Data and Code Sharing in Bioinformatics:
From Bermuda to Toronto to Your Laptop

Victoria Stodden
Department of Statistics, Columbia University

UC Berkeley Statistics and Genomics Seminar
March 13, 2014

International Strategy Meetings on Human DNA Sequencing

- Bermuda 1996
- Fort Lauderdale 2003
- Amsterdam 2008
- Toronto 2009

Reproducible Research

- Principle of Academic Licensing

The 1996 Bermuda Agreement

Primary Genomic Sequence Should be in the Public Domain
It was agreed that all human genomic sequence information, generated by centres funded for large-scale human sequencing, should be freely available and in the public domain in order to encourage research and development and to maximise its benefit to society.

Primary Genomic Sequence Should be Rapidly Released

- Sequence assemblies should be released as soon as possible; in some centres, assemblies of greater than 1 Kb would be released automatically on a daily basis.
- Finished annotated sequence should be submitted immediately to the public databases.

Bermuda 1997 and 1998

Bermuda 1997 provided agreed standards on error rates and details on submission and annotation. Created a one year maximum claim on a sequence.

Bermuda 1998 extended the human data release principles to other organisms. (not adopted by funding agencies as previous agreements had been.)

The 2003 Fort Lauderdale Agreement

About 40 stakeholders reaffirm Bermuda 1996, and recommend further that:

- Bermuda be extended to apply to all sequence data, including both the raw traces and whole genome shotgun assemblies,
- the principle of rapid pre-publication release should apply to other types of data from other large-scale production centers specifically established as "community resource projects" (ie. International Human Genome Sequencing Consortium, the Mouse Genome Sequencing Consortium, the Mammalian Gene Collection, the SNPs Consortium, and the International HapMap Project)
- pre-publication data release requires community-wide support due to the incentive to publish the first analysis of one's own data.

The 2003 Fort Lauderdale Agreement

Introduces the notion of "Tripartite Sharing of Responsibility"
Summary:

- Funding Agencies: require free and unrestricted data release from community projects in central and searchable databases,
- Resource Producers: publish a Project Description, and make immediate availability of well-described, high quality data,
- Resource Users: cite data sources appropriately, possibly through the Project Description.

The 2008 Amsterdam Agreement

Extends the principle of rapid data release to proteomics data.

Since many center and funding agencies outside the the mainstream remain unaware of these agreements, they are affirmed in Toronto in May 2009.

The 2009 Toronto Agreement

Goals:

- continued policy discussions from the Bermuda and Fort Lauderdale agreements,
- endorsed the value of rapid prepublication data release for large reference data sets in biology and medicine that have broad utility,
- prepublication data release should go beyond genomics and proteomics studies to other data sets and annotated clinical resources (a range of project sizes, minimum standard should be data release at publication),

The 2009 Toronto Agreement

Building on Fort Lauderdale 2003,

- Funding Agencies: announce release requirements; peer review includes dataset release plans; provide help to develop appropriate consent, security, access and governance mechanisms; provide long-term support of databases,
- Data Producers: publish a citable marker paper with dataset information; simultaneous release of relevant metadata; create databases with all versions archived, including raw data,
- Resource Users: allow data producers first analysis, cite data sources accurately and completely, be aware early data may be subject to later quality improvements,
- Scientific Journal Editors: provide guidance to authors and reviewers on the third-party use of prepublication data in manuscripts.

Resolving the Toronto Agreement

Switch to Agenda.

Broadening the Impact: New Haven 2009

Switch to Data and Code Sharing Roundtable, Nov 21, 2009.

The Reproducible Research Standard

Goals:

- realign IP rights with scientific norms,
- release of all scientific research, including code and data.

To satisfy the Reproducible Research Standard:

1. CC BY on media such as text, figures,
2. Attribution license on code: such as Apache 2.0, MIT, LGPL,
3. Data under CC0 or Science Commons Open Access Data Protocol,
4. "original selection and arrangement" of the data, under CC BY or attribution open source license.

Principle of Academic Licensing

Principle of Scientific Licensing: Legal encumbrances to the dissemination, sharing, use, and re-use of scientific research compendia should be minimized, each requiring a strong and compelling rationale before application.

