

The Central Role of Geophysics in the Reproducible Research Movement

Victoria Stodden
Department of Statistics
Columbia University

Open-source Exploration and Production Software - Six Years Later
74th European Association of Geoscientists and Engineers Conference
Copenhagen, Denmark
June 8, 2012

The Credibility Crisis in Computational Science

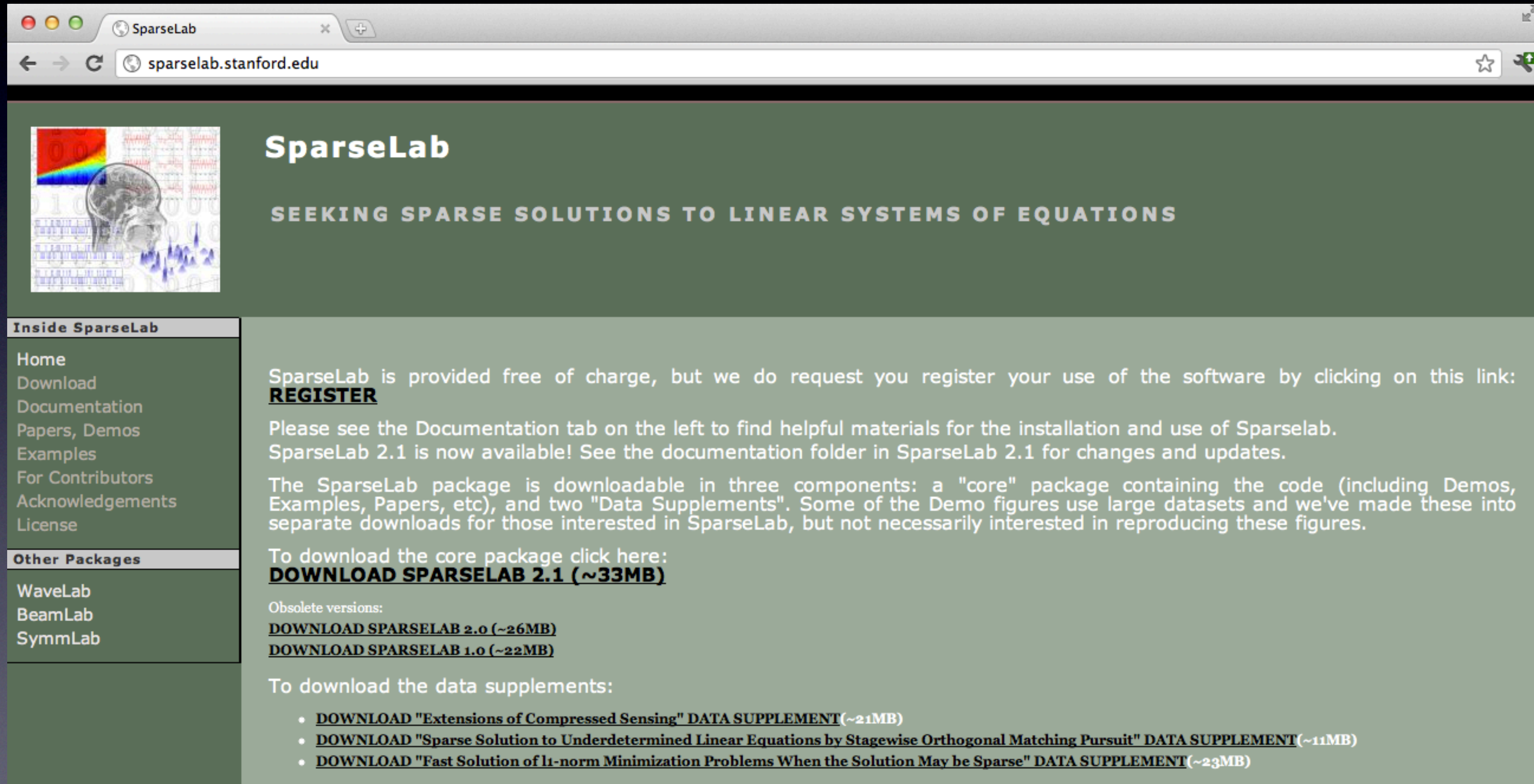
JASA June	Computational Articles	Code Publicly Available
1996	9 of 20	0%
2006	33 of 35	9%
2009	32 of 32	16%
2011	29 of 29	21%

