

Digital Scholarship in Scientific Research: Open Questions in Reproducibility and Curation

Victoria Stodden
Department of Statistics
Columbia University

Open Access Event
University of Wisconsin - Milwaukee
Feb 8, 2013

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***

February 8, 2013

HUFF
POST SCIENCE

Set the Default to "Open": Reproducible Science in the Computer Age

Posted: 02/07/2013 2:48 pm

It has been conventional wisdom that computing is the "third leg" of the stool of modern science, complementing theory and experiment. But that metaphor is no longer accurate. Instead, computing now pervades all of science, including theory and experiment. Nowadays massive computation is required just to reduce and analyze experimental data, and simulations and computational explorations are employed in fields as diverse as climate modeling and research mathematics.

Unfortunately, the culture of scientific computing has not kept pace with its rapidly ascending pre-eminence in the broad domain of scientific research. In experimental research work, researchers are taught early the importance of keeping notebooks or computer-based logs of every detail of their work---experimental design, procedures, equipment used, raw results, processing techniques, statistical methods used to analyze the results, and other relevant details of an experiment.

My own experience (the long tail)

- our group at Stanford practiced “really reproducible research” inspired by Stanford Professor Jon Claerbout:

“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.

Example: Wavelab (1999)



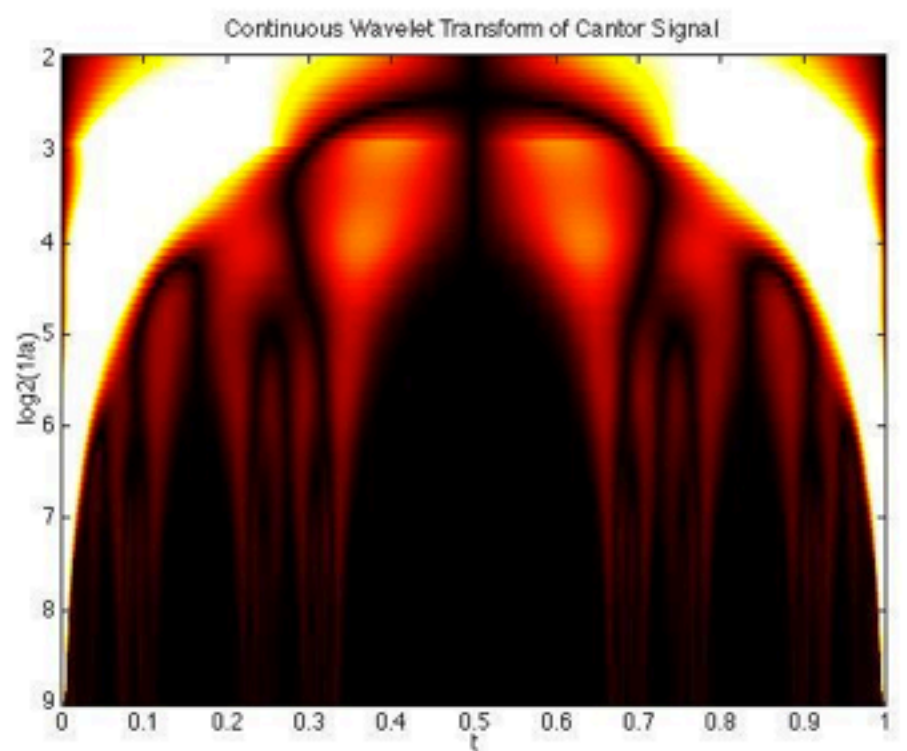
The screenshot shows a web browser window with the title "WaveLab802" and the address bar displaying "www-stat.stanford.edu/~wavelab/". The main heading "WAVELAB 850" is in large, bold, orange letters. Below it is a navigation bar with links: [Home](#), [Download](#), [Documentation](#), [Registration](#), [Links](#), [Contact](#), and [Acknowledgements](#). The main content area features a figure on the left and text on the right. The figure is a "Continuous Wavelet Transform of Cantor Signal" plot, showing a complex, fractal-like pattern of red and yellow on a black background. The x-axis is labeled t and ranges from 0 to 1. The y-axis is labeled $\log_2(1/a)$ and ranges from 2 to 9. To the right of the figure, the text describes WaveLab as a collection of Matlab functions for wavelet analysis, listing several techniques: orthogonal and biorthogonal wavelet transforms, translation-invariant wavelets, interpolating wavelet transforms, cosine packets, wavelet packets, and matching pursuit. It also includes a link to a more detailed introduction. Below this, the "Philosophy--why do it?" section explains that WaveLab implements the concept of *reproducible research*, where the software and instructions are made available alongside the research results.

