The Credibility Crisis in Computational Science: An Information Issue

Victoria Stodden
Department of Statistics
Columbia University
@victoriastodden

Dean’s Lecture
UC Berkeley School of Information
February 1, 2012
Scientific Communication

The Scientific Record
Scientific Research is Changing
Examples
The Credibility Crisis
Survey of the Machine Learning Community

Responses to the Credibility Crisis

Community Responses
Policy Responses
Press Responses

Open Questions
The Concept of a Scientific Fact

In *Opus Tertium* (1267) Roger Bacon distinguishes experimental science by:

1. verification of conclusions by direct experiment,
2. discovery of truths unreachable by other approaches,
3. investigation of the secrets of nature, opening us to a knowledge of past and future.

- described a repeating cycle of observation, hypothesis, experimentation, and the need for independent verification,
- recorded his experiments (e.g. the nature and cause of the rainbow) in enough detail to permit reproducibility by others.
Inductive Scientific Reasoning

In *Novum Organum* (1620) Francis Bacon proposes:

1. the gathering of facts, by observation or experimentation,

2. verification of general principles.

“There are and can be only two ways of searching into and discovering truth. The one flies from the senses and particulars to the most general axioms, and from these principles, the truth of which it takes for settled and immovable. ... The other derives axioms from the senses and particulars, rising by a gradual and unbroken ascent, so that it arrives at the most general axioms last of all. This is the true way, but as yet untried.”
The Scientific Record

- The Royal Society of London founded 1660 (the “Invisible College”),
- members discussed Francis Bacon’s “new science” from 1645,
- Society correspondence reviewed by the first Secretary, Henry Oldenburg,
- Oldenburg became the founder, editor, author, and publisher of *Philosophical Transactions*, launched in 1665.
The “Invisible College” included Robert Boyle, the “father of chemistry,”

Boyle introduced standards for scientific communication: enough information must be included to allow others to independently reproduce the finding.

delineates science, concept of reproducibility permits verification and knowledge transfer,

knowledge in method not in the finding itself.
Scientific Research is Changing

Today, scientific computation is becoming central to the scientific method:

- Changing how research is conducted in many fields,
- Changing the types of the scientific questions we can ask,
- Changing how we learn about our world.

Conjecture: Today’s academic scientist probably has more in common with a large corporation’s information technology manager than with a philosophy or English professor at the same university.
I. Examples of Pervasiveness of Computational Methods

- For example, in statistics:

<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>

- Social network data and the quantitative revolution in social science (Lazier et al. 2009);

- Computation reaches into traditionally nonquantitative fields: e.g. Wordhoard project at Northwestern examining word distributions by Shakespearian play.
1. Climate Simulation: Community Climate Models
2. High Energy Physics: Large Hadron Collider

- 4 LHC experiments at CERN: 15 petabytes produced annually
- Data shared through grid to mobilize computing power
- Director-General of CERN (Heuer): “Ten or 20 years ago we might have been able to repeat an experiment. They were simpler, cheaper and on a smaller scale. Today that is not the case. So if we need to re-evaluate the data we collect to test a new theory, or adjust it to a new development, we are going to have to be able reuse it. That means we are going to need to save it as open data.” Computer Weekly, August 6, 2008
3. Dynamic modeling of macromolecules: SaliLab UCSF

The structural dynamics of macromolecular processes
Daniel Russel¹, Keren Lasker¹,², Jeremy Phillips¹,³, Dina Schneidman-Duhovny¹, Javier A Velázquez-Muriel¹ and Andrej Sali¹

Dynamic processes involving macromolecular complexes are essential to cell function. These processes take place over a wide variety of length scales from nanometers to micrometers, and over time scales from nanoseconds to minutes. As a result, information from a variety of different experimental and computational approaches is required. We review the relevant sources of information and introduce a framework for integrating the data to produce representations of dynamic processes.

No single technique, computational or experimental, is able to span all relevant spatial and temporal scales (Figure 3). For static complexes, for example, X-ray crystallography can generate atomic structures of the components, while single particle cryo-electron microscopy (cryo-EM) can provide average mass density maps of the whole assembly at nanometer resolution for the whole assembly. For processes, computer simulations are beginning to reach the microsecond time scale, while
4. Mathematical “proof” by simulation and grid search
Science Digitized: Four Technological Transformations

1. *Data Deluge*: enormous, and increasing, amounts of data collection:
   - CMS project at LHC: 300 “events” per sec → 780TB/yr → several PB when processed,
   - Sloan Digital Sky Survey: 8th data release (2010), 49.5TB,
   - quantitative revolution in social science due to abundance of social network data (Lazier et al, Science, 2009),
   - Science Magazine survey of peer reviewers: 20% regularly work with datasets >100GB; 7% >1TB (N=1700; Feb 11, 2011).

