Making research reproducible: git, R and org-mode

Philipp Homan, MD, PhD

phoman1@northwell.edu
http://github.com/philipphoman/mrr

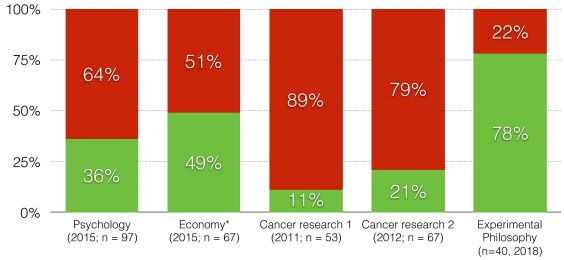
Why should we even care?

It’s not reproducible
if it only runs on your laptop!

http://www.jonzelner.net/docker/reproducibility/

We are faced with a replication crisis

Which part of published findings can be independently replicated?

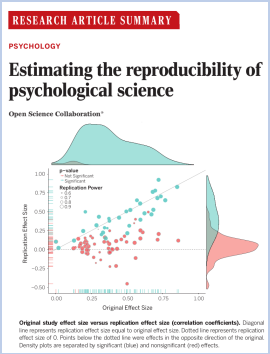


*The data on economics is about *reproducibility*; i.e. the attempt to get the same results if you apply the original data analysis on the original data set.

Open Science Collaboration (2015); Chang & Li (2015); Begley, C. G., & Ellis, L. M. (2012). Prinz, F., Schlange, T., & Asadullah, K. (2011); Cova et al. (2018)

Credit: Felix Schoenbrodt 2018 (@nicebread303)

We are faced with a replication crisis



RESEARCH ARTICLE SUMMARY

PSYCHOLOGY

Estimating the reproducibility of psychological science

Open Science Collaboration*

Original study effect size versus replication effect size (correlation coefficients). Diagonal line represents replication effect size equal to original effect size. Dotted line represents replication effect size of 0. Points below the dotted line were effects in the opposite direction of the original. Density plots are separated by significant (blue) and nonsignificant (red) effects.

Open Science Collaboration 2015, Science

- Ongoing methodological crisis in science
- Results of many scientific studies difficult to replicate
- Involves diverse fields (from psychology to cancer research)

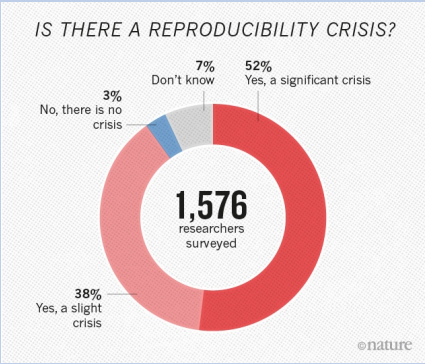
Ioannidis 2005, PLOS Med

In addition

In addition

We are probably also faced with a reproducibility crisis

We are probably also faced with a reproducibility crisis



Baker 2016, Nature

Replicating vs. reproducing

- Replicating: People going out and collecting new data
- Reproducing: People analyzing the same data
- What cannot be replicated often difficult to reproduce

jblevins.org/log/rep

What makes it hard to reproduce our own work?

What makes it hard to reproduce our own work?

DONALD AND BARBARA ZUCKER SCHOOL of MEDICINE AT HOFSTRA/NORTHWELL

We don’t always use reproducible work flows

What does that mean?

- Using just intuition when organizing data, manuscripts, code
- The same goes for analyzing data
- After a couple of months (sometimes weeks) it is hard to remember:

What does that mean?

- Using just intuition when organizing data, manuscripts, code
- The same goes for analyzing data
- After a couple of months (sometimes weeks) it is hard to remember:
  1. What we did

What does that mean?

- Using just intuition when organizing data, manuscripts, code
- The same goes for analyzing data
- After a couple of months (sometimes weeks) it is hard to remember:
  1. What we did
  2. Why we did it

What does that mean?

- Using just intuition when organizing data, manuscripts, code
- The same goes for analyzing data
- After a couple of months (sometimes weeks) it is hard to remember:
  1. What we did
  2. Why we did it
  3. How we did it

What can we do about it?

Three simple rules:

What can we do about it?

Three simple rules:

1. Separate data from analysis

What can we do about it?

Three simple rules:

1. Separate data from analysis
2. Use version control

What can we do about it?

Three simple rules:

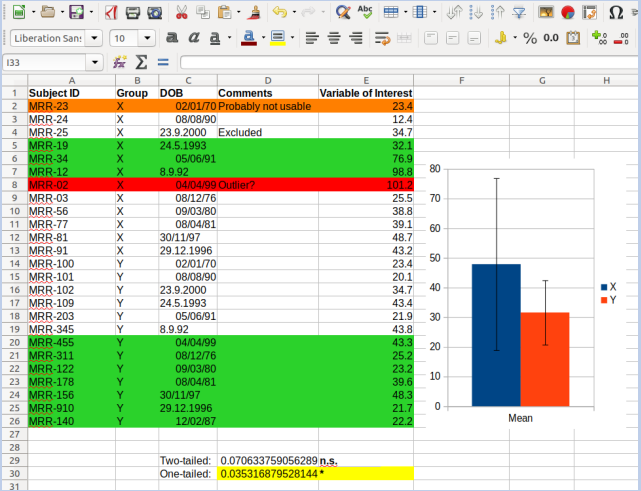
1. Separate data from analysis
2. Use version control
3. Use code to analyze data (not GUIs)

What can we do about it?

