Big Data and Reproducing Scientific Results

Victoria Stodden
Department of Statistics
Columbia University

AALS Midyear Meeting
Berkeley, CA
June 12, 2012
The Credibility Crisis in Computational Science

<table>
<thead>
<tr>
<th>JASA June</th>
<th>Computational Articles</th>
<th>Code Publicly Available</th>
</tr>
</thead>
<tbody>
<tr>
<td>1996</td>
<td>9 of 20</td>
<td>0%</td>
</tr>
<tr>
<td>2006</td>
<td>33 of 35</td>
<td>9%</td>
</tr>
<tr>
<td>2009</td>
<td>32 of 32</td>
<td>16%</td>
</tr>
<tr>
<td>2011</td>
<td>29 of 29</td>
<td>21%</td>
</tr>
</tbody>
</table>

Generally, data and code not made available at publication, insufficient information communicated for verification, replication of results.

⇒ A Credibility Crisis
Data and Code Sharing
Barrier 1: Copyright

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
ShareAlike isn’t for Science

The motivations and goals of the Open Source Software community differ from those of the scientific community:

• industry collaboration and re-use of code,
• different licensing needs in different scientific projects,
• mixing of scientific codes,
• scientific knowledge as a public good.
Barrier 2: Bayh-Dole and Software

• Bayh-Dole (1980) to create incentives for universities to patent, thereby making inventions accessible/licensable,

• computational scientists’ dilemma: patent vs share code for verification and reproducibility,

• incentives distortion: potential code withholding, obfuscation, startups vs science.
Sharing is Happening..

• Federal funding agency requirements: NSF Data Management Plan, America COMPETES Re-authorization,

• Journal publishing requirements: Science, Nature, PNAS,...

• Promotion, tenure, hiring committees,

• Grassroots community sharing initiatives.
Solution Component 3: 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

• 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)
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)
Whitehouse RFIs

‣ “Public Access to Peer-Reviewed Scholarly Publications Resulting From Federally Funded Research”

‣ “Public Access to Digital Data Resulting From Federally Funded Scientific Research”

Comments were due January 12, 2012.

Solution Component 4: Journal Policy

Computational Science Journals (Stodden and Guo, preliminary results)

<table>
<thead>
<tr>
<th>Stated Policy, Summer 2011</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Proportion requiring data</td>
<td>15%</td>
</tr>
<tr>
<td>Proportion requiring code</td>
<td>7%</td>
</tr>
<tr>
<td>Proportion requiring supplemental materials</td>
<td>9%</td>
</tr>
<tr>
<td>Proportion Open Access</td>
<td>58%</td>
</tr>
</tbody>
</table>

N=170; journals classified using Web of Science classifications (Mathematical & Computational Biology, Statistics & Probability, Multidisciplinary Science).
This is a Grassroots Movement

- AMP 2011 “Reproducible Research: Tools and Strategies for Scientific Computing”
- Open Science Framework / Reproducibility Project in Psychology
- 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”
- ...