Reproducibility in Computational Science: Framing the Concept

Victoria Stodden
Department of Statistics
Columbia University

UCLA Department of Information Studies
January 24, 2011

Agenda

- Definitions of reproducibility
- Researchers: Implementation in different fields
- Policy makers
- Journal editors
- Grant giving agencies

Claerbout/Donoho

- *Really Reproducible Research:*
- "An article about computational science in a scientific publication if 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."

Gary King

- *Replication Standard:*
- "The *replication standard* holds that sufficient information exists with which to understand, evaluate, and build upon a prior work if a third party could replicate the results without any additional information from the author."

Theme: Science as Collaboration

- *Horizontal collaboration*: with peers (see NSF report on Grand Challenge Communities)
- *Vertical collaboration*: essential to science, building on others' work, external verification and results bolstering.

Pervasiveness of Irreproducibility

- Scientific computation is becoming central to the scientific method
  - ✦ changing how research is done in many fields
  - ✦ changing the nature of how we learn about our world
- Today's academic scientist probably has more in common with a large corporation's IT manager than with a philosophy or English professor at the same university.

Pervasiveness of Computation in Statistics

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

In different fields..

- Statistics and other heavily computational fields: much of the work is in scripts and code development,
- Biology and bioinformatics, economics, etc: heavy use of prewritten, often commercial, software,
- Problems: massive data (CERN, astrophysics); massive code bases (ICES, NCAR)

Policy: who cares?

- Computational scientists subject to incentives just like anyone else: what are these?
  - journal editors and publication standards
  - grant giving agency requirements
  - tenure and institutional review committees

Journal policy

- requiring code and data for publication?
  - IP issues
  - cost and storage issues
  - inertia
  - concerns about the nature of scientific progress
  - example: biostatistics

Grant giving

- NSF data management plan (cost? impact?)
- NIH PubMedCentral, doi's (Duke scandal and the IOM committee)
- OSTP interest, leadership on regulatory change, open gov?

Regulatory Structure

- Intellectual property law 1: copyright as a barrier to reproducibility: *Reproducible Research Standard* (Stodden)
- Intellectual property law 2: patents as a barrier to reproducibility: Bayh-Dole

Feist and Data in the Law

- raw facts not copyrightable
- original "selection and arrangement" of these facts is copyrightable. (Feist Publns Inc. v. Rural Tel. Serv. Co., 499 U.S. 340 (1991)).
- implies a residual copyright in data
- what is a "raw fact" in science?

Provenance

- My opinion: solutions will be technological
- tracking analysis for sharing, versioning
- negative results and cherry picking
- open licensing or fair use exceptions for academic research, build permissions into tags