Projects, Packrat, & Tidyverse

New ways to do reproducible research in R

Alicia Hofelich Mohr, Ph.D.
University of Minnesota

Open access, freely available online

Essay

Why Most Published Research Findings Are False

John P. A. Ioannidis

Summary

COMMENT | OPEN ACCESS

Gene name errors are widespread in the scientific literature

Mark Ziemann, Yotam Eren and Assam El-Osta

Genome Biology 2016 17:177 | DOI: 10.1186/s13059-016-1044-7 | © The Author(s). 2016

Published: 23 August 2016

RESEARCH ARTICLE SUMMARY

PSYCHOLOGY

Estimating the reproducibility of psychological science

Open Science Collaboration*

INTRODUCTION: Reproducibility is a defin- | viously observed finding and is the means of

POLICY & ETHICS

Is There a Reproducibility Crisis in Science?

By Nature Video on May 28, 2016

Is Economics Research Replicable? Sixty Published Papers from Thirteen Journals Say 'Usually Not'

FEDS Working Paper No. 2015-083

http://dx.doi.org/10.17016/FEDS.2015.083

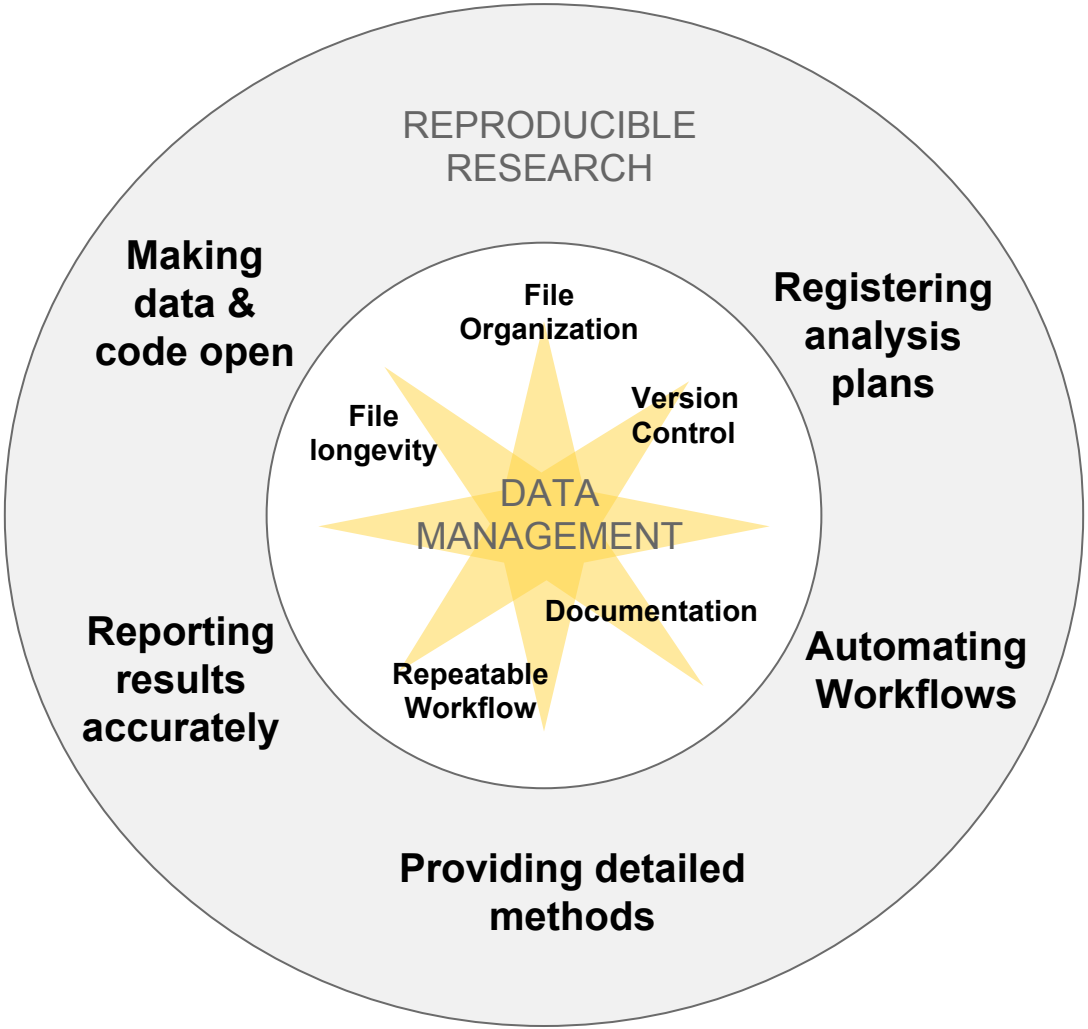
26 Pages • Posted: 6 Oct 2015

Andrew C. Chang

Board of Governors of the Federal Reserve System

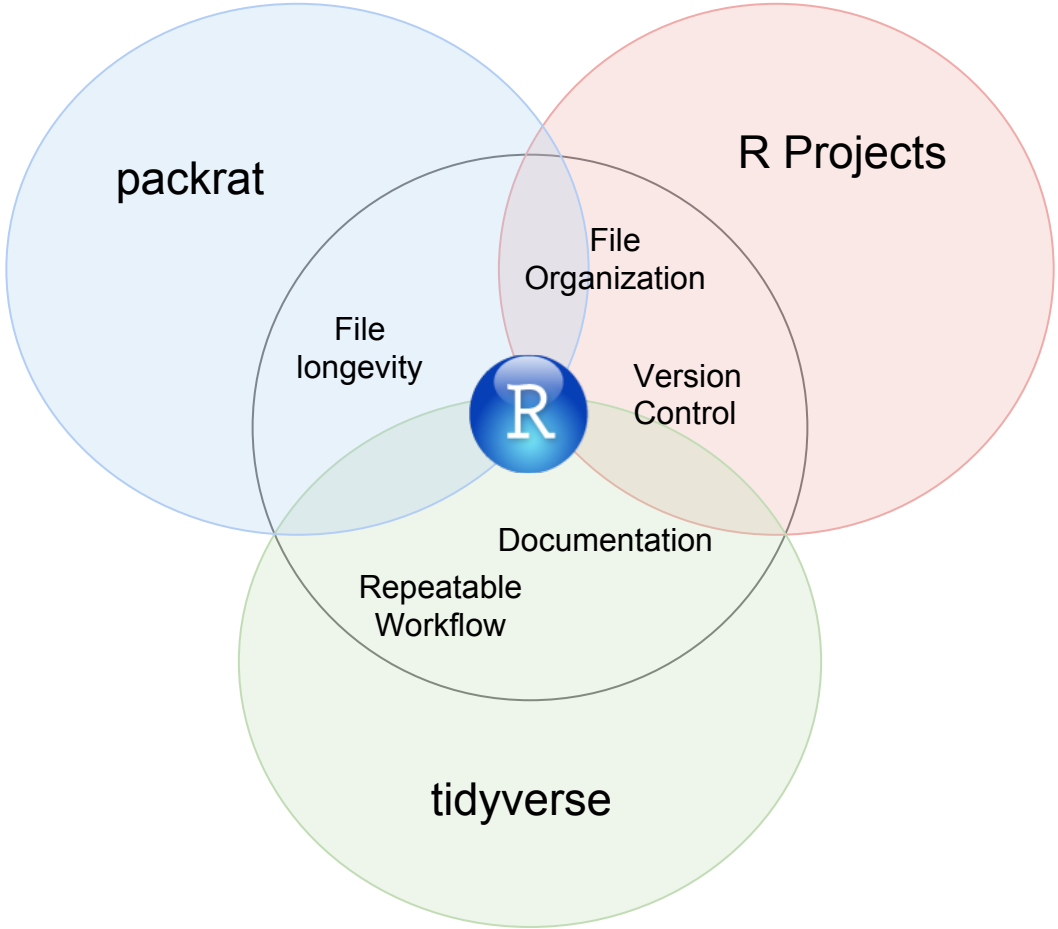
Phillip Li

Office of the Comptroller of the Currency

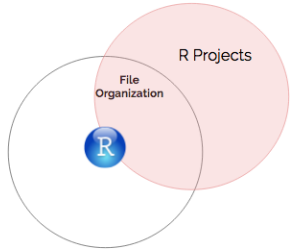

REPRODUCIBLE RESEARCH
Making data & code open
Registering analysis plans
File Organization
Version Control
File longevity
DATA MANAGEMENT
Documentation
Repeatable Workflow
Reporting results accurately
Automating Workflows
Providing detailed methods

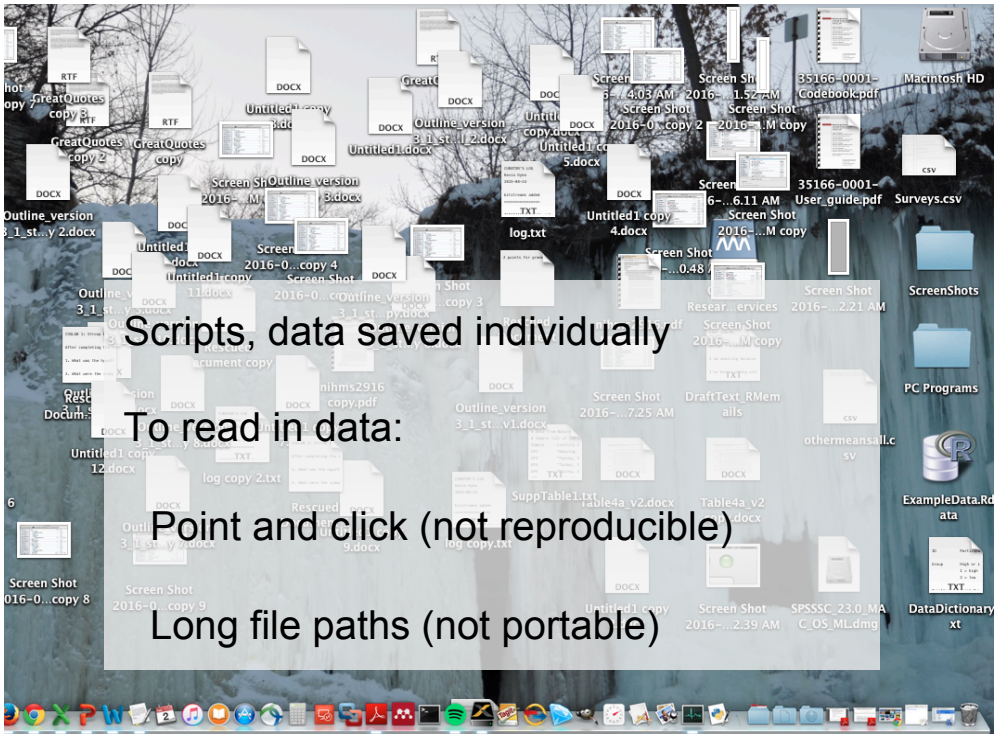
Data management isn't *really* inherently rewarding