2. *High Performance Computing Power*: massive simulations of complete physical systems, systematically varying parameter settings,

3. *Software as Source*: deep intellectual contributions to science encapsulated *only* in code,

4. *Internet*: ease of digital communication as science goes digital.
Relaxed practices regarding the communication of computational details is creating a **credibility crisis** in computational science, not only among scientists, but as a basis for policy decisions and in the public mind.

Recent prominent examples,

- Climategate 2009,
- Microarray-based clinical trials recently terminated at Duke University.
Clinical trials based on flawed genomic studies

Timeline:


- Coombes, Wang, Baggerly at M.D. Anderson Cancer Center cannot replicate, and find simple flaws: genes misaligned by one row, column labels flipped, genes repeated and missing from analysis..

- 2007 correspondence and a supplementary report submitted to the Journal of Clinical Oncology and publication declined; 2008 Nature Medicine declines their correspondence.

- Clinical trials initiated in 2007 (Duke), 2008 (Moffitt).
Clinical trials based on flawed genomic studies

- Duke launches internal investigation Sept 2009; all three trials suspended in Oct 2009,
- Oct 2009: results reported validated, regardless of errors, because data blinded (later found not to be true),
- Jan 2010: Duke clinical trials resume, patients allocated to treatment and control groups. “Neither the review nor the raw data are being made available at this time.”
- July 2010: 33 prominent biostatisticians write to Varmus as head of IOM urging suspension of the trials and an examination of standards of review, including reproducibility.
- Sept 2010: IOM committee “Review of Omics-Based Tests for Predicting Patient Outcomes in Clinical Trials” formed,
- Nov 2010: Potti resigns and the clinical trials are terminated.
Controlling Error is Central to Scientific Progress

“The scientific method’s central motivation is the ubiquity of error - the awareness that mistakes and self-delusion can creep in absolutely anywhere and that the scientist’s effort is primarily expended in recognizing and rooting out error.”

David Donoho et al. (2009)
The Third Branch of the Scientific Method

- Branch 1: Deductive/Theory: e.g. mathematics; logic,
- Branch 2: Inductive/Empirical: e.g. the machinery of hypothesis testing; statistical analysis of controlled experiments,
- Branch 3?? Large scale extrapolation and prediction, using simulation and other data-intensive methods,
- Branch 4?? Data driven discovery.
Toward a Resolution of the Credibility Crisis

- Typical scientific communication doesn’t include sufficient detail for reproducibility i.e. the code and data that generated the findings.

- Most published computational scientific results today are near impossible to replicate.

**Thesis**: Computational science cannot be elevated to a third (or fourth) branch of the scientific method until it generates *routinely verifiable knowledge*. (Donoho, Stodden, et al. 2009)

Sharing of underlying code and data is a necessary part of this solution, enabling *Reproducible Research.*
“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.

(reminiscent of Boyle)
Survey of Machine Learning Community (Stodden 2010)

**Question**: Why isn’t reproducibility practiced more widely? Answer builds on literature of free revealing and open innovation in industry, and the sociology of science.

- Sample: American academics registered at the Machine Learning conference NIPS.
- Respondents: 134 responses from 593 requests (∼23%).
## Top Reasons Not to Share

<table>
<thead>
<tr>
<th>Code</th>
<th>Reason</th>
<th>Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>77%</td>
<td>Time to document and clean up</td>
<td>54%</td>
</tr>
<tr>
<td>52%</td>
<td>Dealing with questions from users</td>
<td>34%</td>
</tr>
<tr>
<td>44%</td>
<td>Not receiving attribution</td>
<td>42%</td>
</tr>
<tr>
<td>40%</td>
<td>Possibility of patents</td>
<td>-</td>
</tr>
<tr>
<td>34%</td>
<td>Legal barriers (ie. copyright)</td>
<td>41%</td>
</tr>
<tr>
<td>-</td>
<td>Time to verify release with admin</td>
<td>38%</td>
</tr>
<tr>
<td>30%</td>
<td>Potential loss of future publications</td>
<td>35%</td>
</tr>
<tr>
<td>30%</td>
<td>Competitors may get an advantage</td>
<td>33%</td>
</tr>
<tr>
<td>20%</td>
<td>Web/Disk space limitations</td>
<td>29%</td>
</tr>
</tbody>
</table>
### Top Reasons to Share

