Intellectual Contributions to Digitized Science: Implementing the Scientific Method

Victoria Stodden
Department of Statistics
Columbia University

SIAM Conference on Computational Science and Engineering
"Verifiable, Reproducible Research and Computational Science" Mini Symposium
Mar 4, 2011

Agenda

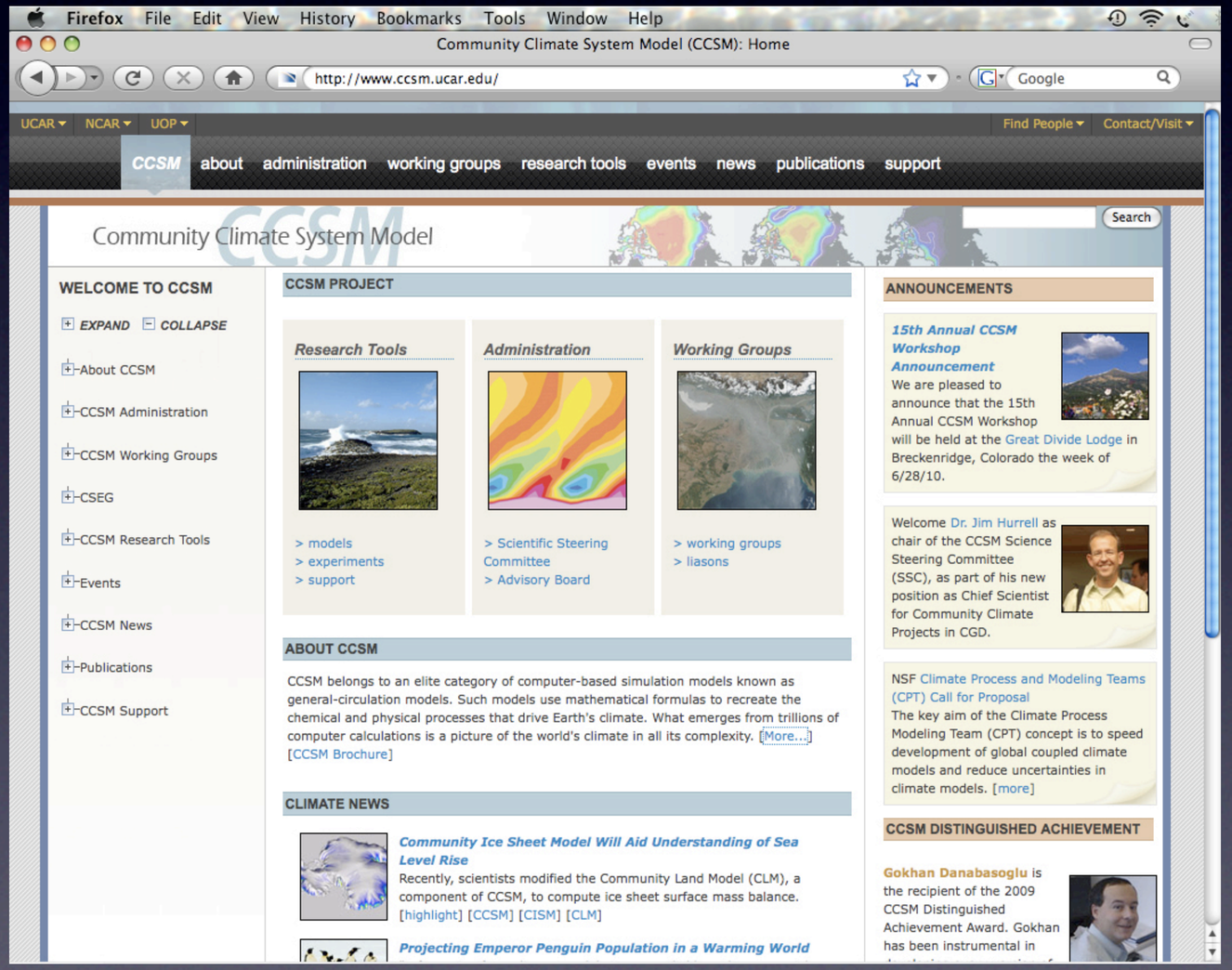
1. Centrality of reproducibility to the scientific method,
2. The practice and tools of scientific investigation are changing,
3. Incentives: how to facilitate code and data sharing?
4. Legal barriers to reproducibility
5. Emergent incentive problems to be addressed: citation, peer review, ex-ivory tower scientific contributions..
6. Community response

