Reproducible Research

Mark Agerton

2019 September 23

3 kinds of reproducibility

- **Methods:** *The ability to implement, as exactly as possible, the experimental and computational procedures, with the same data and tools, to obtain the same results.*
- Results: *obtaining the same results from the conduct of an independent study*
- Inferential: do we interpret results from an independent or reanalysis study?

Goodman, Fanelli, and Ioannidis (2016)

Reproducible research

Make the *entire research process* transparent

Not just regression tables. It includes

- Every step of data-work (including downloading!)
- Your figures
- Sources for any fact you cite

Reproducible research is about **workflow** and **sharing**

Why bother with reproducible research?

1. We are scientists who (ostensibly) care about the truth, and need others to be able to verify it
2. We can build off one another's work.
3. Journals care about it.
4. It'll make your life easier (in the long run).

New AER Reproducibility Guidelines (Methods)

- AER's Data and Code Availability Policy

  For econometric, simulation, and experimental papers, the replication materials shall include

  a. the data set(s),
  b. the programs used to create any final and analysis data sets from raw data,
  c. programs used to run the final models, and
  d. description sufficient to allow all programs to be run.

- Social Science Data Editors Verification guidance and Replication Template

Ten Simple Rules for Reproducible Computational Research

1. For Every Result, Keep Track of How It Was Produced
2. Avoid Manual Data Manipulation Steps
3. Archive the Exact Versions of All External Programs Used
4. Version Control All Custom Scripts
5. Record All Intermediate Results, When Possible in Standardized Formats
6. For Analyses That Include Randomness, Note Underlying Random Seeds
7. Always Store Raw Data behind Plots
8. Generate Hierarchical Analysis Output, Allowing Layers of Increasing Detail to Be Inspected
9. Connect Textual Statements to Underlying Results
10. Provide Public Access to Scripts, Runs, and Results

(Sandve et al. 2013)

How to do this?

- Gold Standard? Create entire paper with one command.
- *Literate programming*: Rmarkdown, Jupyter notebooks, Stata's `dyndoc`
- Usually not practical: too many files!

Mark's strategy

1. Document everything
2. Version control (almost) everything
3. Automate as much as ~~possible~~ practical

Benefits & costs

- On the plus side
  - Keeps me organized
  - Means others can understand what I'm doing
  - Lowers marginal cost of re-running analysis with new data, assumptions, etc
  - Reduces errors
- But
  - Higher fixed costs
  - Co-authors grumpy about new software

Starting a new project

- Everything lives in one folder
- Sublime Text and R projects
- Track everything with Git
- And keep only 1 version (no redundancy!)
- Everything is a text file (except data and .pdfs)
    - “Diff”-able (track changes)
    - Future-proofed
    - Searchable from command line, Sublime Text

Folder structure and filenames

- `README.md` in every folder that might need one
- Save `raw_data`
  - Ideally download programatically code
  - `README.md` documents acquisition
  - Save md5 hash of data
  - Don't ever modify it!
- `intermediate_data`
  - Processed data
- `writeup`
  - `paper/` has .tex files
  - `yyyy-mm-presentation/`
  - `figures/`
  - `tables/`
  - various notes
- `code`
  - `master.do`, `run_all.sh` or `MAKE` script to run all analysis
  - Name files in order of analysis `00a - download prices.R`, `00b - download shapefiles.R`

Staying organized with lengthy jobs on a cluster

- Jobs on cluster get a unique job ID
- Keep a `README.md` log with *metadata* (data about data) listing job IDs
- In the log, print
    - unique “commit hash” to log that identifies current snapshot of the codebase
    - Lots of intermediate output

Resources on project management / organization

- Software Carpentry Data Management lesson
- Other economists' workflows
  - Ryan Kellogg's Lab Wiki
  - Hunt Allcott's Lab Wiki
  - Gentzkow and Shapiro *Lab Wiki*
  - Gentzow and Shapiro *PDF Code and Data for the Social Sciences: A Practitioner's Guide*
  - Knittel and Metaxoglou (2016) *Working with Data: Two Empiricists' Experience*

What is Git?

