Computational Reproducibility is Crucial for Scientific Software Platforms

Victoria Stodden
Department of Statistics
Columbia University

Galaxy Community Conference
University of Oslo
July 1, 2013

Agenda

1. What is “computational reproducibility”? (and who cares?)
2. At the frontiers
   - Policy in Washington
   - Journal policies
   - Tools and software
3. Challenges to Reproducible Research

Defining Reproducible Research

“Really Reproducible Research” pioneered 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 ... set of instructions [and data] which generated the figures.”

paraphrased by David Donoho, 1998.

Computational Reproducibility

- *Argument*: But don’t you undermine science by making it too easy to replicate results? No one will independently verify anymore.

- *Answer*: No. Independent verification is crucial, and will still be recognized as a scientific contribution. However, if the results do not match (and they won’t) we need the complete workflows to understand why they differ (and find errors).

- *Bonus answer*: Computation makes complexity very easy. The traditional paper cannot hope to capture all the steps taken in generating a computational science result. These steps have a crucial impact on findings.

Reproducibility

- not a new concept, rooted in *skepticism*
- Transactions of the Royal Society 1660's
- Transparency, knowledge transfer -> goal to perfect the *scholarly record*. Nothing else.
- Technology has changed the nature of experimentation, data, and communication.

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: 9th data release (SDSS-III 2012), 60TB,
    - 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.

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%

Ioannidis (2011): 9% of authors studied made data available.

Generally, data and code not made available at the time of publication, insufficient information 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.
- Computational science as practiced today does not generate reliable knowledge. See e.g. Ioannidis, “Why Most Published Research Findings are False,” PLoS Med, 2005.

Advances at the Frontiers

1. What is “computational reproducibility”? (and who cares?)
2. At the frontiers
   - Policy in Washington: OSTP and Congress
   - Journal policies: Advances over the last two years
   - Tools and software: exponential rate of advance in tool development to support and communicate computational science
3. Challenges to Reproducible Research

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.

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)

Science Policy in Congress

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

Recall...

- 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)

Executive Memorandum will operate through Data Management Plans

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

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

Journal 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
Encouraged/addressed, may be reviewed and/or hosted	17	21	4
Implied	0	3	3
No mention	141	134	-7

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

Findings

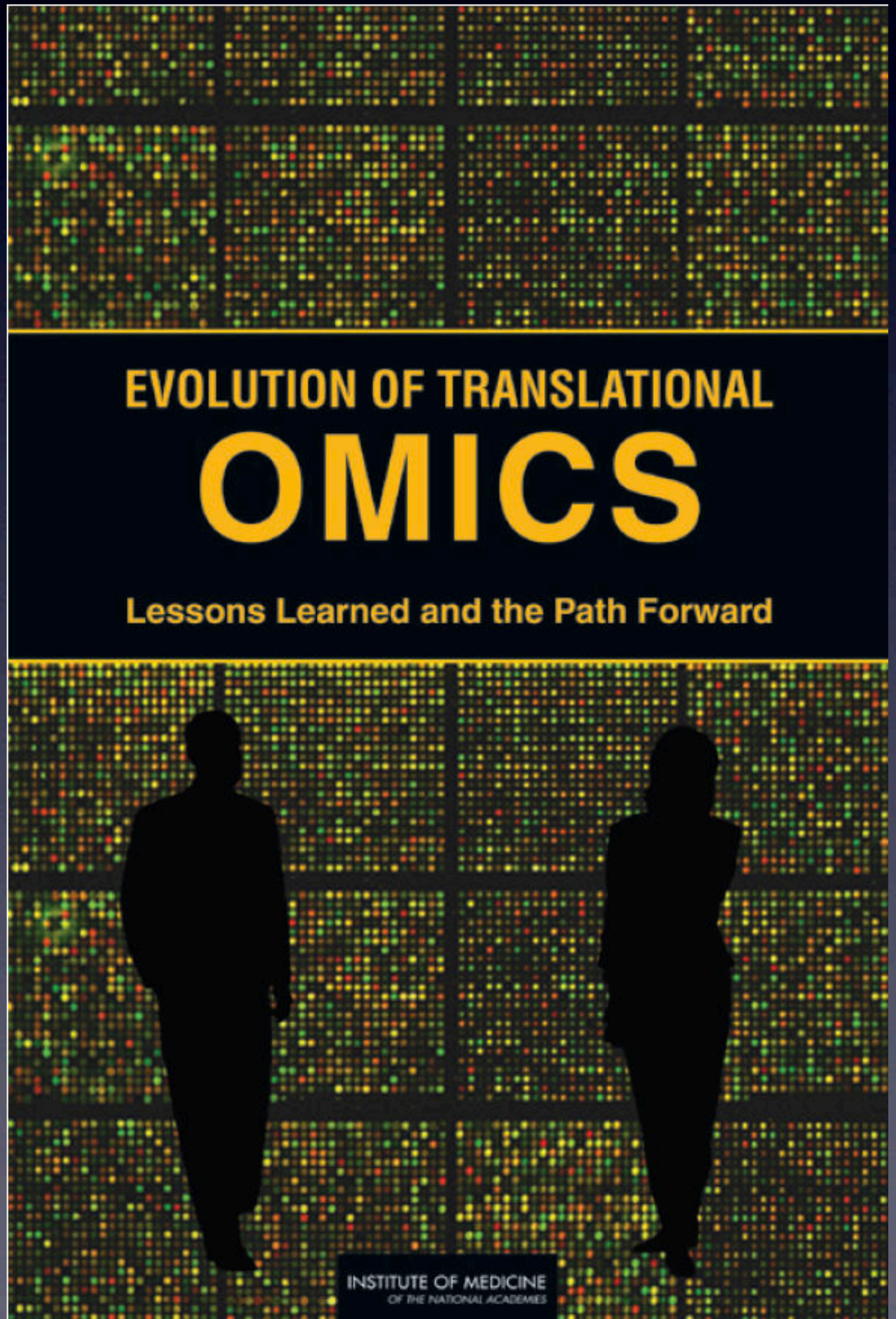
- Changemakers are journals with high impact factors.
- Progressive policies are not widespread, but being adopted rapidly.
- Close relationship between the existence of a supplemental materials policy and a data policy.
- Data and supplemental material policies appear to lead software policy.

