

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

Factors that influence this problem and some correlates thereof.

is characteristic of the
vary a lot depending
field targets highly

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>

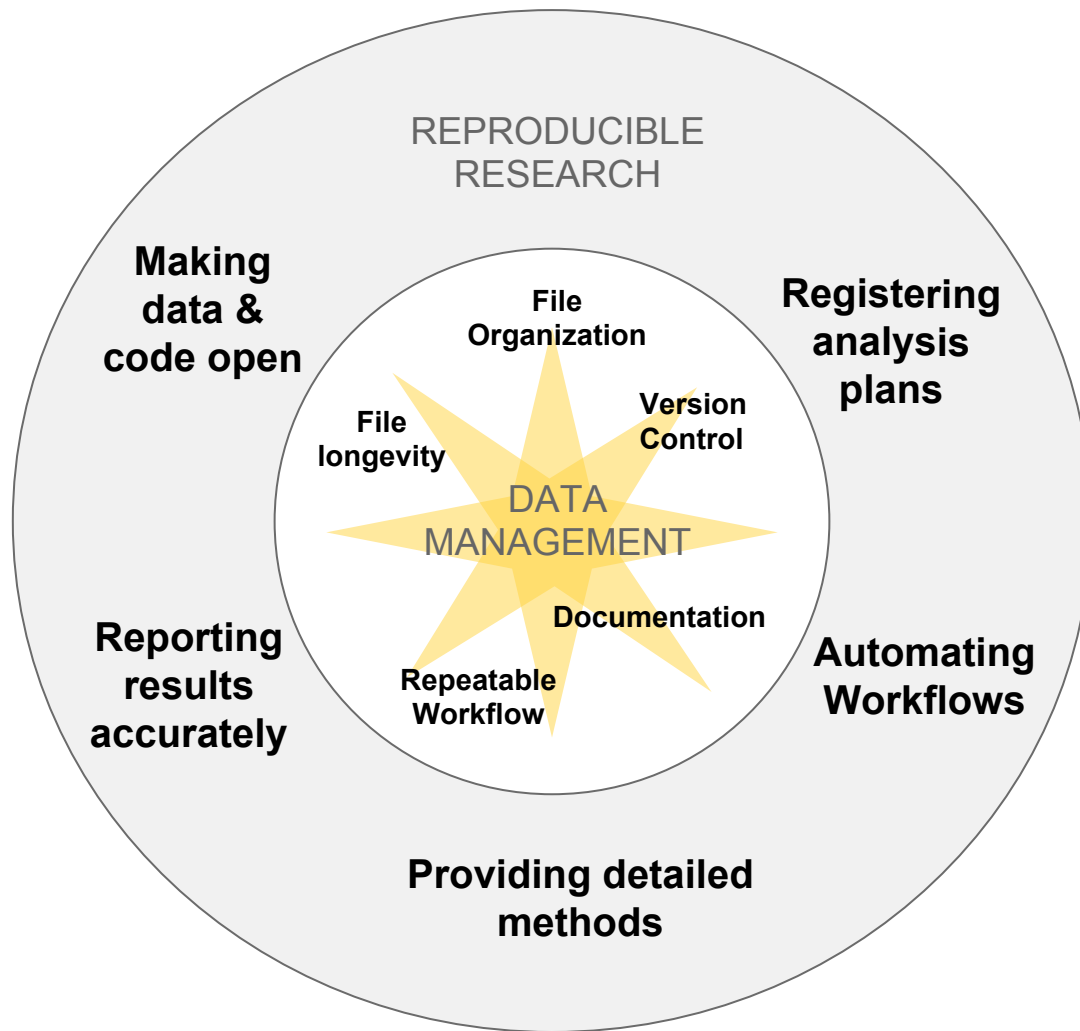
26 Pages • Posted: 6 Oct 2015

[Andrew C. Chang](#)

Board of Governors of the Federal Reserve System

[Phillip Li](#)

