Trust, but Replicate: Evidence, Authority, Reproducibility, and the Scientific Method

Philip B. Stark

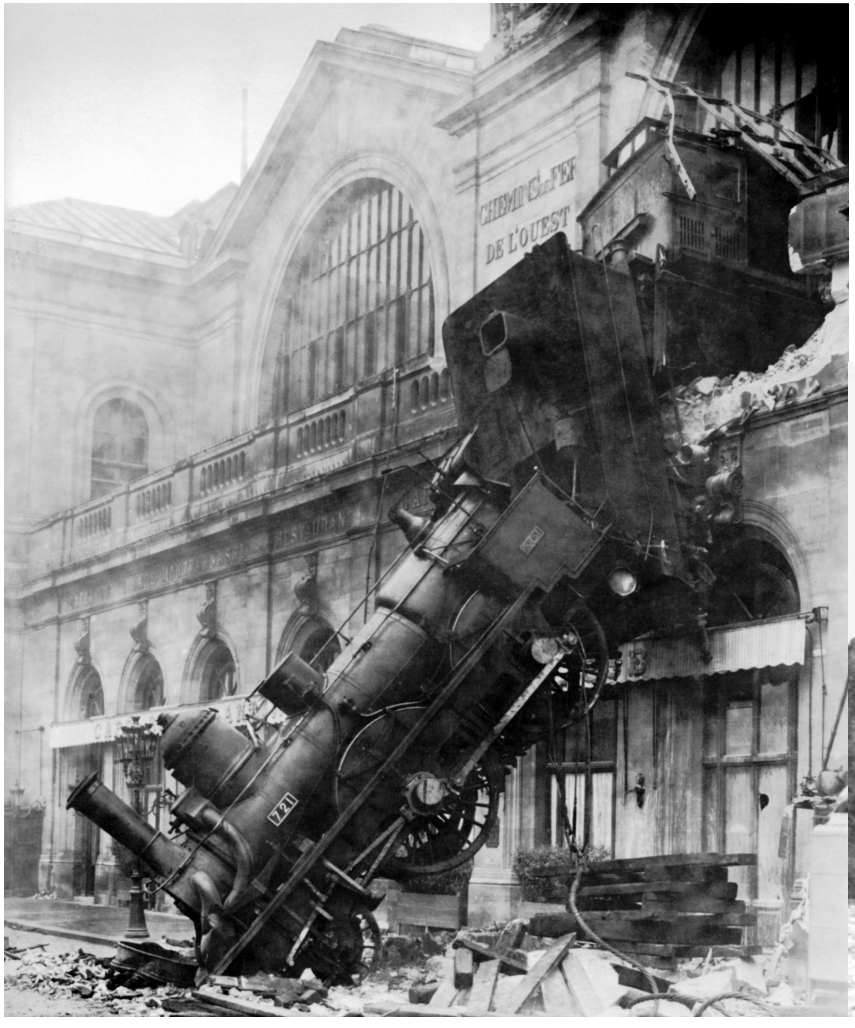
Department of Statistics
University of California, Berkeley

2014 SIAM/ASA Conference on Uncertainty Quantification
Savannah, GA
30 March – 3 April 2014

Abstract

How can we quantify the reliability or uncertainty of a method, analysis, or result, if we don't have the essential details, including the data and a recipe to reproduce the analysis? How can we tell precisely what analysis was performed, to see whether it was appropriate and correct? How can we tell whether a result is artifactual or real, accidental or generalizable? To check a claim might require knowing details of the instrument used to collect data, the raw data, data cleaning, data pre-processing, model selection algorithms, model fitting algorithms, statistical tests, scripts and code, and software package versions; settings and tuning parameters in all those things; and even the software build environment. And results can be extremely sensitive to small changes to any of those things. Computationally driven research and data intensive research have largely abandoned the scientific method, and now rely more on trusting authority than on direct evidence. It is crucial to rectify this situation by developing tools, processes, and habits that allow our collaborators, referees, and others to check, use, and extend our work. Some of these exist already, and science has much to learn from current practice in software development. I will discuss a pilot course on Reproducible and Collaborative Statistical Data Science aimed to help "the next generation" develop better habits than my generation of scientists learned.

UQ and Reproducibility



??

James Bashford / AP

UQ and Reproducibility

NASA

Reuters / Japan TSB

There's a problem

- Science
  http://www.sciencemag.org/content/343/6168/229.summary
  "Recently, the scientific community was shaken by reports that a troubling proportion of peer-reviewed preclinical studies are not reproducible." McNutt, 2014
- Nature http://www.nature.com/nature/focus/reproducibility/
  "Over the past year, Nature has published a string of articles that highlight failures in the reliability and reproducibility of published research (collected and freely available at go.nature.com/huhbyr)." 2013
- Reproducibility Project
  "Do normative scientific practices and incentive structures produce a biased body of research evidence? . . . sampling from the 2008 issues of three prominent psychology journals - *Journal of Personality and Social Psychology*, *Psychological Science*, and *Journal of Experimental Psychology: Learning, Memory, and Cognition*." 2011
- G. Johnson, 2014. New Truths That Only One Can See
  http://www.nytimes.com/2014/01/21/science/new-truths-that-only-one-can-see.html?ref=todayspaper&_r=0

Hoist with our own petard!

$$\text{(Data-Driven Science)} - \text{(public data)} \neq \text{Science}$$

$$\text{(Computationally Driven Science)} - \text{(public code)} \neq \text{Science}$$

Claims without evidence are not Science.

