Data Access and Ownership

Victoria Stodden
School of Information Sciences
University of Illinois at Urbana-Champaign

CI Workshop for NSF Large Facilities
Arlington, VA
December 1, 2015
“Setting the Default to Reproducible” in Computational Science Research

By Victoria Stodden, Jonathan M. Borwein, David H. Bailey

Following a late-2012 workshop at the Institute for Computational and Experimental Research in Mathematics, a group of computational scientists have proposed a set of standards for the dissemination of reproducible research.

Computation is now central to the scientific enterprise, and the emergence of powerful computational hardware, combined with a vast array of computational software, presents novel opportunities for researchers. Unfortunately, the scientific culture surrounding computational work has evolved in ways that make it difficult to verify findings, efficiently build on past research, or even apply the basic tenets of the scientific method to computational procedures.

As a result, computational science is facing a credibility crisis [1,2,4,5]. The enormous scale of state-of-the-art scientific computations, using tens or project results, with little attention to reproducibility. It penalizes those who devote the time needed to produce really reproducible research. It is regrettable that software development is often discounted. It has been compared to, say, constructing a telescope, rather than doing real science. Thus, scientists are discouraged from writing or testing code. Sadly, NSF-funded projects on average remain accessible on the web only about a year after funding ends. Researchers are busy with new projects and lack the time or money to preserve the old. With the ever-increasing importance of computation and software, such attitudes and practices must change.
Data Access

Some (additional) rationales:

• Verification of findings
• Unanticipated new findings as technologies and methods develop
• Combination with other datasets for unanticipated new discoveries
• Extension/improvement on existing findings as data improved
Evolving Data Access

• Data decreasingly “static”: streaming data, instrument improvements, method improvements / calibration / data integrity improvements, compression/denoising, etc

• Increasing use of advanced statistical and machine learning methods

• Increasing need for CI: data infrastructure AND software/tools/workflows creation and infrastructure integration and support

• Reproducibility demanding direct connection from data to the scholarly record, includes code/workflows
Data Ownership: Legal and Normative Questions

- Data producer / PI
- PI’s institution
- Repositories / Governance and stewardship models / Cost-benefit analysis
- Public
- Subjects
- Data contributors
- Funding Agencies
- Industry partners, publishers

Citation?

Data post processing value add?
Open Data => Reproducibility

2020 Vision for LT DP in HEP

- **Long-term – e.g. FCC timescales:** disruptive change
  - By 2020, all archived data – e.g. that described in DPHEP Blueprint, including LHC data – easily findable, fully usable by designated communities with clear (Open) access policies and possibilities to annotate further
  - Best practices, tools and services well run-in, fully documented and sustainable; built in common with other disciplines, based on standards

- **DPHEP portal,** through which data / tools accessed
  - “HEP FAIRport”: Findable, Accessible, Interoperable, Re-usable

- **Agree with Funding Agencies clear targets & metrics**
Unpacking “Reproducibility”

“Empirical Reproducibility”

“Computational Reproducibility”

“Statistical Reproducibility”

V. Stodden, IMS Bulletin (2013)
1. Empirical Reproducibility

Sorting Out the FACS: A Devil in the Details

William C. Hineq,1,2,∗ Ying Su,3,4,5,6 Irene Kuhn,3 Kornelia Polyak,3,4,5 and Mina J. Bissell1,2
1Life Sciences Division, Lawrence Berkeley National Laboratory, Mailstop 977R225A, 1 Cyclotron Road, Berkeley, CA 94720, USA
2Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, MA 02215, USA
3Department of Medicine, Brigham and Women’s Hospital, Boston, MA 02115, USA
4Department of Medicine, Harvard Medical School, Boston, MA 02115, USA
5These authors contributed equally to this work
6Correspondence: chines@lbl.gov (W.C.H.), ying_su@dfci.harvard.edu (Y.S.)
http://dx.doi.org/10.1016/j.cell.2014.02.021

The reproduction of results is the cornerstone of science; yet, at times, reproducing the results of others can be a difficult challenge. Our two laboratories, one on the East and the other on the West Coast of the United States, decided to collaborate on a problem of mutual interest—namely, the heterogeneity of the human breast. Despite using seemingly identical methods, reagents, and specimens, our two laboratories quite reproducibly were unable to replicate each other’s fluorescence-activated cell sorting (FACS) profiles of primary breast cells. Frustration of studying cells close to their context in vivo makes the exercise even more challenging.