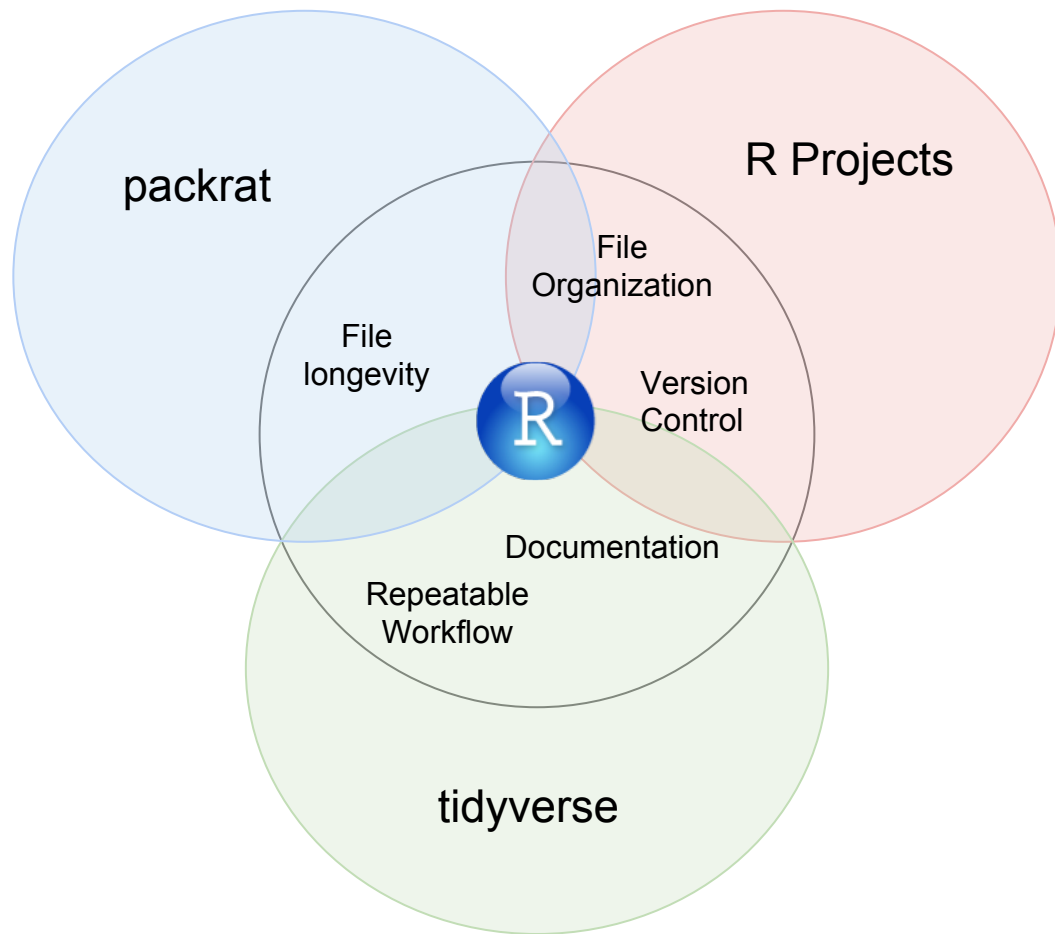
Office of the Comptroller of the Currency



