Tools for Academic Research: Resolving the Credibility Crisis in Computational Science

Victoria Stodden
Department of Statistics
Columbia University

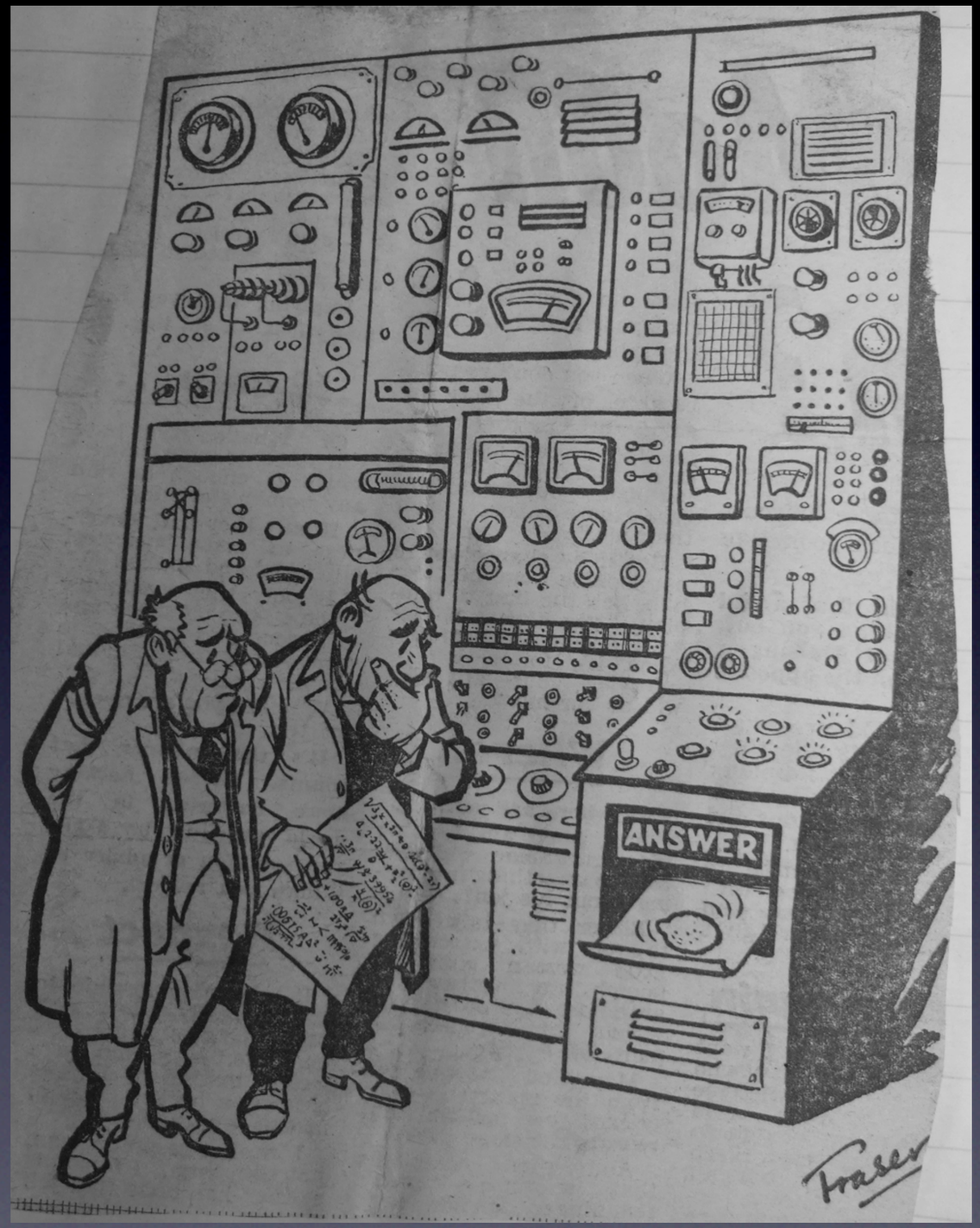
Computer Science and Engineering Colloquia
University of California at Riverside
May 16, 2011

Agenda

1. The onslaught of bits: How is the scientific method changing?
2. Reproducibility: barriers and solutions
3. Legal Barriers 1: *Reproducible Research Standard*
4. Legal Barriers 2: Incentivizing Tech Transfer
5. Tools to facilitate scientific communication
6. Giving credit for contributions in the digital age
7. Brave New World: Challenges to Open Science

Computational Methods Emerging as Central to the Scientific Enterprise

- enormous, and increasing, amounts of data collection,
  - ~3TB/yr genome sequence data: ~1000 sequencers running full time producing 600GB each run (HiSeq 2000, 11 days per run),
  - 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.
- massive simulations of the complete evolution of a physical system, systematically varying parameters,
- deep intellectual contributions now encoded in software.