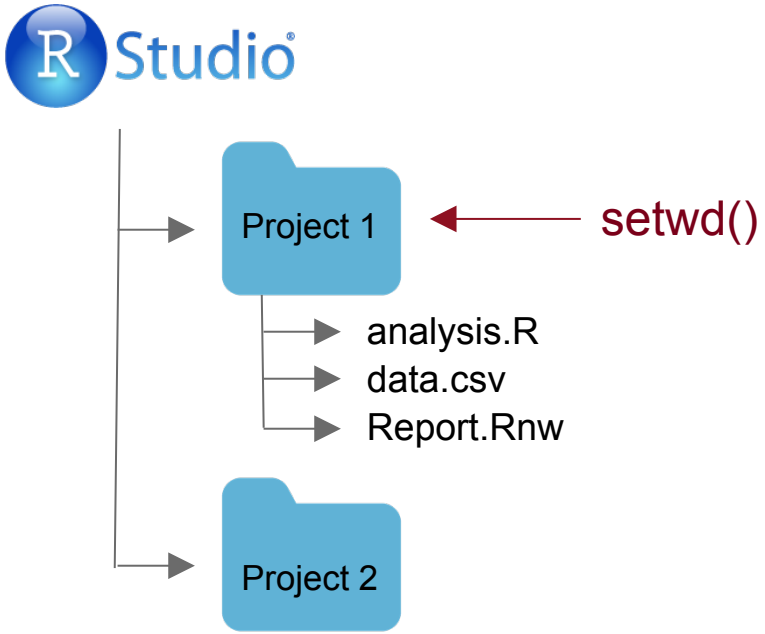
Researchers don't like having to learn new tools


packrat
R Projects
tidyverse
File
Organization
File
longevity
Version
Control
Documentation
Repeatable
Workflow
R

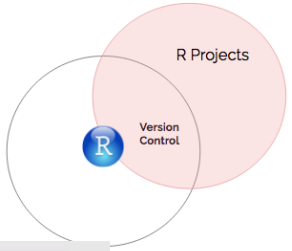
Every file on its own vs Project containers

