

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?

@goodman2016what

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

[@sandve2013ten]

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 programmatically 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*
 - ▶ Gentzkow 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
- ▶ The biggest-baddest “Undo” button ever — roll back to any *committed* file
- ▶ Like Blockchain lite
 - ▶ Everyone gets the entire project history
 - ▶ Versions identified with cryptographic hash
- ▶ Searchable
- ▶ Can merge text file versions

Git

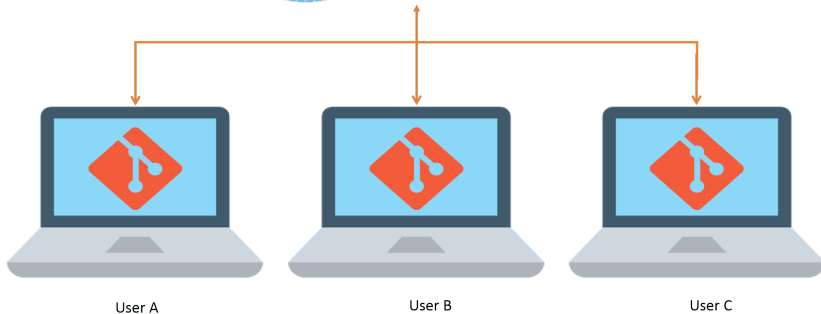


Figure 1: *If that doesn't fix it, git.txt contains the phone number of a friend of mine who understands git. Just wait through a few minutes of 'It's really pretty simple: just think of branches as ...' and eventually you'll learn the commands*

Git vs Dropbox

- ▶ Dropbox is better than nothing
 - ▶ Limits on how far back history saved
 - ▶ Can't choose what snapshots (*commits*) get saved
 - ▶ No metadata (*commit messages*) with saved versions
 - ▶ No tool to *merge* conflicting versions
 - ▶ Searching history is hard
 - ▶ Who changed what? (technical term: *blame*)
- ▶ But Git has higher fixed costs

Git vs Github



Repositories vs commits

- ▶ **Repository:** folder where the entire history is tracked (*repo*)
 - ▶ Can be on your computer (*local*)
 - ▶ Or hosted online (*remote*)
 - ▶ Puts a folder `.git` in your directory
- ▶ **Commit**
 - ▶ “snapshot” of your repository at a point in time. 3 pieces of info
 1. A unique ID (*SHA-1 hash*)
 2. Content: pointer to *parent commit* + *diff* (changes to it)
 3. Metadata: *commit message*, author, timestamp
- ▶ **Github:** online service hosts repos
 - ▶ Can have git without Github
 - ▶ Public vs private
 - ▶ Adds extra functionality, project management
 - ▶ Great [education discount](#)

What I'll show you today

- ▶ Git GUIs
 - ▶ **Github Desktop** (what we do today)
 - ▶ Git Kraken
 - ▶ Sublime Merge
- ▶ Can also use command line, esp. for complicated situations

Workflow: Fork + Clone a repo

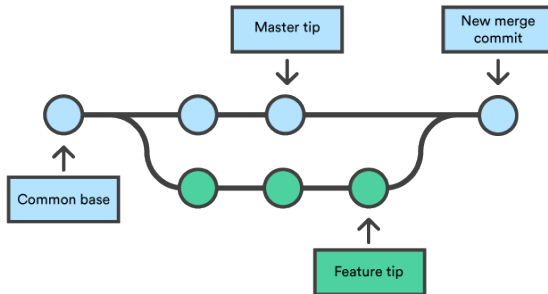
- ▶ Go to
<https://github.com/magerton/research-computing-bootcamp.git>
- ▶ *Fork* the repo on the remote
 - ▶ Creates a *clone* of the repo in YOUR Github account
 - ▶ A *fork* keeps ability to sync original repo and yours
- ▶ *Clone* the repo from remote to local

(makes new folder)

```
git clone https://github.com/[USERNAME]/research-computing-b
```

Branches

- ▶ Branches are sets of commits
- ▶ Keep master branch with functioning code
- ▶ Develop new features on *branches*
- ▶ When done, *merge* branches back to master



Workflow: Switching branches

```
# change working directory to new repo  
cd my-first-test-repo
```

```
# add `-b` flag to create new branch  
git checkout marks-test-branch
```

Workflow: staging changes + adding commit

make a new file

```
echo "do Actually Relevant Econ" > note-to-self.md
```

see changes

```
git status
```

```
git diff
```

stage changes

```
git add *.md      # markdown files
```

```
git add -u        # tracked files
```

```
git add --all     # all files
```

```
git commit -m "Note to self" -m "About how econ is great"
```

