Disseminating Reproducible Computational Research: Tools, Innovations, and Best Practices

Victoria Stodden
Department of Statistics
Columbia University

SIAM Computing in Science and Engineering
Boston, MA
Feb 28, 2013

News: Obama's EM

- On Feb 22, 2013 the Office of Science and Technology Policy released an Executive Memorandum instructing Federal Agencies with more than $100m in research expenditures to devise plans to:

1. Make peer reviewed research publications openly available within 12 months of publication,

2. Make digitally formatted data arising from federal grants should be stored and publicly accessible to search, retrieve, and analyze.

Each Public Access Plan Shall...

a) Maximize access, by the general public and without charge, to digitally formatted scientific data created with Federal funds, while:

i) protecting confidentiality and personal privacy,

ii) recognizing proprietary interests, business confidential information, and intellectual property rights and avoiding significant negative impact on intellectual property rights, innovation, and U.S. competitiveness, and

iii) preserving the balance between the relative value of long-term preservation and access and the associated cost and administrative burden;

b) Ensure that all extramural researchers receiving Federal grants and contracts for scientific research and intramural researchers develop data management plans, as appropriate, describing how they will provide for long-term preservation of, and access to, scientific data in digital formats resulting from federally funded research, or explaining why longterm preservation and access cannot be justified;

c) Allow the inclusion of appropriate costs for data management and access in proposals for Federal funding for scientific research;

d) Ensure appropriate evaluation of the merits of submitted data management plans;

e) Include mechanisms to ensure that intramural and extramural researchers comply with data management plans and policies;

f) Promote the deposit of data in publicly accessible databases, where appropriate and available;

g) Encourage cooperation with the private sector to improve data access and compatibility, including through the formation of public-private partnerships with foundations and other research funding organizations;

h) Develop approaches for identifying and providing appropriate attribution to scientific data sets that are made available under the plan;

i) In coordination with other agencies and the private sector, support training, education, and workforce development related to scientific data management, analysis, storage, preservation, and stewardship; and

j) Provide for the assessment of long-term needs for the preservation of scientific data in fields that the agency supports and outline options for developing and sustaining repositories for scientific data in digital formats, taking into account the efforts of public and private sector entities.

Congress: America COMPETES

- America COMPETES Re-authorization (2011):
  - § 103: Interagency Public Access Committee:

    "coordinate Federal science agency research and policies related to the dissemination and long-term stewardship of the results of unclassified research, *including digital data* and peer-reviewed scholarly publications, supported wholly, or in part, by funding from the Federal science agencies." (emphasis added)

  - § 104: Federal Scientific Collections: OSTP "shall develop policies for the management and use of Federal scientific collections to improve the quality, organization, *access, including online access*, and long-term preservation of such collections for the benefit of the scientific enterprise." (emphasis added)

Whitehouse RFIs

- "Public Access to Peer-Reviewed Scholarly Publications Resulting From Federally Funded Research"
- "Public Access to Digital Data Resulting From Federally Funded Scientific Research"

Comments were due January 12, 2012.

President Obama's first executive memorandum stressed transparency in government, ie. http://data.gov

Credibility Crisis

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

Generally, data and code not made available at the time of publication, insufficient information captured in the publication for verification, replication of results.

➡ **A *Credibility Crisis***

Updating the Scientific Method

Argument: computation presents only a *potential* third branch of the scientific method (Stodden et al 2009):

- Branch 1 (deductive): mathematics, formal logic,
- Branch 2 (empirical): statistical analysis of controlled experiments,
- Branch 3,4? (computational): large scale simulations / data driven computational science.

The Ubiquity of Error

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

Sharing Incentives

Code		Data
91%	Encourage scientific advancement	81%
90%	Encourage sharing in others	79%
86%	Be a good community member	79%
82%	Set a standard for the field	76%
85%	Improve the calibre of research	74%
81%	Get others to work on the problem	79%
85%	Increase in publicity	73%
78%	Opportunity for feedback	71%
71%	Finding collaborators	71%

Survey of the Machine Learning Community, NIPS (Stodden 2010)

Barriers to Sharing

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%

Survey of the Machine Learning Community, NIPS (Stodden 2010)

Tools for Computational Science

- Dissemination Platforms:

  RunMyCode.org | IPOL | Madagascar
  MLOSS.org | thedatahub.org | nanoHUB.org
  Open Science Framework

- Workflow Tracking and Research Environments:

