Applying the Creative Commons Philosophy to Scientific Innovation

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

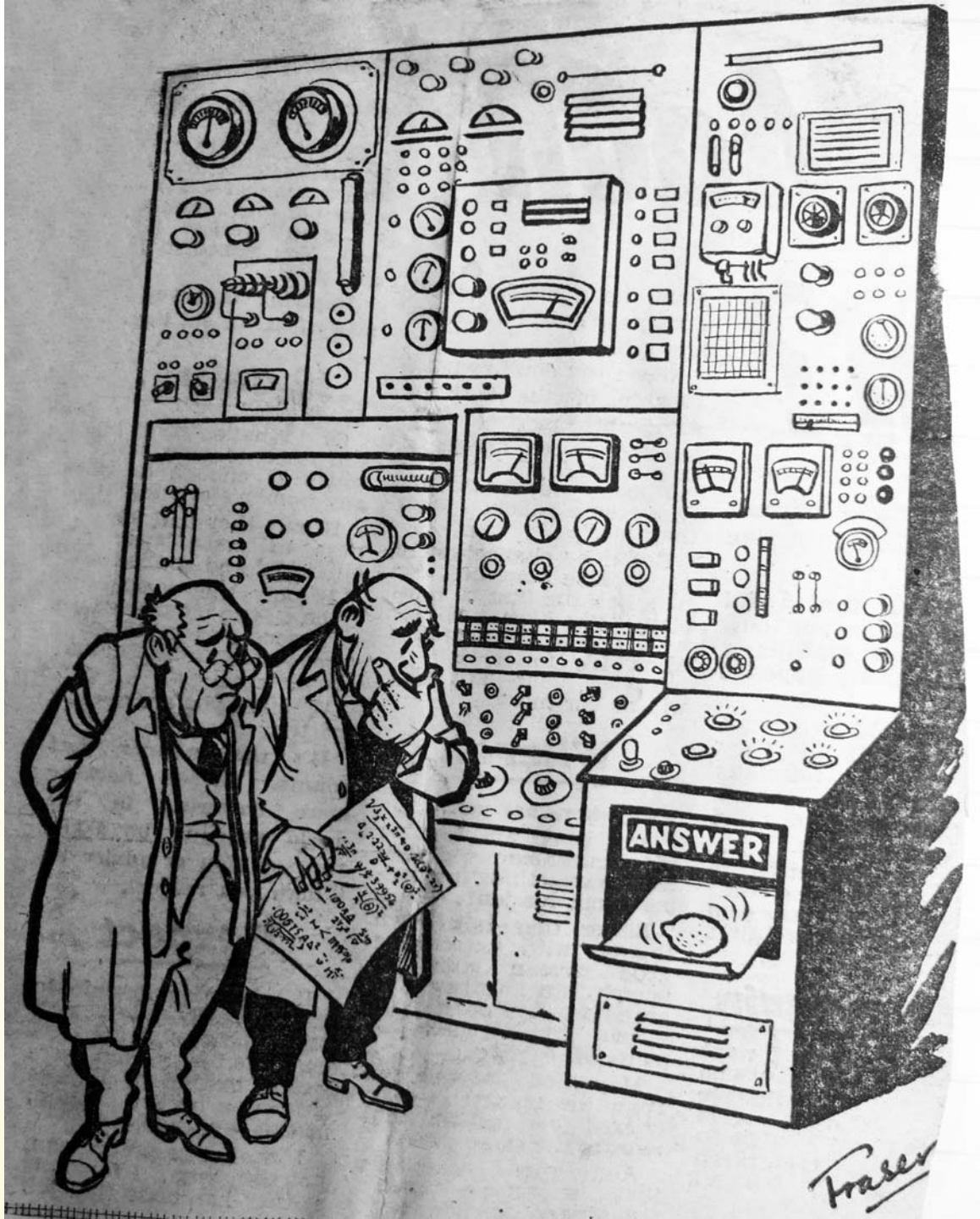
Acesso Livre à Informação Científica
Reitoria UNL - Campolide, Lisbon
February 11, 2010

Agenda

1. Creative Commons Ideas
2. Facilitating Reproducibility in Computational Science
3. Untangling Intellectual Property Issues

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


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)

Reproducibility

- Computational science: researcher works with code or data in generating published results.
- Reproducibility: the ability of others to recreate and verify computational results, given appropriate software and computing resources.

Science is Changing

- A transformation of the scientific enterprise through massive computation, in scale, scope, and pervasiveness, is currently underway..