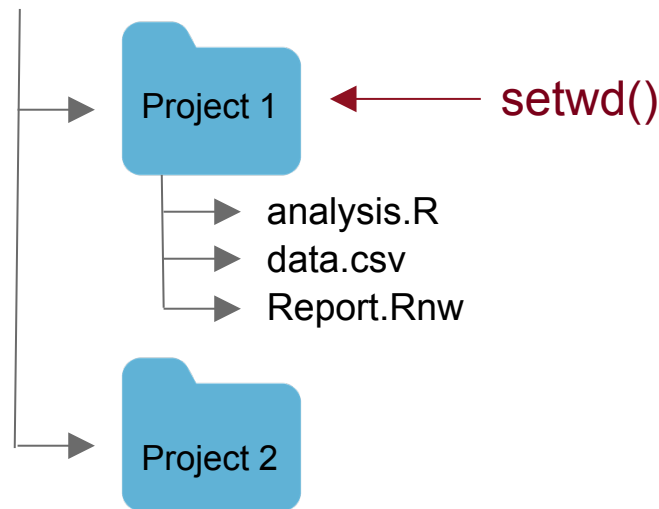
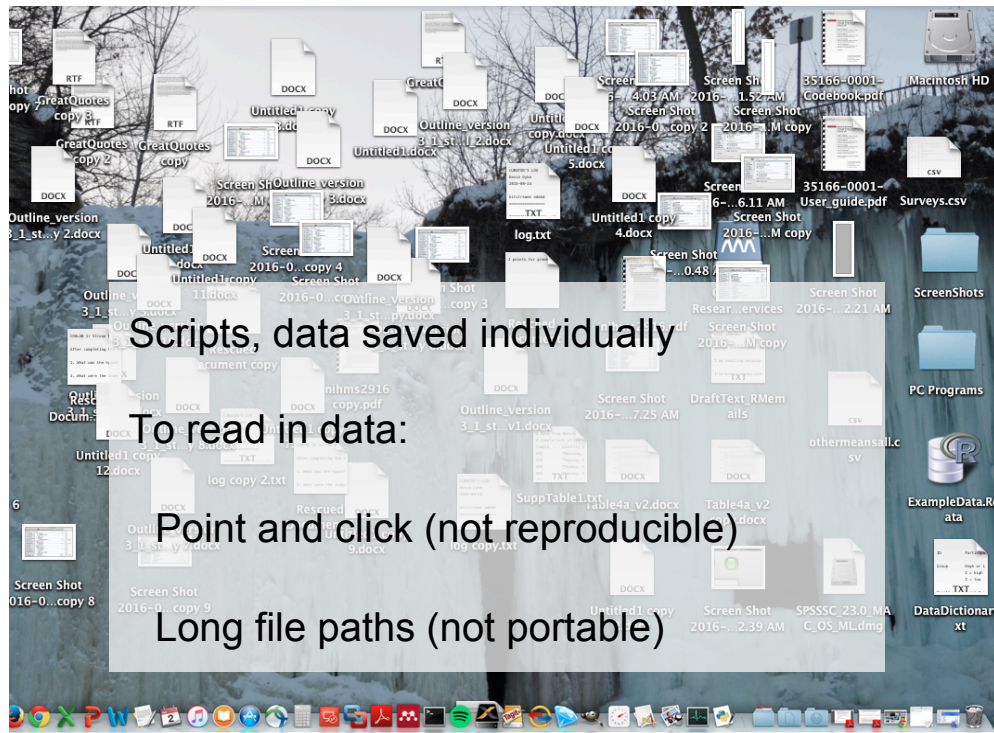
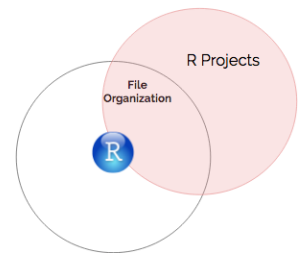
Data management isn't *really* inherently rewarding

Researchers don't like having to
learn new tools

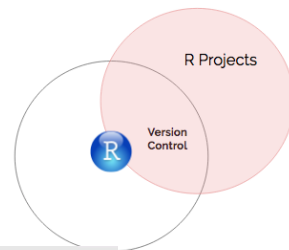




Every file on its own vs Project containers



Without Projects: DIY Discipline



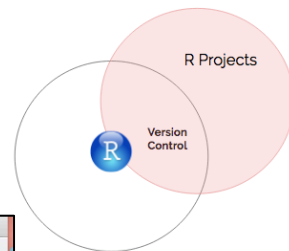
The screenshot shows a macOS file explorer window titled 'olddatafiles'. The left sidebar lists 'Favorites' including 'All My Files', 'Google Drive', 'Dropbox', 'iCloud Drive', 'AirDrop', 'Applications', and 'Desktop'. The main pane displays a list of files with columns for 'Name', 'Date Modified', and 'Size'. The files listed are:

Name	Date Modified	Size
1434743_PreprocessingScript	Oct 7, 2016, 3:02 PM	7 KB
1434743_PreprocessingScript_20170419	Apr 25, 2017, 11:51 AM	7 KB
1434743_PreprocessingScript_20170419		
1434743_PreprocessingScript_20170419		
1434743_PreprocessingScript_20170419		
1434743_sample_data_files_20161008		
1434743_sample_data_files_20170421		
1434743_sample_data_files_20170421		
1434743_sample_data_files_20170509		

