Moore / Sloan Data Science Environment

Working Group on Reproducibility and Open Science

Randy LeVeque
David Beck
Joe Hellerstein
Bill Howe
Dan Halperin
Stephanie Wright
Many others... You??

http://escience.washington.edu/reproducible

http://uwescience.github.io/reproducible

What does Reproducible Research mean?

Ability to determine exactly how scientific results were obtained.

- Basis of scientific method.
- Required for confidently building on past results.
- Critical for accountability in engineering analysis / decision making.

Standards and best practices in computational/data science are not yet well codified.

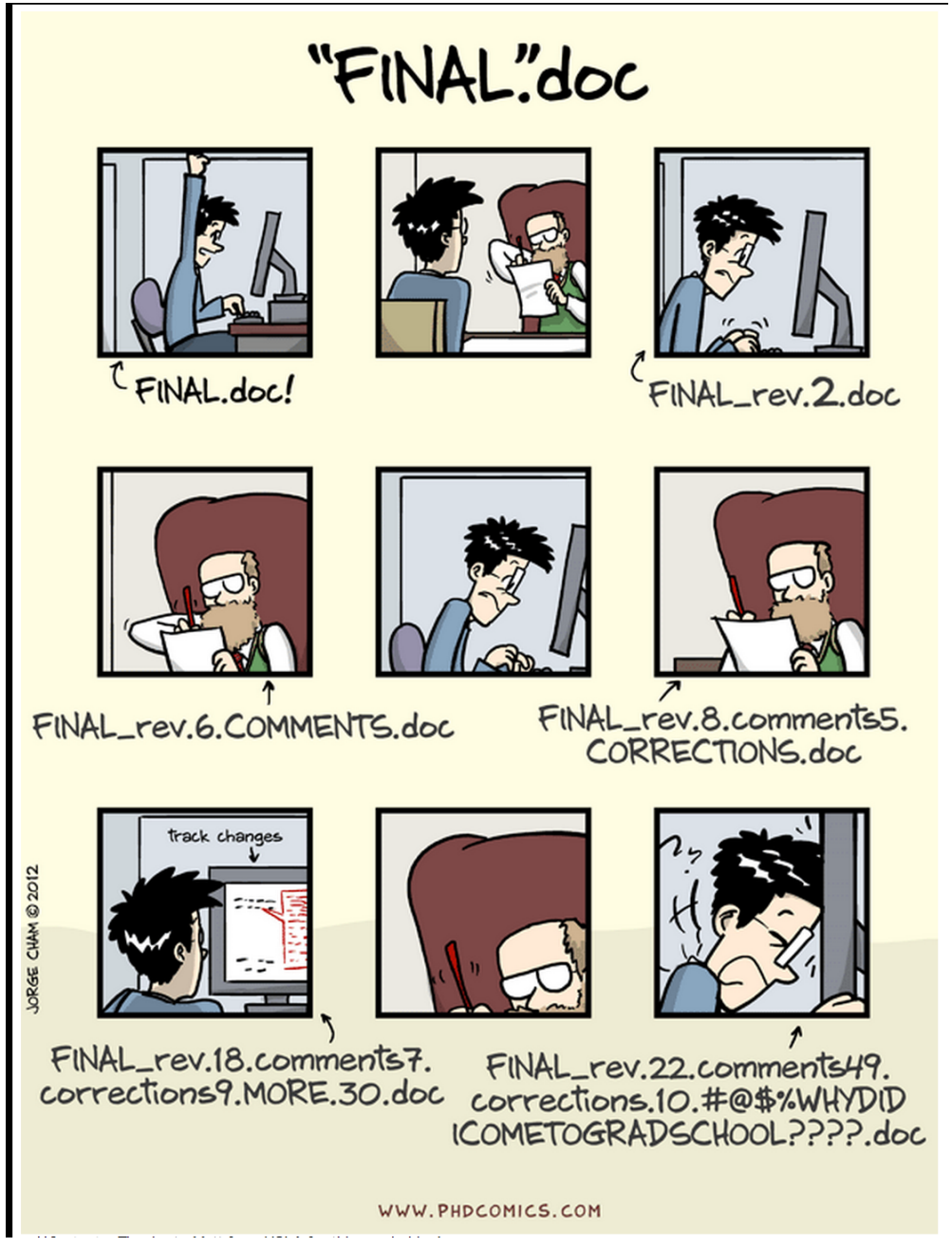
- Experimental science: Lab notebooks, methodology section of publications, etc.
- Mathematics: Proofs are required in publications.

Quote from Reproducible Research: A Cautionary Tale

By David Crotty, March 26, 2014 on **the scholarly kitchen blog**

If your experiment consists of running numerical data through an algorithm, then releasing your data and your code allows others to quickly verify that you've done what you've said you've done. But when it comes to other types of research, wet bench experiments or observational work for example, reproduction is not quite so simple.

