Interface for dynamic simulation
- mid 1930's: calculate the motion of cosmic rays in Earth's magnetic field..

Example: Proofs

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



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


ANSWER
Fraser

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)

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?

Emerging Credibility Crisis in Computational Science

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

- Thesis: Accession to 3rd branch of the scientific method involves the production of *routinely verifiable knowledge*.

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)

Barriers to Sharing: Survey

Hypotheses:

1. Scientists are motivated to share or not share work by perceptions of personal gain or loss.
2. The willingness to reveal work reflects a scientist's desire to belong to a community and gain feedback on work.

Survey of Computational Scientists

- *Subfield*: Machine Learning
- *Sample*: American academics registered at top Machine Learning conference (NIPS).
- *Respondents*: 134 responses from 593 requests.

Top Reasons Not to Share

Code		*Data*
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

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

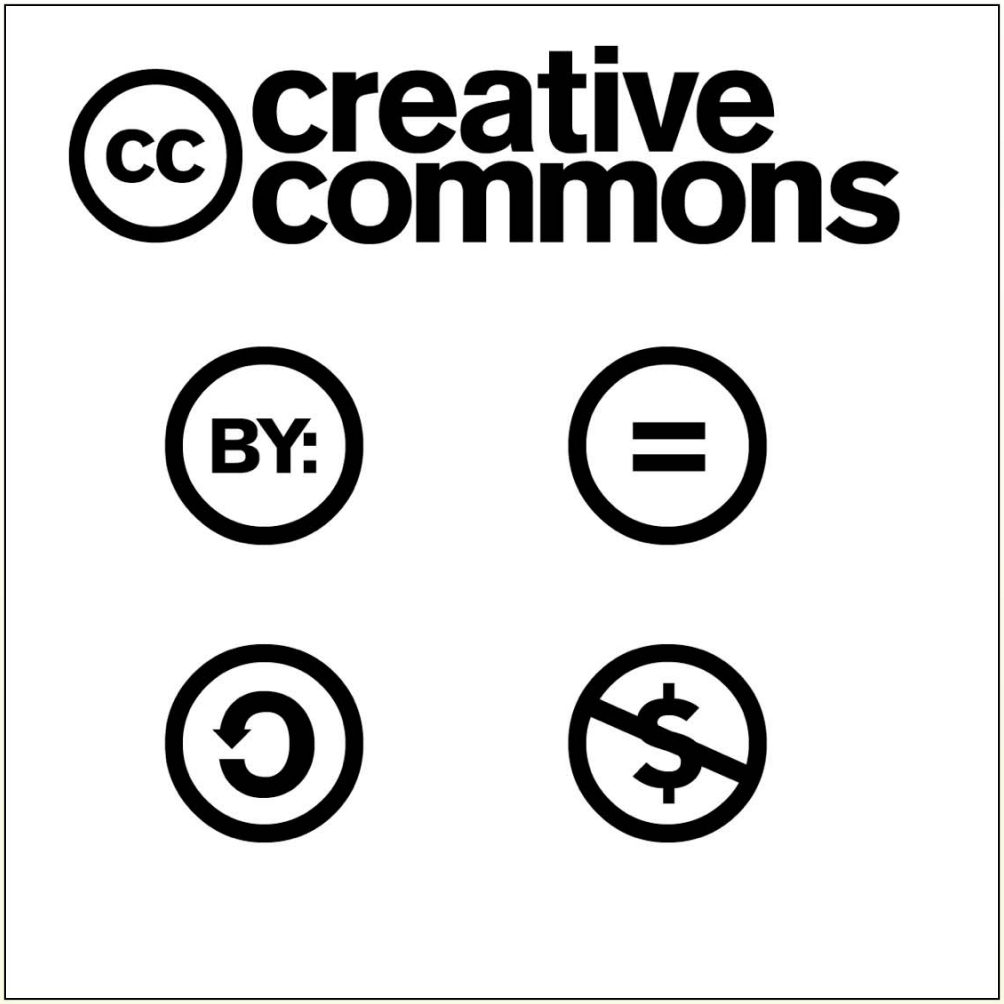
Creative Commons

- 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 provision.
- GPL not suitable for scientific code.