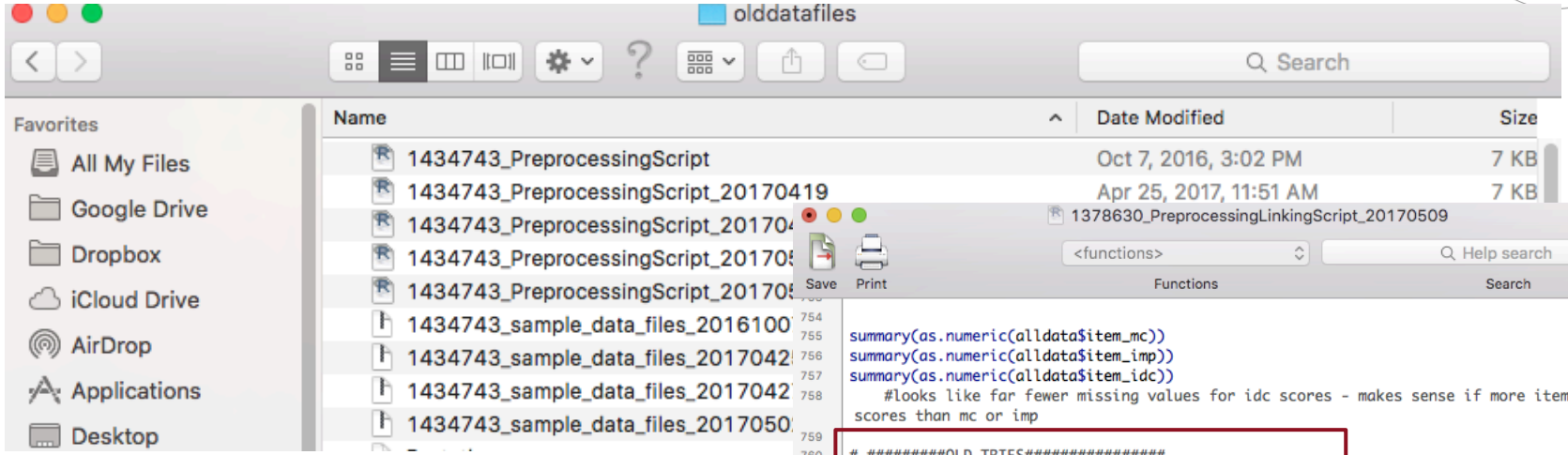


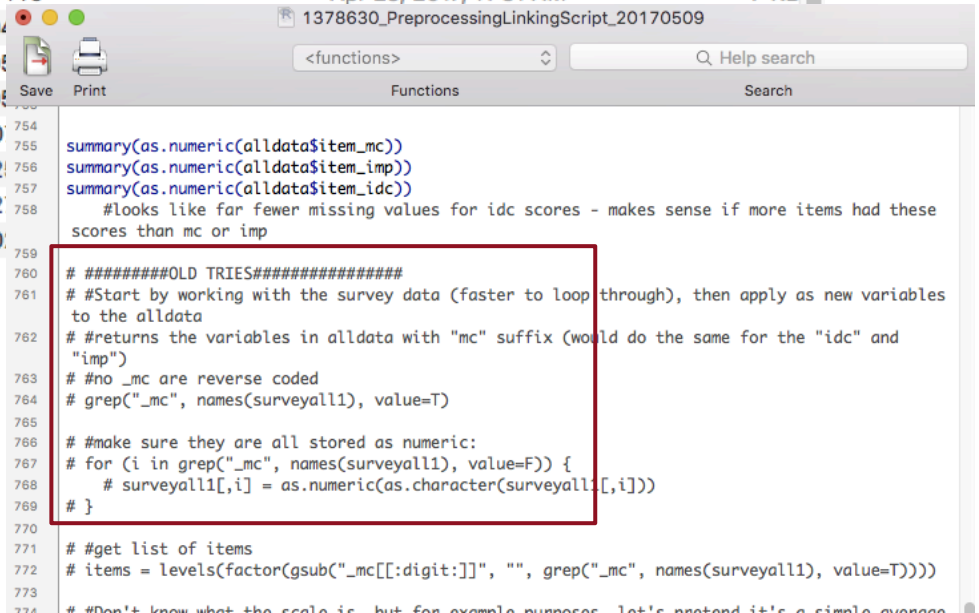




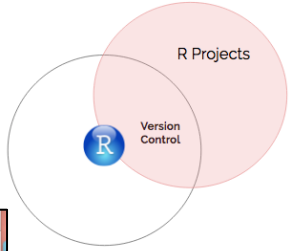
Without Projects: DIY Discipline

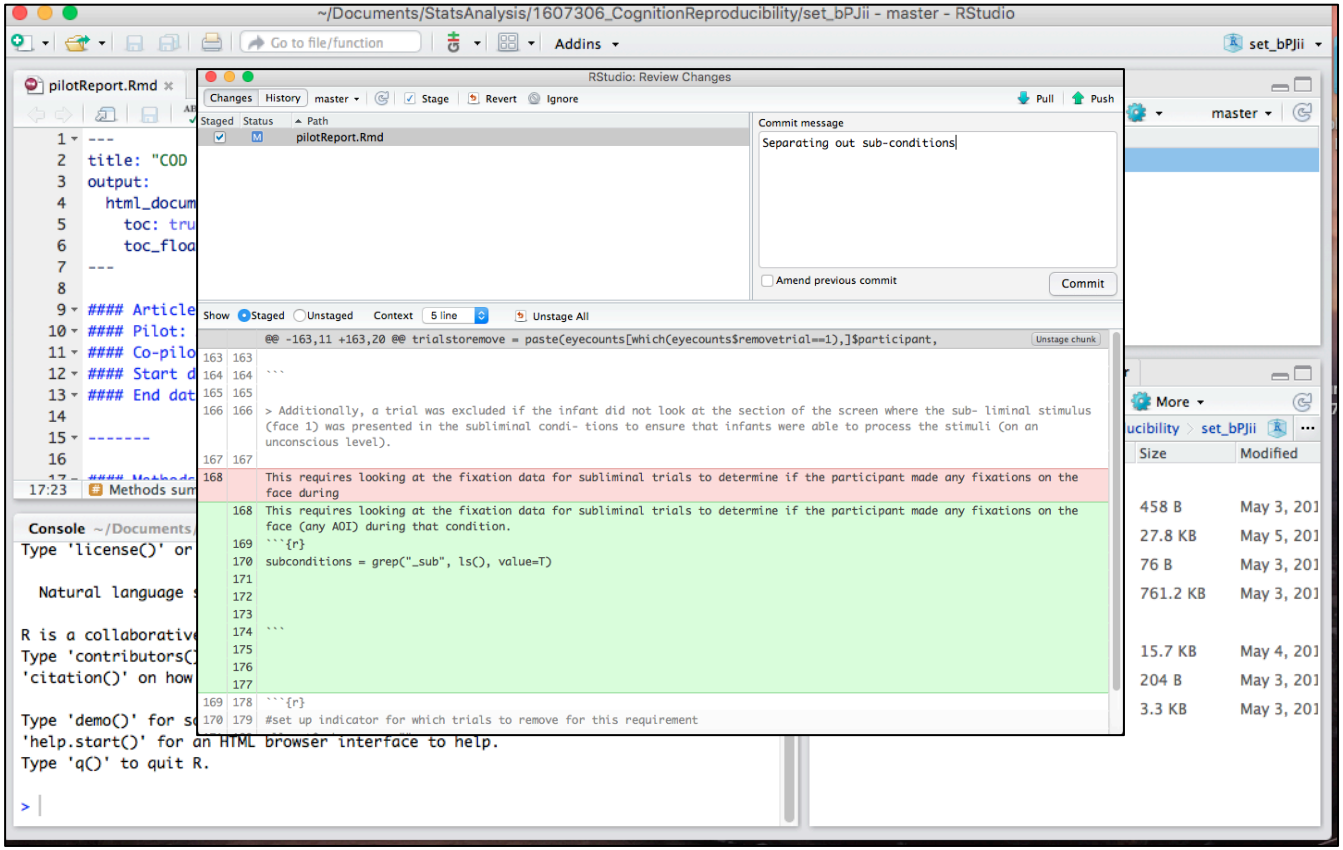




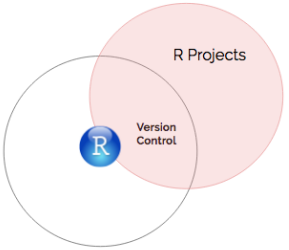


With Projects: Git integration





Without Projects: Creative in-script tracking



R_example_code_AliciaComments_20170424

Save Print <functions> Functions Help search Search

```
151  ## EMAIL APRIL 27
152  #bringing in data
153  te = read.csv("~/Documents/Spring2
      data.csv")
154
155  #gnet_2a graphing!
156  #Alicia: So in your case, you'll w
      R knows 0 is 0
157  E(gnet_2a)$edgewidth = as.numeric(
158
159  #graphing gnet_2a
160  #plot from net2a, making vertex si
      the layout where largest is cente
      color
161  #Alicia: You can do this with "del
      delete, and you can do that with
      subset the data
162  gnet_2a_edrm = delete_edges(gnet_2
163
```