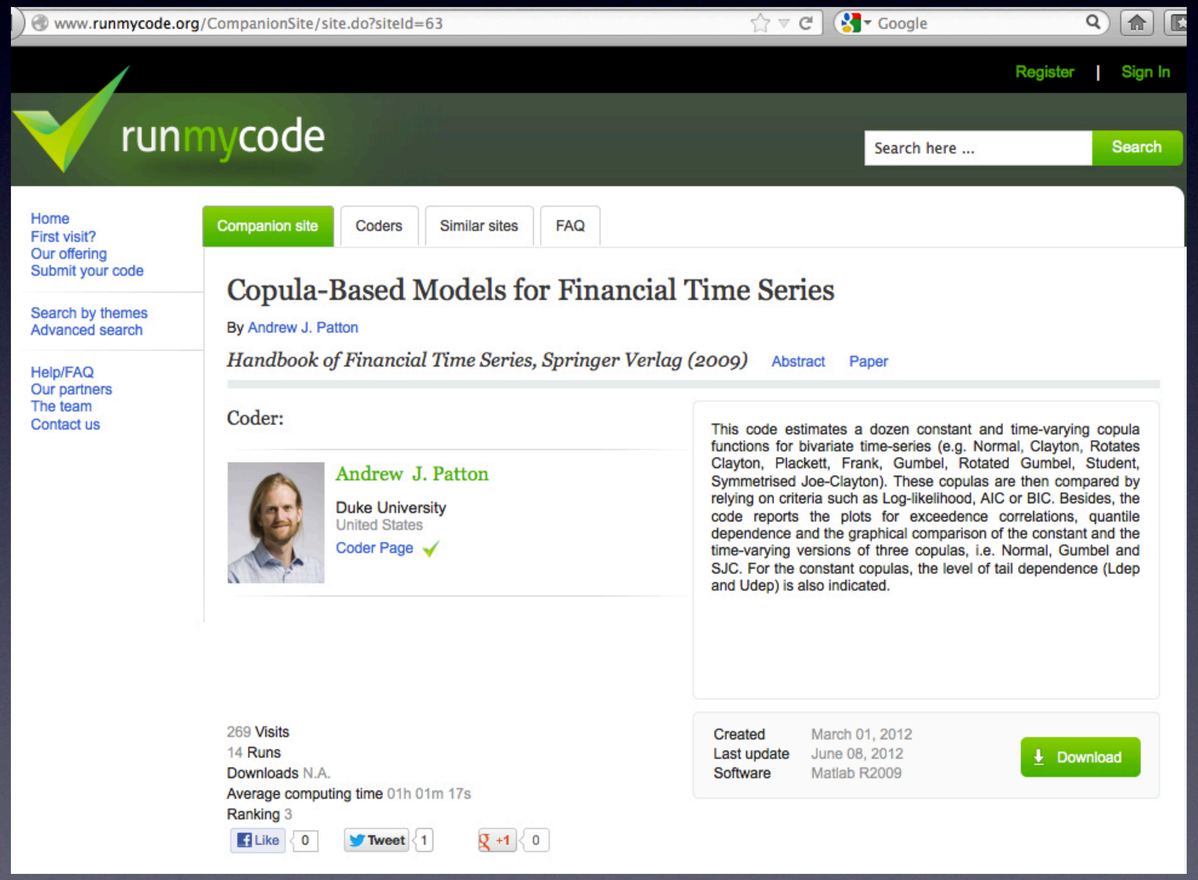
  VisTrails | Kepler | CDE
  Galaxy | GenePattern | Paper Mâché
  Sumatra | Taverna | Pegasus

