

Opening Government Funded Knowledge Creation: Increasing Transparency and Scientific Integrity

Victoria Stodden

Information Society Project @ Yale Law School

<vcs@stanford.edu>

Gov 2.0 Expo

Washington, DC

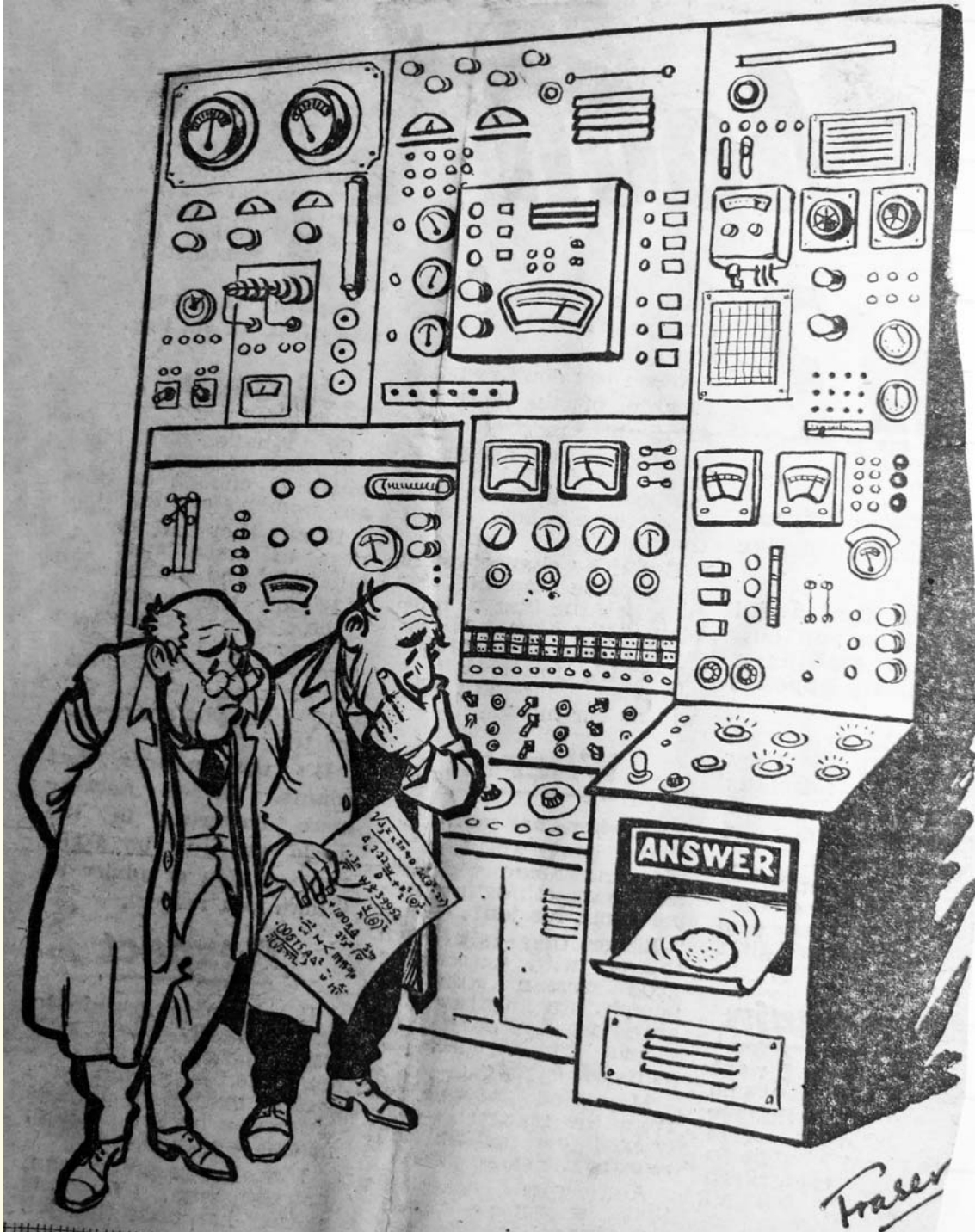
May 26, 2010

Agenda

1. Public Policy Crisis: Evidence-based Policy and Massive Computation
2. Reproducibility of Scientific Results a Necessary Response
3. Barriers to Open Science
4. Government Action on Transparency in Science
5. Where we need to go

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...



Tracy

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)

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

<i>Code</i>		<i>Data</i>
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

<i>Code</i>		<i>Data</i>
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%

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



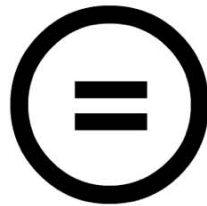
- Founded by Larry Lessig to make it easier for artists to share and use creative works
- A suite of licenses that allows the author to determine terms of use attached to works

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

 **creative
commons**



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.

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...

Government Response

1. Federal Research Public Access Act (Open Access)
2. Data plans for NSF grants
3. NSF Task Force on Cyberscience and Engineering, Grand Challenge Communities, and Virtual Organizations Report

Future Directions

Open Science involves balancing incentives:

- Funding Agency Requirements
- Tenure/University/Award Requirements
- Journal Requirements
- Scientific Requirements

Computational research output as a
compendium of code, data, and paper.

Papers and Links

[<vcs@stanford.edu>](mailto:vcs@stanford.edu)

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