Streamlining your research through

Data Management and

Reproducible Research

Andrew Johnson, CU Boulder

Tobin Magle, Colorado State University

What is data management?

Depends who you ask...

- Researchers
- Data managers
- Librarians
- Institutional Review Boards
- Research funders
- Journal publishers

Data management is a process

Actually, it’s a bunch of processes...

- Storage
- Backup
- Organization
- Documentation
- Sharing
- Preservation

~~What is~~ *Why* data management?

Do you need to?

- Keep data secure?
- Organize data?
- Store data for a long time?
- Share data? (With whom?)
- Make data reusable?
- Meet a funder/publisher requirement?
- All of the above?

Let’s focus on reuse

In order to enable reuse...

- Others need to access your data
- Others need to understand your data
- Simple, right?

So, what do you need?

To allow for access...

- A repository or data sharing platform
- Many options:
  - General (figshare, Dryad, Open Science Framework)
  - Disciplinary (re3data.org)
  - Institutional (CU Scholar, Digital Collections of Colorado - CSU)
- Again, the “why” is important:
  - Complying with funder/journal policy?
  - Preserving for the long-term?
  - Enabling reuse?

To allow for understanding...

- Documentation (metadata)
- Readme file(s): https://cornell.app.box.com/v/ReadmeTemplate
  - Project/dataset-level: Title, authors/creators (contact info), date(s), location(s), related works, licenses/restrictions, recommended citation
  - Methodological information: Processing steps, calibration, environmental conditions, instrument- or software-specific information
  - File-level: Names (and naming convention), descriptions
  - Data-level: Variables, codes, units, missing values

Example

Open Science Framework (https://osf.io/)

- Free tool for organizing, managing, sharing data (and all other research objects)
- Allows for all levels of access
- Extremely flexible with regard to documentation
- Shameless self-promotion:
  - “Workshop for Increasing Openness and Reproducibility in Quantitative Research”, CRDDS in Norlin Library (Room E206), September 19, 2017, 1-4pm, RSVP at https://goo.gl/YJvVng

Managing data is just one part of reproducible research...

Reproducible research

is the practice of distributing all data, software source code, and tools required to reproduce the results discussed in a research publication.

https://www.ctspedia.org/do/view/CTSpedia/ReproducibleResearchStandards

Reproducible research

=

Data

+

Analysis instructions

Reproducible research

=

Transparency

Replication vs. Reproducibility

Replication: Same conclusion new study (gold standard)

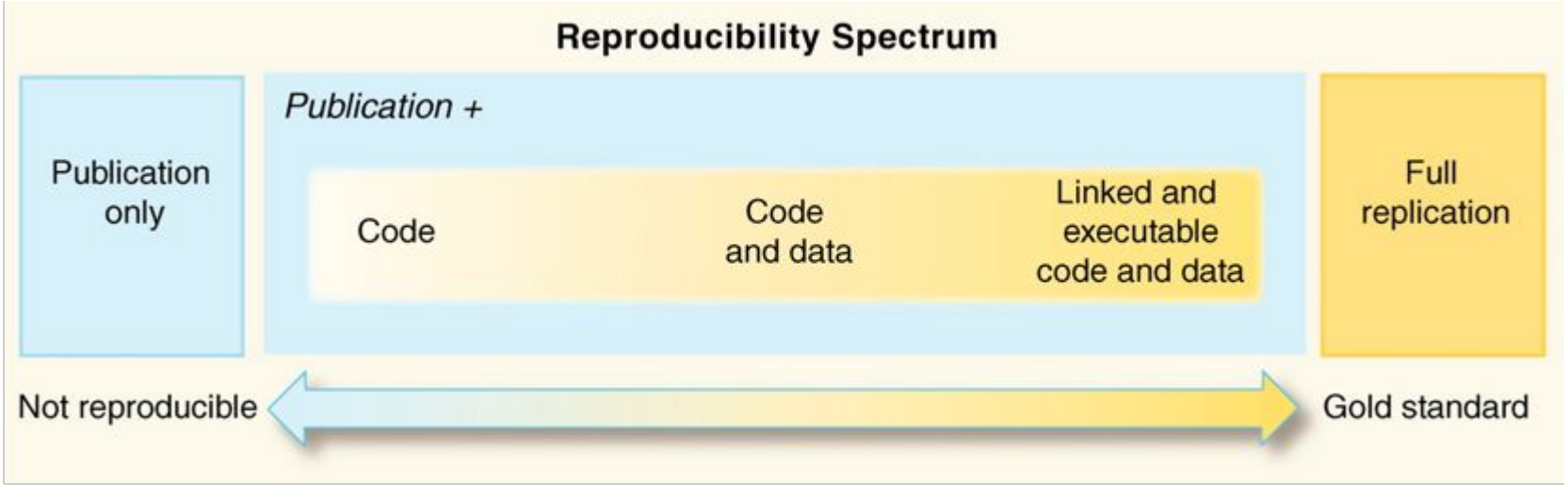
"Again, and Again, and Again …" **BR Jasny et. al.** Science, 2011. 334(6060) pp. 1225 DOI: 10.1126/science.334.6060.1225