Stodden, *Enabling Reproducible Research: Open Licensing for Scientific Innovation*, 2009, p. 14.

Defining *Reproducibility*

- Empirical Reproducibility: in meatspace
  - traditional notion of reproducibility (Boyle): sufficient description in the research paper for others in the field to replicate
  - access to reagents, cell lines, and other physical materials required to replicate the research
  - Begley, Nature "Try Harder"
- Computational Reproducibility: in silica
  - access to digital data and code used to generate findings
  - execution?
  - algorithm descriptions

Communication Standards: Data and Code

- Gentleman and Temple Lang proposed the *Research Compendia* (2003): computational scholarly communication as a triple of narrative, data, and code.
- Claerbout (1994), Donoho et al (2009): advocated publishing *really reproducible research* including data and code.
- Data sharing in genomics beginning in 1996.
- Gary King (1998) advocated reproducibility in political science and the social sciences.

Rationale

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 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.

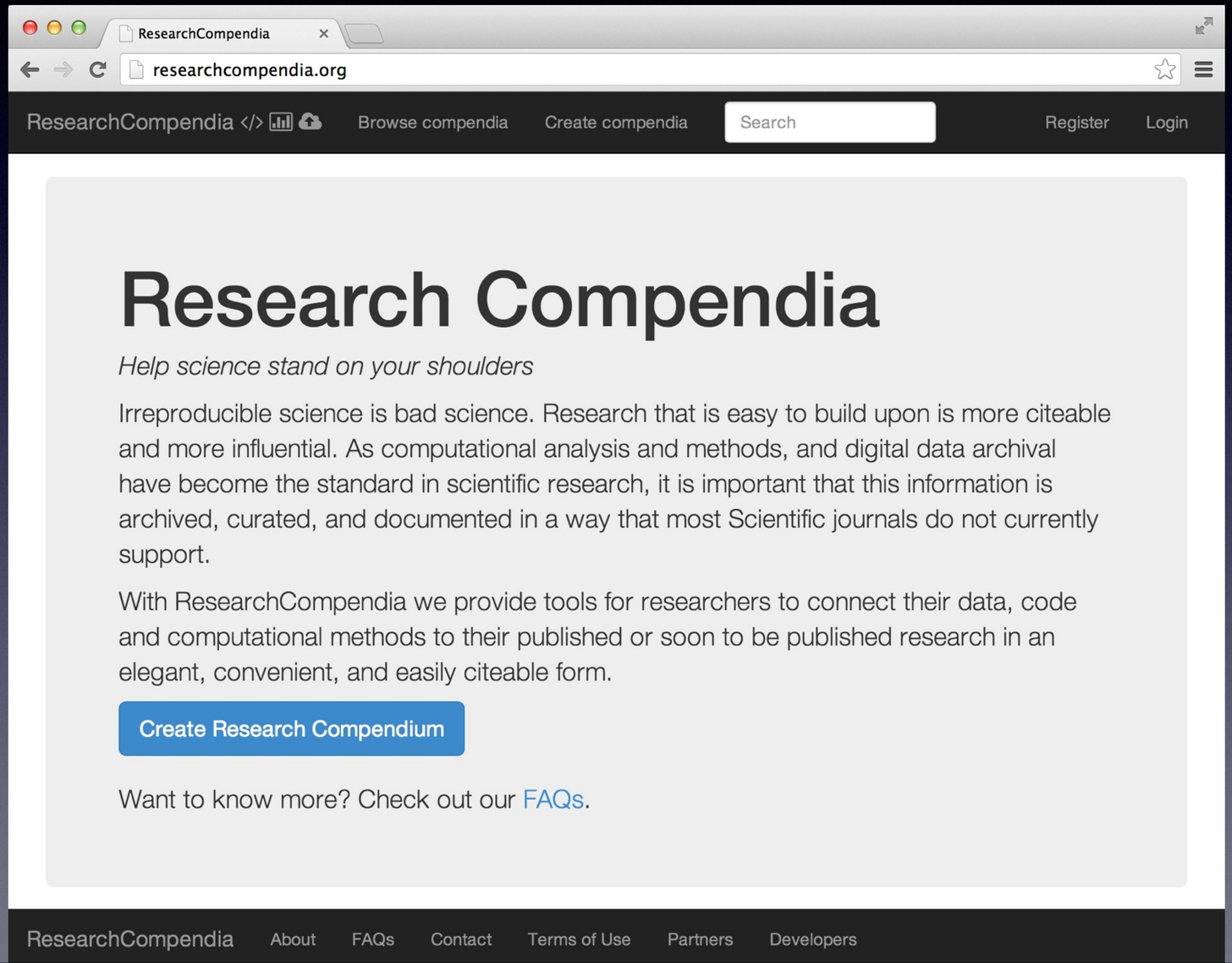
ResearchCompendia.org

Persistently connect:

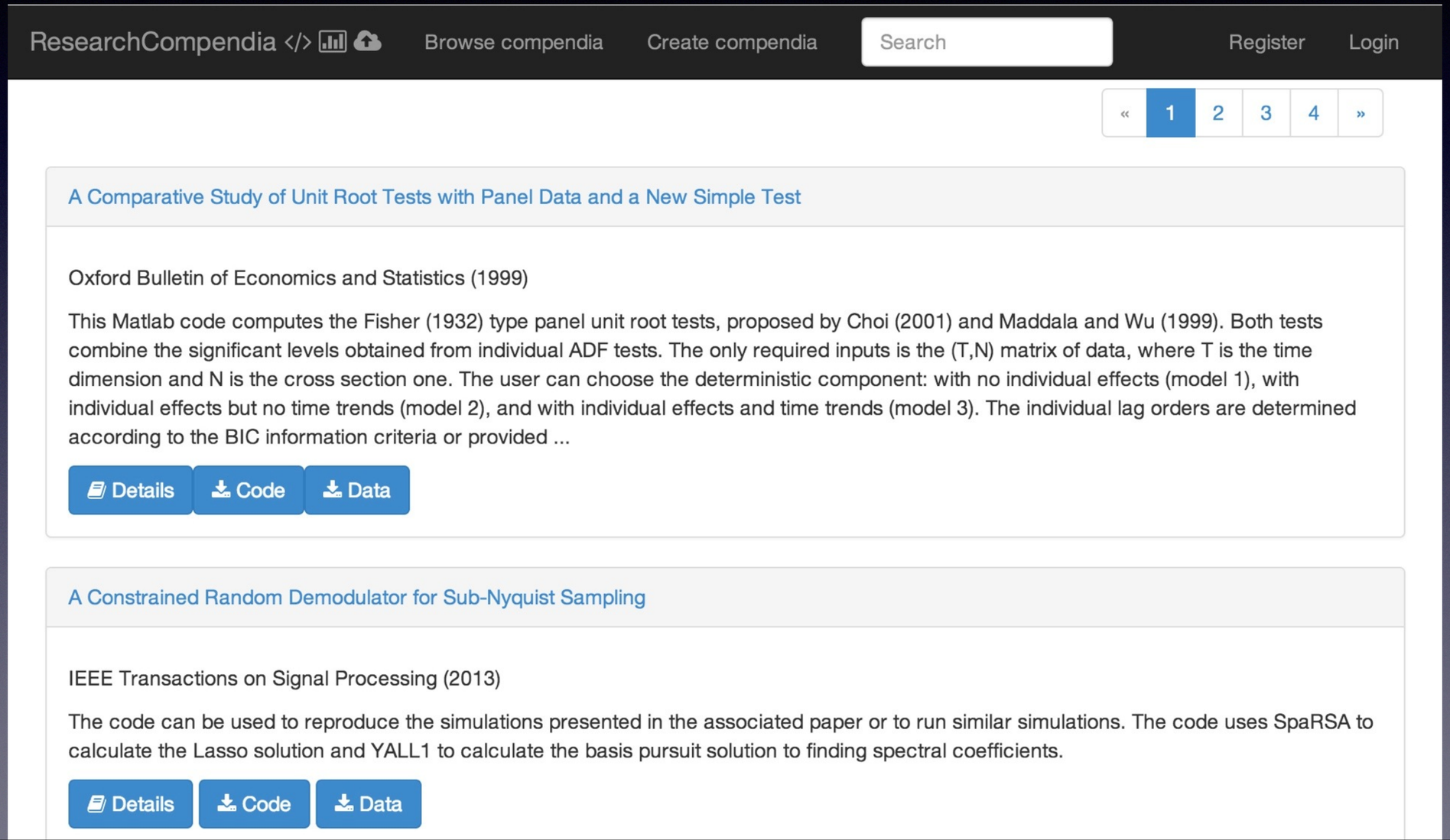
- data output,
- code,
- narrative/publications.