Barriers to Journal Policy Making

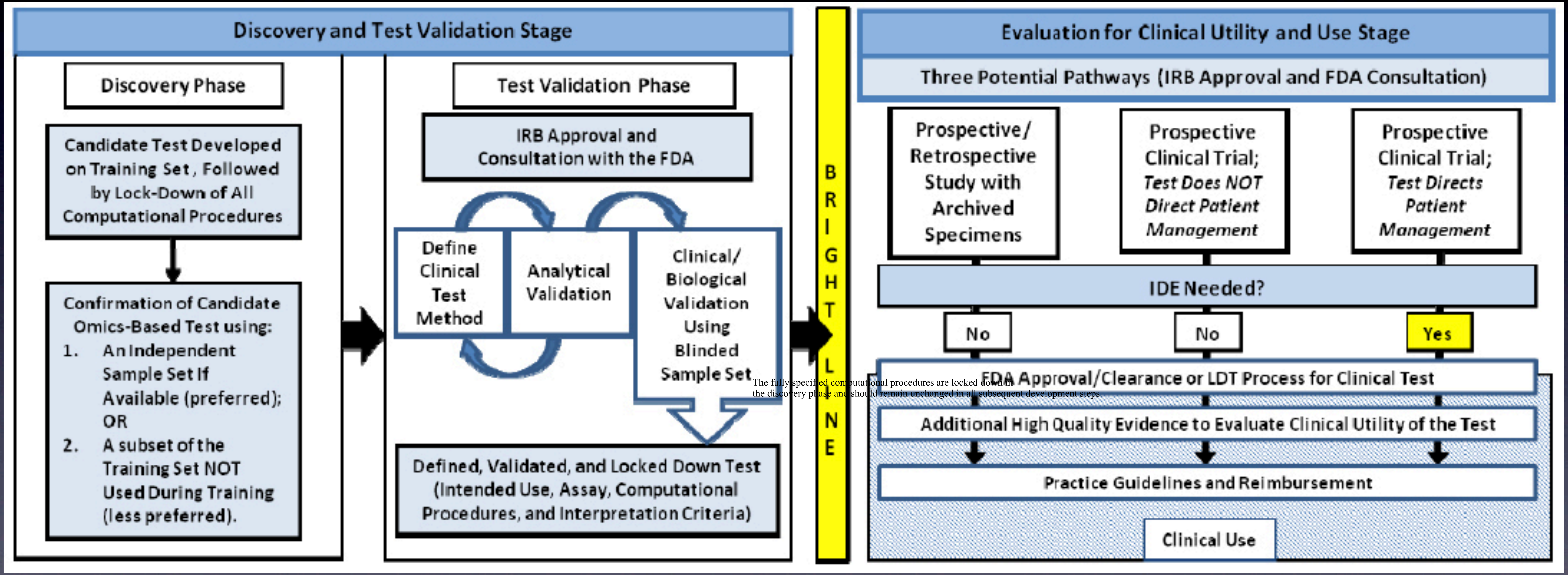
- Standards for code and data sharing,
- Meta-data, archiving, re-use, documentation, sharing platforms, citation standards,
- Review, who checks replication, if anyone,
- Burdens on authors, especially less technical authors,
- Evolving, early research; affects decisions on when to publish,
- Business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