PUSH changes to remote

```
git push
```

Workflow: checkout branch and merge branches

Switch back to master

```
git checkout master
```

Merge in changes from `marks-test-branch` to current branch

```
git merge marks-test-branch
```


Workflow: Create a repo

- ▶ Make a new *private* repo on Github (remote): github.com/new
- ▶ Create a *local* repo and then push changes to remote

```
# new directory
```

```
mkdir my-first-test-repo
```

```
# cd into it
```

```
cd my-first-test-repo
```

```
# create readme
```

```
echo "# my-first-test-repo" >> README.md
```

```
# tells git to track new repo
```

```
git init
```

```
# STAGES the file
```

```
git add README.md
```

```
# COMMITS staged changes
```

```
git commit -m "first commit"
```

Authentication with Github Desktop & command line

- ▶ Github Desktop uses https-based authentication
- ▶ Command line works best with ssh
 - ▶ More from Arnon tomorrow=
 - ▶ Github [Authentication help](#)
 - ▶ To switch from https to ssh
 - ▶ See Github help on [changing a remote's URL](#)

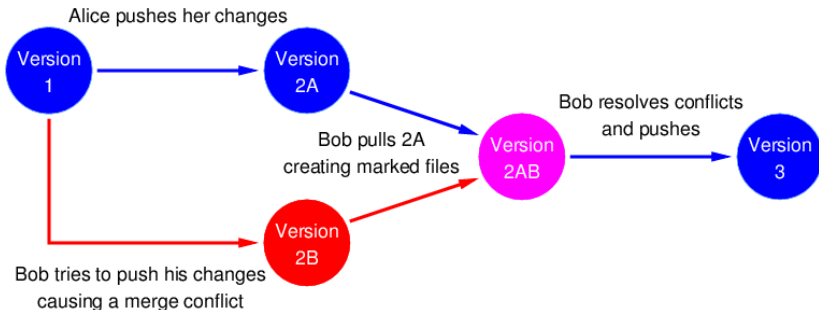
```
git remote -v
```

```
git remote set-url origin git@github.com:USERNAME/repository.git
```

Syncing between computers

- ▶ push changes to the remote
- ▶ pull changes from remote to local
 - ▶ `pull = fetch + git merge FETCH_HEAD`

merge conflicts



- Resolve in text editor ([Atom](#) is great for this!)

Important files

- ▶ `README.md` is for readers
 - ▶ Automatically displayed as HTML by Github
- ▶ `.gitignore` tells Git which files not to track
 - ▶ LaTeX temp files, big binary files, private keys
 - ▶ `.gitignore` has an example
- ▶ `LICENSE.md` tells how others can use, cite, modify, etc your work

What to do about binary files?

- ▶ Git is fantastic for tracking text files (.txt, .tex, .md, .R, .do, .py, .jl, .m)
 - ▶ But not made for big binary files (.zip, .dta, .RData, .rda, .jld, .gz)
 - ▶ Use git-lfs to handle binary files $\leq 2\text{Gb}$
 - ▶ Store binary files on a separate server
 - ▶ Commits a text *pointer file* with unique *hash* and URL to repo
 - ▶ Way faster
 - ▶ 1 Gb for free, \$5/mo for 50 Gb
- ```
git lfs track my-big-file.zip # single file
git lfs track "*.zip" # all files ending in .zip
```

# Project management on Github

- ▶ Permissions: who has the right to push to the repo?
  - ▶ repo > Settings > Collaborators
- ▶ Issue-tracking
  - ▶ Can reference issues in commit messages
  - ▶ Can tag other users

# Organizing data

- ▶ What I've started doing
  - ▶ Save raw data into `./raw_data/`, with README.md on how to get it
  - ▶ Ingest data *programmatically*
    - ▶ *Normalize* data as much as feasible. Redundancy is evil!
    - ▶ Save to `./intermediate_data/`
  - ▶ Further processing
    - ▶ Ideally, save to `./final_data/` (I don't... 🙄)
- ▶ If data are huge, interesting guide on [Software Carpentry](#)



## Normalizing data: Suppose you get this Excel spreadsheet

| LocID | Loc. nm.     | PlantID | Plant nm.    | Soil | Soil desc           |
|-------|--------------|---------|--------------|------|---------------------|
| 11    | Kirstenb. G. | 431     | Leucadendron | A    | Sandstone           |
|       |              | 446     | Protea       | B    | Sandstone/limestone |
|       |              | 482     | Erica        | C    | Limestone           |
| 12    | Karbon. M.   | 431     | Leucadendron | A    | Sandstone           |
|       |              | 449     | Restio       | B    | Sandstone/limestone |