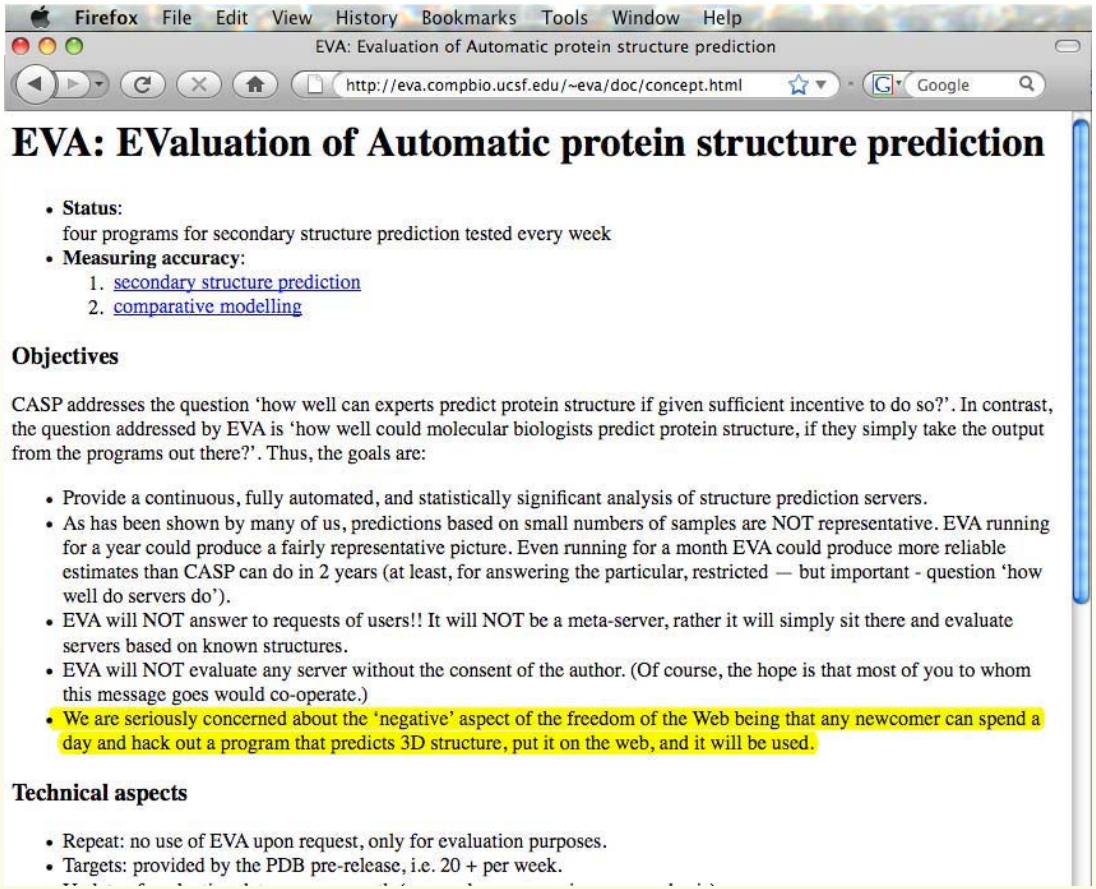
Releasing Data?

- 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))
- (what is a “raw fact”?)

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…

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

Legacy Issues

- Leadership in open science from bioinformatics, esp genome community
- Open Data Declarations from 1996
- Most recent: “Prepublication Data Sharing,” The Toronto International Data Release Workshop Authors, Nature, 461(10), Sept 2009.
- Yale Workshop on Data and Code Sharing, Nov 2009.

Papers and Links

- "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

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