Scientific Reproducibility: First Steps and Guiding Questions

Victoria Stodden
Department of Statistics
Columbia University

Expert Panel Discussion
A Workshop of the National Science Board
March 28, 2011

A Crisis in Computational Science

- Computational methods becoming central to the scientific enterprise:
  - enormous, and increasing, amounts of data collection,
  - intellectual contributions now encoded in software,
  - typical scientific results rely on both data and code.
- Data and code typically not made available, rendering published results unverifiable, not reproducible.

➡ A Credibility Crisis

Reproducibility is Central to the Scientific Method

- Other branches of science incorporate reproducibility of results:
  - deductive branch (mathematics, formal logic): the well-defined concept of the proof,
  - inductive branch (experimental sciences): machinery of hypothesis testing, structured communication of methods and protocols.
- *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.

Computation Emerging as Central to the Scientific Endeavor

For example, in statistics,

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

Framing Principle for Scientific Communication: *Reproducibility*

- Open Data is a natural corollary of reproducibility,
- Open Code is included in the open science discussion,
- Facilitates community-level decision making,
- Gives guidance on what and how to share,
- Encourages adoption of openness by scientists,
- Is a scientific imperative demanding action,
- Gives clarity in the definition of a computational fact,
- Wider and deeper communication of scientific knowledge.

What and how to share

- Share data, code such that the results can be replicated.
- Open questions:
  - where do these files reside? large files?
  - defining reproducibility: start from the same dataset the investigator started with.
  - incentives?

"Establish best practices for the release of science and engineering applications and data as well as the workflows involved in their creation to ensure the reproducibility of computational results."
NSF-ACCI Task Force Report, Dec 2010.

Groundswell from across the Computational Sciences

- 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"

Funding Agencies

- **NSF must take a leadership role**:
    - grassroots movements across many disciplines
    - multifaceted collective action problem

What to do? 1: Pilot Projects

- Fund a small number of projects from different areas to be fully reproducible,
- Permit the grantees to describe their reproducibility plan (like data release plan),
- Grantees propose their additional needs.

- Creates an experiment to understand the requirements of reproducible research:
    - repository? extra coders? support for locally hosted code and data? publication mechanisms for reproducible work?

What to do? 2: "PubCentral"

- Expand PubMedCentral (create PubCentral) to include NSF-funded manuscripts and their replication data and code,
- Or, expand arXiv to include replication data and code.
- Assign unique DOI's to papers, data, and code.
- (pilot this with some well-chosen projects?)

1. Do we need a study of the legal issues? Identify any specific barriers (international collaboration?),
2. Analyze NSF peer-reviewed Data Management plan submissions to discover costs and other barriers.

What to do? 2: Recognize data and code contributions

- Enforce a standard of unique identification for the data and code associated with NSF-funded published results,
- In NSF bio, grant applicants list their data and code contributions as well as publications.

Incentives: Citation and Contributions

- Collaborative efforts in database building?
  - differential citation? (web vs article citation, microcitation)
  - database versioning (e.g. King and Altman 2007, Donoho and Gavish 2011)
  - citizen contributions? (Galaxy Zoo, Open Dinosaur Project)
- Code development? review?
- Code maintenance for reproducibility, scientific reuse?
  - platform building (DANSE, Wavelab, Sparselab)
  - open source software as a model?

What to do? 3: Support Tool Development

- workflow tracking and provenance ie. Vistrails.org and many others,
- automatic cloud repository and unique identifiers for published results (Donoho and Gavish 2011),
- collaborative tools ie. colwiz,
- versioning of datasets and code used for replication.

- Another area for well-chosen pilot projects.

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”

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

Supplemental Slides

1. barriers to sharing
2. problems introduced or enhanced by openness
3. journal policies

Barriers to Data and Code Sharing in Computational Science

Survey of Machine Learning Community (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%

Challenges to Open Science

- “Taleb Effect” - scientific discoveries as (misused) black boxes,
- nefarious uses?
- black boxes and opacity in software (why the traditional methods section is inadequate, massive codebases),
- lock-in: calcification of ideas in software?
- independent replication discouraged?
- policy maker engagement: finding support for our norms,
- Commercial incentives for the scientist/university (Bayh-Dole).

Error Correction and Review

- Different approaches by journals:
  - may offer unreviewed “supplemental materials” section,
  - may require data and/or code to be provided upon request (Science as of Feb 11 2011),
  - may employ an Associate Editor for Reproducibility (Biostatistics, Biometrical Journal) or replicate results (ACM SIGMOD),
  - may publish correspondence from the review process (Molecular Systems Biology, The European Molecular Biology Organization Journal),
  - new journals, ie. Open Research Computation, BMC Data Notes
  - ignore the issue..