IOM “Evolution of Translational Omics: Lessons Learned and the Path Forward”

- March 23 2012, IOM releases report,
- Recommends new standards for omics-based tests, including a fixed version of the software, expressly for verification purposes.


EVOLUTION OF TRANSLATIONAL
OMICS
Lessons Learned and the Path Forward
INSTITUTE OF MEDICINE
OF THE NATIONAL ACADEMIES

IOM Report: Figure S-1



Discovery and Test Validation Stage

Discovery Phase

Candidate Test Developed on Training Set, Followed by Lock-Down of All Computational Procedures

Confirmation of Candidate Omics-Based Test using:

1. An Independent Sample Set If Available (preferred); OR
2. A subset of the Training Set NOT Used During Training (less preferred).

Test Validation Phase

IRB Approval and Consultation with the FDA

Define Clinical Test Method

Analytical Validation

Clinical/ Biological Validation Using Blinded Sample Set

Defined, Validated, and Locked Down Test (Intended Use, Assay, Computational Procedures, and Interpretation Criteria)

BRIGHT LINE

Evaluation for Clinical Utility and Use Stage

Three Potential Pathways (IRB Approval and FDA Consultation)

Prospective/ Retrospective Study with Archived Specimens

Prospective Clinical Trial; *Test Does NOT Direct Patient Management*

Prospective Clinical Trial; *Test Directs Patient Management*

IDE Needed?

No

No

Yes

FDA Approval/Clearance or LDT Process for Clinical Test

Additional High Quality Evidence to Evaluate Clinical Utility of the Test

Practice Guidelines and Reimbursement

Clinical Use

The fully specified computational procedures are locked down in the discovery phase and should remain unchanged in all subsequent development steps.

"The fully specified computational procedures are locked down in the discovery phase and should remain unchanged in all subsequent development steps."

NAS Data Sharing Report

SHARING
PUBLICATION-RELATED
DATA AND MATERIALS

- Sharing Publication-Related Data and Materials: Responsibilities of Authorship in the Life Sciences, (2003)

- “Principle 1. Authors should include in their publications the data, algorithms, or other information that is central or integral to the publication—that is, whatever is necessary to support the major claims of the paper and would enable one skilled in the art to verify or replicate the claims.”

Tools for Computational Science

- Dissemination Platforms:
  - RunMyCode.org
  - IPOL
  - Madagascar
  - MLOSS.org
  - thedatahub.org
  - nanoHUB.org
  - Open Science Framework
- Workflow Tracking and Research Environments:
  - Galaxy
  - Kepler
  - CDE
  - VisTrails
  - GenePattern
  - Paper Mâché
  - Sumatra
  - Taverna
  - Pegasus
- Embedded Publishing:
  - Verifiable Computational Research
  - Sweave
  - Collage Authoring Environment
  - SHARE

Challenges to Reproducible Research

1. What is “computational reproducibility”? (and who cares?)
2. At the frontiers
   - Policy in Washington
   - Journal policies
   - Tools and software
3. Challenges to Reproducible Research: Opportunities and caveats

Openness in Science

- Science Policy must support scientific ends: Reliability and accuracy of the scientific record.
- Facilitate Reproducibility - the ability to regenerate published computational results (data and code availability, alongside results).
- Need infrastructure to (minimally) facilitate (1):
    1. deposit/curation of data and code,
    2. link to published article,
    3. permanence of link.

Science Policy

- “Open Data” is not well-defined. Scope: Share data and code that *permit others in the field to replicate published results*. (traditionally done by the publication alone).

- Data and code availability at the time of publication.

- Public access. “With many eyeballs, all bugs are shallow.” Recall: primary goal of the scientific method to root out error.

- Need infrastructure/software tools to facilitate (2): Data/code suitable for sharing, created *during the research process*.