WaveLab802

www-stat.stanford.edu/~wavelab/

WAVELAB 850

[Home](#) [Download](#) [Documentation](#) [Registration](#) [Links](#) [Contact](#) [Acknowledgements](#)



Continuous Wavelet Transform of Cantor Signal

WaveLab is a collection of Matlab functions that have been used by the authors and collaborators to implement a variety of algorithms related to wavelet analysis. A partial list of the techniques made available:

- orthogonal and biorthogonal wavelet transforms,
- translation-invariant wavelets,
- interpolating wavelet transforms,
- cosine packets,
- wavelet packets,
- matching pursuit,

and a lot more; [Here is a more detailed introduction.](#)

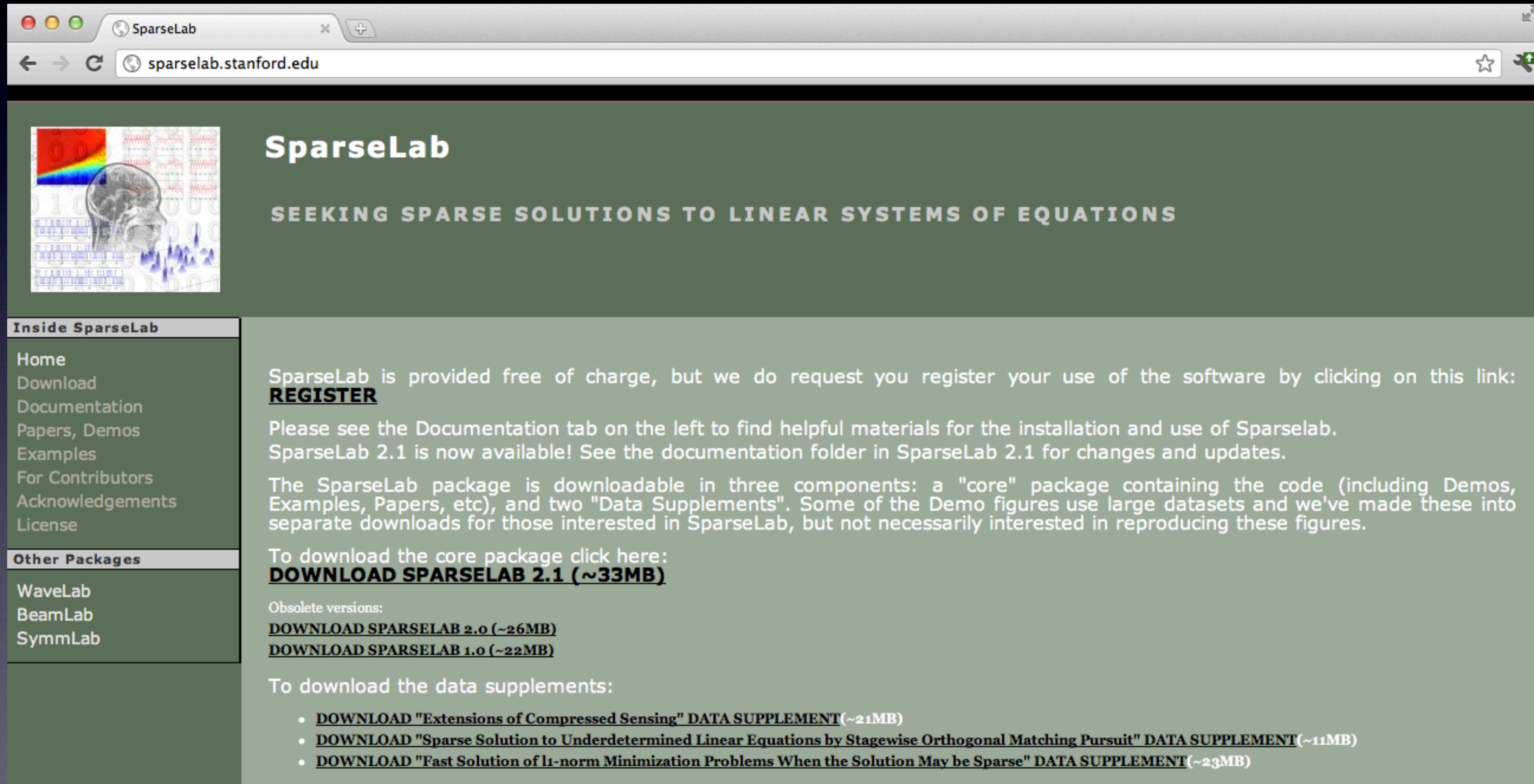
Philosophy--why do it?

