

# Data-Intensive Science: Methods for Reproducibility and Dissemination

Victoria Stodden  
Department of Statistics  
Columbia University

IGERT Brown Bag Lunch Seminar  
Columbia University  
November 13, 2012



# Agenda

1. Examples of Computational Science
2. The Lifecycle of Code and Data
3. Principled Sharing of Scientific Output
4. Sharing Tools and Modalities
5. Open Questions in Open Science



# Computation is Becoming Central to Scientific Research

1. enormous, and increasing, amounts of data collection:
  - CMS project at LHC: 300 “events” per second, 5.2M seconds of runtime per year, .5MB per event = 780TB/yr => several PB when data processed,
  - Sloan Digital Sky Survey: 8th data release (2010), 49.5TB,
  - quantitative revolution in social science due to abundance of social network data (Lazier et al, *Science*, 2009)
  - Science survey of peer reviewers: 340 researchers regularly work with datasets >100GB; 119 regularly work with datasets >1TB (N=1700, Feb 11, 2011, p. 692)
2. massive simulations of the complete evolution of a physical system, systematically varying parameters,
3. deep intellectual contributions now encoded in software.



# Conjecture: degrees of digitization exist in most research today

- simple example: Researcher uses his or her laptop to store data and do simple calculations. e.g. social sciences, lab work, field work.
- more complex example: computer scripts written to implement algorithms. e.g. image processing, Wordhoard project, computational mathematics.
- even more complexity: massive data (e.g. genomics or medical claims data) and/or massive code bases.
- most complex: multicore parallel processing for real data or simulations. e.g. geophysical modeling, astrophysics, high energy physics.



# My own experience

- our lab 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.



# 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

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? (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.



# Legal Barriers Affecting Scientists



# 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.





# 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



# 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,
- Incentives to patent and commercialize,
- Collaboration agreements with industry,
- Hiring agreements, institutional rules,
- National security.



