Open data and scientific reproducibility

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

Data Science @ LHC 2015 Workshop
CERN
Nov 13, 2015

Closing Remarks: Open Data and Roundtable on Data Access

1. Goal: by 2020 all experiments who have declared they will share data, will be able to do it; and it will be discoverable (hard!)
2. why? there are several use cases
3. open data is undefined: access to closed or more open formats.
4. right now DMPs are required in the experiments

History of LEP

- when LEP was approved, card readers were still being used…
- July 14 1989 (first beams), software wasn't yet ready
- in 1992, technology was the cray, then unix machines, then the grid. cards were obsolete!
- How is important in the short term but tech changes mean we focus on the what and the why.

ALICE

- goal of data preservation is reproducibility, allow reprocessing of full chain of analysis (not public but AOD provided)
- and allow reanalysis by others
- There are 4 levels to data release: data available on 3rd party platform; 2: simplified formats made publicly available; 3: data with high levels of abstraction made available (10% after 5 years (starting now), 100% after 10 years). 4: raw data made available (only members of collaboration)

ALICE

- attribution important
- no liability
- data released with tools for analysis

ATLAS

- 3 messges:
- from management: Executive Board approved proposal to released about 1 fb-1 of the 2012 data in a limited format, along with simple tools on the Data Portal, sometime early 2016.
- goals is education and outreach, not really to carry out new science

ATLAS

- Open Data group started about a year about and used for eduction (students use data, extract a signal, etc)
- 4.10^6 events, 7GB
- most studies are monte carlo and most users dont need the full raw data.
- discussions underway for longterm preservation. also for bit-level preservation.
- documentation? data comes with tools but needs more documentation which may be an iterative process with expression of user needs. A forum can help.

CMS

- Similar 4 levels: 1: open accès publication and additional numerical data; 2: similfied data for outreach and education; 3: reconstructed data and tools to analyze. 4: full raw data.
- Now: data, tools, instructions, examples.
- Challenge is knowledge preservation and meta-data, especially context at the time of data analysis.
- building open data benchmarks (highlevel validation code) to compare with other results later.

LHCb

- Level 1: results are public. data associated with results made available; 2: outreach/education (samples for masterclass exercises); 3: reconstructed data (50% of data 5 years after the data are collected; 100% 10 years after). 4: not permitting access to raw data because of the complexity of data processing and data size.
- Challenge data to be included in open data

BaBar

- goal: preserving raw data through computing structure for analysis.
- a wiki for real-time documentation of data usage and analysis, framework.
- data stored on tape.
- must join collaboration to access data, propose your new theory.. export framework, review for simplification, create a data portal.
- funding through 2018 - after that?

Open Data @CERN

- announcement of portal about a year ago made a big impact
- Reddit AMA
- extending code for new analysis

Recast

- Saving parameters and estimates of machine learning models. e.g. Higgs discovery.
- not enough information in the papers for reproduibility.

My Questions

- Workflows and tools to capture context during analysis.
- Links to publications
- versions (bit-level preservation?)
- ambitious plans for raw data access (except LHCb; CMS a subset of the data): driven by funding agencies and institutes, incremental approach to avoid catastrophes.. Monte Carlo produces much greater amounts of data.

My Questions

- feedback loop: how can users report bugs, or contribute anything substantial?
- presumably non-collaboration users won't have access to hardware.
- long term support? post project?
- is there any coordination across the projects? (should there be?)
- time lag between publication and 5 year embargo period - so how to link figures to raw data/software? Data DOIs in the publication? What abt snapshots of tools, with DOIs?

My Questions

- transparency in the process of creating data access? a model to other large collaborations.
- links between github and the open data portal? snapshots/versioning?
- incentives for preservation? integration of librarians into the discovery process? Could also be a model for other projects.

My Questions

- Open Data @CERN as a prototype for discovering how the public uses the data, what is most useful for longterm preservation.
- integration between Open Data @CERN tools back into the research pipeline. Connection between analysis tools and pipelines, and availability in Open Data @CERN. Comparisons of results by independent efforts (even within CERN).
- Can these pipelines be shared? for example including parameters and model fitting information. Zenodo/Open Data @CERN?

Really Reproducible Research

- "Really Reproducible Research" inspired by Stanford Professor Jon Claerbout:

  "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 ... set of instructions [and data] which generated the figures."
  David Donoho, 1998.