My own experience

- 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: Sparselab (2006)



The screenshot shows a web browser window with the URL `sparselab.stanford.edu`. The page features a header with the title "SparseLab" 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 l_1 -norm Minimization Problems When the Solution May be Sparse" DATA SUPPLEMENT (~23MB)**

Updating the Scientific Method

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

- 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.

Survey of the Machine Learning Community, NIPS (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%

Solutions are interlocking..

1. Tools
2. Intellectual Property Barriers
3. Funding Agency Policy / Federal Regulations
4. Journal Policy
5. Institutional Expectations

Solution Component I: Tools

- *Dissemination Platforms:*

Madagascar

RunMyCode.org

HUBzero.org

MLOSS.org

thedatahub.org

nanoHUB.org

- *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

Solution Component 2: IP

- Software is both copyrighted and possibly patentable. Data is copyright in Europe and may have copyright in the US.
 - Copyright: author sets terms of use using an open license.
 - Patents: Bayh-Dole (1980) vs reproducible research.

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.



Responses Outside the Sciences 2: Creative Commons

- Creative Commons provides a suite of licensing options for digital artistic works:
 - BY: if you use the work attribution must be provided,
 - NC: the work cannot be used for commercial purposes,
 - ND: no derivative works permitted,
 - SA: derivative works must carry the same license as the original

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 Kultura Award 2008

ShareAlike isn't for Science

The motivations and goals of the Open Source Software community differ from those of the scientific community:

- industry collaboration and re-use of code,
- different licensing needs in different scientific projects,
- mixing of scientific codes,
- scientific knowledge as a public good.

➔ The extra restrictions of ShareAlike cost more than they benefit.

Copyright and Data

- Copyright adheres to raw facts in Europe.
- In the US raw facts are not copyrightable, but the original “selection and arrangement” of these facts is copyrightable. (Feist Publns Inc. v. Rural Tel. Serv. Co., 499 U.S. 340 (1991)).
 - ➔ the possibility of a residual copyright in data (attribution licensing or public domain certification).
 - ➔ Law doesn't match reality on the ground: What constitutes a “raw” fact anyway?

Other Legal Barriers

- HIPAA (Health Information Portability and Accountability Act) and privacy regulations,
- Collaboration agreements with industry / trade secrets,
- Hiring agreements, institutional rules,
- National security.

Bayh-Dole and Software Patents

- Bayh-Dole (1980) to create incentives for universities to patent, thereby making inventions accessible,
- computational scientists: patents vs share code for verification and reproducibility,
- incentives distortion (e.g. In Re Bilski): potential code withholding, obfuscation, startups vs science.

Solution Component 3: 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

- 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)

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>

Solution Component 4: Journal Policy

Computational Science Journals (Stodden and Guo, preliminary results)

Stated Policy, Summer 2011

Proportion requiring data	15%
Proportion requiring code	7%
Proportion requiring supplemental materials	9%
Proportion Open Access	58%

N=170; journals classified using Web of Science classifications (Mathematical & Computational Biology, Statistics & Probability, Multidisciplinary Science).

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.

Solution Component 5: Institutional Expectations



"Behind one door is tenure - behind the other is flipping burgers at McDonald's."

Challenges to Open Science

(even if it was easy)

- “Taleb Effect” - scientific discoveries as (misused) black boxes,
- nefarious uses / public deception
- black boxes and opacity in software (why the traditional methods section is inadequate: massive codebases, industry hosted platforms),
- lock-in: calcification of ideas in software?
- independent replication discouraged?
- policy maker engagement: advocacy within the scientific community.

This is a Grassroots Movement

- 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, Workshop, July 2011
- Reproducible Research in Computational Science: What, Why and How, Community Forum, July 2011
available at <http://www.stodden.net>