ANSWER
Fraser

Updating the Scientific Method

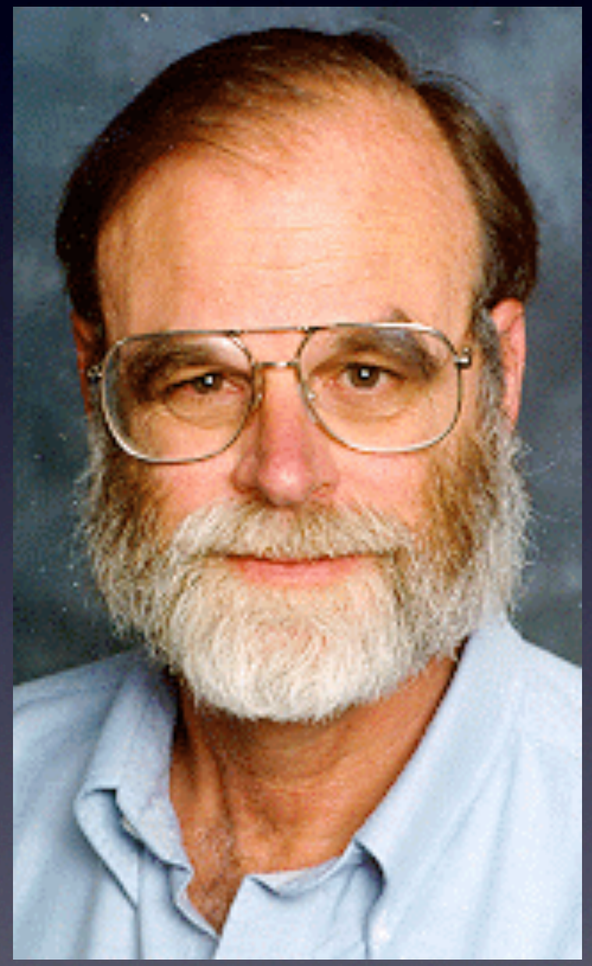
Many have argued (Gray) that data-driven discovery, engenders a "fourth paradigm" of science:

1: theory,

2: experimentation,

3: large scale computational simulation,

4: data-driven scientific discovery.



Updating the Scientific Method

Others (Donoho) have argued 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.

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.
- *Computational science must develop standards for reproducibility before it can be considered a third branch of the scientific method*,

  ➡ Data and Code Sharing with publication.

Computation Emerging as Central to the Scientific Endeavor

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

- Data and code typically not made available in scientific publishing, rendering results unverifiable, not reproducible.

  ➡ A *Credibility Crisis* (ClimateGate, Duke Clinical Trials,...)

Reproducibility is Necessary

Framing Principle for Scientific Communication: *Reproducibility*

- "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

Groundswell from across the Computational Sciences

- 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"
- ...

Barriers to Data and Code Sharing in Computational Science

Survey of 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%

The Legal and Policy Framework

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
  - limited time: generally life of the author +70 years

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.

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 Kaltura Award 2008

Data as Raw Facts

- “raw facts” not copyrightable,
- original “selection and arrangement” of these facts is copyrightable. (Feist Publns Inc. v. Rural Tel. Serv. Co., 499 U.S. 340 (1991)),
- implies a residual copyright in data,
- lack of clarity for scientists: what is a “raw fact” in science??

Benefits of the *RRS*

- Promotion of (legal) reproducible research,
- Focus becomes release of entire research compendium,
- Hook for funders, journals, institutional policy makers,
- Standardization avoids license incompatibilities,
- Clarity of rights, beyond Fair Use.

