Open Challenges to Open Science

Victoria Stodden
Department of Statistics
Columbia University

Sustaining the Digital Research Enterprise
Data Policies Summit
National Institutes for Health
October 18, 2011

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*

Computational Methods Emerging as Central to the Scientific Enterprise

1. enormous, and increasing, amounts of data collection,
   - ~3TB/yr genome sequence data: ~1000 sequencers running full time producing 600GB each run (HiSeq 2000, 11 days per run),
   - 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.
2. massive simulations of the complete evolution of a physical system, systematically varying parameters,
3. deep intellectual contributions now encoded in software.

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.

The Challenges of Implementation

Interlocking set of incentives that influence scientific output:

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

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,
- Burdens on authors, especially less technical authors,
- Evolving, early research; when to publish,
- Business concerns, attracting the best papers.

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)

Activity

Review of Omics-Based Tests for Predicting Patient Outcomes in Clinical Trials

Type:	Consensus Study
Topics:	Biomedical and Health Research, Health Services, Coverage, and Access
Boards:	Board on Health Care Services

Activity Description

An IOM committee will review the published literature to identify appropriate evaluation criteria for tests based on "omics" technologies (e.g. genomics, epigenomics, proteomics, metabolomics) that are used as predictors of clinical outcomes. The committee will recommend an evaluation process for determining when predictive tests based on omics technologies are fit for use as a basis for clinical trial design, including stratification of patients and response to therapy in clinical trials. The committee will identify criteria important for the analytical validation, qualification, and utilization components of test evaluation.

The committee will apply these evaluation criteria to predictive tests used in three cancer clinical trials conducted by Duke University investigators (NCT00509366, NCT00545948, NCT00636441). For example

Excerpt: Letter to Varmus

“We strongly urge that the clinical trials in question ... be suspended until a fully independent review is conducted of both the clinical trials and of the evidence and predictive models being used to make cancer treatment decisions. For this to happen, sufficiently detailed data and annotation must be made available for review. The data should be sufficiently documented for provenance to be assessed (as both gene and sample mislabeling have been documented in these data), and the computer code used to predict which drugs are suitable for particular patients must be made available to allow an independent group of expert genomic data analysts to assess its validity and reproducibility using the data supplied.” (July, 2010)

Patents: 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,
    - (aside from altering a scientist's incentives toward commercial ends).

Institutional Expectations

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

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

Barriers to Data and Code Sharing in Computational Science

Survey of 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%

The Reproducibility Movement: Broad Grassroots Efforts

- 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"
- ...

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