A brief history of the reproducibility movement

Victoria Stodden
Department of Statistics
Columbia University

Reproducibility in Computational
and Experimental Mathematics
ICERM, Brown University
Dec 10, 2012

- The Changing Concept of a Scientific Fact
  - The Scientific Record
  - Computational Science
  - Examples
- International Strategy Meetings on Human DNA Sequencing
  - Bermuda 1996
  - Fort Lauderdale 2003
  - Amsterdam 2008
  - Toronto 2009
  - Public Debate
- Duke Clinical Trial Experience
- Other Experiences
  - Geophysics Experience
  - Statistics
  - Other Efforts

The Concept of a Scientific Fact

In *Opus Tertium* (1267) Roger Bacon distinguishes experimental science by:

1. verification of conclusions by direct experiment,
2. discovery of truths unreachable by other approaches,
3. investigation of the secrets of nature, opening us to a knowledge of past and future.

- described a repeating cycle of observation, hypothesis, experimentation, and the need for independent verification,
- recorded his experiments (e.g. the nature and cause of the rainbow) in enough detail to permit reproducibility by others.

Inductive Scientific Reasoning

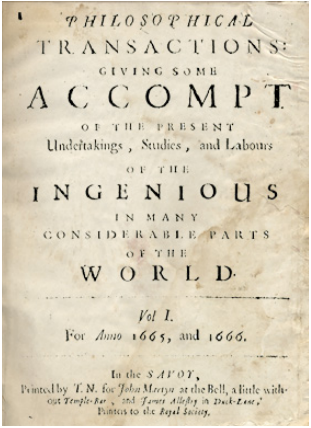
In *Novum Organum* (1620) Francis Bacon proposes:

1. the gathering of facts, by observation or experimentation,
2. verification of general principles.

"There are and can be only two ways of searching into and discovering truth. The one flies from the senses and particulars to the most general axioms, and from these principles, the truth of which it takes for settled and immoveable. ... The other derives axioms from the senses and particulars, rising by a gradual and unbroken ascent, so that it arrives at the most general axioms last of all. This is the true way, but as yet untried."

The Scientific Record

- The Royal Society of London founded 1660 (the "Invisible College"),
- members discussed Francis Bacon's "new science" from 1645,
- Society correspondence reviewed by the first Secretary, Henry Oldenburg,
- Oldenburg became the founder, editor, author, and publisher of *Philosophical Transactions*, launched in 1665.



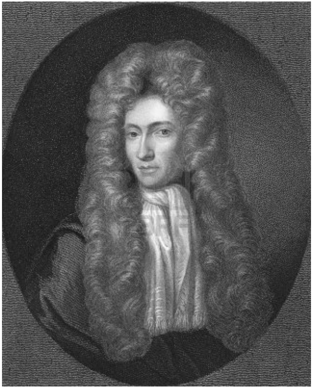
PHILOSOPHICAL
TRANSACTIONS:
GIVING SOME
ACCOMPT
OF THE PRESENT
Undertakings, Studies, and Labours
OF THE
INGENIOUS
IN MANY
CONSIDERABLE PARTS
OF THE
WORLD.

Vol I.
For *Anno* 1665, and 1666.

In the *SAVOY*,
Printed by T. N. for *John Martyn* at the Bell, a little without *Temple-Bar*, and *James Alleſtry* in *Duck-Lane*,
Printers to the *Royal Society*.

The Last Update to the Scientific Method: 1665

- The "Invisible College" included Robert Boyle, the "father of chemistry,"
- Boyle introduced *standards* for scientific communication: enough information must be included to allow others to independently reproduce the finding.
- delineates science, concept of reproducibility permits verification and knowledge transfer,
- knowledge in **method** not in the **finding** itself.