Policy: Bayh-Dole Act

- Bayh-Dole Act (1980), designed to promote the transfer of academic discoveries for commercial development, via licensing of patents.
- Legislators blind to the coming digital revolution, impact on software and algorithm patenting. Tech Transfer Offices and code release.
- Implications for science as a disruptor of openness norms:
    - patents => delay in revealing code, or closed code,
    - I assert Bilski => obfuscation of methods submitted for patents,
    - (aside from altering a scientist's incentives toward commercial ends).

Policy: 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)

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."

- NSF peer-reviewed Data Management Plan, 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)

Tools and Infrastructure

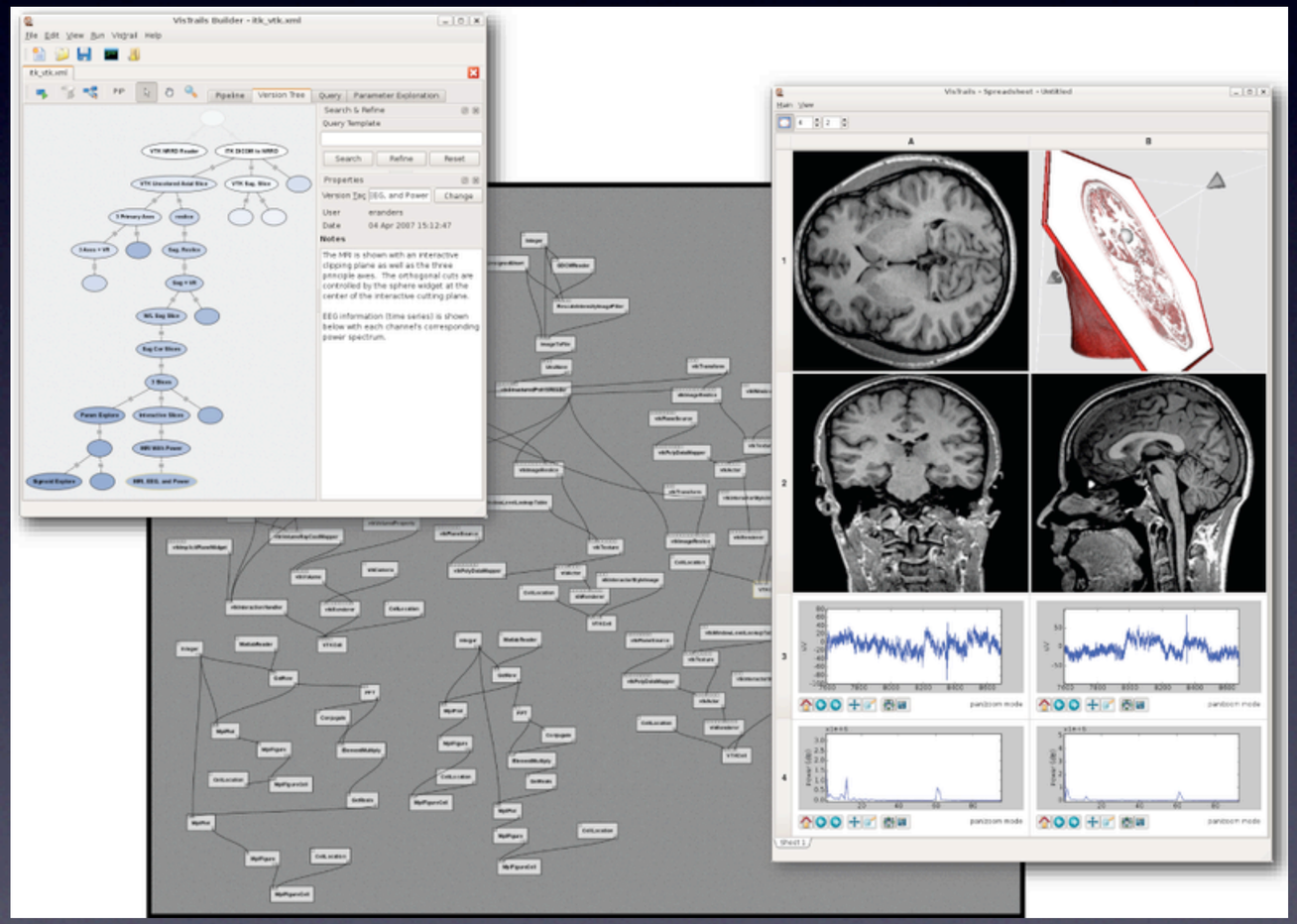
Barriers from Computational Infrastructure