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

Example: Community Climate System Model (CCSM)

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

Example: High Energy Physics

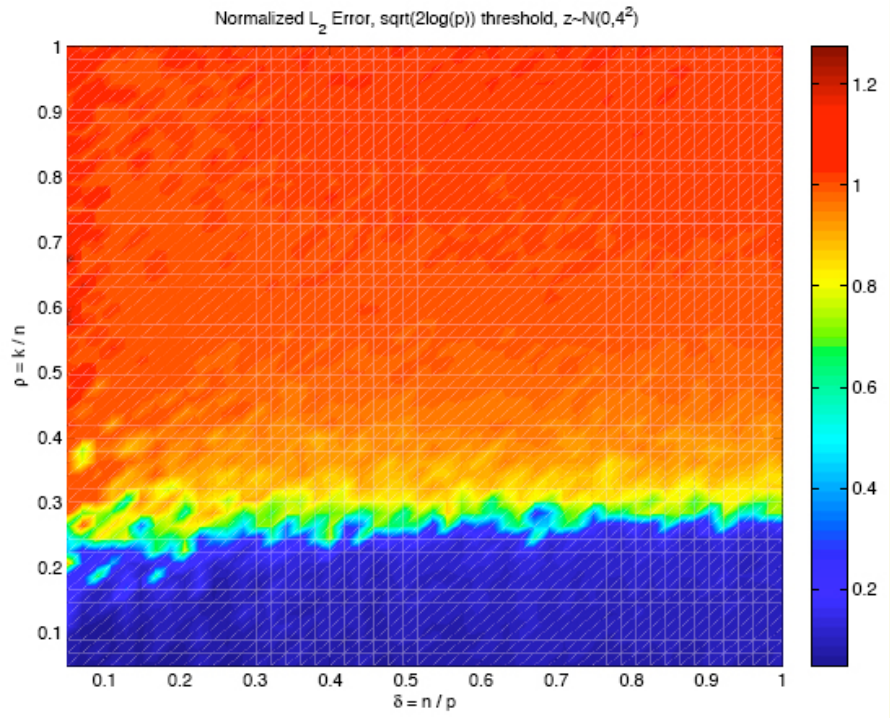
- 4 LHC experiments at CERN: 15 petabytes produced annually
- Data shared through grid to mobilize computing power
- Director of CERN (Heuer): "Ten or 20 years ago we might have been able to repeat an experiment.They were simpler, cheaper and on a smaller scale. Today that is not the case. So if we need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

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.

- 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)

Legal Barriers to Reproducibility

- In the US, 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

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 (and scale matters)

- Simple case: open data and small scripts. Suits simple definition.
- Hard case: Inscrutable code, organic programming.
- Harder case: massive computing platforms, streaming data.

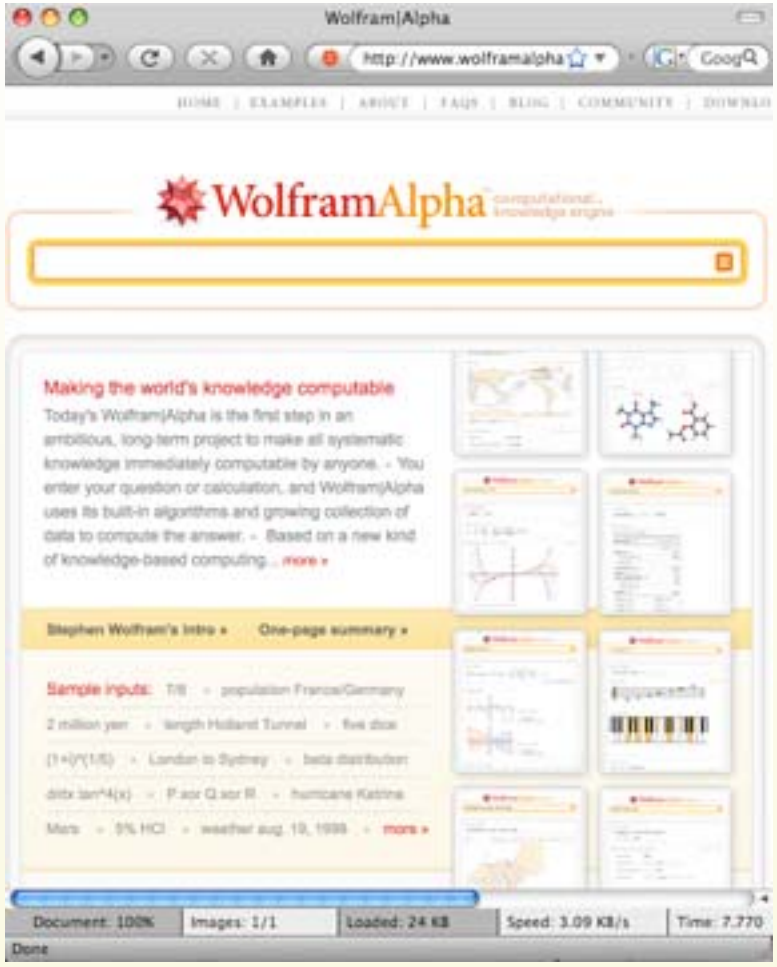
- Can we have reproducibility in the hard cases?