# Funding Agency Policy



# 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>)

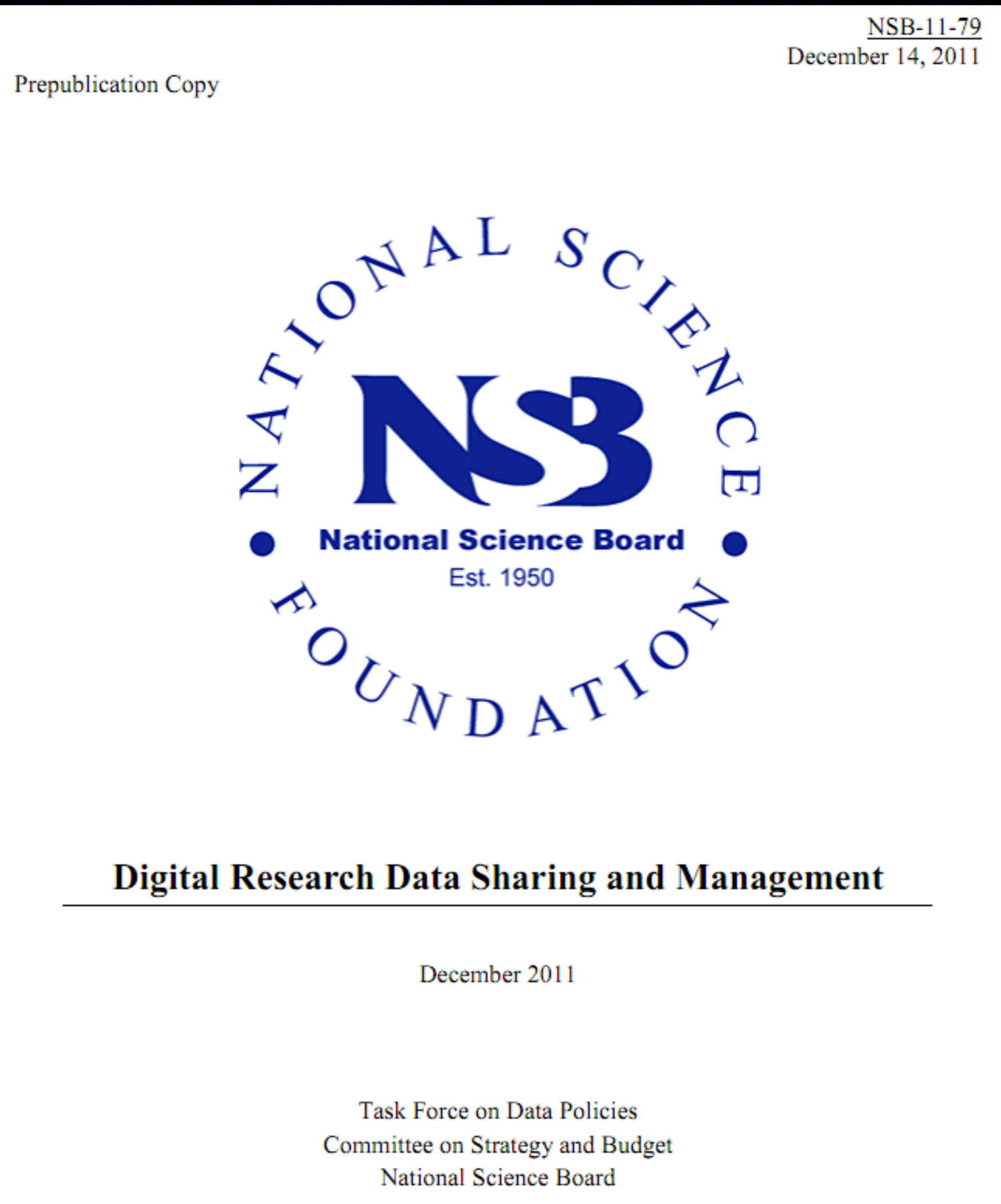


# NSF Data Management Plan

- 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](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)



# 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.



# Barriers Facing Scientists



# 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%



# Access to Scientific Knowledge



# Journal Requirements

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.