Example from [MariaDB.com](https://mariadb.com)

## Trying to create a table. . . .

| LocID | Loc. nm.     | PlantID | Plant nm.    | Soil | Soil desc           |
|-------|--------------|---------|--------------|------|---------------------|
| 11    | Kirstenb. G. | 431     | Leucadendron | A    | Sandstone           |
| NULL  | NULL         | 446     | Protea       | B    | Sandstone/limestone |
| NULL  | NULL         | 482     | Erica        | C    | Limestone           |
| 12    | Karbon. M.   | 431     | Leucadendron | A    | Sandstone           |
| NULL  | NULL         | 449     | Restio       | B    | Sandstone/limestone |

- ▶ Rows 1–3 from one group
- ▶ Rows 4–5 from another
- ▶ How do you reference a row?

## Each record needs to stand alone

| LocID | Loc. nm.     | PlantID | Plant nm.    | Soil | Soil desc           |
|-------|--------------|---------|--------------|------|---------------------|
| 11    | Kirstenb. G. | 431     | Leucadendron | A    | Sandstone           |
| 11    | Kirstenb. G. | 446     | Protea       | B    | Sandstone/limestone |
| 11    | Kirstenb.    | 482     | Erica        | C    | Limestone           |
| 12    | Karbon. M.   | 431     | Leucadendron | A    | Sandstone           |
| 12    | Karbon. M.   | 449     | Restio       | B    | Sandstone/limestone |

- ▶ This is what Stata likes
- ▶ *Primary Key*
  - ▶ “specific choice of a minimal set of attributes (columns) that uniquely specify a tuple (row) in a relation (table)” [Wikipedia](#)
  - ▶ Here: Loc. AND Plant
  - ▶ Called a *natural* primary key
  - ▶ Could also be an ID number (row number?)

This is better, but

| LocID | Loc. nm.     | PlantID | Plant nm.    | Soil | Soil desc           |
|-------|--------------|---------|--------------|------|---------------------|
| 11    | Kirstenb. G. | 431     | Leucadendron | A    | Sandstone           |
| 11    | Kirstenb. G. | 446     | Protea       | B    | Sandst/limestone    |
| 11    | Kirstenb.    | 482     | Erica        | C    | Limestone           |
| 12    | Karbon. M.   | 431     | Leucadendron | A    | Sandstone           |
| 12    | Karb. M.     | 449     | Restio       | B    | Sandstone/limestone |

1. Table stores redundant information: *LocID* AND *Loc nm*
2. Misspellings!

## Removing the fields not dependent on the entire key

| LocID | PlantID |
|-------|---------|
| 11    | 431     |
| 11    | 446     |
| 11    | 482     |
| 12    | 431     |
| 12    | 449     |

| PlantID | Plant nm.     |
|---------|---------------|
| 431     | Leaucadendron |
| 446     | Protea        |
| 482     | Erica         |
| 449     | Restio        |

| LocID | Loc. nm.     |
|-------|--------------|
| 11    | Kirstenb. G. |
| 12    | Karbon. M.   |

| Soil | Soil desc           |
|------|---------------------|
| A    | Sandstone           |
| B    | Sandstone/limestone |
| C    | Limestone           |

- ▶ No redundancy
- ▶ Join using *Foreign Key*: “foreign key is a set of attributes that references a candidate key” [Wikipedia](#)
- ▶ *Factor variables* in Stata, R or *Categorical Arrays* in Julia

# Software tools I like

- ▶ Everything is cross-platform
- ▶ Everything is a text file  $\implies$  need a good text editor
  - ▶ Syntax highlighting
  - ▶ Compile LaTeX & Markdown/Pandoc
  - ▶ Searchable with [Regular Expressions](#)
- ▶ Minimize redundancy  $\implies$  one text format to rule them all!
  - ▶ Markdown/Pandoc
    - ▶ Compile to Word, LaTeX, PPT, Beamer, Reveal.js, HTML
    - ▶ Auto-formatted on Github
    - ▶ Lighter than LaTeX
- ▶ Git, Github Desktop
- ▶ [Zotero](#) for bibliography (can share with people via [Groups](#))
- ▶ [Sublime Text 3](#) for everything else

# References

- ▶ [https://github.com/kellogggrk/Kellogg\\_RA\\_Manual](https://github.com/kellogggrk/Kellogg_RA_Manual)