Experimental Bias

Experimental biases:

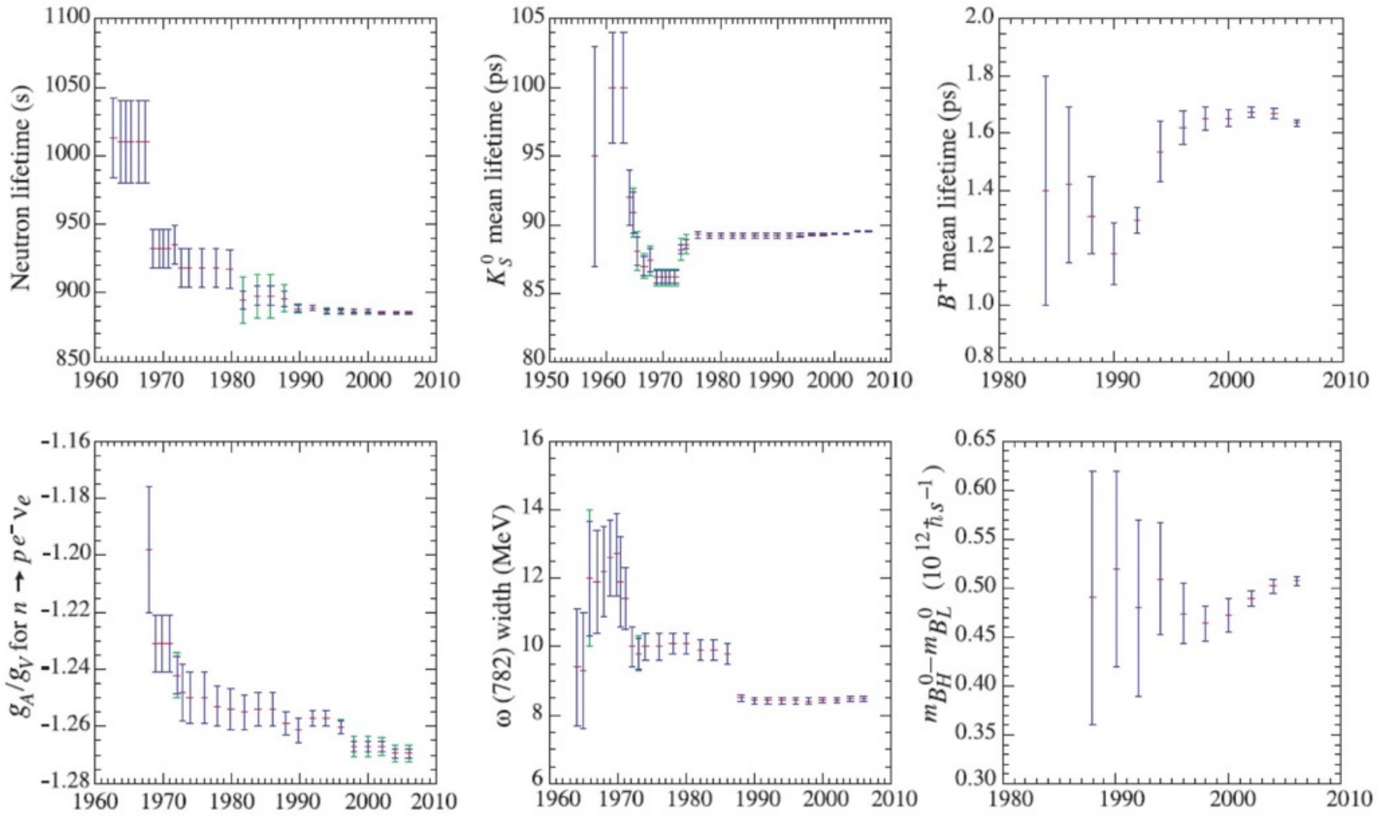


Figure 2: Historical record of values of some particle properties published over time, with quoted error bars (Particle Data Group).

Figure courtesy of James Berger

www.nsf.gov/mps/perspectives/reliable_science_sep2015.jsp

HOME | FUNDING | AWARDS | DISCOVERIES | NEWS | PUBLICATIONS | STATISTICS | ABOUT NSF | FASTLANE

National Science Foundation
Directorate for Mathematical & Physical Sciences (MPS)