Three simple rules:

1. **Separate data from analysis**
2. Use version control
3. Use code to analyze data (not GUIs)

Separating data from analysis

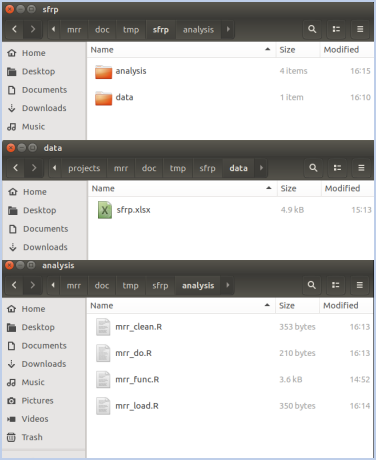


	Subject ID	Group	DOB	Comments	Variable of Interest
2	MRR-23	X	02/01/70	Probably not usable	23.4
3	MRR-24	X	08/08/90		12.4
4	MRR-25	X	23.9.2000	Excluded	34.7
5	MRR-19	X	24.5.1993		32.1
6	MRR-34	X	05/06/91		76.9
7	MRR-12	X	8.9.92		98.8
8	MRR-02	X	04/04/99	Outlier?	101.2
9	MRR-03	X	08/12/76		25.5
10	MRR-56	X	09/03/80		38.8
11	MRR-77	X	08/04/81		39.1
12	MRR-81	X	30/11/97		48.7
13	MRR-91	X	29.12.1996		43.2
14	MRR-100	Y	02/01/70		23.4
15	MRR-101	Y	08/08/90		20.1
16	MRR-102	Y	23.9.2000		34.7
17	MRR-109	Y	24.5.1993		43.4
18	MRR-203	Y	05/06/91		21.9
19	MRR-345	Y	8.9.92		43.8
20	MRR-455	Y	04/04/99		43.3
21	MRR-311	Y	08/12/76		25.2
22	MRR-122	Y	09/03/80		23.2
23	MRR-178	Y	08/04/81		39.6
24	MRR-156	Y	30/11/97		48.3
25	MRR-910	Y	29.12.1996		21.7
26	MRR-140	Y	12/02/87		22.2
29			Two-tailed:	0.070633759056289	n.s.
30			One-tailed:	0.035316879528144	*

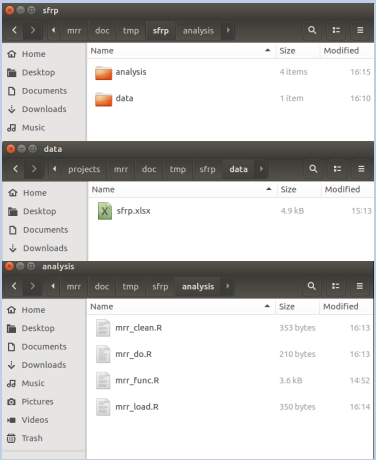
Was this done here?

Was this done here?

Separating data from analysis



Separating data from analysis



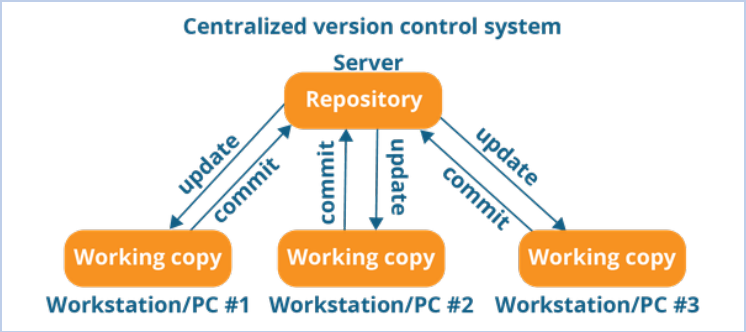
- We want one and only one data set to work with
- Once finalized (cleaned etc.), it is never touched again
- Any analysis reads from but never writes to this data set

What can we do about it?

Three simple rules:

1. Separate data from analysis
2. **Use version control**
3. Use code to analyze data (not GUIs)

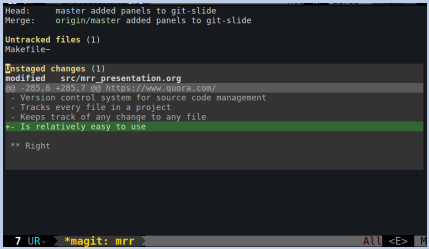
Use a version control system (= use git)



https://www.quora.com/

What is git and why should I use it?