Goals:

- link research outputs for reproducibility and reuse,
- citation standards,
- open source.



Browsing the Research Compendia



Scientific Research Varies Widely

- Different research questions call for different tools, solutions, and implementations to reach "really reproducible research."
- Questions can be solely data-driven research to empirical research contained entirely in software (simulations).
- "Data" has very different meanings depending on the question behind the research.
- Empower communities to reach clearly specified goals that support science, with funds, deadlines, and enforcement (and community engagement in the process).

Openness in Science

- Need infrastructure to facilitate, at the time of publication, (at least):
  1. deposit/curation of versioned data and code,
  2. link to published article,
  3. permanence of link.
- Need infrastructure/software tools to facilitate:
  4. data/code suitable for sharing, created *during the research process*
  5. Public access. "With many eyeballs, all bugs are shallow."

Mandates: 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 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)

Software management plans appearing.. (BigData joint NSF/NIH solicitation)

2013: Open Science in DC

- Feb 22: Executive Memorandum directing federal funding agencies to develop plans for public access to data and publications.
- May 9: Executive Order directing federal agencies to make their data publicly available.

Science Policy in Congress

- America COMPETES due to be reauthorized, drafting underway.
- Sensenbrenner introduced "Public Access to Science," Sept 19, 2013.
- Hearing on Research Integrity and Transparency by the House Science, Space, and Technology Committee (March 5).
- Reproducibility cannot be an unfunded mandate.

Tools for Computational Science

- Dissemination Platforms:

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

- Workflow Tracking and Research Environments:

VisTrails	Kepler	CDE	IPython Notebook
Galaxy	GenePattern	Paper Mâché	
Sumatra	Taverna	Pegasus	

- Embedded Publishing:

Verifiable Computational Research	Sweave
Collage Authoring Environment	SHARE

References

- "Toward Reproducible Computational Research: An Empirical Analysis of Data and Code Policy Adoption by Journals," PLoS ONE, June 2013
- "Reproducible Research," guest editor for Computing in Science and Engineering, July/August 2012.
- "Reproducible Research: Tools and Strategies for Scientific Computing," July 2011.
- "Enabling Reproducible Research: Open Licensing for Scientific Innovation," 2009.

available at http://www.stodden.net

Sharing: Journal Policy

- Journal Policy snapshots June 2011 and June 2012:
- 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.

Journal Data Sharing Policy

	2011	2012
Required as condition of publication, barring exceptions	10.6%	11.2%
Required but may not affect editorial decisions	1.7%	5.9%
Encouraged/addressed, may be reviewed and/or hosted	20.6%	17.6%
Implied	0%	2.9%
No mention	67.1%	62.4%

Source: Stodden, Guo, Ma (2013) PLoS ONE, 8(6)

Journal Code Sharing Policy

	2011	2012
Required as condition of publication, barring exceptions	3.5%	3.5%
Required but may not affect editorial decisions	3.5%	3.5%
Encouraged/addressed, may be reviewed and/or hosted	10%	12.4%
Implied	0%	1.8%
No mention	82.9%	78.8%

Source: Stodden, Guo, Ma (2013) PLoS ONE, 8(6)

Findings

- 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 pre-publication, if anyone,
- Burdens on authors, especially less technical authors,
- Evolving, early research; affects decisions on when to publish,
- Business concerns, attracting the best papers.

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"
- ...