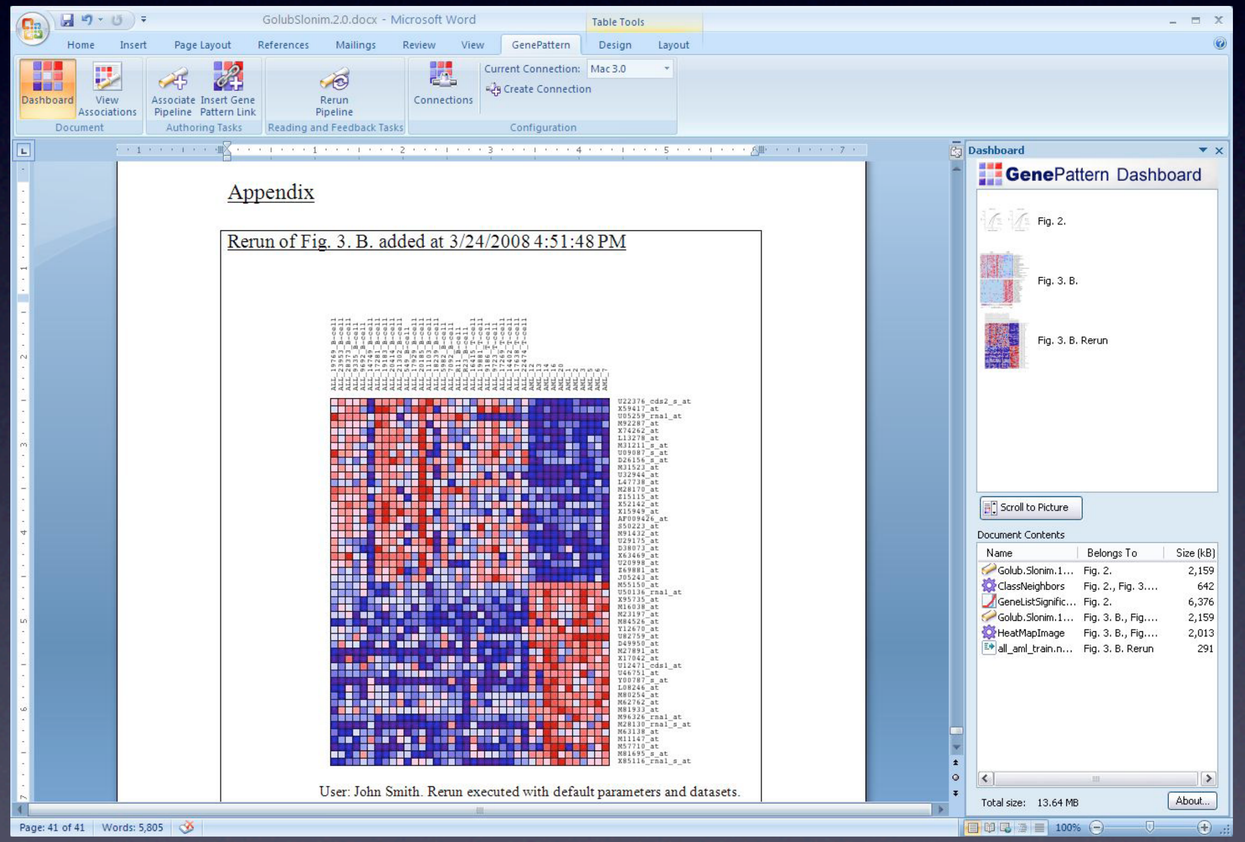
- typical scientific software a dialog, not envisioned as collaboration,
- tools to facilitate (later) sharing *during* the research process: worlflow tracking, data provenance,
- testing for code: unit tests, regression tests,
- systems for opening code collaboration and community building (Github?),

Scientific Research Software Development

- workflow tracking and provenance ie. Vistrails.org, GenePattern, Taverna, Trident, and many others,
- automatic cloud repository and unique identifiers for published results (Donoho and Gavish 2011, Altman and King 2007),
- collaborative tools ie. colwiz, Mendeley,
- versioning of datasets and code used for replication, as a standard.

