Funding Agency Policy and the Credibility Crisis in Computational Science

Victoria Stodden
Department of Statistics
Columbia University

Applied Mathematics Perspectives
Reproducible Research Workshop: Community Forum
Renaissance Hotel, Vancouver, British Columbia
July 16, 2011

Funding Agency Policy

- NSF grant guidelines: "NSF ... expects investigators to share with other researchers, at no more than incremental cost and within a reasonable time, the data, samples, physical collections and other supporting materials created or gathered in the course of the work. It also encourages grantees to share software and inventions or otherwise act to make the innovations they embody widely useful and usable."

- NSF peer-reviewed Data Management Plan, January 2011.

- NIH (2003): "The NIH endorses the sharing of final research data to serve these and other important scientific goals. The NIH expects and supports the timely release and sharing of final research data from NIH-supported studies for use by other researchers." (>$500,000, include data sharing plan)

Open Questions

- What is the next step for data management plans?

- Are there other policies that funding agencies should consider?

- Aside from classified work, should facilitative policies differ by funding agency, or any other criteria?

Updating the Scientific Method

Donoho and others have argued that computation presents only a *potential* third branch of the scientific method:

- Branch 1 (deductive): mathematics, formal logic,

- Branch 2 (empirical): statistical analysis of controlled experiments,

- Branch 3? (computational): large scale simulations.

The Ubiquity of Error

- The central motivation for the scientific method is to root out error:

  - Deductive branch: the well-defined concept of the proof,

  - Empirical branch: the machinery of hypothesis testing, structured communication of methods and protocols.

- Computational science as practiced today does not generate reliable knowledge.

- *Computational science must develop standards for reproducibility before it can be considered a third branch of the scientific method,*

  ➡ Data and Code Sharing with publication.

Supplemental Slides

Policy: Bayh-Dole Act

- Bayh-Dole Act (1980), designed to promote the transfer of academic discoveries for commercial development, via licensing of patents.

- Legislators blind to the coming digital revolution, impact on software and algorithm patenting. Tech Transfer Offices and code release.

- Implications for science as a disruptor of openness norms:

  - patents => delay in revealing code, or closed code,
  - I assert Bilski => obfuscation of methods submitted for patents,
  - altering a scientist's incentives toward commercial ends.

Policy: America COMPETES

- America COMPETES Re-authorization (2011):

    - § 103: Interagency Public Access Committee:

      "coordinate Federal science agency research and policies related to the dissemination and long-term stewardship of the results of unclassified research, *including digital data* and peer-reviewed scholarly publications, supported wholly, or in part, by funding from the Federal science agencies." (emphasis added)

    - § 104: Federal Scientific Collections: OSTP "shall develop policies for the management and use of Federal scientific collections to improve the quality, organization, *access, including online access*, and long-term preservation of such collections for the benefit of the scientific enterprise." (emphasis added)

Computation Emerging as Central to the Scientific Endeavor

JASA June	Computational Articles	Code Publicly Available
1996	9 of 20	0%
2006	33 of 35	9%
2009	32 of 32	16%
2011	29 of 29	21%

- Data and code typically not made available at the time of scientific publication, rendering results unverifiable, not reproducible.

  ➡ A *Credibility Crisis* (ClimateGate, Duke Clinical Trials,...)

Response from Within the Sciences

The *Reproducible Research Standard (RRS)* (Stodden, 2009)

- A suite of license recommendations for computational science:

    - Release media components (text, figures) under CC BY,
    - Release code components under Modified BSD or similar,
    - Release data to public domain or attach attribution license.

  ➡ Remove copyright's barrier to reproducible research and,

  ➡ Realign the IP framework with longstanding scientific norms.

Winner of the Access to Knowledge Kaltura Award 2008

Yale Data and Code Sharing Roundtable 2009

- Roundtable on Data and Code Sharing in computational science Nov 21, 2009:

  - gathered 30 computational scientists from a variety of fields, funding agency folks, publishers, librarians, university policy makers, lawyers...

  - Draft Position Statement (published in IEEE Computing in Science and Engineering, Sep/Oct 2010)

  - recommendations for stakeholders: scientists, journal editors, funding agencies, universities.

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

References

- "Enabling Reproducible Research: Open Licensing for Scientific Innovation"

- "The Scientific Method in Practice: Reproducibility in the Computational Sciences"

- "Open Science: Policy Implications for the Evolving Phenomenon of User-led Scientific Innovation"

- Reproducible Research: Tools and Strategies for Scientific Computing, July 2011

- Reproducible Research in Computational Science: What, Why and How, Community Forum, July 2011

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