Tools are crucial..

- but typically unrewarded by the established (and prestigious) funding structures
- “isn’t that something a private company should do?” (no!)
- even if you get the money:
  - salaries uncompetitive
  - positions short term (length of grant)
  - attitudes toward software contributions outdated (“we haven’t traditionally rewarded the lens grinders”)
- software dev environment unsophisticated in general

A Grassroots Movement

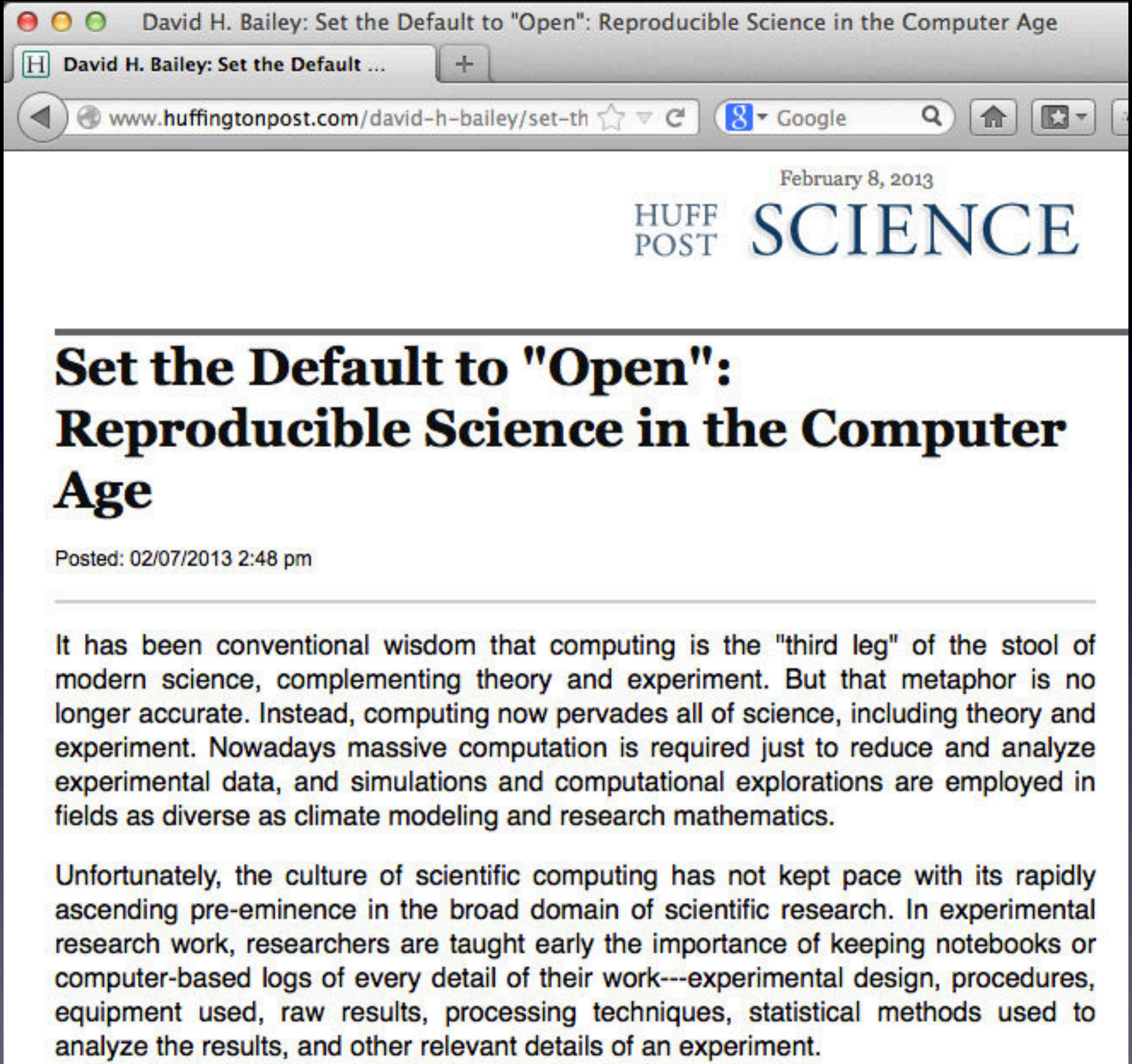
- AMP 2011 “Reproducible Research: Tools and Strategies for Scientific Computing”
- Open Science Framework / Reproducibility Project in Psychology
- AMP / ICIAM 2011 “Community Forum on Reproducible Research Policies”
- SIAM Geosciences 2011 “Reproducible and Open Source Software in the Geosciences”
- 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

- "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

Supplemental Slides


David H. Bailey: Set the Default to "Open": Reproducible Science in the Computer Age
David H. Bailey: Set the Default ...
www.huffingtonpost.com/david-h-bailey/set-th
Google

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.