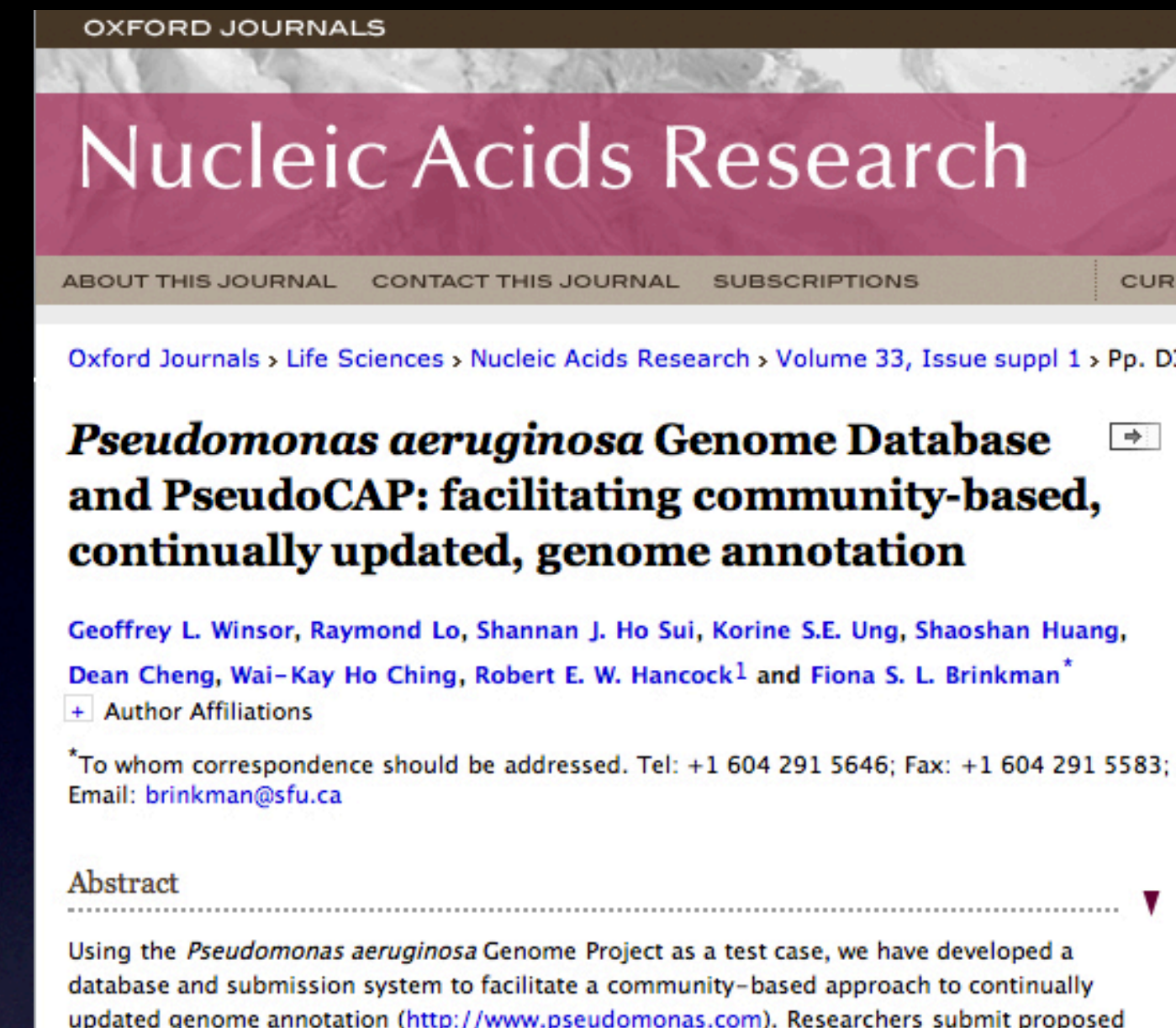
# Barriers to Journal Policy Making

- Standards for code and data sharing,
- Meta-data, archiving, re-use, documentation, sharing platforms,
- 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.



# Citation and Contributions

- Evaluation standards: citation and publication record
- Collaborative efforts in database building?
- Differential citation? (web vs articles, microcitation)
- Database versioning (e.g. King & Altman 2007, Donoho & Gavish 2011)
- Citizen contributions? (Galaxy Zoo, Open Dinosaur Project)
- Code development? pre-publication review?
- Code maintenance for post-publication reproducibility, scientific reuse?
  - platform building (DANSE, Madagascar, Wavelab, Sparselab)





# Tools and Platforms



# Barriers from Computational Infrastructure

- software typically used in science is a dialog, not envisioned as collaboration,
- tools to facilitate (later) sharing *during* the research process: workflow tracking, data provenance,
- testing for code: unit tests, regression tests,
- versioning: what tool in which platform did you use to get that result?
- systems for opening code collaboration and community building (Github, HUBzero, ...),