Reproducibility is Central to the Scientific Method

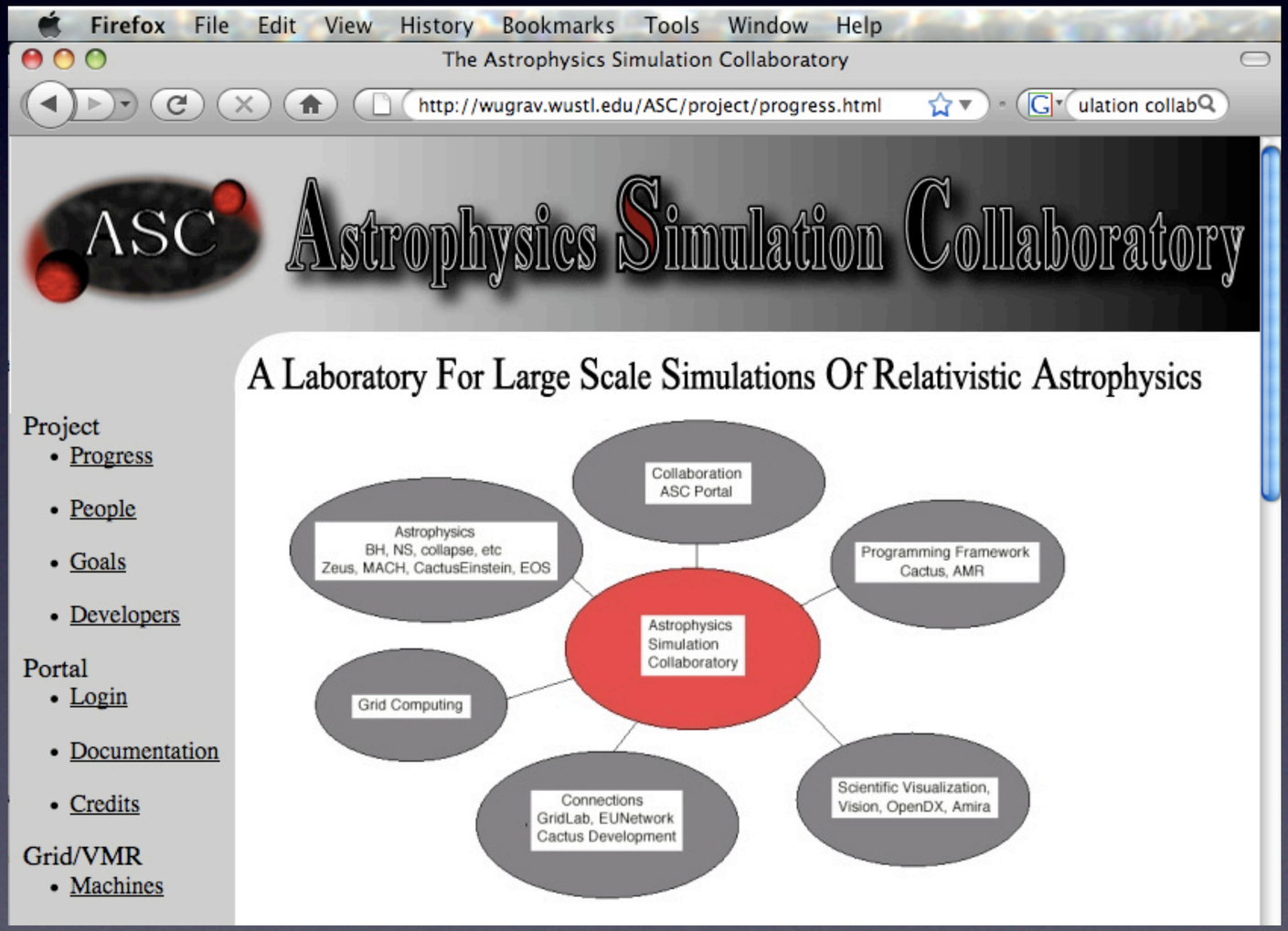
- Other branches of science incorporate reproducibility of results:
  - deductive branch (mathematics, formal logic): the well-defined concept of the proof,
  - inductive branch (experimental sciences): machinery of hypothesis testing, structured communication of methods and protocols.
- Computational Science must develop standards for reproducibility before it can be considered a third branch of the scientific method,

➡ Data and Code Sharing, with publication.