Overlaid on the file explorer is a code editor window titled '1378630_PreprocessingLinkingScript_20170509'. The editor shows R code with a red box highlighting a specific section:

```
# #####OLD TRIES#####  
# #Start by working with the survey data (faster to loop through), then apply as new variables  
# #to the alldata  
# #returns the variables in alldata with "mc" suffix (would do the same for the "idc" and  
# #"imp")  
# #no _mc are reverse coded  
# #grep("_mc", names(surveyall1), value=T)  
  
# #make sure they are all stored as numeric:  
# #for (i in grep("_mc", names(surveyall1), value=F)) {  
#   # surveyall1[,i] = as.numeric(as.character(surveyall1[,i]))  
# }  
  
# #get list of items  
# #items = levels(factor(gsub("_mc[[:digit:]]", "", grep("_mc", names(surveyall1), value=T))))  
  
# #Don't know what the scale is, but for example purposes, let's pretend it's a simple average
```

With Projects: Git integration



~/Documents/StatsAnalysis/1607306_CognitionReproducibility/set_bPJii - master - RStudio

Go to file/function Addins set_bPJii

pilotReport.Rmd

```
1 ---
2 title: "COD
3 output:
4   html_document
5   toc: true
6   toc_float:
7     ---
8
9   #### Article
10  #### Pilot:
11  #### Co-pilot:
12  #### Start d
13  #### End dat
14
15  -----
16
17  #### Methods sum
18
```

RStudio: Review Changes

Changes History master Pull Push

Staged Status Path

Commit message

Separating out sub-conditions

Amend previous commit Commit

Show Staged Unstaged Context 5 line Unstage All

```
@@ -163,11 +163,20 @@ trialstoremove = paste(eyecounts[which(eyecounts$removetrials==1),]$participant,
163 163
164 164
165 165
166 166
167 167
168 168
169 169
170 170
171 171
172 172
173 173
174 174
175 175
176 176
177 177
178 178
179 179
```

> Additionally, a trial was excluded if the infant did not look at the section of the screen where the sub- liminal stimulus (face 1) was presented in the subliminal condi- tions to ensure that infants were able to process the stimuli (on an unconscious level).

This requires looking at the fixation data for subliminal trials to determine if the participant made any fixations on the face during

This requires looking at the fixation data for subliminal trials to determine if the participant made any fixations on the face (any AOI) during that condition.

```
```{r}
subconditions = grep("_sub", ls(), value=T)
```
```

169 178 ```{r}

170 179 #set up indicator for which trials to remove for this requirement

Console ~/Documents

Type 'license()' or

Natural language s

R is a collaborative

Type 'contributors()

'citation()' on how

Type 'demo()' for s

'help.start()' for an HTML browser interface to help.

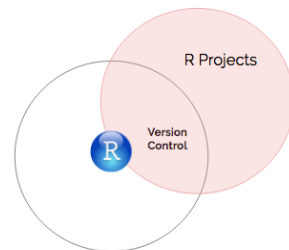
Type 'q()' to quit R.

> |

More

ucibility > set_bPJii

| Size | Modified |
|----------|------------|
| 458 B | May 3, 201 |
| 27.8 KB | May 5, 201 |
| 76 B | May 3, 201 |
| 761.2 KB | May 3, 201 |
| 15.7 KB | May 4, 201 |
| 204 B | May 3, 201 |
| 3.3 KB | May 3, 201 |



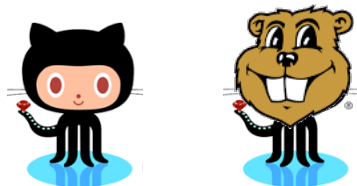
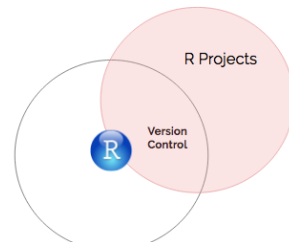
Without Projects: Creative in-script tracking

```
150
151 ## EMAIL APRIL 27
152 #bringing in data
153 te = read.csv("~/Documents/Spring
data.csv")
154
155 #gnet_2a graphing!
156 #Alicia: So in your case, you'll v
R knows 0 is 0
157 E(gnet_2a)$edgewidth = as.numeric
158
159 #graphing gnet_2a
160 #plot from net2a, making vertex s
the layout where largest is cente
color
161 #Alicia: You can do this with "de
delete, and you can do that with
subset the data
162 gnet_2a_edrm = delete_edges(gnet_2
163
```

```
180
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
```

