

Transparency in Scientific Discovery: Innovation and Knowledge Dissemination

Victoria Stodden
Department of Statistics
Columbia University

Open Science Summit
Mountain View, CA
October 22, 2011

Why is Science Open?

For hundreds of years, science has required the sharing of methodologies such that other researchers can independently verify published results.

Open science facilitates:

- 1.reproducibility,
- 2.innovation in academia and industry,
- 3.access to scientific knowledge.

Reproducibility

Computational Methods Emerging as Central to the Scientific Enterprise

1. enormous, and increasing, amounts of data collection,

- CMS project at LHC: 300 “events” per second, 5.2M seconds of runtime per year, .5MB per event = 780TB/yr => several PB when data processed,
- Sloan Digital Sky Survey: 8th data release (2010), 49.5TB,
- quantitative revolution in social science due to abundance of social network data (Lazer et al, *Science*, 2009)

2. massive simulations of the complete evolution of a physical system, systematically varying parameters,

3. deep intellectual contributions now encoded in software.

Computation 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***

Updating the Scientific Method

Donoho and others argue 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.
- See e.g. Ioannidis, “Why Most Published Research Findings are False,” PLoS Med, 2005.

Framing Principle for Scientific Communication: *Reproducibility*

- “The idea is: 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.” David Donoho, 1998.
- *(simple) definition*: a result is reproducible if a member of the field can independently verify the result.
- *Side Effect*: Reproducibility is a scoping mechanism for data and code sharing.

Implementation Challenges

Interlocking set of incentives that influence scientific output:

- grant and funding agency requirements,
- patents and financial incentives,
- intellectual property constraints,
- institutional expectations (hiring, promotion, awards),
- journal and publication requirements,
- requirements of scientific integrity.

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.” (2005 and earlier)

- NSF peer-reviewed Data Management Plan (DMP), 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)

NSF Data Management Plan

“Proposals submitted or due on or after January 18, 2011, must include a supplementary document of no more than two pages labeled ‘Data Management Plan.’ This supplementary document should describe how the proposal will conform to NSF policy on the dissemination and sharing of research results.” (<http://www.nsf.gov/bfa/dias/policy/dmp.jsp>)

NSF Data Management Plan

- No requirement or directives regarding data openness specifically.
- But, “Investigators are expected to share with other researchers, at no more than incremental cost and within a reasonable time, the primary data, samples, physical collections and other supporting materials created or gathered in the course of work under NSF grants. Grantees are expected to encourage and facilitate such sharing. Privileged or confidential information should be released only in a form that protects the privacy of individuals and subjects involved.” (http://www.nsf.gov/pubs/policydocs/pappguide/nsf11001/aag_6.jsp#VID4)

Congress: 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)

Innovation and Technology Transfer

Bayh-Dole Act (1980)

- Promote the transfer of academic discoveries for commercial development, via licensing of patents (Tech Transfer Offices),
- Legislators blind to the coming digital revolution, and the impact on software patents and code release for reproducibility.
- 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,
 - alters a scientist's incentives toward commercial ends, instead of the production of science as a public good.

Barriers Facing Scientists

Survey of the Machine Learning Community, NIPS (Stodden 2010)

Code		Data
77%	Time to document and clean up	54%
52%	Dealing with questions from users	34%
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%
30%	Competitors may get an advantage	33%
20%	Web/disk space limitations	29%

Institutional Expectations



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

Copyright © 2003 David Farley, d-farley@ibiblio.org

Access to Scientific Knowledge

Journal Requirements

Computational Science Journals (Stodden and Guo)

Stated Policy, Summer 2011

Proportion requiring data	13.5%
Proportion requiring code	6.5%
Proportion requiring supplemental materials	8.8%
Proportion Open Access	21.8%

N=170; journals classified using Web of Science classifications.

Barriers to Journal Policy Making

- Standards for code and data sharing,
- Meta-data, archiving, re-use, documentation, sharing platforms,
- Review, who checks replication, if anyone,
- Burdens on authors, especially less technical authors,
- Evolving, early research; affects decisions on when to publish,
- Business concerns, attracting the best papers.

This is a Grassroots Movement

- AMP 2011 “Reproducible Research: Tools and Strategies for Scientific Computing”
- AMP / ICIAM 2011 “Community Forum on Reproducible Research Policies”
- SIAM Geosciences 2011 “Reproducible and Open Source Software in the Geosciences”
- ENAR International Biometric Society 2011: Panel on Reproducible Research
- AAAS 2011: “The Digitization of Science: Reproducibility and Interdisciplinary Knowledge Transfer”
- SIAM CSE 2011: “Verifiable, Reproducible Computational Science”
- Yale 2009: Roundtable on Data and Code Sharing in the Computational Sciences
- ACM SIGMOD conferences
- NSF/OCI report on Grand Challenge Communities (Dec, 2010)
- IOM “Review of Omics-based Tests for Predicting Patient Outcomes in Clinical Trials”
- ...

Principles for Advocacy

- work within existing scientific norms as much as possible,
 - reproducible research => open data and code
- scientific integrity => openness for verifiability,
 - required for establishing scientific facts
- data deluge deeply affecting many fields, many researchers have questions,
 - engagement => answer these questions, myth-bust

References

- “The Scientific Method in Practice: Reproducibility in the Computational Sciences”
- “Open Science: Policy Implications for the Evolving Phenomenon of User-led Scientific Innovation”
- “Enabling Reproducible Research: Open Licensing for 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.stodden.net>