Data and Code Sharing in Computational Science

Victoria Stodden

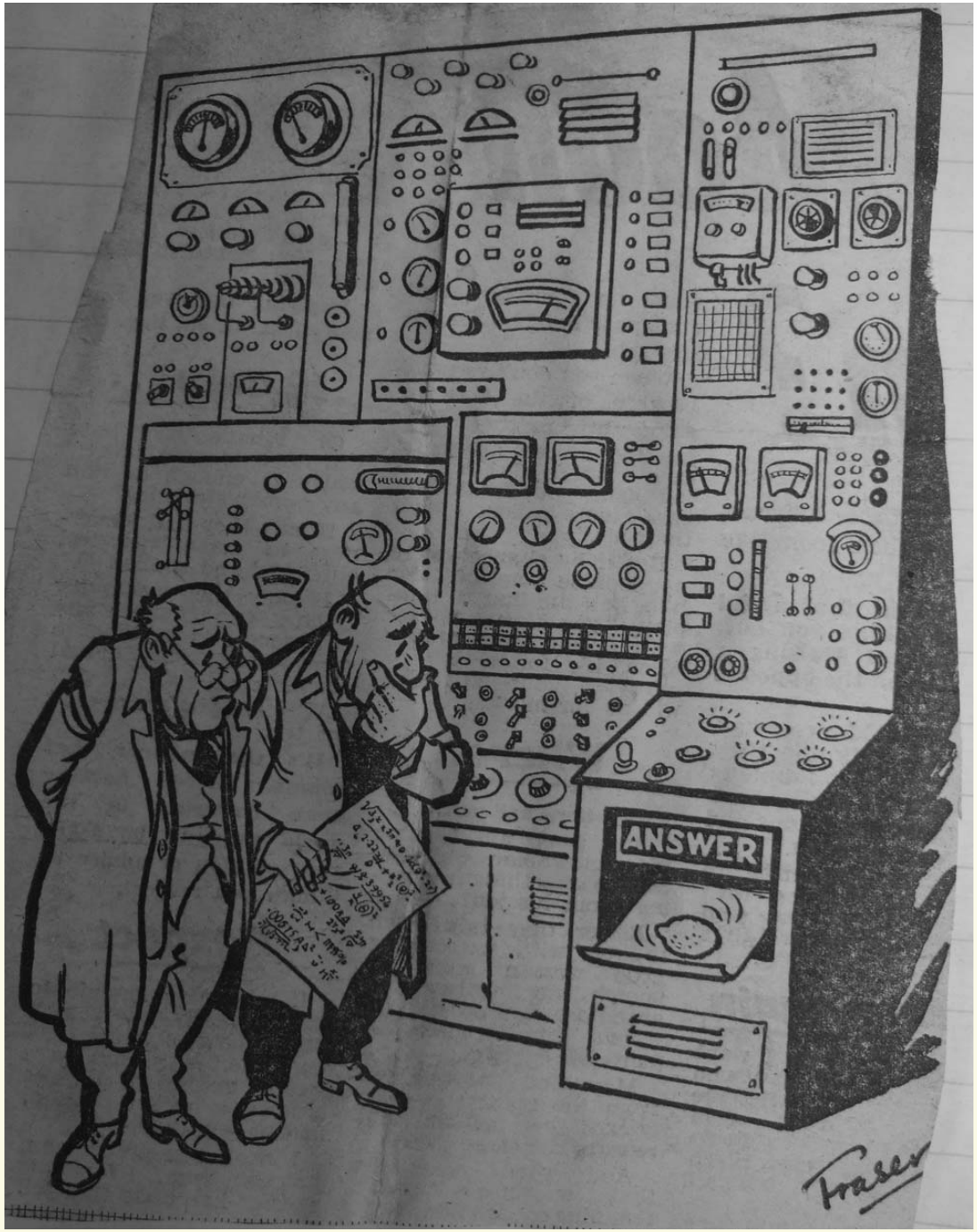
Information Society Project,

Yale Law School

vcs@stanford.edu

Yale Law School

November 21, 2009


ANSWER
Fraser

Reproducibility

- (Simple) 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.

Survey of Computational Scientists

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

Have you been scooped?

Idea Theft	Count	Proportion
At least one publication scooped	53	0.51
2 or more scooped	31	0.30
No ideas stolen	50	0.49

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.

Barriers to Sharing: Legal

- 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:
  - 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 or similar.
- Both licenses free the scientific work of copying and reuse restrictions and have an attribution component.

Releasing Data?

- Raw facts alone generally 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…

Papers

- "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