# Facilitation by 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

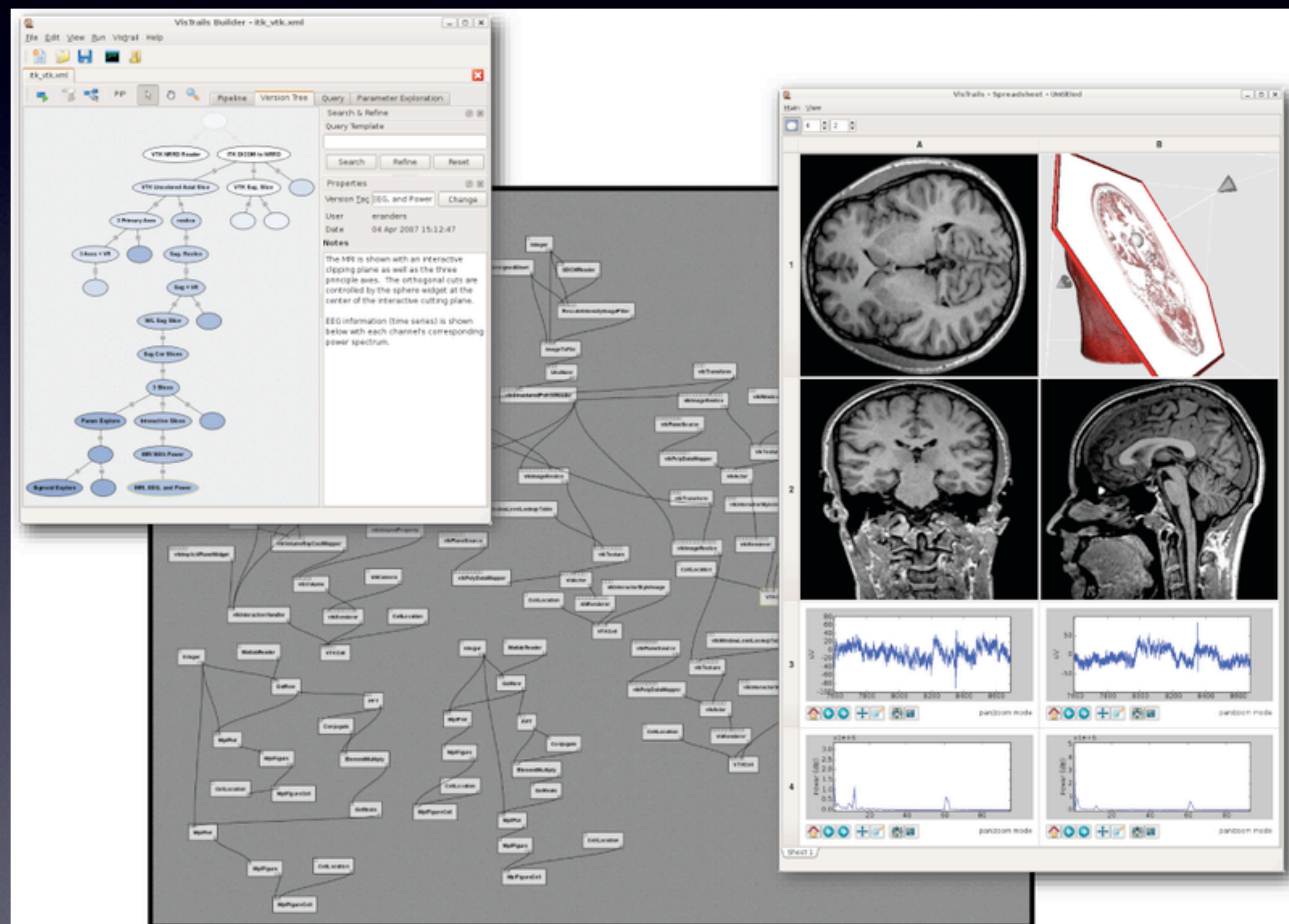
Sweave

Collage Authoring Environment

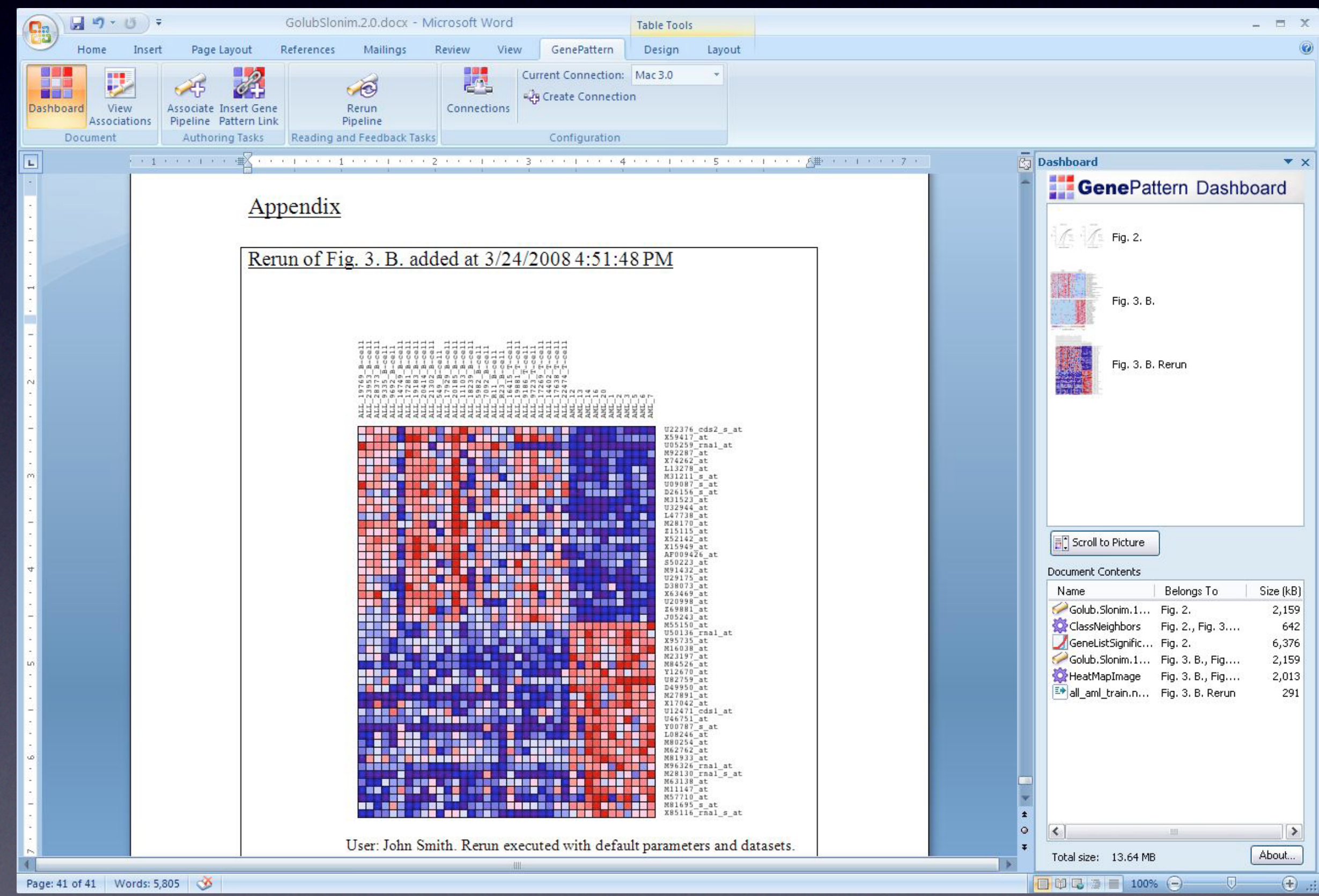
SHARE



# Vistrails and GenePattern Examples



The left side of the image shows the Vistrails Builder interface with a pipeline graph. The right side shows a Vistrails Spreadsheet with a grid of visualizations: four brain MRI slices, four time-series plots, and four power spectrum plots.



The image shows a Microsoft Word document with the following content:

**Appendix**  
**Rerun of Fig. 3. B. added at 3/24/2008 4:51:48 PM**

The heatmap displays gene expression data for various genes (e.g., ALL-13749, U22376) across samples (e.g., U22376\_cds2\_s\_at, X59417\_at).

At the bottom, it says: User: John Smith. Rerun executed with default parameters and datasets.

On the right is the GenePattern Dashboard with a 'Document Contents' table:

Name	Belongs To	Size (kB)
Golub.Slonim.1...	Fig. 2.	2,159
ClassNeighbors	Fig. 2., Fig. 3...	642
GeneListSignific...	Fig. 2.	6,376
Golub.Slonim.1...	Fig. 3. B., Fig...	2,159
HeatMapImage	Fig. 3. B., Fig...	2,013
all_aml_train.n...	Fig. 3. B. Rerun	291



# This is a Grassroots Movement

- 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”
- ...



# 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: finding support for our norms, advocacy within the scientific community.



# 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>