Tracking changes

Initial co

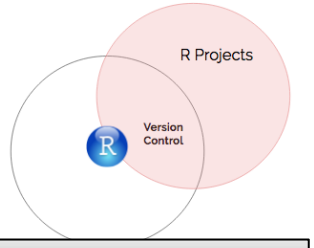
Collaborator

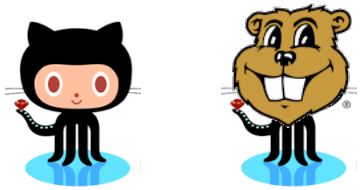
```
181  #First, find the variables that are duplicated
182  grep("\\.1", names(s1a), value=T)
183  grep("\\.1", names(s2pre1), value=T)
184  grep("\\.1", names(s2post1), value=T)
185  ###~Does "\\" here mean "whatever two characters happen to precede .1", "who
      characters happen to precede .1" or something else?
186  ###~# Here they are escape characters so that it searches literally for ".1"
      interpreting the "." as a special character meaning search for anything tha
      before a 1. Because I'm using grep, it will return anything that has ".1" i
      just ".1" alone (which would be names(s1a)==".1" to find an exact match to
      exactly named ".1").
187
188  #Then, replace ".1" with ".0"
189  names(s2pre1) = gsub("\\.1", "\\.0", names(s2pre1))
190  names(s2post1) = gsub("\\.1", "\\.0", names(s2post1))
191
```