WaveLab implements the concept of *reproducible research*.

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.

We make WaveLab available to make the full content of our scholarship available, enabling others to understand and reproduce our work.

Example: Sparselab (2006)



The screenshot shows a web browser window with the address bar displaying "sparselab.stanford.edu". The page has a dark green header with the "SparseLab" logo and the tagline "SEEKING SPARSE SOLUTIONS TO LINEAR SYSTEMS OF EQUATIONS". A sidebar on the left contains navigation links under "Inside SparseLab" and "Other Packages". The main content area provides information about the software, including a registration link, documentation, and download links for the core package and data supplements.

SparseLab

SEEKING SPARSE SOLUTIONS TO LINEAR SYSTEMS OF EQUATIONS

Inside SparseLab

- Home
- Download
- Documentation
- Papers, Demos
- Examples
- For Contributors
- Acknowledgements
- License

Other Packages

- WaveLab
- BeamLab
- SymmLab

SparseLab is provided free of charge, but we do request you register your use of the software by clicking on this link: **REGISTER**

Please see the Documentation tab on the left to find helpful materials for the installation and use of Sparselab. SparseLab 2.1 is now available! See the documentation folder in SparseLab 2.1 for changes and updates.

The SparseLab package is downloadable in three components: a "core" package containing the code (including Demos, Examples, Papers, etc), and two "Data Supplements". Some of the Demo figures use large datasets and we've made these into separate downloads for those interested in SparseLab, but not necessarily interested in reproducing these figures.

To download the core package click here: **DOWNLOAD SPARSELAB 2.1 (~33MB)**

Obsolete versions:

- DOWNLOAD SPARSELAB 2.0 (~26MB)**
- DOWNLOAD SPARSELAB 1.0 (~22MB)**

To download the data supplements:

- **DOWNLOAD "Extensions of Compressed Sensing" DATA SUPPLEMENT (~21MB)**
- **DOWNLOAD "Sparse Solution to Underdetermined Linear Equations by Stagewise Orthogonal Matching Pursuit" DATA SUPPLEMENT (~11MB)**
- **DOWNLOAD "Fast Solution of l1-norm Minimization Problems When the Solution May be Sparse" DATA SUPPLEMENT (~23MB)**

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.
- 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.

Digital Scientific Transparency

- raises information issues:
 - ▶ incentives for sharing, barriers to data and code availability,
 - ▶ lifecycle of data/code, stewardship of digital scholarly objects,
 - ▶ metadata, provenance, curation issues.
- accelerates scientific discovery:
 - ▶ broad validation of scientific findings,
 - ▶ facilitating dataset recombination and linking, avoiding duplication of code.

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%

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%

Intellectual Property Barriers

- 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 I: 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

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

[Register](#) | [Sign In](#)



Search here ...

Search

[Home](#)
[First visit?](#)
[Our offering](#)
[Submit your code](#)

[Search by themes](#)
[Advanced search](#)

[Help/FAQ](#)
[Our partners](#)
[The team](#)
[Contact us](#)

The concept

As simple as 1,2,3

1. A researcher has an **idea**.
2. The researcher writes a **paper** based on this idea.
3. Using RunMyCode, the researcher creates a **companion website** associated with this paper. The companion website allows people to implement the methodology presented in the paper.

[Learn more >>](#)



[About](#)

[Concept](#)

[Purpose](#)

[Create your own companion website >>](#)

The Companion Page

www.runmycode.org/CompanionSite/site.do?siteId=63

runmycode

Register | Sign In

Search here ... Search

Home
First visit?
Our offering
Submit your code

Search by themes
Advanced search

Help/FAQ
Our partners
The team
Contact us


Companion site Coders Similar sites FAQ

Copula-Based Models for Financial Time Series

By Andrew J. Patton

Handbook of Financial Time Series, Springer Verlag (2009) Abstract Paper

Coder:

 **Andrew J. Patton**
Duke University
United States
[Coder Page](#) ✓

This code estimates a dozen constant and time-varying copula functions for bivariate time-series (e.g. Normal, Clayton, Rotates Clayton, Plackett, Frank, Gumbel, Rotated Gumbel, Student, Symmetrised Joe-Clayton). These copulas are then compared by relying on criteria such as Log-likelihood, AIC or BIC. Besides, the code reports the plots for exceedence correlations, quantile dependence and the graphical comparison of the constant and the time-varying versions of three copulas, i.e. Normal, Gumbel and SJC. For the constant copulas, the level of tail dependence (Ldep and Udep) is also indicated.

