How Technology is (Rapidly) Expanding the Scope of the Law in Statistics

Victoria Stodden
Department of Statistics
Columbia University

Law and Statistics Session
International Chinese Statistical Association Symposium
New York City
June 27, 2011

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, systematically varying parameters,
- deep intellectual contributions now encoded in software.

Updating the Scientific Method

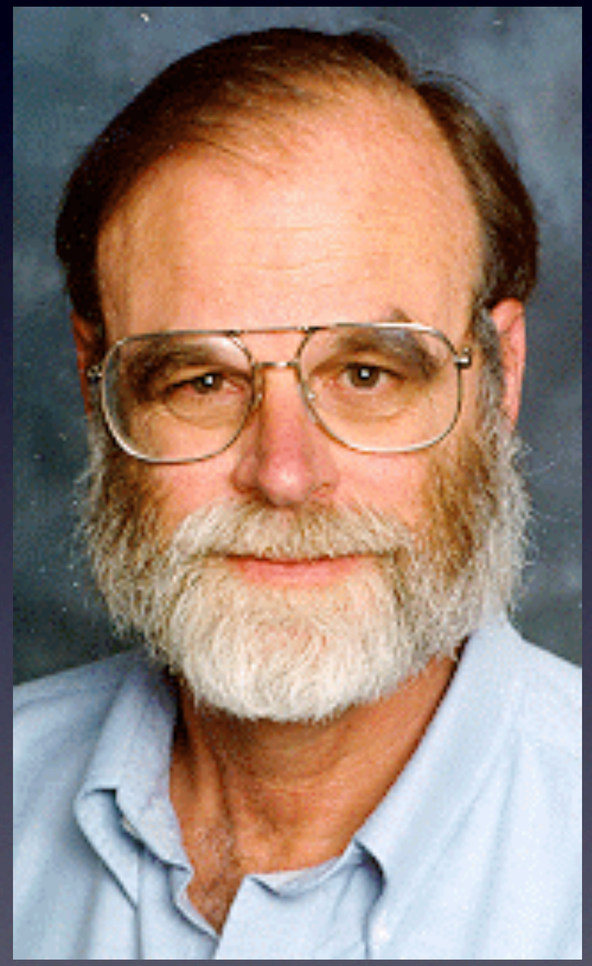
Many have argued (Gray) that data-driven discovery, engenders a "fourth paradigm" of science:

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,...)

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

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.

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

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"
- ...

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).

Bilski and Incentives

- Bilski et al v. Kappos (No. 08-964) 561 US __
- In one of the principal results, the Court found that Bilski's inclusion of a mathematical description of his invention (risk hedging calculations) *weakened* the patent application,
- conjecture: this creates an incentive for patent applicants to obscure any mathematical foundations of the invention.

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