Trust Your Science? Open Your Data and Code

Victoria Stodden

Reproducibility in the computational sciences seems to be capturing everyone's attention. Movements to address the reliability of published computational results are arising in fields as disparate as geophysics, political science, fluid dynamics, computational harmonic analysis, fMRI research, and bioinformatics. Open data and code in climate modeling has taken on a new priority since ClimateGate in 2009 (i.e., *www.nature.com/news/2010/101013/full/467753a.html*), and *Amstat News* has recounted efforts to ensure reproducibility in genomics research in the wake of the termination of clinical trials at Duke University in December of 2010 (see *http://magazine.amstat.org/blog/2011/01/01/scipolicyjan11*).

These efforts are essential to addressing the "credibility crisis" in science. It is impossible to believe most of the computational results presented at conferences and in published papers today. Even mature branches of science, despite all their efforts, suffer severely from the problem of errors in final published conclusions. Traditional scientific publication is incapable of finding and rooting out errors in scientific computation, and standards of verifiability must be developed.

A smattering of these efforts will give a sense of the scope at which the community is addressing this issue. The Institute of Medicine of the National Academies is undertaking a consensus study titled "Review of Omics-Based Tests for Predicting Patient Outcome in Clinical Trials," (see *www.iom.edu/Activities/Research/OmicsBasedTests.aspx*), and sessions on reproducibility were held at SIAM Geosciences 2011, this year's AAAS annual meeting, and SIAM Computing in Science and Engineering 2011. Also, a three-day workshop is to be held at Applied Mathematics Perspectives this month.

In 2009, stakeholders from biology, computational chemistry, geophysics, law, astronomy, and other fields collectively drafted a declaration on data and code sharing in the computational sciences (*www.computer.org/portal/web/csdl/doi/10.1109/MCSE.2010.113* and *www.stanford.edu/~vcs/Conferences/RoundtableNov212009*). Since January of this year, the National Science Foundation has required data management plans to be peer reviewed with every grant application.

Open access to data and software are relevant to advancing trustworthy science and are discussed in the 2010 reauthorization of the America Competes Act (see *http://blog.stodden.net/2011/05/27/regulatory-steps-toward-open-science-and-reproducibility-we-need-a-science-cloud*). As we embrace and tackle this issue across the computational sciences, concepts are inevitably labeled with different terms and different concepts are emphasized. I will touch on the semantic and substantive differences in the various approaches to reliability in computational and data-enabled sciences.

Reproducibility, Replicability, and Repeatability

I learned from my advisor, David Donoho, that reproducibility meant releasing data and code such that others may regenerate your results on their own systems (i.e., releasing the full computational environment that produces a result). Donoho paraphrases Stanford professor Jon Claerbout, an early pioneer in reproducible research, as follows:

> 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. *(http://sepwww.stanford.edu/doku.php?id=sep:research:reproducible:seg92)*

In our case, this typically meant releasing MATLAB scripts and data files, often along with a MATLAB GUI we wrote that permitted the user to select the figure he or she wished to regenerate, adjust parameter settings, and view the source code (*www-stat.stanford.edu/~wavelab*, *http://sparselab.stanford.edu*). Unless we can inspect the code and data, we cannot resolve differences in output between independent methods or independent implementations of even purportedly identical methods.

Replication, using author-provided code and data, and independent reproduction work hand-in-hand. We can reserve the term "replicability" for the regeneration of published results from author-provided code and data. Gary King, a Harvard professor, proposed the Replication Standard in 1995: "[T]hat sufficient information exists with which to

Victoria Stodden is assistant professor of statistics at Columbia University. She completed both her PhD in statistics and her law degree at Stanford University. Her current research focuses on how pervasive and large-scale computation is changing our practice of the scientific method: reproducibility of computational results, understanding factors underlying code and data-sharing among researchers, and the role of legal framing for scientific advancement.

understand, evaluate, and build upon a prior work if a third party could replicate the results without any additional information from the author." (*http://gking.harvard.edu/files/abs/replication-abs.shtml*)

Reproducibility is a more general term, implying both replication and the regeneration of findings with at least some independence from the code and/or data associated with the original publication. Both refer to the analysis that occurs after publication. A third term, "repeatability," is sometimes used in place of reproducibility, but this is more typically used as a term of art referring to the sensitivity of results when underlying measurements are retaken.

To summarize, we need replicability, in part, to resolve differences in outcomes that arise from reproduced computational results, regardless of whether the experiments have been repeated. We loosely say that results have been verified if we have reproduced them, but as standards for code quality in computational and data-enabled science increase, we should supplant this with the more precise definitions of verification, validation, and error quantification developed in scientific computing.

Data Evaluation Standards

The movement toward openness in computational and data-enabled science has a long and successful history in genomic research, due to pioneering efforts in response to the public/private race to decode the human genome in the 1990s. That community gathered in Bermuda in 1996 to develop a cooperative strategy for both genome decoding and managing the resulting data. Biologist and Nobel Prize winner John Sulston said, "The principle of data availability had to be endorsed at the Bermuda meeting or else mutual trust would have been impossible."

The meeting resulted in the "Bermuda Principles," which shaped the data-sharing practices among the researchers, ensuring rapid open release of human genome sequence data. The convention of releasing community statements continued as these principles were reaffirmed and extended three more times (most recently in July 2009: *Nature*).

The nature of the underlying research and the technology involved meant this discussion centered on open data and rarely mentioned code. Accordingly, a vocabulary was developed within this data-oriented context, such as the term "data provenance." What data provenance means depends on whether you understand data as a community resource and hence are interested in tracking modifications and updates to the data set, or as a local entity, and thus are interested in recording filtering and other data operations that ready it for analysis in a particular project. It also can refer to both, and both concepts are essential for effective reproducibility.

The term "research workflow" incorporates the latter definition (the changes made to data to prepare it for analysis), but also includes the analysis steps that generated the published results and other procedures that affect interpretation of the findings, such as otherwise unreported hypothesis tests.

It is very easy to underrate the importance of clarity in conceptualizing the role of data in open science. A quick glance at discussions in the blogosphere might lead a casual observer to think all that mattered is the openness of data. This stems from the framing dialogue in the pioneering days in human genome sequencing, combined with today's vastly increased capacity for data collection, but leads to a conclusion that is too simple. Transparency in the communication of scientific methodology arises from the notion of reproducibility in science, not the other way around. Open data is a prerequisite for verifiable research; reproducibility is not a convenient mechanism in support of the notion of big open data as if sometimes promulgated. Science has never been about open data per se, but openness is something hard fought and won in the context of reproducibility.

Scientists have scarce resources, and changing the scientific method to include open data for its own sake—untethered to our age-old concept of reproducibility—requires a deeper justification and understanding of the trade-offs involved. Open data in support of reproducibility is an enormous challenge in itself, and can best be accomplished in a principled way within our current system of scientific norms.

An Open Call to Computational Scientists

Making both the data and code underlying scientific findings conveniently available in such a way that permits reproducibility is of urgent priority for the credibility of the research and the elevation of computational and data-enabled research to a bona fide branch of the scientific method. The independent efforts occurring today in many disciplines and subdisciplines can inform each other and provide a guide for computational fields just starting to grapple with the issue of reproducibility. ■