If only it were so easy in computational/data science!



http://www.phdcomics.com/comics/archive.php?comicid=1531

Private reproducibility...

Use scripts, not GUIs, for data analysis and visualization.

Use version control / provenance tracking tools.

Archive code and data used for published results.

Why?

- Ability to check results in prior publication,
- Ability to build on your own past research of your own (or students / collaborators).
- Easily modify tables/figures to satisfy referees, etc.

Auditable Research: Even if code and data are not shared, there should be a permanent record that can be checked.

Analogous to lab notebooks.

Public Reproducibility...

Allowing others to reproduce your results.
(Readers, referees, researchers down the hall...)

Why?

- Verifying scientific integrity of results.
- Aids in understanding ideas, implementing methods
- Increases impact of work.

What does Reproducible Research mean?

What does it mean that others can reproduce your results?

Possible answers...

- Download the code and type `make plots`, see identical plots appear.
- Be able to implement the algorithm from description in paper and other archived sources, and get essentially the same results.
- Various things in between.

Terms such as replicable or repeatable are sometimes used in addition to reproducible.

Science Code Manifesto

Manifesto Discussion Endorse Resources About

Software is a cornerstone of science. Without software, twenty-first century science would be impossible. Without better software, science cannot progress.

But the culture and institutions of science have not yet adjusted to this reality. We need to reform them to address this challenge, by adopting these five principles:

Code	All source code written specifically to process data for a published paper must be available to the reviewers and readers of the paper.
Copyright	The copyright ownership and license of any released source code must be clearly stated.
Citation	Researchers who use or adapt science source code in their research must credit the code's creators in resulting publications.
Credit	Software contributions must be included in systems of scientific assessment, credit, and recognition.
Curation	Source code must remain available, linked to related materials, for the useful lifetime of the publication.

Endorsements at: `http://sciencecodemanifesto.org/`

Tools to facilitate reproducibility

- Version control systems (VCS)

  CVS, Subversion (server-client),
  Git, Mercurial, Bazaar, etc. (distributed).

- Public hosting sites for VCS repositories

  Github, Bitbucket, Google code, sourceforge, etc.

  Collaboration on open source projects,
  Archiving code used for publications.

- Other archives with stable URLs, DOIs

  Institutional or public data repositories,
  journal supplementary materials,
  figshare.com

Tools to facilitate reproducibility

- Workflow Management Systems

  VisTrails, Madagascar, Sumatra, Taverna, Galaxy, etc.

  Capture the workflow used to generate figures, tables, etc.

  Facilitate tracking the provinance of individual results.
  Data, code, compilers, graphics tools, etc.

  Often work together with VCS for source code.

Tools to facilitate reproducibility

- Literate Programming tools

  CWEB, Doxygen, Sphinx, Sweave, etc.

- Notebooks / Publishing tools

  Mathematica, Maple, Matlab,

  Sage, IPython, knitr, RStudio, etc.

Tools to facilitate reproducibility

- Virtualization

  Package code along with complete environment
  (OS, compilers, graphics tools, etc.)

  E.g., VirtualBox, VMware, etc.

- Cloud computing

  E.g., Amazon EC2, Windows Azure, etc. + VM

- Web platforms for running code

  E.g., RunMyCode.org, wakari.io,

  cloud.sagemath.com

Policy Issues

Should journals require data/code sharing?

Some already do, e.g. Science:

> **Data and materials availability.**
>
> *All data necessary to understand, assess, and extend the conclusions of the manuscript must be available to any reader of* Science*.*
>
> *All computer codes involved in the creation or analysis of data must also be available to any reader of* Science*.*
>
> *After publication, all reasonable requests for data and materials must be fulfilled.*

http://www.sciencemag.org/site/feature/contribinfo/prep/

Certification of reproduciblity?

Stamp of reproducibility on journal papers.

E.g. *Biostatistics* awards “kite marks”,

- **D** for data availability,
- **C** for code availability,
- **R** for full reproduciblity

Checked by “reproducibility editor”.

Campus-wide certification of labs or research groups following “best practices?” (What are these?)

Reward structure

What are the rewards / penalties for attempting to do reproducible research?

Often much more to be gained by moving on to next project than cleaning up and posting code.

Little recognition available vs. many potential downsides of sharing.

How can we encourage more recognition and better support?

Concerns for young researchers in particular.

Altmetrics: Measuring contributions other than traditional journal publications, e.g., **ImpactStory.org**

Guidelines for Reproducibility & Open Science

See: **http://uwescience.github.io/reproducible**

Goals:

- To achieve greater scientific validity and integrity by making it easier to verify published results.
- To increase productivity of current and future researchers on funded projects.
- To increase the impact of the research performed, software developed, and papers published.
- To help promote data and code as first class research products.
- To increase access to and usability of research products by other researchers.
- To use the DSE as a test bed for developing and promoting tools and cultural changes across a broad spectrum of academic disciplines.