Replication isn't always feasible: too big, too costly, too time consuming, one time event, rare samples

Reproducibility: Same results from same data and code (minimum standard for validity)

Reproducible Research in Computational Science". **RD Peng** Science, 2011. 334 (6060) pp. 1226-1227 DOI: 10.1126/science.1213847

Reproducibility spectrum



"Reproducible Research in Computational Science". **RD Peng** Science, 2011. 334 (6060) pp. 1226-1227 DOI: 10.1126/science.1213847

Reproducible research considerations

- Documentation
  - Automation
- Version Control
- Reproducible reports

How do you document your process?

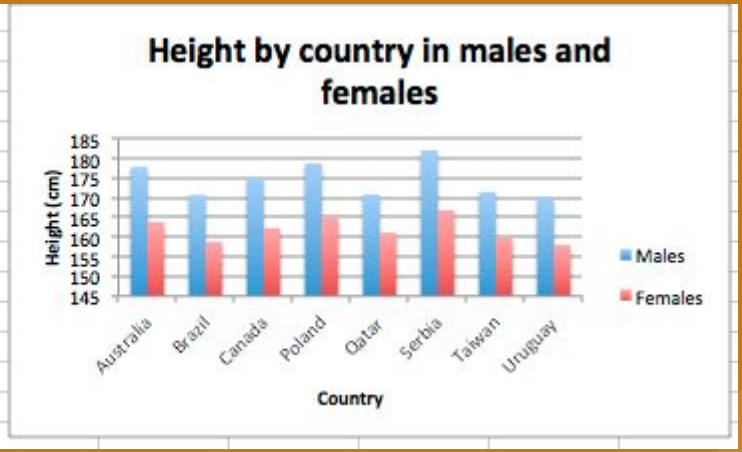
Document data cleaning and analysis

- **Minimum**: Written instructions

- **Optimal**: Automation via scripts

Example: Excel

	Height in cm	
	Males	Females
Australia	177.8	163.8
Brazil	170.7	158.8
Canada	175.1	162.3
Poland	178.7	165.1
Qatar	170.8	161.1
Serbia	182	166.8
Taiwan	171.4	159.9
Uruguay	170	158



By hand, documentation is...