- A program that keeps track of file histories
- Like blockhain lite — everyone gets the entire project history
- The biggest-baddest "Undo" button ever — roll back to any *commit*ted file
- Searchable
- Able to merge text file versions

Git

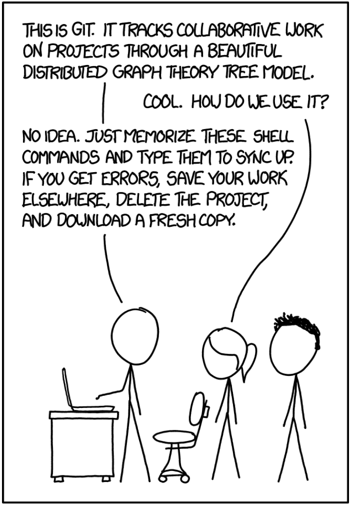
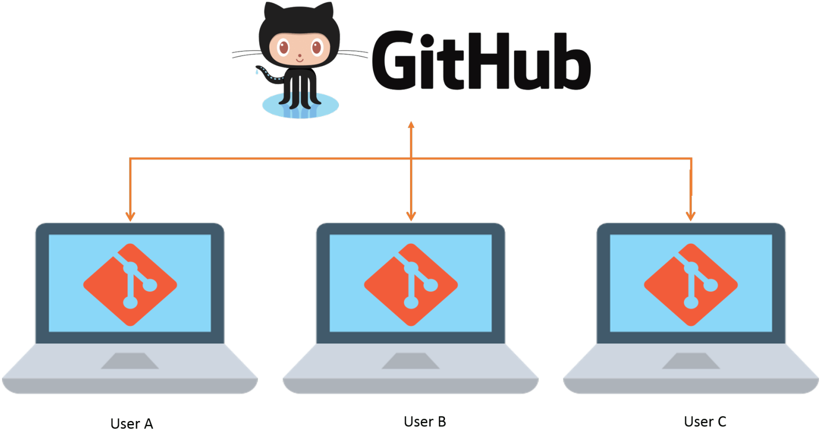

THIS IS GIT. IT TRACKS COLLABORATIVE WORK ON PROJECTS THROUGH A BEAUTIFUL DISTRIBUTED GRAPH THEORY TREE MODEL.
COOL. HOW DO WE USE IT?
NO IDEA. JUST MEMORIZE THESE SHELL COMMANDS AND TYPE THEM TO SYNC UP. IF YOU GET ERRORS, SAVE YOUR WORK ELSEWHERE, DELETE THE PROJECT, AND DOWNLOAD A FRESH COPY.

Figure 1: *If that doesn't fix it, git.txt contains the phone number of a friend of mine who understands git. Just wait through a few minutes of 'It's really pretty simple, just think of branches as. . . ' and eventually you'll learn the commands that will fix everything.*

Git vs Github



Git

https://hackernoon.com/a-gentle-introduction-to-git-and-github-the-eli5-way-43f0aa64f2e4
https://www.slideshare.net/HubSpot/git-101-git-and-github-for-beginners

- Resources: https://lectures.quantecon.org/jl/more_julia/version_control.html and
- Why version control?
- Version control is “undo”
- Everything is text!!
- Github vs git
- Education discounts
- Private vs public
- Issues

Structuring data

https://v4.software-carpentry.org/data/index.html
https://mariadb.com/kb/en/library/database-normalization-overview/

References

- https://github.com/kelloggrk/Kellogg_RA_Manual

Goodman, Steven N., Daniele Fanelli, and John P. A. Ioannidis. 2016. "What Does Research Reproducibility Mean?" *Science Translational Medicine* 8 (341): 341ps12–341ps12. doi:10.1126/scitranslmed.aaf5027.

Sandve, Geir Kjetil, Anton Nekrutenko, James Taylor, and Eivind Hovig. 2013. "Ten Simple Rules for Reproducible Computational Research." *PLOS Computational Biology* 9 (10): e1003285. doi:10.1371/journal.pcbi.1003285.