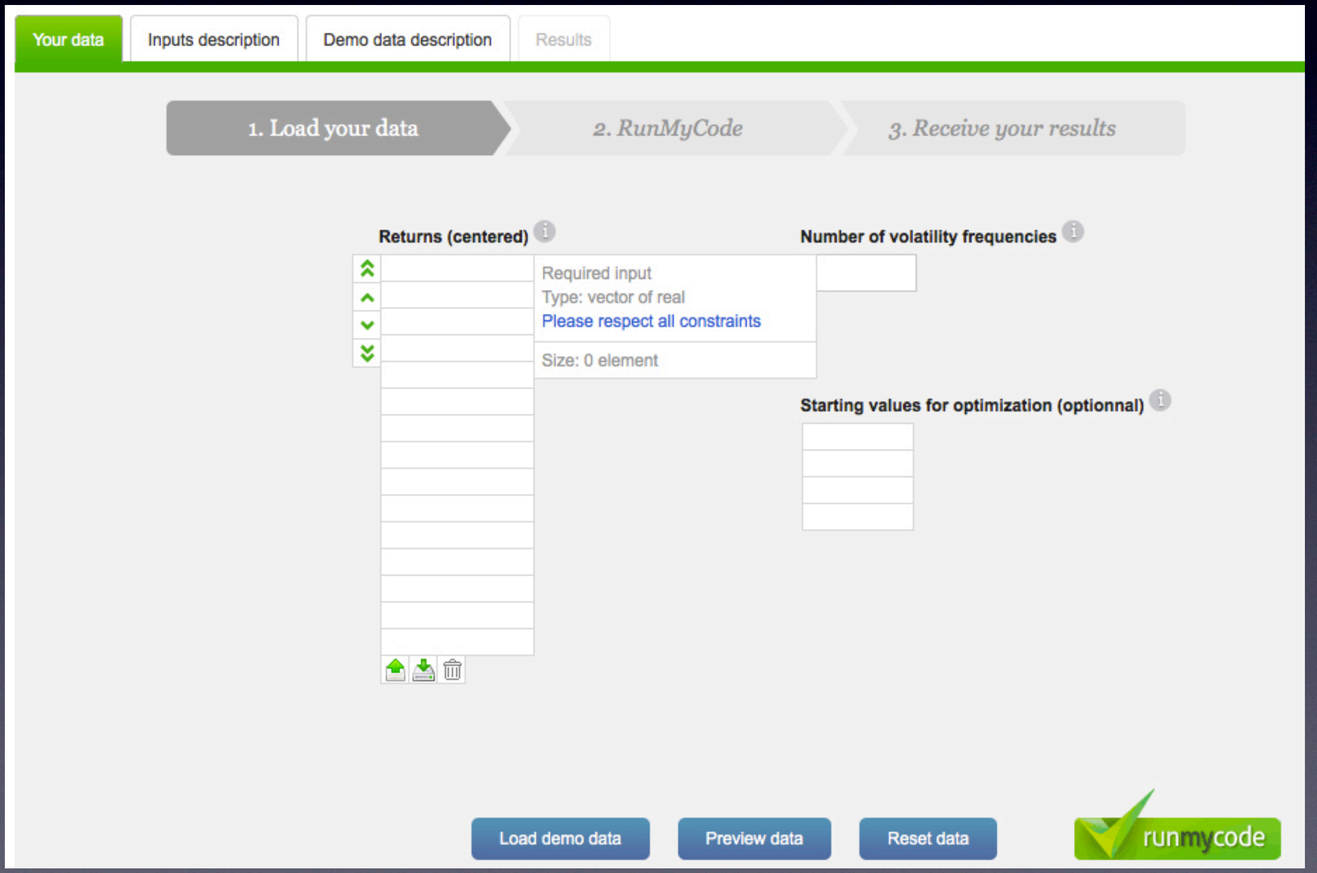
- Embedded Publishing:

  Verifiable Computational Research | Sweave
  Collage Authoring Environment | SHARE

RunMyCode.org

The Companion Page





Best Practices in Licensing

- Software is both copyrighted (by default) and patentable.
- Copyright: author sets terms of use using an open license:
  - Attribution only (ie. Modified BSD, MIT license, LGPL)
  - *Reproducible Research Standard (Stodden 2009)*
- Patents: Bayh-Dole (1980) vs reproducible research (Stodden 2012)
  - delays, barriers to software access
  - *Bilski v Kappos (2011)*

Legal Barriers: Copyright

"To promote the Progress of Science and useful Arts, by securing for limited Times to Authors and Inventors the exclusive Right to their respective Writings and Discoveries." (U.S. Const. art. I, §8, cl. 8)

- Original expression of ideas falls under copyright *by default* (papers, code, figures, tables..)
- Copyright secures exclusive rights vested in the author to:
  - reproduce the work
  - prepare derivative works based upon the original

Exceptions and Limitations: Fair Use.

Responses Outside the Sciences 1: Open Source Software

- Software with licenses that communicate alternative terms of use to code developers, rather than the copyright default
- Hundreds of open source software licenses:
  - GNU Public License (GPL)
  - (Modified) BSD License
  - MIT License
  - Apache 2.0 License
  - ... see http://www.opensource.org/licenses/alphabetical

Responses Outside the Sciences 2: Creative Commons

- Founded in 2001, by Stanford Law Professor Larry Lessig, MIT EECS Professor Hal Abelson, and advocate Eric Eldred.
- Adapts the Open Source Software approach to artistic and creative digital works.

Response from Within the Sciences

The *Reproducible Research Standard* (*RRS*) (Stodden, 2009)

- A suite of license recommendations for computational science:
    - Release media components (text, figures) under CC BY,
    - Release code components under Modified BSD or similar,
    - Release data to public domain or attach attribution license.

➡ Remove copyright's barrier to reproducible research and,

➡ Realign the IP framework with longstanding scientific norms.

Winner of the Access to Knowledge Kaltura Award 2008

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

ICERM Workshop 2012

"Reproducibility in Computational and Experimental Mathematics,"

- 6 organizers, held December 10-14, 2012.
- ~70 participants; talks, demos, lightning talks, breakout groups. (See ICERM webpage: http://icerm.brown.edu/tw12-5-rcem)
- Workshop Report: "Setting the Default to Reproducible"
- Workshop wiki: http://wiki.stodden.net

“Setting the Default to Reproducible”

- Workshop report distills discussion and breakout group feedback into 3 main recommendations:

1. It is important to promote a culture change that will integrate computational reproducibility into the research process.

2. Journals, funding agencies, and employers should support this culture change.

3. Reproducible research practices and the use of appropriate tools should be taught as standard operating procedure in relation to computational aspects of research.