Created March 01, 2012
Last update June 08, 2012
Software Matlab R2009

269 Visits
14 Runs
Downloads N.A.
Average computing time 01h 01m 17s
Ranking 3

[Like](#) 0 [Tweet](#) 1 [+1](#) 0

[Download](#)

Your data Inputs description Demo data description Results

1. Load your data 2. RunMyCode 3. Receive your results

Returns (centered) *i*

Number of volatility frequencies *i*

Starting values for optimization (optional) *i*

Required input
Type: vector of real
Please respect all constraints
Size: 0 element

Load demo data Preview data Reset data

runmycode

RunMyCode.org

- inform research on sharing, scientific transparency, impact of computation on discovery and validation:
 - ▶ facilitate code and data sharing, alongside published articles,
 - ▶ longevity and persistence of digital scholarly objects - 10 year guarantee (via partnerships) including metadata,
 - ▶ recognize data, code, and reimplementations contributions,
 - ▶ execution of code in the cloud, or locally,
 - ▶ public interaction/access, community engagement, large scale validation, acceleration of discoveries,
 - ▶ understand the data lifecycle, reuse, best practices.

Sharing: Journal Policy

- Journal Policy setting study design:
- Select all journals from ISI classifications “Statistics & Probability,” “Mathematical & Computational Biology,” and “Multidisciplinary Sciences” (this includes Science and Nature).
- $N = 170$, after deleting journals that have ceased publication.
- Create dataset with ISI information (impact factor, citations, publisher) and supplement with publication policies as listed on journal websites, in June 2011 and June 2012.

Data Sharing Policy

	2011	2012	Change
Required as condition of publication, barring exceptions	18	19	1
Required but may not affect editorial decisions	3	10	7
Explicitly encouraged/addressed, may be reviewed and/or hosted	35	30	-5
Implied	0	5	5
No mention	114	106	-8

Code Sharing Policy

	2011	2012	Change
Required as condition of publication, barring exceptions	6	6	0
Required but may not affect editorial decisions	6	6	0
Explicitly encouraged/addressed, may be reviewed and/or hosted	17	21	4
Implied	0	3	3
No mention	141	134	-7

Findings

- Journals generally not hosting data/code.
- Changemakers are journals with high impact factors.
- Progressive policies are not widespread, but being adopted rapidly.
- Close relationship between the existence of a supplemental materials policy and a data policy.
- Data and supplemental material policies appear to lead software policy.

Barriers to Journal Policy Making

- Standards for code and data sharing,
- Meta-data, archiving, re-use, documentation, sharing platforms, citation standards,
- Review, who checks replication, if anyone,
- Burdens on authors, especially less technical authors,
- Evolving, early research; affects decisions on when to publish,
- Business concerns, attracting the best papers.

Sharing: Funding Agency Policy

- NSF grant guidelines: “NSF ... expects investigators to share with other researchers, at no more than incremental cost and within a reasonable time, the data, samples, physical collections and other supporting materials created or gathered in the course of the work. It also encourages grantees to share software and inventions or otherwise act to make the innovations they embody widely useful and usable.” (2005 and earlier)
- NSF peer-reviewed Data Management Plan (DMP), January 2011.
- NIH (2003): “The NIH endorses the sharing of final research data to serve these and other important scientific goals. The NIH expects and supports the timely release and sharing of final research data from NIH-supported studies for use by other researchers.” (>\$500,000, include data sharing plan)

