The Future of Computational Science:
Information Sharing and Reproducibility

Victoria Stodden
vcs@stanford.edu

Berkman Center for Internet and Society
March 31, 2009

Agenda

1. Scientific research is being transformed by massive computation
2. Credibility crisis through lack of reproducibility
3. Copyright: a barrier to sharing of scientific work
4. Solution: the *Reproducible Research Standard*

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.

The Third Branch of the Scientific Method

- Computation transforming scientific enterprise; analysis beyond idealized cases
- Branch 1: *Deductive/Theory*: e.g. mathematics
- Branch 2: *Empirical*: e.g. statistical data analysis of controlled experiments

Emerging Credibility Crisis in Computational Science

- Error control seems to be forgotten
- Published computational science typically 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.
  - Established in first two branches.

2nd Transformation: Changes in Scientific Communication

- Internet: communication of all computational research details/data is now possible
- Scientists often post papers but not their complete body of research

- Changes coming…

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.

Problem: Legal Barriers to Sharing

- 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, for example:
  - 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

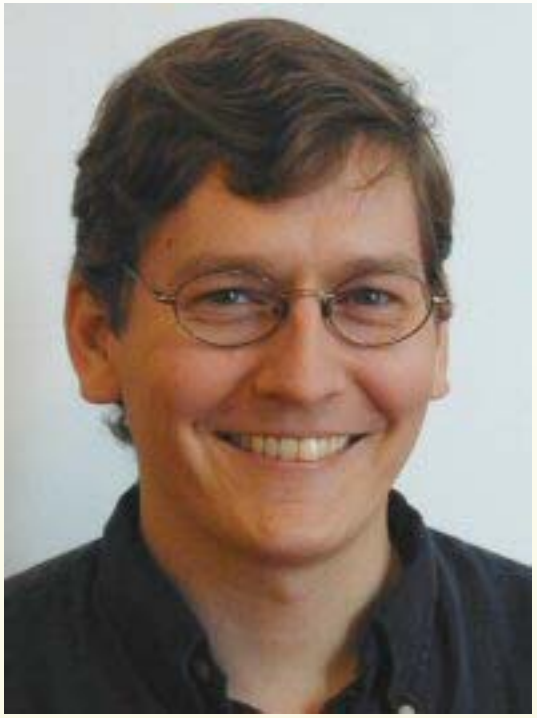
Realignment of legal rights with scientific norms:

- Release media components (text, figures) under CC BY.
- Release code components under Modified BSD, MIT, Apache 2.0 or LGPL.
- These licenses free the scientific work of copying and reuse restrictions and have an attribution component.

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))

The RRS and Science Commons



- Science Commons, a Creative Commons project, is headed by John Wilbanks
- Joint work to establish the RRS as a Science Commons standard (webpage, logo).
- Researchers can “brand” their work as reproducible

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

Real and Potential Wrinkles

- Attribution:
  - Legal attribution and academic citation not isomorphic
- Need for individual scientist to act
- Platforms for revealing?
- Openness standards

and...

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

Draft of this work

"Enabling Reproducible Research: Open Licensing for Scientific Innovation"

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

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… .”