Technology and the Scientific Method: Tools and Policies for Addressing the Credibility Crisis in Computational Science

Victoria Stodden
Department of Statistics
Columbia University

Computational Methods Emerging as Central to the Scientific Enterprise

- enormous, and increasing, amounts of data collection,

    - ~3TB sequence data produced each year: ~1000 sequencers running full time producing 600GB each run (HiSeq 2000, 11 days per run),

    - LHC grid expected to produce ~10-15PB each year,

- massive simulations of the complete evolution of a physical system, while systematically varying parameters,

- data mining for subtle patterns in vast databases,

- intellectual contributions now encoded in software,

- typical scientific results published today rely on data and code.

Computation Emerging as Central to the Scientific Endeavor

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

- Data and code typically not made available in scientific publishing, rendering results unverifiable, not reproducible.

  ➡ A *Credibility Crisis* (ClimateGate, Duke Clinical Trials,...)

Reproducibility is Central to the Scientific Method

- Other branches of science incorporate reproducibility of results:

  - deductive branch (mathematics, formal logic): the well-defined concept of the proof,

  - inductive branch (experimental sciences): machinery of hypothesis testing, structured communication of methods and protocols.

- *Computational Science must develop standards for reproducibility before it can be considered a third branch of the scientific method,*

  ➡ Data and Code Sharing, with publication.

Framing Principle for Scientific Communication: *Reproducibility*

- "The *replication standard* holds that sufficient information exists with which to understand, evaluate, and build upon a prior work if a third party can replicate the results without any additional information from the author." Gary King, 1995.

- "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.

Implication of Reproducibility as a Framing Principle

- Open Data is a natural corollary of reproducibility,
- Open Code is included in the open science discussion,
- Facilitates community-level decision making,
- Gives guidance on what and how to share,
- Encourages adoption of openness by scientists,
- Is a scientific imperative demanding action,
- Gives clarity in the definition of a computational fact,
- Wider and deeper communication of scientific knowledge.

Groundswell from across the Computational Sciences

- 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

- NSF/OCI report on Grand Challenge Communities (Dec, 2010)

- IOM "Review of Omics-based Tests for Predicting Patient Outcomes in Clinical Trials"

Barriers to Data and Code Sharing in Computational Science

Survey of Machine Learning Community (Stodden, 2010):

Code		Data
77%	Time to document and clean up	54%
52%	Dealing with questions from users	34%
44%	Not receiving attribution	42%
40%	Possibility of patents	-
34%	Legal Barriers (ie. copyright)	41%
-	Time to verify release with admin	38%
30%	Potential loss of future publications	35%
30%	Competitors may get an advantage	33%
20%	Web/disk space limitations	29%

Citation and Contributions

- Collaborative efforts in database building?

- differential citation? (web vs article citation, microcitation)

- database versioning (e.g. King & Altman 2007, Donoho & Gavish 2011)

- citizen contributions? (Galaxy Zoo, Open Dinosaur Project)

- Code development? review?

- Code maintenance for reproducibility, scientific reuse?

  - platform building (DANSE, Madagascar, Wavelab, Sparselab)

  - open source software as a model?

Scientific Research Software Development

- workflow tracking and provenance ie. Vistrails.org and many others,

- automatic cloud repository and unique identifiers for published results (Donoho and Gavish 2011, Altman and King 2007),

- collaborative tools ie. colwiz,

- versioning of datasets and code used for replication.

Incentives

- Journal policies - data and code requirements,

- Funding agency policy,

- University and institutional policy

- Bayh-Dole Act / America Competes Act,

- what about... exceptionally large datasets or code bases? black box code and proprietary software?

References

- "Enabling Reproducible Research: Open Licensing for Scientific Innovation"

- "The Scientific Method in Practice: Reproducibility in the Computational Sciences"

- "Open Science: Policy Implications for the Evolving Phenomenon of User-led Scientific Innovation"

available at http://www.stanford.edu/~vcs

Supplemental Slides

Challenges to Open Science

- "Taleb Effect" - scientific discoveries as (misused) black boxes,

- nefarious uses?

- black boxes and opacity in software (why the traditional methods section is inadequate, massive codebases),

- lock-in: calcification of ideas in software?

- independent replication discouraged?

- policy maker engagement: finding support for our norms,

- Commercial incentives for the scientist/university (Bayh-Dole).

Error Correction and Review

- Different approaches by journals:
  - may offer unreviewed "supplemental materials" section,
  - may require data and/or code to be provided upon request (Science as of Feb 11 2011),
  - may employ an Associate Editor for Reproducibility (Biostatistics, Biometrical Journal) or replicate results (ACM SIGMOD),
  - may publish correspondence from the review process (Molecular Systems Biology, The European Molecular Biology Organization Journal),
  - new journals, ie. Open Research Computation, BMC Data Notes
  - ignore the issue..