Controlling Error is Central to Scientific Progress

"The scientific method's central motivation is the ubiquity of error - the awareness that mistakes and self-delusion can creep in absolutely anywhere and that the scientist's effort is primarily expended in recognizing and rooting out error."
David Donoho et al. (2009)

The Third Branch of the Scientific Method

- Branch 1: Deductive/Theory: e.g. mathematics; logic,
- Branch 2: Inductive/Empirical: e.g. the machinery of hypothesis testing; statistical analysis of controlled experiments,
- Branch 3? 4? Large scale extrapolation and prediction, using simulation and other data-intensive methods.

Scientific Research is Changing

Scientific computation emerging as central to the scientific method:

- Simulation of the complete evolution of a physical system, systematically changing parameters,
- (Massive) data driven research, machine-generated hypotheses.

Thesis: Computational science cannot be elevated to a third branch of the scientific method until it generates *routinely verifiable knowledge*. (Donoho, et al. 2009)

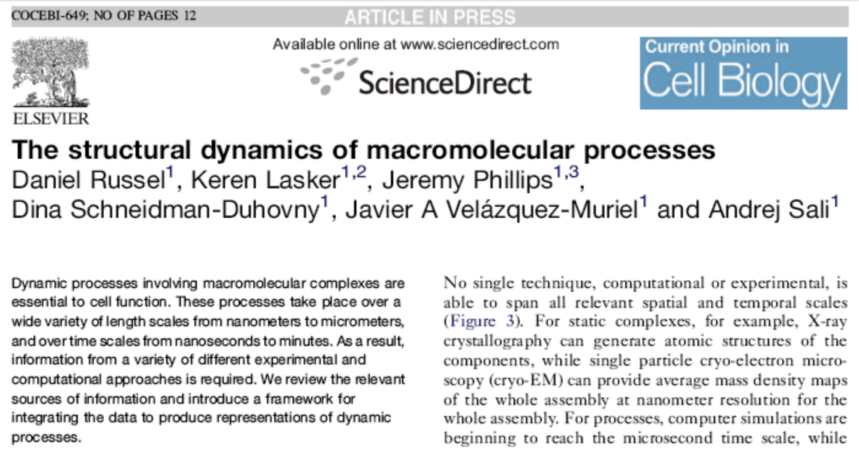
I. Examples of Pervasiveness of Computational Methods

- 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%
2011	29 of 29	21%

- Social network data and the quantitative revolution in social science (Lazer et al. 2009);
- Computation reaches into traditionally nonquantitative fields: e.g. Wordhoard project at Northwestern examining word distributions by Shakespearian play.

2. Dynamic modeling of macromolecules: SaliLab UCSF



COCEBI-649; NO OF PAGES 12

ARTICLE IN PRESS

ELSEVIER

Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Cell Biology

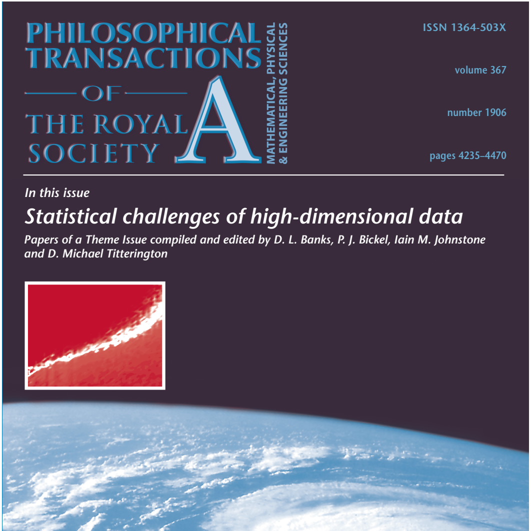
The structural dynamics of macromolecular processes

Daniel Russel[1], Keren Lasker[1,2], Jeremy Phillips[1,3], Dina Schneidman-Duhovny[1], Javier A Velázquez-Muriel[1] and Andrej Sali[1]

Dynamic processes involving macromolecular complexes are essential to cell function. These processes take place over a wide variety of length scales from nanometers to micrometers, and over time scales from nanoseconds to minutes. As a result, information from a variety of different experimental and computational approaches is required. We review the relevant sources of information and introduce a framework for integrating the data to produce representations of dynamic processes.