Bring back Science

We've traded the Scientific Method for "trust me!"
We should insist on evidence.

UQ and Reproducibility

- How can we quantify uncertainty if we don't know what was done, or what it was done to?
- Guessing what was done seems, umm, inadequate.
- Many well-documented failures show the consequences.
- How do we restore the Scientific Method to Science?
- Reproducibility, replicability, repeatabiity, verifiability, auditability, stability, confirmability, reusability, extensibility
- Need fresh, non-overloaded terminology: *same-lab do-over-able*, *independent-lab do-over-able*, *from-data do-over-able*, *analysis re-code-able*, *analysis do-over-able*, *from-data code-reused do-over-able*, . . . ?
- Looking under the hood, versus an end-run on correctness or *ab initio*, independent do-over.
- Nb: "the data" are not the data, especially in Big Science.

- What's the underlying experiment?
- What are the raw data? How were they collected/selected?
- How were the raw data processed to get the "data"?
- What analysis was reported to have been done on the processed data?
- Was that analysis the right analysis to do?
- Was that analysis done correctly?
- Were the results reported correctly?
- Were there ad hoc aspects to the analysis?
- How many analyses were done before arriving at that one? What were they? What were the results? How was multiplicity treated?
- What would happen if different choices were made?
- Can someone else use the tools?

Personal failure stories

Multitaper spectrum estimation for time series with gaps: lost *C* source for MEX files; old MEX files not compatible with some systems.

> *Unfortunately I was not able to find my code for multitapering. I am pretty sure I saved them after I finished my thesis, along with all the documentation, but it seems like I lost them through one of the many computer moves and backups since. I located my floppy (!) disks with my thesis text and figures but not the actual code.*

Poisson tests of declustered catalogs: current version of code does not run.

Why work reproducibly?

Cornford, 1908. *Microcosmographia Academica*

There is only one argument for doing something; the rest are arguments for doing nothing.

The argument for doing something is that it is the right thing to do.

Incentives, disincentives, moral hazard

- it's the right thing to do: check, reuse, extend, share, collaborate w/ others & your future self
- greater impact
- greater scientific throughput overall
- no *direct* academic credit
- requires changing one's habits, tools, etc.
- fear of scoops, tipping one's hand, exposure of flaws
- IP issues, data moratoria in "big science," etc.
- *may be* slower to publish a single project
- systemic friction: lack of tools & training
- lack of infrastructure to host runnable code, big data
- lack of support from journals, length limits, etc.
- lack of standards?

When and how?

- Built-in or bolt-on?
- Tools
- Training
- Developing good habits
- Changing academic criteria for promotions:
  How nice that you advertised your work in *Science*, *Nature*, *NEJM*, etc.! Where's the actual work? Where's the evidence that it's right? That it's useful to others?

Narrow replicability and reproducibility

- If something only works under *exactly* the same circumstances, shrug.
- If you can push a button and regenerate the figures and tables but you can't confirm what the code does, shrug.

Obfuscation, trust, and UQ

CERN Large Hadron Collider (LHC) ATLAS and CMS: Both use COLLIE for confidence limits, code proprietary to the team.

Software environments for reproducible research & teaching

- Teaching, research labs, multi-PI and multi-institute collaborations.
- In computational courses, can take two weeks to get everyone "on the same page" w/ software, VMs, etc.
  OS matters, versions matter, build environments matter, ...
- Work done by one PhD student is rarely usable by the advisor or the next PhD student—much less by the rest of the world.
  Claerbout's experience.
- BCE: reproducible recipe to (re)create software environment that fosters reproducible work.

> Any project involving 2 or more computers—that is to say, almost any current research project—requires some level of reproducibility. ... existing solutions rarely scale beyond the immediate collaborators on the project. ... the development of virtual environments both for semester-long university courses as well as brief workshops and trainings ranging from hours to about a week. ... the real challenge is in simplifying the *use* of these environments. ... Berkeley Common Environment (BCE), an easy-to-install, standardized virtual environment that supports a broad range of instructional needs.

Teaching reproducible computational research

- Statistics 157, fall 2013:
  Reproducible and Collaborative Statistical Data Science
- Project: improved earthquake forecasts for Southern CA
- Syllabus includes introduction to virtual machines, GitHub, IPython, SCEC data
- `http://youtu.be/Bq71Pqdukeo`, `Git_That_Data.mp4`