Solutions for Harder Cases

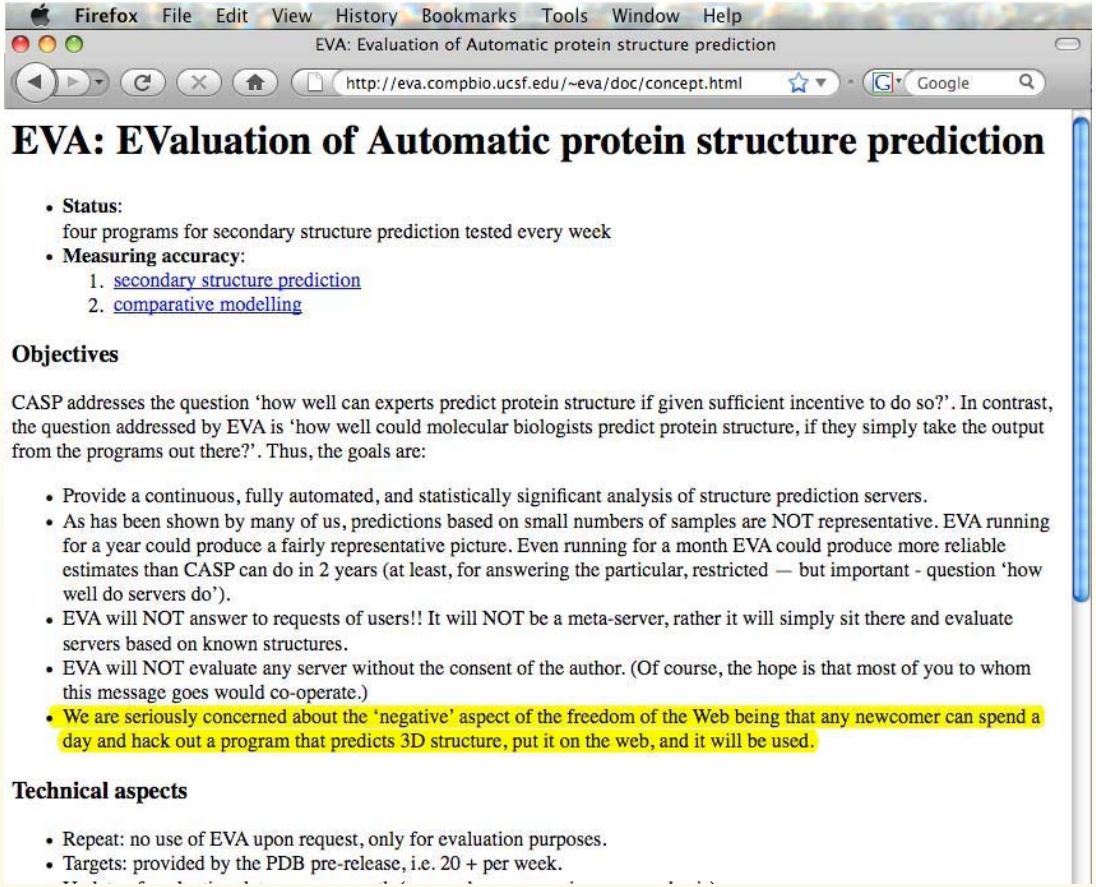
- Tools for reproducibility:
  - Standardized testbeds, automatic code checking
  - Open code for continuous data processing, flags for “continuous verifiability”
  - Standards and platforms for data sharing
  - Provenance and workflow tracking tools (Mesirov)
- Tools for attribution:
  - Generalized contribution tracking
  - Legal attribution/license tracking tracking and search (RDFa)

Case Study: Wolfram|Alpha

- Obscure code - testbeds for verifiability
- Dataset construction methods opaque
- (claims copyright over outputs)



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
- 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

Papers and Links

- "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

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

Appendix: Attribution

- Legal attribution and academic citation not isomorphic.

- Minimize administrative burden
- Evolving norms / field specific norms / technology

- CC-BY: “keep intact all copyright notices for the Work and provide, reasonable to the medium or means You are utilizing… .”