No single technique, computational or experimental, is able to span all relevant spatial and temporal scales (Figure 3). For static complexes, for example, X-ray crystallography can generate atomic structures of the components, while single particle cryo-electron microscopy (cryo-EM) can provide average mass density maps of the whole assembly at nanometer resolution for the whole assembly. For processes, computer simulations are beginning to reach the microsecond time scale, while

3. Mathematical “proof” by simulation and grid search



PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A

MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

ISSN 1364-503X

volume 367

number 1906

pages 4235–4470

In this issue

Statistical challenges of high-dimensional data

Papers of a Theme Issue compiled and edited by D. L. Banks, P. J. Bickel, Iain M. Johnstone and D. Michael Titterington

Toward Transparency in Computational Science

Examples of influential steps toward transparency in dissemination of results:

- data sharing standards in bioinformatics,
- Institute of Medicine's recommendation for open (and fixed) code requirements for the FDA,
- geophysics and statistics.

A complete accounting is impossible in this talk...

The 1996 Bermuda Agreement

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

Primary Genomic Sequence Should be Rapidly Released

- Sequence assemblies should be released as soon as possible; in some centers, 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.

The Bioinformatics experience frames public understanding

Conjecture: Much of the public (Congressional and Whitehouse) understanding of scientific transparency stems from the experience in bioinformatics: the focus is on **open data**, rather than reproducibility or transparency.

Clinical trials based on flawed genomic studies

Timeline:

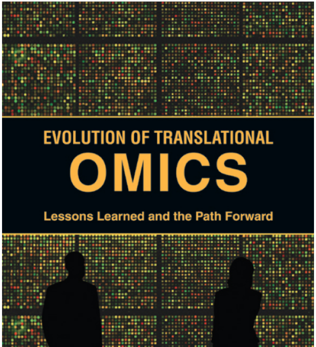
- Potti et al (2006), Nature Medicine; (2006) NEJM; (2007) Lancet Oncology; (2007) Journal of Clinical Oncology: evidence of genomic signatures to guide use of chemotheraputics (*all since retracted*),
- Coombes, Wang, Baggerly at M.D. Anderson Cancer Center cannot replicate, and find flaws: genes misaligned by one row, column labels flipped, genes repeated and missing from analysis..
- 2007 correspondence and a supplementary report submitted to the Journal of Clinical Oncology and publication declined; 2008 Nature Medicine declines their correspondence.
- Clinical trials initiated in 2007 (Duke), 2008 (Moffitt).

Clinical trials based on flawed genomic studies

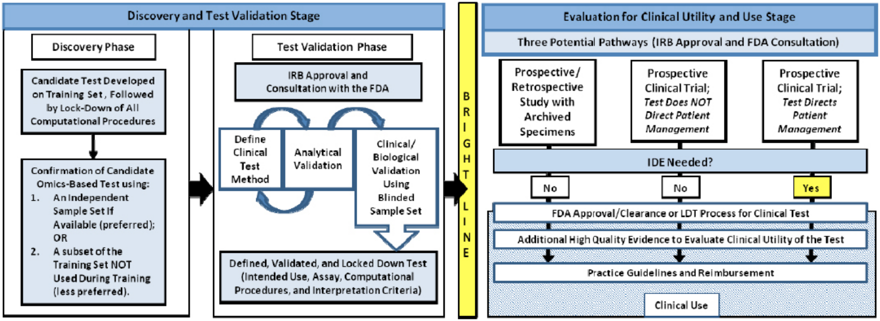
- Duke launches internal investigation Sept 2009; all three trials suspended in Oct 2009,
- Oct 2009: results reported validated, regardless of errors, because data blinded (later found not to be true),
- Jan 2010: Duke clinical trials resume, patients allocated to treatment and control groups. "Neither the review nor the raw data are being made available at this time."
- July 2010: 33 prominent biostatisticians write to Varmus as head of IOM urging suspension of the trials and an examination of standards of review, including reproducibility.
- Sept 2010: IOM committee "Review of Omics-Based Tests for Predicting Patient Outcomes in Clinical Trials" formed,
- late 2010: Potti resigns, Nevins removed from position, and the clinical trials are terminated.

