Irreproducible Musings on Reproducibility

Philip B. Stark

Department of Statistics
University of California, Berkeley

2014 UW / Moore-Sloan Workshop on Reproducibility
University of Washington
Seattle, WA
8 May 2014

Mongo like "Reproducibility"!

- Reproducibility good. But what is it?
- Reproducibility good. But why?
- Reproducibility hard. But why?
- Reproducibility hard to sell. But why?
- Hard to teach an old dog new tricks.
- Solution: Work with puppies.

Why do we like reproducibility?

- Provides (a way to generate) evidence of correctness
- Enables re-use, modification, extension, ...
- Exposes methods, which might be interesting and instructive
- Gut feeling that transparency and openness are good

Claim: Reproducibility is a tool, not a primary goal.

Might accomplish some of those goals without it, but it's a Very Powerful Tool.

- 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.

Software environments for reproducible/collaborative 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.

BCE first-cut ingredients

- a version of linux, perhaps Ubuntu
- docker
- lxc
- git, a git gui, gitlabhq, gitannex assistant
- Python + IPython + Numpy + Scipy + Matplotlib + Pandas + Cython + other libraries
- R and various libraries for statistics and machine learning
- mySQL, MariaDB, SQLite
- LaTeX, BibTeX + AMS, Beamer, & other styles
- some stack for distributed computing
- test suites for all the software

Teaching reproducible and collaborative 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`