- Slow
- Error prone
- Not easily replicated
- Not suited to replicates

Example: R script

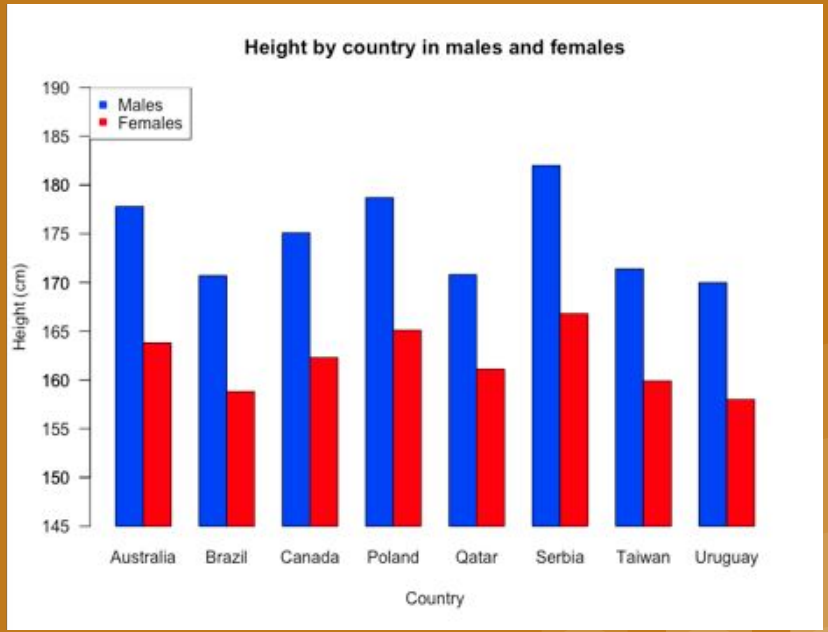
```r
#download the file
download.file(url = "http://libguides.colostate.edu/ld.php?content_id=27156359",
              destfile = "ex1.csv",
              method="libcurl")

#Load the data from the file into an R variable
height<-read.csv("ex1.csv", row.names="Country")

#Now let's plot the data:

counts<-t(as.matrix(height)) #converts the variable height to a format that
#can be plotted
counts<-counts-145            #transforms the data so it looks like the excel plot
barplot(counts,               #the height of the bar
        beside = TRUE,        #put cols next to eachother
        main="Height by country in males and females",  #plot title
        xlab="Country",       #X axis label
        ylab="Height (cm)",   #Y axis label
        col=c("blue", "red"), #bar colors
        offset=145,           #shifts the axis to make it look like excel
        ylim=c(145,190),      #y axis limits
        las=1)                #horizontal text
axis(side=2,                  #marks on the left of axis
     at=c(145,150,155,160,165,170,175,180,185),  #where you want ticks
     las=1) #horizontal text

legend(x=0, y=190, #coordinates of where you want the legend to go
       legend=c("Males", "Females"), #legend text label
       col=c("blue", "red"),         #colors
       pch=15)                       #shape of legend
```



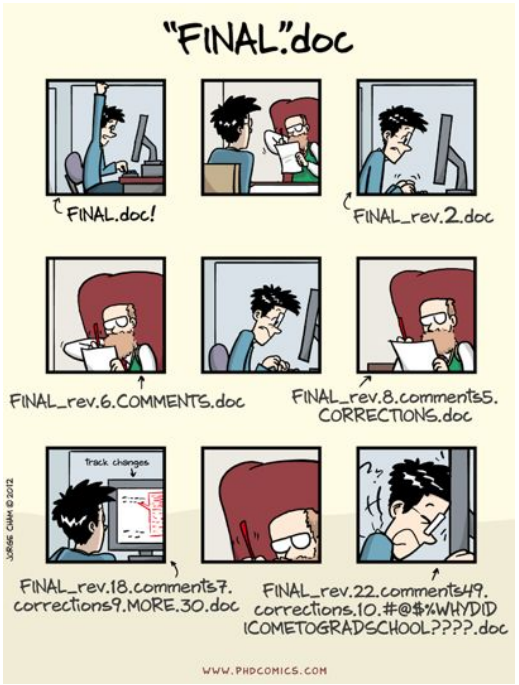
Automation makes documentation...

- **Done by default**: doing the analysis is 90% of the documentation
- Easily replicated
- Makes analyzing replicates easy

How do you track changes?

Version control

- **Intuitive:** saving V1, V2...
- **Formalized**: A system that records changes to a file or set of files over time so that you can recall specific **versions** later
- Examples: git, svn
- Have a high learning curve



How do you create reproducible reports?

Literate programming

=

Human readable text

+

Machine readable code

Examples of literate programming

- Markdown
- R Markdown
- Jupyter notebooks (supports Julia, R, and python)
- R notebooks

Questions?

Andrew Johnson

andrew.m.johnson@colorado.edu

@prezseventeen

Tobin Magle

tobin.magle@colostate.edu

@tobinmagle