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)

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.

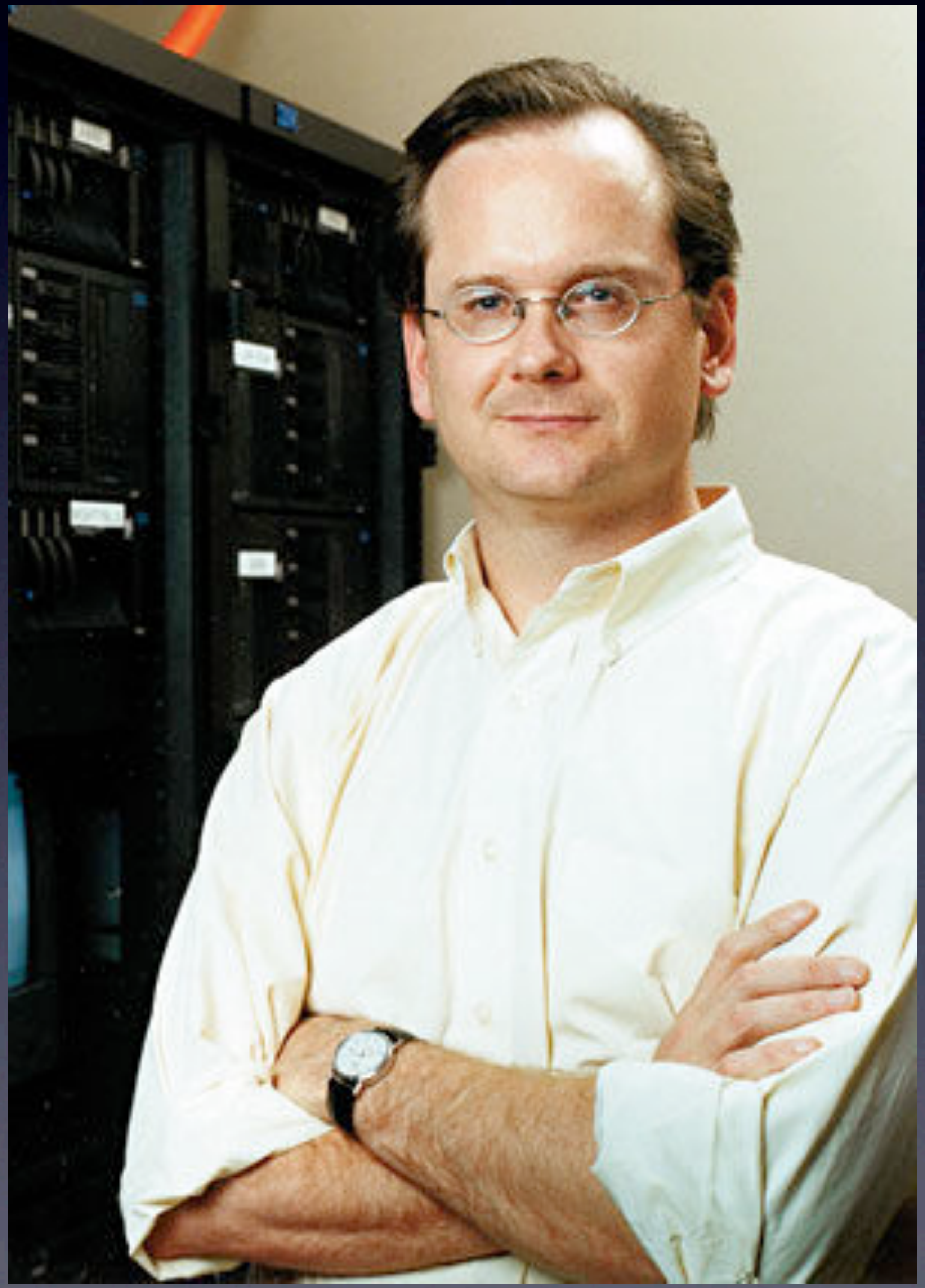
Responses Outside the Sciences 1: Open Source Software

- 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

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