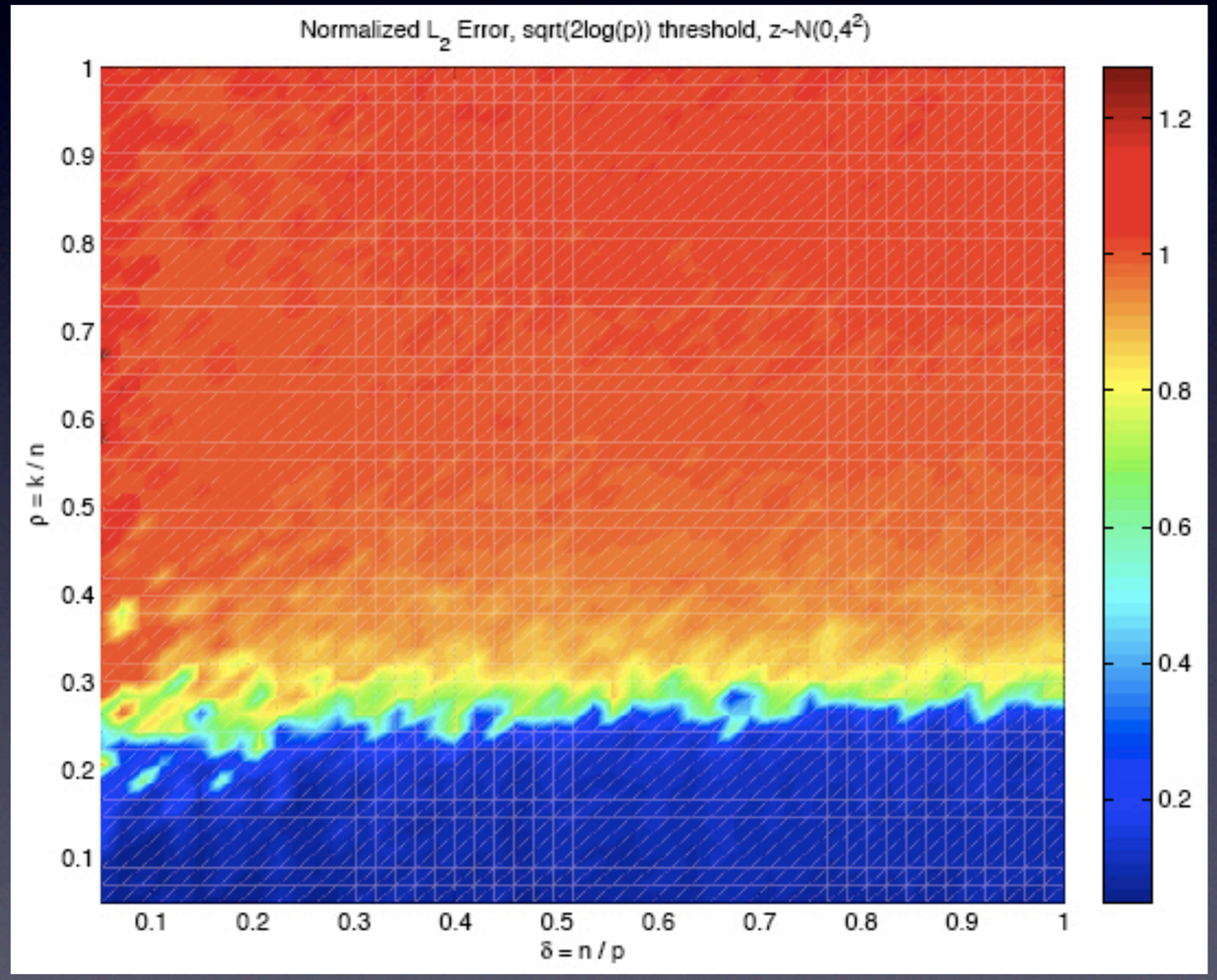
New Computational Science 1



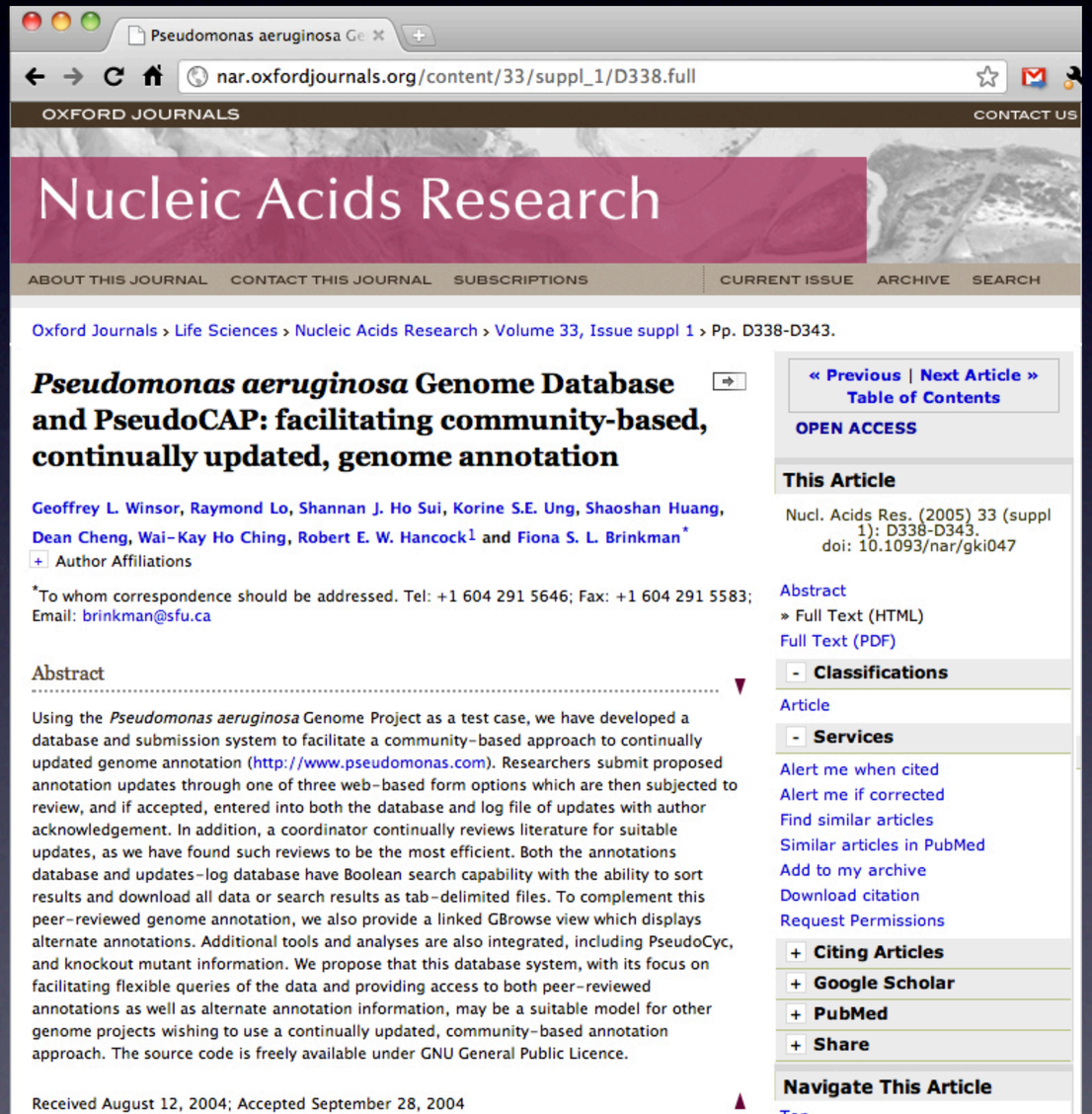
New Computational Science 2



New Computational Science 3



New Computational Science 4



Computation Central to the Scientific Endeavor

For example, in statistics,

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

A Crisis in Computational Science

- Computational methods becoming central to the scientific enterprise:
  - enormous, and increasing, amounts of data collection,
  - intellectual contributions now encoded in software,
  - typical scientific results rely on both data and code.
- Data and code typically not made available, rendering published results unverifiable, not reproducible.

➡ A Credibility Crisis

Barriers to Data and Code Sharing in Computational Science

Survey of Machine Learning Community (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%

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

- Adapts the Open Source Software approach to artistic and creative digital works
- Provides a suite of licensing options:
    - 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

Incentives and Open Questions: Citation and Contributions

- Collaborative efforts in database building?
  - differential citation? (web vs article citation, microcitation)
  - database versioning (e.g. King and Altman 2007, Donoho and Gavish 2011)
  - citizen contributions? (Galaxy Zoo, Open Dinosaur Project)
- Code development? review?
- Code maintainance for reproducibility, scientific reuse?
  - platform building (DANSE, Wavelab, Sparselab)
  - open source software as a model?

Challenges to Open Science

- “Taleb Effect” - scientific discoveries as (misused) black boxes,
- nefarious uses?
- 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?
- policy maker engagement: finding support for our norms,
- Commercial incentives for the scientist/university (Bayh-Dole).

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