<table>
<thead>
<tr>
<th>Code</th>
<th>Reason</th>
<th>Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>91%</td>
<td>Encourage scientific advancement</td>
<td>81%</td>
</tr>
<tr>
<td>90%</td>
<td>Encourage sharing in others</td>
<td>79%</td>
</tr>
<tr>
<td>86%</td>
<td>Be a good community member</td>
<td>79%</td>
</tr>
<tr>
<td>82%</td>
<td>Set a standard for the field</td>
<td>76%</td>
</tr>
<tr>
<td>85%</td>
<td>Improve the caliber of research</td>
<td>74%</td>
</tr>
<tr>
<td>81%</td>
<td>Get others to work on the problem</td>
<td>79%</td>
</tr>
<tr>
<td>85%</td>
<td>Increase in publicity</td>
<td>73%</td>
</tr>
<tr>
<td>78%</td>
<td>Opportunity for feedback</td>
<td>71%</td>
</tr>
<tr>
<td>71%</td>
<td>Finding collaborators</td>
<td>71%</td>
</tr>
</tbody>
</table>
Grassroots Efforts in Many Fields

Independent efforts by researchers:

- 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
- ...
Efforts in Policy

Policy changes:

▶ NSB report “Digital Research Data Sharing and Management” (Dec 2011),
▶ Whitehouse ‘Requests for Information’ on Open Access and Data Policies (Nov 2011),
▶ NSF report “Changing the Conduct of Science in the Information Age” (Aug 2011),
▶ NSF Data Management Plan requirement (Jan 2011),
▶ America COMPETES Re-authorization (Jan 2011),
▶ Journal policy movement toward code and data requirements (ie. Science Feb 2011),
▶ NSF/OCI report on Grand Challenge Communities (Dec 2010),
▶ NIH, NSF requests for information on policy (Dec 2010),
▶ ...

...
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)

NSF peer-reviewed Data Management Plan, Jan 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).
"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 specific requirements or directives regarding data sharing.
- 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.”
  
In Congress

- 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 Requests for Information

- “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.
Scoping the Scientific Information Sharing Problem

Discussions of open data could be more usefully reframed as implementing reproducibility. Science is not about open data (directly):

**Implication 1:** Open Data is a natural corollary of Reproducible Research,

**Implication 2:** Open Code is included in the open science discussion,

**Implication 3:** What to share and how to share is clear,

**Implication 4:** Adoption of openness by scientists,

**Implication 5:** We can establish scientific facts,

**Implication 6:** Scientific communication is augmented, on an internet scale.
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.
Popular Press

- “The Truth Wears Off,” *New Yorker Magazine*, Dec 2010: asserts the ‘discovery’ of a mysterious effect by which replicated experiments decrease in significance level.
- “it appears that nature often gives us different answers”
- evidence provided in the article:
  - tests on three schizophrenia drugs,
  - Professor Schooler’s inability to replicate his own research results,
  - his colleagues’ assurances that this happens ‘all this time,’
  - ESP experiments from the 1930’s,
  - tests for symmetry in sex selection,
  - temporal trends in hundreds of ecology papers.

Question: why bias the publication of results towards ones that agree with previously published results? (Merton’s proposed *Universalism* scientific norm)
Popular Press

- Profile of the work of John Ioannidis, Stanford University School of Medicine.
  - Exposure of bias and flawed statistical reasoning in medical research,
  - Decline effect due to initial ‘exaggerations’ of the results and researcher error,
  - Misinterpretation of p-values, artificial lowering of p-values.
Open Questions for Open Data and Code

- Under what terms do scientists share data and code?
- Characteristics of the scientific problem that makes sharing easier or harder.
- Software Development:
  - ease of implementation ie. data provenance and workflow, ("progress depends on artificial aids becoming so familiar they are regarded as natural" I.J. Good, 1958),
  - platforms ie. EarthCube,
- Underlying changes in the scientific method and philosophy of science - what counts as science?
- Standards and guides for information sharing practices: simulation information, data descriptors.
- Use cases and examples.