- Version control system for source code management
- Tracks every file in a project
- Keeps track of any change to any file
- Is relatively easy to use
- Downside: it works best with text



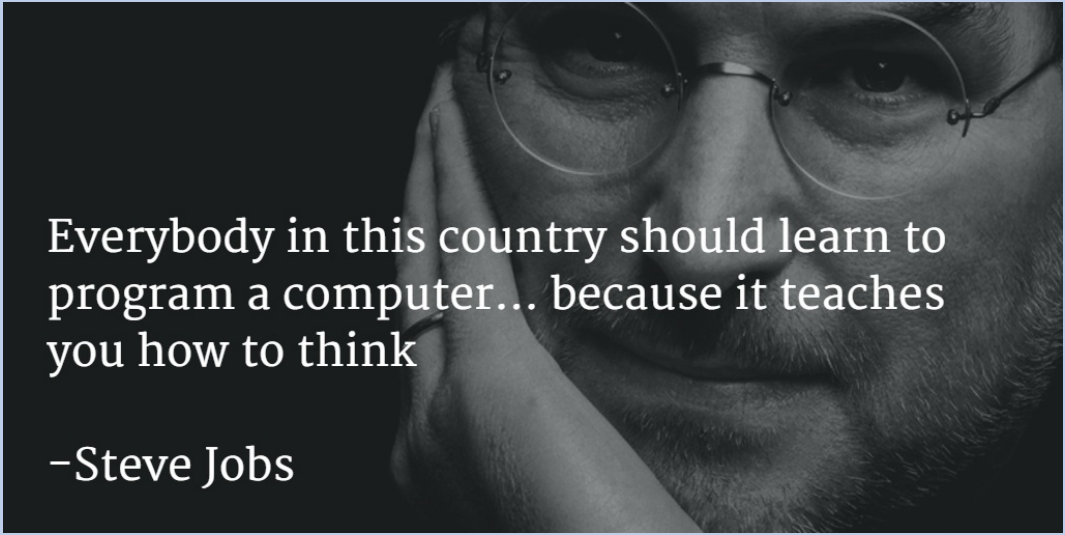
Example: git

What can we do about it?

Three simple rules:

1. Separate data from analysis
2. Use version control
3. **Use code to analyze data (not GUIs)**

Why code?


Everybody in this country should learn to program a computer… because it teaches you how to think
–Steve Jobs

Why code?

- **To keep track of the workflow**
- To make the analysis transparent
- To improve your skills and get more efficient as you code

```
parse_msd <- function(m, sd) {
  #
  # this function will
  # produce a nicely formatted string of
  # mean and sd to be used inline in text
  #
  print(paste("M = ", round(m, 2),
              ", SD = ", round(sd, 2),
              sep=""))
}
```

Without code your analysis won't be reproducible

Options:

- R or RStudio (it's free!), ideally also Python (it's free!)
- Alternatively, Matlab (great, but commercial)
- SAS (has been the market leader in commercial analytics, and it does include a free University Edition now)

Without code your analysis won't be reproducible

Options:

- **R or R studio (it's free!), ideally also Python (it's free!)**
- Alternatively, Matlab (great, but commercial)
- SAS (has been the market leader in commercial analytics, and it does include a free University Edition now)

Example: R

```r
#-----------------------------------------------------------------
# This is a simple R program
# 9/18/18, PH
#-----------------------------------------------------------------
#
# 1. Load and visualize data
#-----------------------------------------------------------------
dat <- read.csv("../data/mrr.csv")

# Histogramms
hist(dat$y[dat$group=="X"], col="blue")
hist(dat$y[dat$group=="Y"], col="blue")

# 2. Compute linear model, adjusted for  age
#-----------------------------------------------------------------
lmfit <- lm(y ~ group + age, data=dat)

# 3. Visualize residuals to check model assumptions
#-----------------------------------------------------------------
plot(density(resid(lmfit)))

# 4. Print coefficients
#-----------------------------------------------------------------
summary(lmfit)
```

Coding: the good news

- It is easier than you think
- Once one language is learned, it's easy to learn another one

Summary: How to make research reproducible

Essential:

1. Separate data and analysis
2. Use git to keep track of changes
3. Use R to keep track of your workflow

Optional:

1. Combine coding and writing to produce manuscripts
2. Use Make to build your project

Summary: How to make research reproducible

Essential:

1. Separate data and analysis
2. Use git to keep track of changes
3. Use R to keep track of your workflow

Optional:

1. **Combine coding and writing to produce manuscripts**
2. Use Make to build your project

Combining coding and writing

Several Options:

- knitr (RStudio)
- **org-mode**
- sweave

Example: org-mode

Summary: How to make research reproducible

Essential:

1. Separate data and analysis
2. Use git to keep track of changes
3. Use R to keep track of your workflow

Optional:

1. Combine coding and writing to produce manuscripts
2. **Use Make to build your project**

Example: Makefile

Conclusion

- We need transparent and reproducible workflows
- Efficient way to improve analyses and writing
- Sharing data, code, workflows may become a requirement

Acknowledgments

- Joe Zellner (jonzellner.net/docker/reproducibility/)
- Andrew Gelman (andrewgelman.com)
- Papaja package in R (crsh.github.io/papaja$_{man}$/)