Recommendations from the Institute of Medicine

- March 23, 2012, IOM releases report, “Evolution of Translational Omics: Lessons Learned and the Path Forward”
- 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

IOM Report: Figure S-1

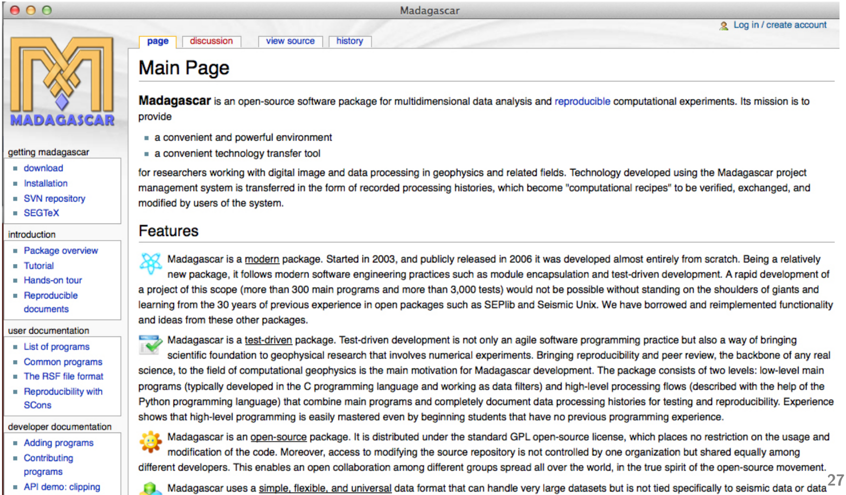

Discovery and Test Validation Stage
Discovery Phase
Candidate Test Developed on Training Set, Followed by Lock-Down of All Computational Procedures
Confirmation of Candidate Omics-Based Test using:
1. An Independent Sample Set If Available (preferred); OR
2. A subset of the Training Set NOT Used During 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."

Experience in Geophysics and Statistics

- 1991: Stanford Professor Jon Claerbout requires theses to conform to standard of reproducibility,
- reduces "startup time" for new students from years to weeks,
- his vision adopted and adapted by many others, e.g. Sergey Fomel, David Donoho.

Madagascar (Sergey Fomel and collaborators)



Main Page

Madagascar is an open-source software package for multidimensional data analysis and reproducible computational experiments. Its mission is to provide

- a convenient and powerful environment
- a convenient technology transfer tool

for researchers working with digital image and data processing in geophysics and related fields. Technology developed using the Madagascar project management system is transferred in the form of recorded processing histories, which become "computational recipes" to be verified, exchanged, and modified by users of the system.

Features

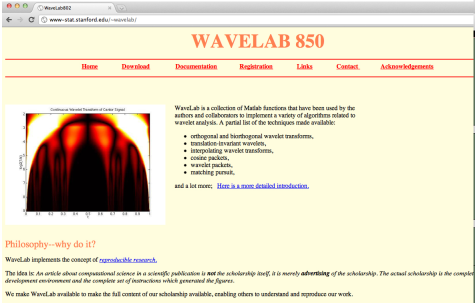
Madagascar is a modern package. Started in 2003, and publicly released in 2006 it was developed almost entirely from scratch. Being a relatively new package, it follows modern software engineering practices such as module encapsulation and test-driven development. A rapid development of a project of this scope (more than 300 main programs and more than 3,000 tests) would not be possible without standing on the shoulders of giants and learning from the 30 years of previous experience in open packages such as SEPlib and Seismic Unix. We have borrowed and reimplemented functionality and ideas from these other packages.