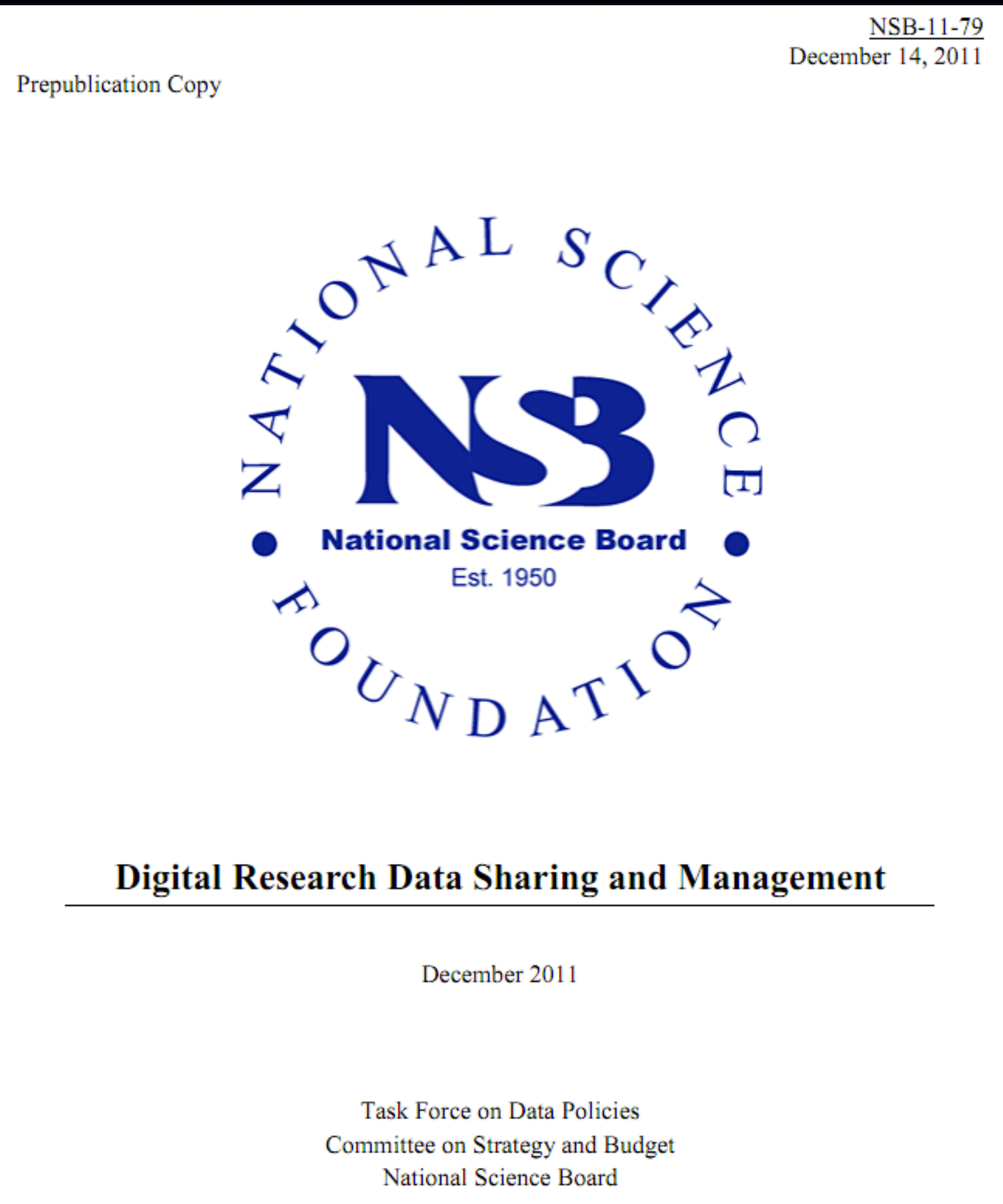
NSF Data Management Plan

“Proposals submitted or due on or after January 18, 2011, must include a supplementary document of no more than two pages labeled ‘Data Management Plan.’ This supplementary document should describe how the proposal will conform to NSF policy on the dissemination and sharing of research results.” (<http://www.nsf.gov/bfa/dias/policy/dmp.jsp>)

NSF Data Management Plan

- No requirement or directives regarding data openness specifically.
- But, “Investigators are expected to share with other researchers, at no more than incremental cost and within a reasonable time, the primary data, samples, physical collections and other supporting materials created or gathered in the course of work under NSF grants. Grantees are expected to encourage and facilitate such sharing. Privileged or confidential information should be released only in a form that protects the privacy of individuals and subjects involved.” (http://www.nsf.gov/pubs/policydocs/pappguide/nsf11001/aag_6.jsp#VID4)

National Science Board Report



“Digital Research Data Sharing and Management,”
December 2011.

[http://www.nsf.gov/nsb/publications/2011/
nsb1124.pdf](http://www.nsf.gov/nsb/publications/2011/nsb1124.pdf)

Rethinking Discovery in Big Data

- The changing role of statistics within modern scientific discovery:
- August 2012: a Subcommittee of the Mathematical and Physical Sciences Advisory Committee, 'Support for the Statistical Sciences at NSF' formed to understand "the growing role of statistics in all areas of science and engineering, including the changing character of research across the spectrum of 'individual investigator' and 'group' science."
- opportunity for integrated thinking regarding research modalities and dissemination

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>

A Grassroots Movement

- ICERM 2012 “Reproducibility in Computational and Experimental Mathematics”
- AMP 2011 “Reproducible Research: Tools and Strategies for Scientific Computing”
- Open Science Framework / Reproducibility Project in Psychology
- 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” ...

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>