QUICK LINKS

SEARCH

MPS HOME | MPS FUNDING | MPS AWARDS | MPS DISCOVERIES | MPS NEWS | ABOUT MPS

Math & Physical Sciences

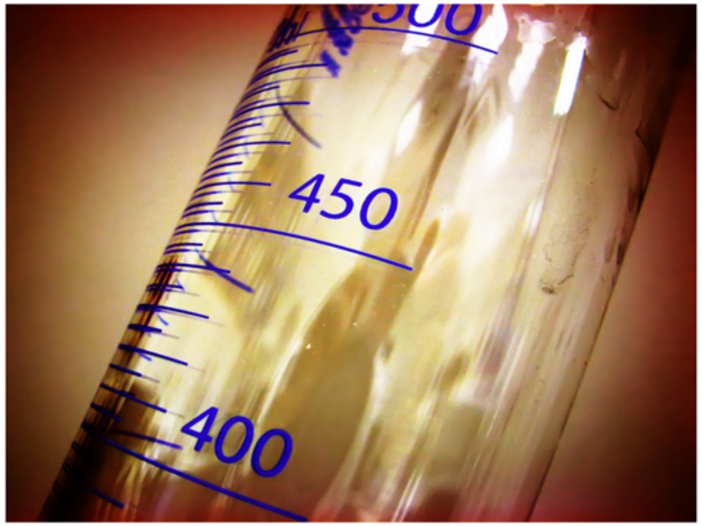
- MPS Home
- About MPS
- Funding Opportunities
- Awards
- News
- Events
- Discoveries
- Publications
- Advisory Committee
- Career Opportunities
- 2013-2014 Distinguished Lecture Series
- View MPS Staff

Search MPS Staff

MPS Organizations

- Astronomical Sciences (AST)
- Chemistry (CHE)
- Materials Research (DMR)
- Mathematical Sciences (DMS)
- Physics (PHY)
- Office of Multidisciplinary

Reliable Science: The Path to Robust Research Results



September 8, 2015

These days, much discussion about the reproducibility of scientific results seems driven by critiques of research in biomedicine and psychology. Most recently, an article in *Science* concluded that 60 percent of a collection of studies were not replicable. This result along with similar analyses of cancer research results have stimulated strong commentary. For example, the *New York Times* print edition headline about the *Science* article was "Psychology's Fears Confirmed: Rechecked Studies Don't Hold Up," coverage that prompted a strong op-ed rebuttal titled, "Psychology Is Not in Crisis."

Issues that arise with human subjects or with other complex living systems do not plague physical science to the same degree. However, the notion of measuring the same value of a physical quantity or the same behavior of a physical system in different laboratories at different times is central to our concept of a valid scientific result. Often the approach is not simply to replicate an experiment, but rather to get at the same quantity via different paths. For example, we can measure the gravitational constant, G, with approaches ranging from a torsional pendulum to atom interferometry.

grants.nih.gov/reproducibility/index.htm

GRANTS & FUNDING

NIH National Institutes of Health
Office of Extramural Research

Grants Policy

- Policy & Guidance
- Compliance & Oversight
- Research Involving Human Subjects
- Office of Laboratory Animal Welfare (OLAW)
- Animals in Research
- Peer Review Policies & Practices
- Guidance for Reviewers
- Intellectual Property Policy
- Publishing NIH Funding
- Invention Reporting (iEdison)
- NIH Public Access
- Research Integrity
- Rigor and Reproducibility
  - Goals
  - News
  - Guidance
  - Timeline
  - Stakeholder Input
  - References

Rigor and Reproducibility

Enhancing reproducibility through rigor and transparency: the information provided on this website is designed to assist the extramural community in addressing rigor and reproducibility in grant applications due on January 25, 2016, and beyond.

On This Page:

- News
- Goals
- Guidance: Rigor and Reproducibility in Grant Applications
- Timeline
- Stakeholder Input
- References and Resources
- Previous Events

News

On October 13, 2015, the NIH published guide notices outlining updates to form instructions for applications due in 2016, including an overview (NOT-OD-16-004), as well as details on Implementing Rigor and Transparency in NIH & AHRQ Research Grant Applications (NOT-OD-16-011) and Implementing Rigor and Transparency in NIH & AHRQ Career Development Award Applications (NOT-OD-16-012).

On October 30, 2015, NIH Deputy Director of Extramural Research Dr. Mike Lauer published an Open Mike blog post on Bolstering Trust in Science through Rigorous Standards. NIH OER has also released a staff training module that provides a General Policy Overview on enhancing reproducibility through rigor and transparency.

www.nih.gov/research-training/rigor-re

NIH National Institutes of Health

Home » Research & Training

RIGOR AND REPRODUCIBILITY

Two of the cornerstones of science advancement are rigor in designing and performing scientific research and the ability to reproduce biomedical research findings. The application of rigor ensures robust and unbiased experimental design, methodology, analysis, interpretation, and reporting of results. When a result can be reproduced by multiple scientists, it validates the original results and readiness to progress to the next phase of research. This is especially important for clinical trials in humans, which are built on studies that have demonstrated a particular effect or outcome.

Johns Hopkins University students in a laboratory. *Johns Hopkins University*

In recent years, however, there has been a growing awareness of the need for rigorously designed published preclinical studies, to ensure that such studies can be reproduced. This webpage provides information about the efforts underway by NIH to enhance rigor and reproducibility in scientific research.