

# Reproducible Research: A Digital Curation Agenda

Victoria Stodden  
@victoriastodden  
Department of Statistics  
Columbia University

7th International Digital Curation Conference  
Bristol, U.K.  
Dec 6, 2011

# Why Curate?

- Because science is open, but why is science open?
- The primary goal of the scientific method is to root out error.
- Today's relaxed practices of science in the digitized world seems to have forgotten this... breezy demos of implementations..

*“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.

# Reproducibility as a Framework

- Open Data is a natural corollary of Reproducible Research,
- Open Code is then included in the Open Science discussion,
- What to share and how to share is scoped,
- Adoption of openness by scientists,
- Scientific facts can be established(!),
- Scientific communication and access are augmented, at internet scale.

# Implementing Reproducibility

- Requires tools, infrastructure, incentives..
- Requires long term access
- Scientific knowledge is not embedded in raw data, it is embedded in what we've done with the data,
- Deep intellectual contributions are often captured only in the code,
- Our scientific culture must be made so that such knowledge is not lost over time.

# Including Open Code in the Open Science Discussion

- The existence of digital datasets necessarily implies code,
- In the vast majority of cases, the complete details of the generation of results are not in the paper but in the code,
- Independent replication? important, but so is reconciling differences between independent replications,
- As important or more important than data: captures the methodology, the novel thinking behind the discovery.

# What Does it Mean for Data to be Open?

- “Open Data” not defined:
  - Are all data shared? all versions? when do I share it? What if I don't use it in my research? What counts as data, anyway?
  - What kinds of meta-data or documentation should appropriately be attached? What archiving is appropriate?
  - To whom should I provide access? How much support and clarification am I responsible for?
- Reproducibility provides guidance on these questions, open data does not.

# The Credibility Crisis

JASA June	Computational Articles	Code Availability Mentioned
1996	9 of 20	0%
2006	33 of 35	9%
2009	32 of 32	16%
2011	29 of 29	21%

Typically data and code are not made available at the time of publication, rendering results unverifiable, not reproducible.

➡ A Credibility Crisis

# Updating the Scientific Method

Donoho and others argue that computation presents only a potential third branch of the scientific method:

1. Branch 1 (deductive): mathematics, formal logic,
2. Branch 2 (empirical): statistical analysis of controlled experiments,
3. Branch 3? (computational): large scale simulations.



# The Ubiquity of Error

- The central motivation for the scientific method is to root out error:
  1. *Deductive branch*: the well-defined concept of the proof,
  2. *Empirical branch*: the machinery of hypothesis testing, structured communication of methods and protocols.
- Computational science as practiced today does not generate reliable knowledge. “breezy demos”

See e.g. Ioannidis, “Why Most Published Research Findings are False,” PLoS Med, 2005.

# A Grassroots Movement

- Science, Dec 2, 2011: special issue on “Data Replication and Reproducibility”
- 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 ongoing: “Review of Omics-based Tests for Predicting Patient Outcomes in Clinical Trials”
- ...

# A Call to Action

- Data and code must be open and long term access assured as an issue of scientific imperative.
- Without the ability to conveniently verify published results, computational science cannot be considered a branch of the scientific method.
- *Reproducible Research Standard* (Stodden 2009): open licensing for computational science publication.
- Reproducible Research provides a path to achieve both these goals.

# References

- “The Scientific Method in Practice: Reproducibility in the Computational Sciences”
- “Open Science: Policy Implications for the Evolving Phenomenon of User-led Scientific Innovation”
- “Enabling Reproducible Research: Open Licensing for Scientific Innovation”
- Reproducible Research: Tools and Strategies for Scientific Computing, July 2011
- Reproducible Research in Computational Science: What, Why and How, Community Forum, July 2011

available at <http://www.stodden.net>