COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

Victoria Stodden
Assistant Professor
Department of Statistics
Columbia University
1255 Amsterdam Avenue, 10th fl.
New York, NY 10027
vcs2115@columbia.edu

March 18, 2011

White Paper for
EXPERT PANEL DISCUSSION ON DATA POLICIES
A Workshop of the National Science Board
March 27-29, 2011

In our workshop charge we were invited to read three reports[1] that formed the basis for the NSB-approved Data Policies Task Force's "Statement of Principles," providing the starting point for this workshop. I take a contrarian perspective and challenge the assumption in all these documents that open data is a foundational component of the scientific endeavor. Instead, I argue that the framing principle should be the reproducibility of computational results, from which open data (along with open code) falls as a natural corollary. In this note I highlight six implications of the framing of reproducible research as a guiding principle for science policy in the digital age.

SCIENCE IS NOT ABOUT OPEN DATA (DIRECTLY)

Scientific computation is emerging as absolutely central to the scientific method, but the prevalence of very relaxed practices is leading to a credibility crisis affecting many scientific fields.[2] It is impossible to verify most of the results that computational scientists present at conferences and in papers today. The principle of reproducibility has been a key pillar of the scientific method since the 1660's but without making the details of the computations - code and data - conveniently available to others, this pillar is lost.

The framing of openness in computational science as an issue of reproducibility gives a number of benefits that do not accrue when the issue is considered as one of open data alone, both in terms of policy recommendations and the ease of

[1] NSTC Interagency Working Group on Digital Data, *Harnessing the Power of Digital Data for Science and Society* (2009); NRC's *Ensuring the Utility and Integrity of Research Data in a Digital Age* (2009); and the NSB report *Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century* (2005).

[2] See e.g. *Climategate*, ("E-mail Fracas Shows Peril of Trying to Spin Science," New York Times, Dec 1, 2009); and the *Duke Clinical Trials Scandal* ("Duke Accepts Resignation of Dr. Anil Potti," Office of News & Communication, Duke University, Nov 19, 2010).

adoption of data and code sharing as a community practice. A partial list of the implications of using reproducible research as a guiding framework, rather than open data directly, follow.

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

For computational scientific results to be reproducible, the data underlying the discoveries much be made available for the purposes of verification. This implies open data, and raises the same questions (versioning, citation and unique identifiers, archiving, repository creation, meta-data and data standards, release requirements and sequestration, among others).

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

For computational science to become reproducible, not only is the data required but communicating the steps taken in analyzing the data is necessary. In fact, intellectual contributions to science are increasingly embedded in the scripts, code, and software used to analyze data. It is no longer possible to encapsulate these innovations in a written description in a paper alone – parameter settings, function invocation sequences, and computational details necessary to replicate the results will be lost. The framing of reproducible research necessarily includes the communication of these steps through code sharing.

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

Questions such as "what data to share? what aspects should be documented and communicated? what aspects need to be standardized and what additional features would promote data reuse?" can be answered within the context of reproducible research. Making data and code conveniently available such that computational results can be replicated provides guidance on these and many other questions, as well as delineating the role for policy. The goal of the verification of results indicates appropriate licensing choices for code and data, for example.[3]

Implication 4: Adoption of openness by scientists

Scientists have long understood the importance of reproducibility to the scientific practice and its role in determining of a scientific fact. The intellectual framework for advancing reproducibility in computational science is already in place in the scientific community, and provides the easiest mechanism for the communication of the importance of openness in computational science.

Implication 5: Scientific fact establishment can be achieved

[3] See V. Stodden, "Enabling Reproducible Research: Open Licensing for Scientific Innovation," International Journal of Communications Law and Policy, 13, 2009.

Open data alone does not sufficiently address the credibility crisis in computational science. Replication is the hallmark of the establishment of a scientific fact, for which both data and code are required.

Implication 6: Scientific communication is augmented, on an internet scale

The imperative of reproducible computational research means making the data and code underlying results conveniently available to others, typically on the internet. Not only does this permit access to scientific knowledge by those outside the ivory tower, scientific communication with the idea of reproducible research in mind means that the full methodology and know-how for generating the results is made available. This has the corollary effect cutting across disciplinary boundaries and international borders and offering an unprecedented opportunity to share knowledge widely.[4]

REPRODUCIBLE RESEARCH IS EMERGING AS A GRASSROOTS MOVEMENT

Workshops, symposia, discussions, and forums are emerging from scientific fields as diverse as bioinformatics, computational mathematics, geophysics, statistics, and neuroscience. The examples are too many to list in totality so I provide a mere sampling:

Mar 2009	"Methods for Reproducible Research," ENAR 2009
Nov 2009	Yale Law School Roundtable on Data and Code Sharing, New Haven, CT
Oct 2010	National Academies Committee on "The Impact of Copyright Policy on Innovation in the Digital Era."
Nov 2010	NSF workshop, "Changing the Conduct of Science in the Information Age," Washington, D.C.
Dec 2010	Institute of Medicine Committee on "Review of Omics-based Tests for Predicting Patient Outcomes in Clinical Trials."
Feb 2011	"The Digitization of Science: Reproducibility and Interdisciplinary Knowledge Transfer," AAAS Annual Meeting, Washington, D.C.
Mar 2011	National Academies Workshop on "The Future of Scientific Knowledge Discovery in Open Networked Environments," Washington, D.C.
Mar 2011	"Verifiable, reproducible research and computational science," SIAM Conference on Computational Science and Engineering, Reno, NV
July 2011	"Reproducible Research: Tools and Strategies for Scientific Computing," Applied Mathematics Perspectives, Vancouver, BC
July 2011	"Community Forum on Reproducible Research Policies," Vancouver, BC

…and many other publications, meetings, and efforts to advance reproducible research.[5]

[4] See "The Digitization of Science: Reproducibility and Interdisciplinary Knowledge Transfer," AAAS Annual Meeting, 2011. See http://www.stanford.edu/~vcs/AAAS2011 for slides and audio.

[5] See e.g. http://www.stanford.edu/~vcs/Conferences/RoundtableNov212009/References.html