Even weirder strategies for responses to responses

Changes aren't easily identifiable, paths break for individual files

With Projects: Easy Github integration

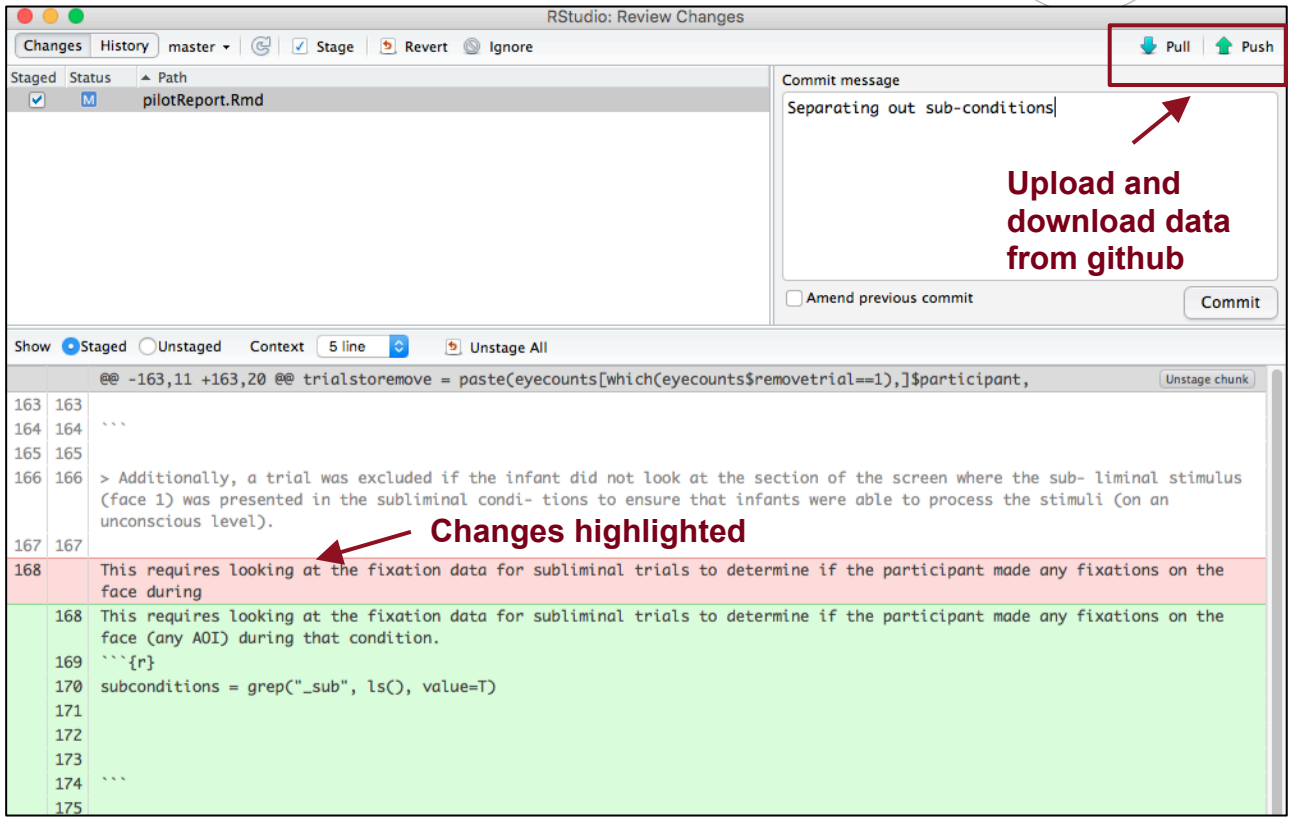




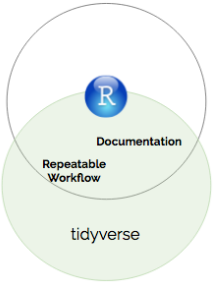
github
SOCIAL CODING

Relative paths to projects

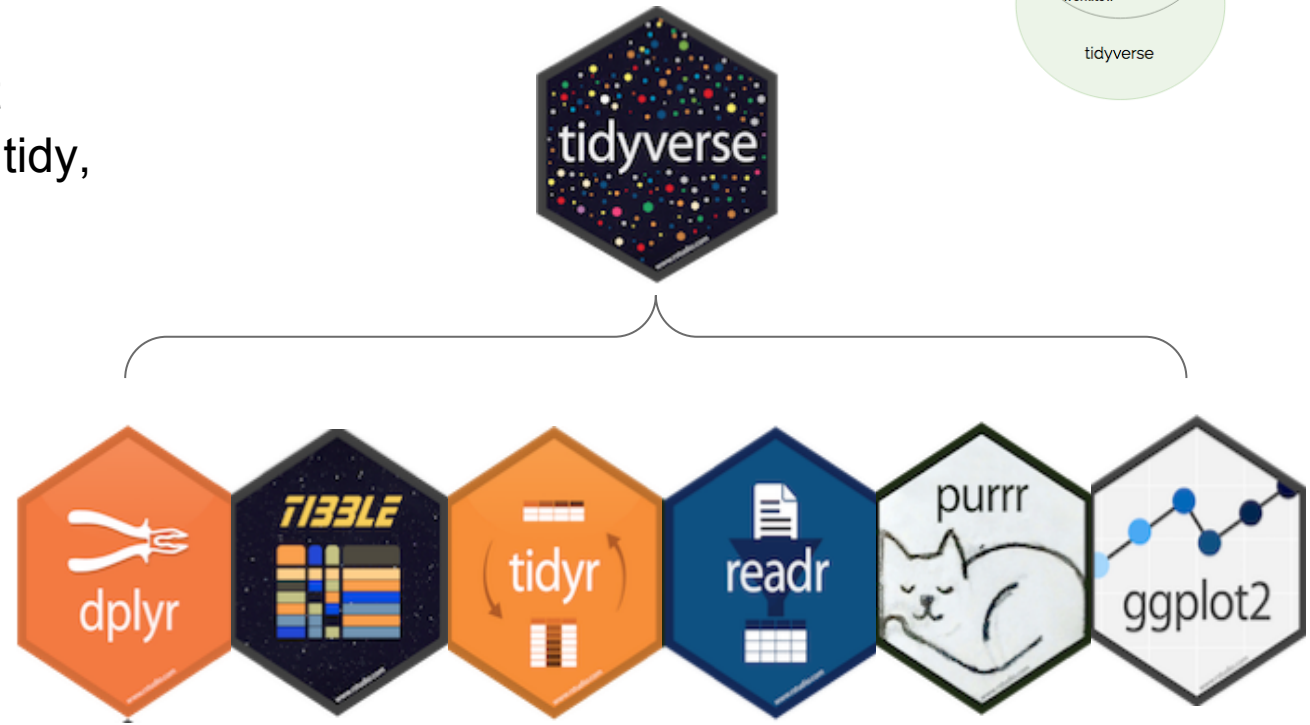
Share in private or public cloud repository



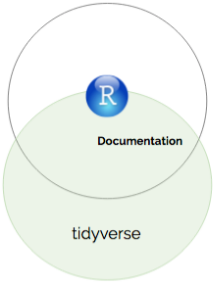
Tidyverse



Umbrella package that contains core tools for tidy, documented, and reproducible analysis



Increasing "self-documentation" of code

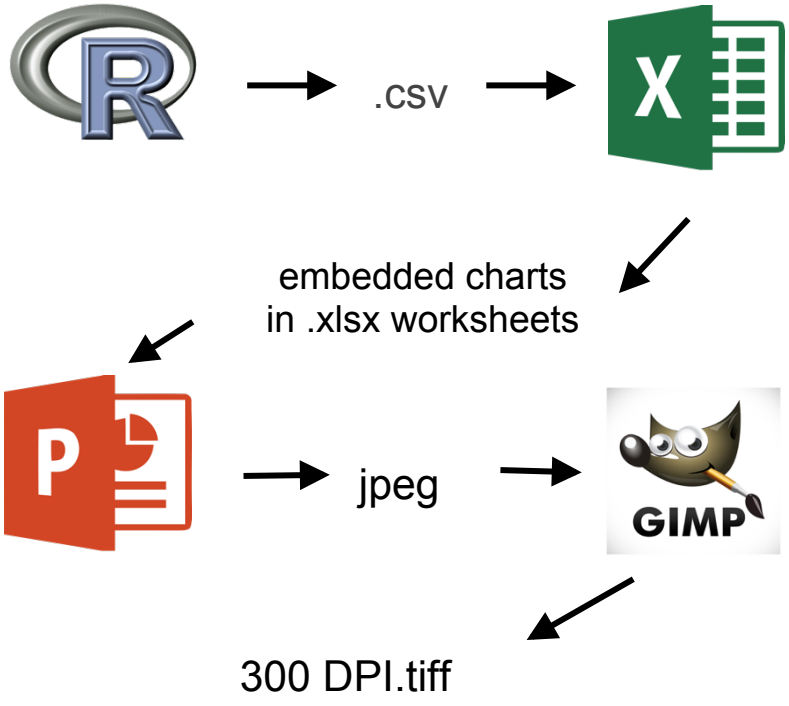


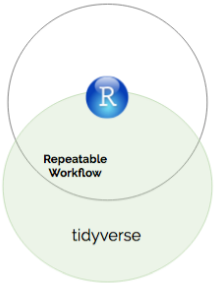
Packages such as tidyr, dplyr, and purr help make code more human readable.

```r
#Base R
data <-
 aggregate(subset(as.data.frame(Titanic),
 as.data.frame(Titanic)$Sex=="Female")
 $Freq,
 list(subset(as.data.frame(Titanic),
 as.data.frame(Titanic)$Sex=="Female")
 $Class, subset(as.data.frame(Titanic),
 as.data.frame(Titanic)$Sex=="Female")
 $Survived), sum)

reshape(data, timevar="Group.2",
 idvar="Group.1", direction="wide")
```

```r
#Dplyr
library(tidyverse)
data <- as.data.frame(Titanic) %>%
            filter(Sex=="Female") %>%
            group_by(Survived, Class) %>%
            summarize(count = sum(Freq))

data %>%
    spread(key=Survived, value=count)
```