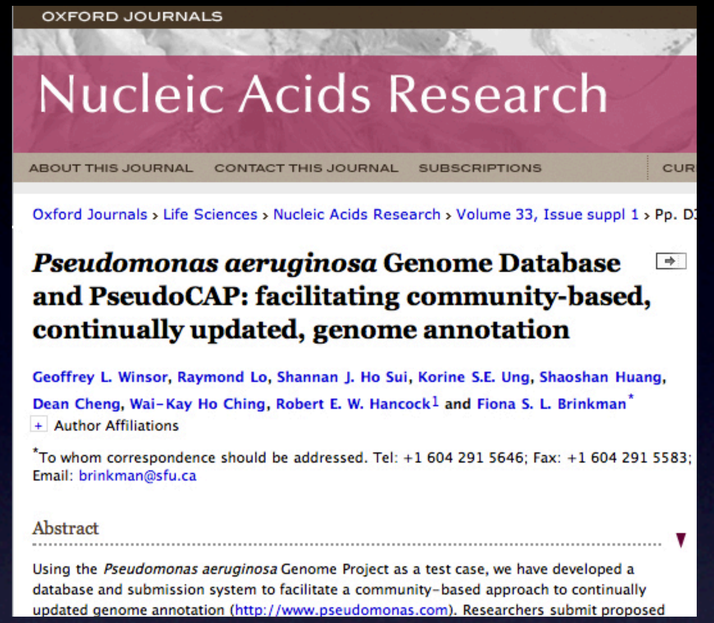
Vistrails and GenePattern Examples





Giving Credit

Citation and Contributions


OXFORD JOURNALS

Nucleic Acids Research

ABOUT THIS JOURNAL CONTACT THIS JOURNAL SUBSCRIPTIONS CUR

Oxford Journals › Life Sciences › Nucleic Acids Research › Volume 33, Issue suppl 1 › Pp. D

***Pseudomonas aeruginosa* Genome Database and PseudoCAP: facilitating community-based, continually updated, genome annotation**

Geoffrey L. Winsor, Raymond Lo, Shannan J. Ho Sui, Korine S.E. Ung, Shaoshan Huang, Dean Cheng, Wai-Kay Ho Ching, Robert E. W. Hancock[1] and Fiona S. L. Brinkman*

+ Author Affiliations

*To whom correspondence should be addressed. Tel: +1 604 291 5646; Fax: +1 604 291 5583; Email: brinkman@sfu.ca

Abstract

Using the *Pseudomonas aeruginosa* Genome Project as a test case, we have developed a database and submission system to facilitate a community-based approach to continually updated genome annotation (http://www.pseudomonas.com). Researchers submit proposed

- 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)
  - open source software as a model?

Error Correction and Review

- Different approaches by journals:
  - may offer unreviewed “supplemental materials” section,
  - may require data and/or code to be provided upon request (Science as of Feb 11 2011),
  - may employ an Associate Editor for Reproducibility (Biostatistics, Biometrical Journal) or replicate results (ACM SIGMOD),
  - may publish correspondence from the review process (Molecular Systems Biology, The European Molecular Biology Organization Journal),
  - new journals, ie. Open Research Computation, BMC Data Notes
  - ignore the issue entirely..

Open Problems

Challenges to Open Science

- “Taleb Effect” - scientific discoveries as (misused) black boxes,
- nefarious uses? public misinterpretation?
- black boxes and opacity in software (why the traditional methods section is inadequate, massive codebases),
- lock-in: calcification of ideas in software?
- independent replication discouraged?
- exceptionally large datasets?
- policy maker engagement: finding support for our norms.

Yale Data and Code Sharing Roundtable 2009

- Roundtable on Data and Code Sharing in computational science Nov 21, 2009:
  - gathered 30 computational scientists from a variety of fields, funding agency folks, publishers, librarians, university policy makers, lawyers...
  - Draft Position Statement (published in IEEE Computing in Science and Engineering, Sep/Oct 2010)
  - recommendations for stakeholders: scientists, journal editors, funding agencies, universities.
- http://www.stanford.edu/~vcs/Conferences/RoundtableNov212009/

References

- “Enabling Reproducible Research: Open Licensing for Scientific Innovation”
- “The Scientific Method in Practice: Reproducibility in the Computational Sciences”
- “Open Science: Policy Implications for the Evolving Phenomenon of User-led 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.stanford.edu/~vcs