Madagascar is a test-driven package. Test-driven development is not only an agile software programming practice but also a way of bringing scientific foundation to geophysical research that involves numerical experiments. Bringing reproducibility and peer review, the backbone of any real science, to the field of computational geophysics is the main motivation for Madagascar development. The package consists of two levels: low-level main programs (typically developed in the C programming language and working as data filters) and high-level processing flows (described with the help of the Python programming language) that combine main programs and completely document data processing histories for testing and reproducibility. Experience shows that high-level programming is easily mastered even by beginning students that have no previous programming experience.

Madagascar is an open-source package. It is distributed under the standard GPL open-source license, which places no restriction on the usage and modification of the code. Moreover, access to modifying the source repository is not controlled by one organization but shared equally among different developers. This enables an open collaboration among different groups spread all over the world, in the true spirit of the open-source movement.

Madagascar uses a simple, flexible, and universal data format that can handle very large datasets but is not tied specifically to seismic data or data

Donoho Lab, Stanford


WaveLab802
www-stat.stanford.edu/~wavelab/

WAVELAB 850

Home Download Documentation Registration Links Contact Acknowledgements

Continuous Wavelet Transform of Cantor Signal
log2(1/a)
2
3
4
5
6
7
8
9
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
t

WaveLab is a collection of Matlab functions that have been used by the authors and collaborators to implement a variety of algorithms related to wavelet analysis. A partial list of the techniques made available:

- orthogonal and biorthogonal wavelet transforms,
- translation-invariant wavelets,
- interpolating wavelet transforms,
- cosine packets,
- wavelet packets,
- matching pursuit,

and a lot more; Here is a more detailed introduction.

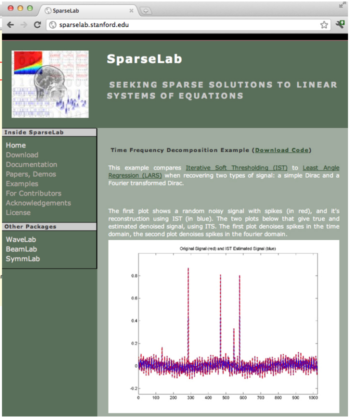
Philosophy--why do it?

WaveLab implements the concept of *reproducible research.*

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 complet development environment and the complete set of instructions which generated the figures.*

We make WaveLab available to make the full content of our scholarship available, enabling others to understand and reproduce our work.

“WaveLab (1999)”



SparseLab

SEEKING SPARSE SOLUTIONS TO LINEAR SYSTEMS OF EQUATIONS

Inside SparseLab

Home
Download
Documentation
Papers, Demos
Examples
For Contributors
Acknowledgements
License

Other Packages

WaveLab
BeamLab
SymmLab

Time Frequency Decomposition Example (Download Code)

This example compares Iterative Soft Thresholding (IST) to Least Angle Regression (LARS) when recovering two types of signal: a simple Dirac and a Fourier transformed Dirac.

The first plot shows a random noisy signal with spikes (in red), and it's reconstruction using IST (in blue). The two plots below that give true and estimated denoised signal, using ITS. The first plot denoises spikes in the time domain, the second plot denoises spikes in the fourier domain.

“Sparselab (2006)”

Grassroots Efforts in Many Fields, Policies

Independent efforts by researchers:

- ICERM 2012 “Reproducibility in Computational and Experimental Mathematics”
- 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
- ...

Policy changes:

- NSF/OCI report on Grand Challenge Communities (Dec 2010)
- NSF report “Changing the Conduct of Science in the Information Age” (Aug 2011)
- IOM “Review of Omics-based Tests for Predicting Patient Outcomes in Clinical Trials” (2012)
- NIH, NSF multiple requests for input on data policies
- Journal policy movement toward code and data requirements (ie. *Science* Feb 2011)
- ...