Better visualization workflows





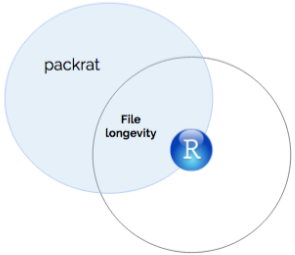
"Grammar of Graphics" plotting

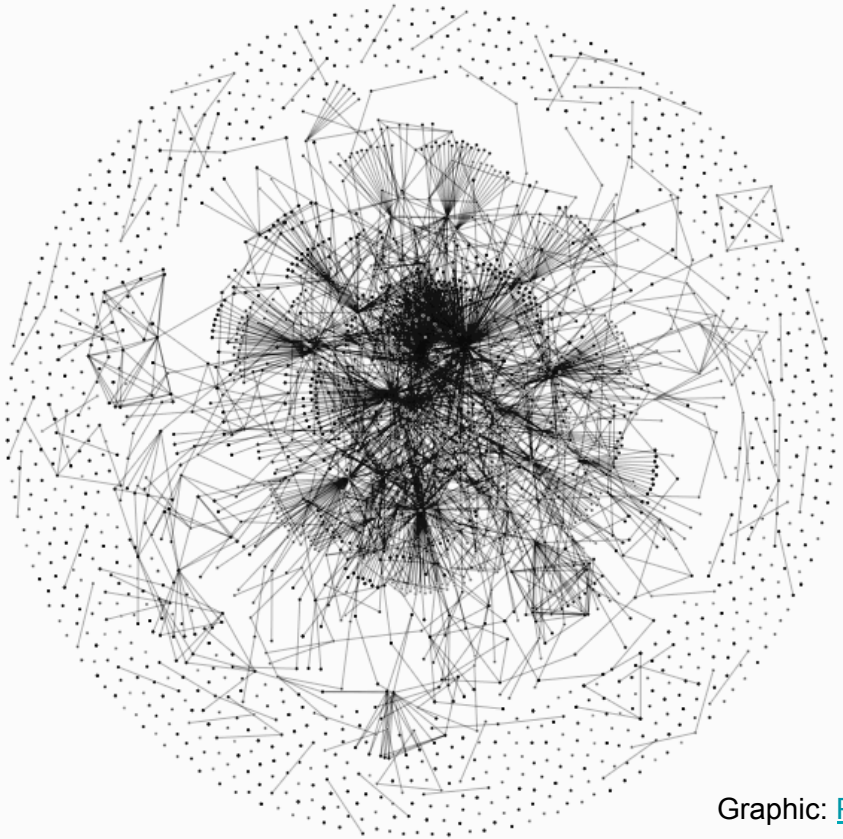
Highly flexible

Easy to customize layouts & aesthetics

Export options

Packrat - dependency management





31 versions of ggplot2 in last 10 yrs

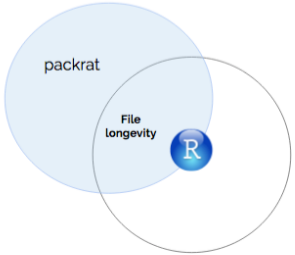
13 versions of dplyr in last 3 yrs

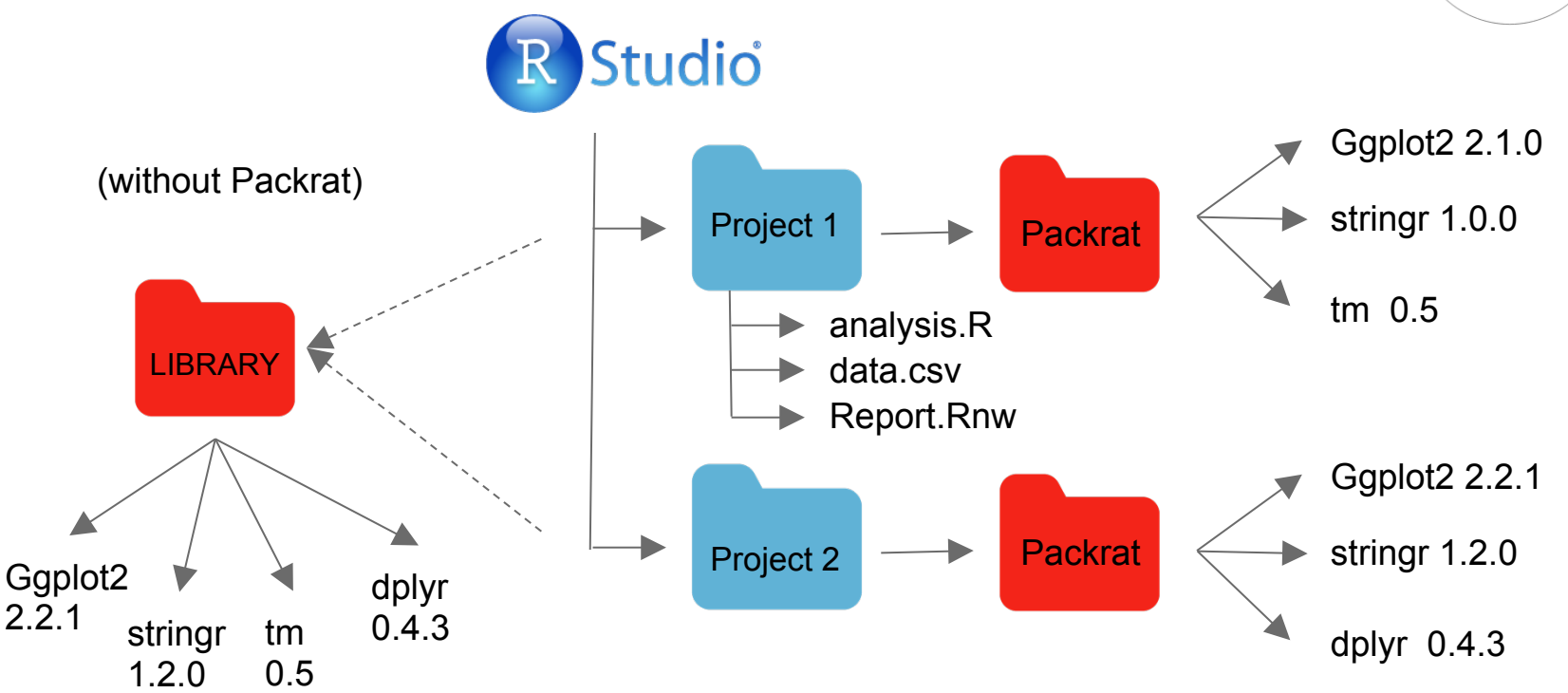
12 versions of tidyr in last 3 yrs