Paired with in situ characterizations, FACS has emerged as the technology most suitable for distinguishing diversity among different cell populations in the mammary gland. Flow instruments have evolved from being able to detect only a few parameters to those now capable of measuring up to—and beyond—an astonishing 50 individual markers per cell [Cheung and Utz, 2011]. As with any exponential increase in data complexity, breast reduction mammoplasties. Molecular analysis of separated fractions was to be performed in Boston (K.P.’s laboratory, Dana-Farber Cancer Institute, Harvard Medical School), whereas functional analysis of separated cell populations grown in 3D matrices was to take place in Berkeley (M.J.B.’s laboratory, Lawrence Berkeley National Lab, University of California, Berkeley). Both our laboratories have decades of experience and established protocols for isolating cells from primary normal breast tissues as well as the capabilities required for

Reproducibility Issues in Research with Animals and Animal Models

The missing "R": Reproducibility in a Changing Research Landscape

A workshop of the Roundtable on Science and Welfare in Laboratory Animal Use

National Academy of Sciences, NAS 125 2100 C Street NW, Washington DC
June 4-5, 2014

The ability to reproduce an experiment is one important approach that scientists use to gain confidence in their conclusions. Studies that show that a number of significant peer-reviewed studies are not reproducible has alarmed the scientific community. Research that uses animals and animal models seems to be one of the most susceptible to reproducibility issues.

Evidence indicates that there are many factors that may be contributing to scientific irreproducibility, including insufficient reporting of details pertaining to study design and planning; inappropriate interpretation of results; and author, reviewer, and editor abstracted reporting, assessing, and accepting studies for publication.

In this workshop, speakers from around the world will explore the many facets of the issue and potential pathways to reducing the problems. Audience participation portions of the workshop are designed to facilitate understanding of the issue.
2. Computational Reproducibility

Traditionally two branches of the scientific method:

Branch 1 (deductive): mathematics, formal logic,

Branch 2 (empirical): statistical analysis of controlled experiments.

Many claim the emergence of new branches:

Branch 3, 4? (computational): large scale simulations / data driven computational science.
Supporting Computational Science

- **Dissemination Platforms:**
  - ResearchCompendia.org
  - MLOSS.org
  - Open Science Framework
  - IPOL
  - thedatahub.org
  - Madagascar
  - nanoHUB.org
  - RunMyCode.org

- **Workflow Tracking and Research Environments:**
  - VisTrails
  - Galaxy
  - Sumatra
  - Kepler
  - GenePattern
  - Taverna
  - CDE
  - Jupyter / IPython Notebook
  - Pegasus
  - Chameleon
  - Flywheel

- **Embedded Publishing:**
  - Verifiable Computational Research
  - Collage Authoring Environment
  - SOLE
  - knitR
  - SHARE
  - Sweave
Research Compendia

Goal: improve understanding of reproducible computational science, trace sources of error.

- link data/code to published claims,
- enable re-use,
- sharing guide for researchers,
- certification of results,
- large scale validation of findings,
- stability, sensitivity checks.
Data / Code Sharing Practices

Survey of the NIPS community:

- 1,758 NIPS registrants up to and including 2008,
- 1,008 registrants when restricted to .edu registration emails,
- After piloting, the final survey was sent to 638 registrants,
- 37 bounces, 5 away, and 3 in industry, gave a final response rate was 134 of 593 or 23%.
- Queried about reasons for sharing or not sharing data/code associated with their NIPS paper.
Editorial: ACM TOMS Replicated Computational Results Initiative

MICHAEL A. HEROUX, Sandia National Laboratories

The scientific community relies on the peer review process for assuring the quality of published material, the goal of which is to build a body of work we can trust. Computational journals such as the ACM Transactions on Mathematical Software (TOMS) use this process for rigorously promoting the clarity and completeness of content, and citation of prior work. At the same time, it is unusual to independently confirm computational results.

ACM TOMS has established a Replicated Computational Results (RCR) review process as part of the manuscript peer review process. The purpose is to provide independent confirmation that results contained in a manuscript are replicable. Successful completion of the RCR process awards a manuscript with the Replicated Computational Results Designation.

This issue of ACM TOMS contains the first [Van Zee and van de Geijn 2015] of what we anticipate to be a growing number of articles to receive the RCR designation, and the related RCR reviewer report [Willenbring 2015]. We hope that the TOMS RCR process will serve as a model for other publications and increase the confidence in and value of computational results in TOMS articles.

- provide independent confirmation that results contained in a manuscript are correct and replicated.

- Successful completion of the RCR process gives the manuscript a Replicated Computational Results designation on the first page of the published article.
3. Statistical Reproducibility

- False discovery, chasing significance, p-hacking (Simonsohn 2012), file drawer problem, overuse and mis-use of p-values, lack of multiple testing adjustments.

- Low power, poor experimental design,

- Data preparation, treatment of outliers, re-combination of datasets, insufficient reporting/tracking practices,

- Poor statistical methods (nonrandom sampling, inappropriate tests or models, model misspecification..)

- Model robustness to parameter changes and data perturbations,

- Investigator bias toward previous findings; conflicts of interest.