3 versions of tidyverse in last yr

Graphic: Package dependencies; tlvincent, Rbloggers; Data: CRAN archive

Single library vs project library



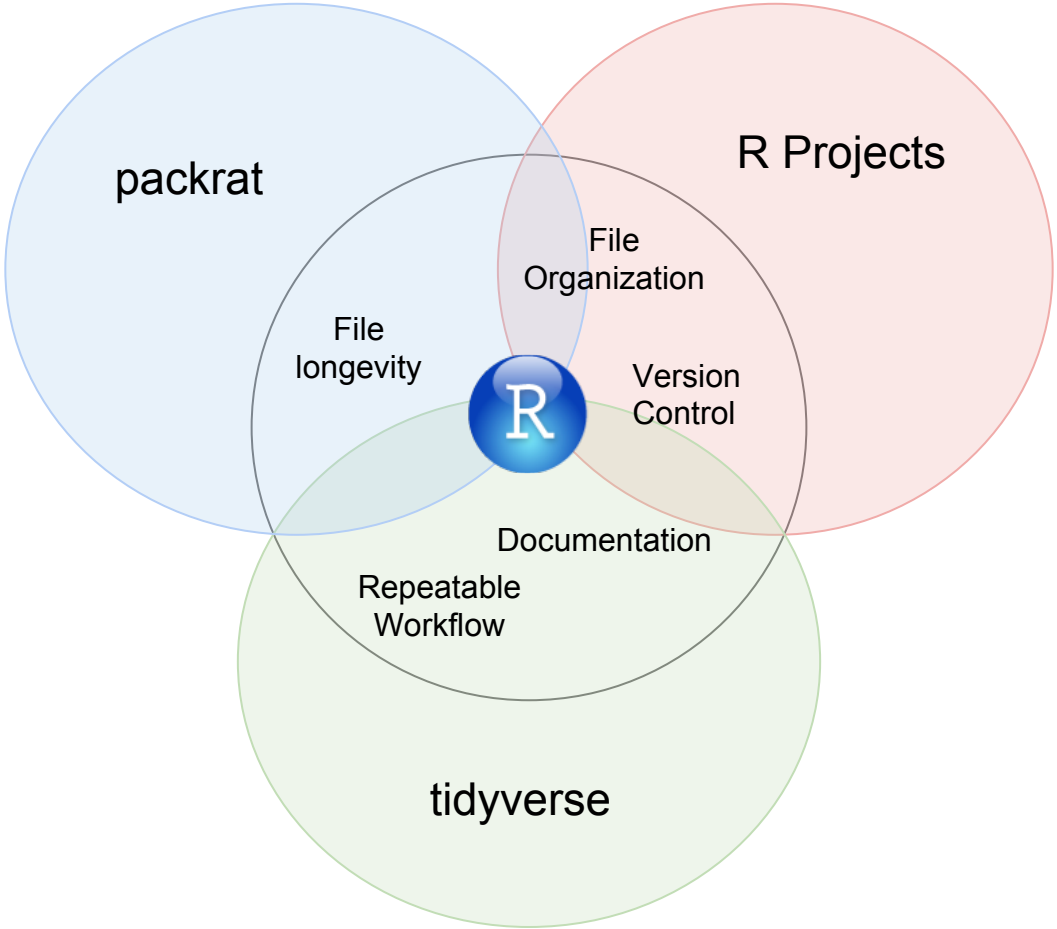


Summary

Reproducibility depends on good data management

Integrating DM actions into the analysis workflow makes it less of a pain

R is pretty awesome



Thank you! Questions?

Alicia Hofelich Mohr

hofelich@umn.edu

http://animaliertorino.com/gatti/