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

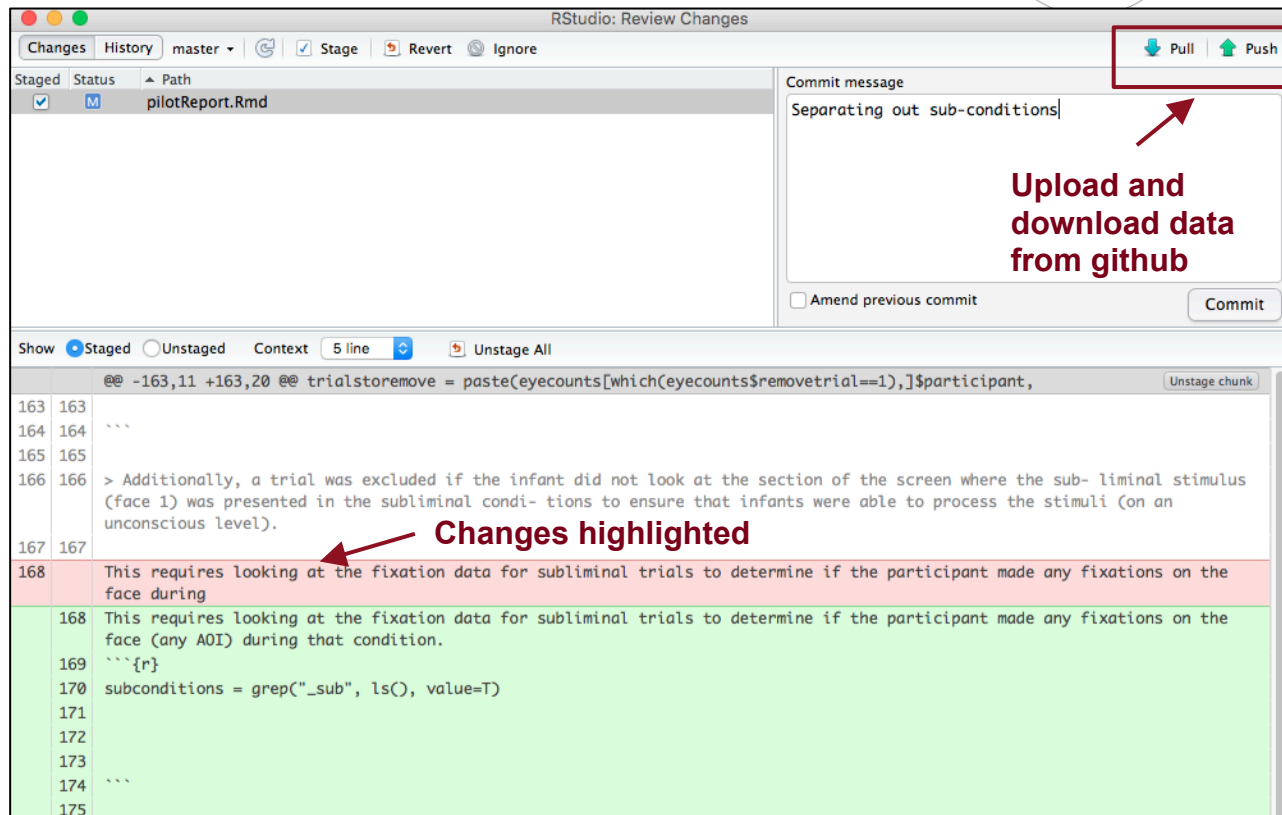
With Projects: Easy Github integration



github
SOCIAL CODING

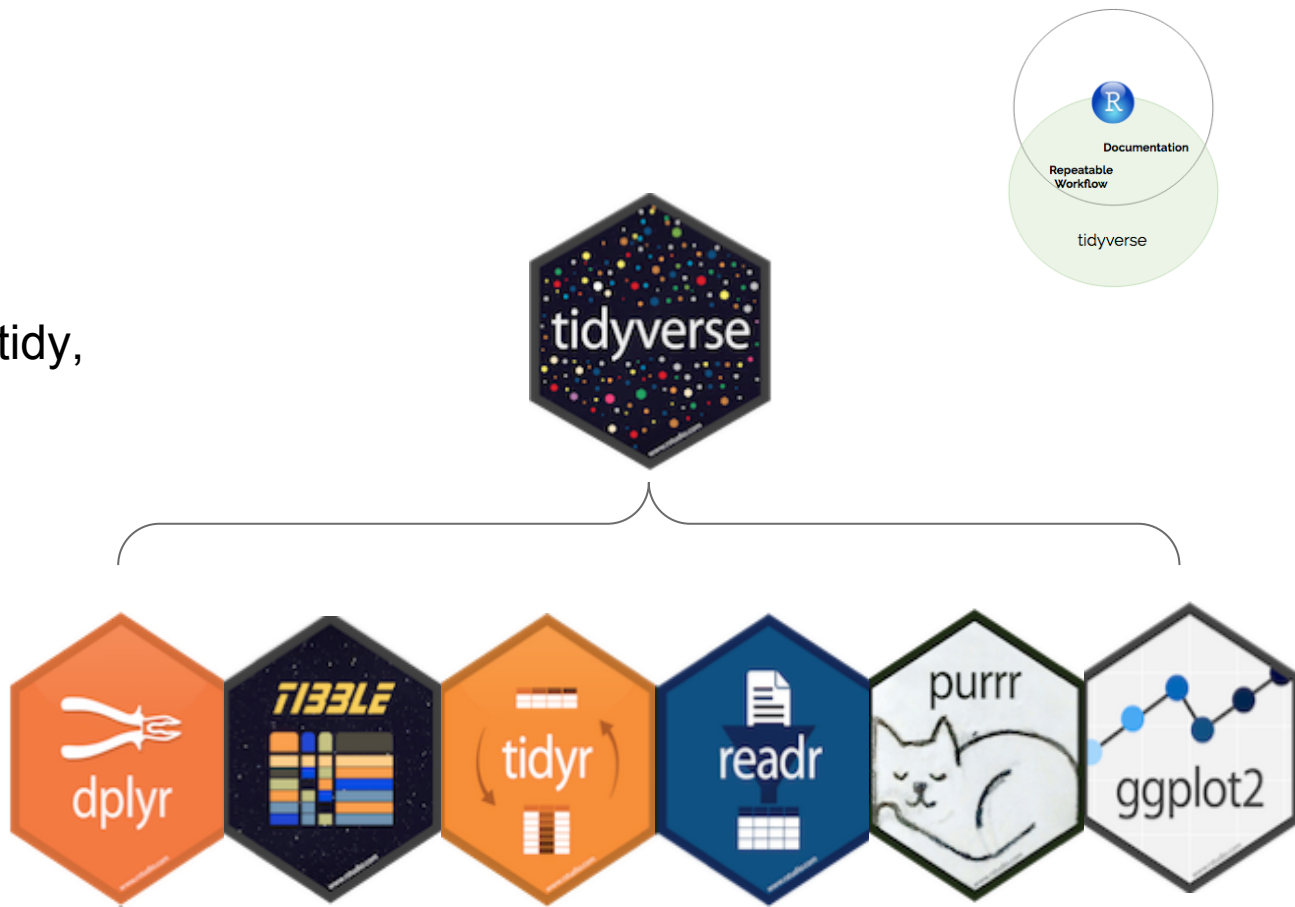
Relative paths to
projects

Share in private or
public cloud
repository

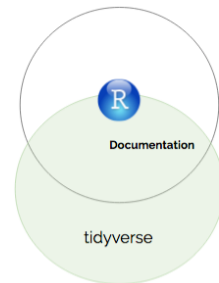


Tidyverse

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.

```
#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")
```

```
#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



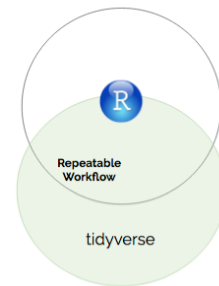
embedded charts
in .xlsx worksheets



→ jpeg →



300 DPI.tiff



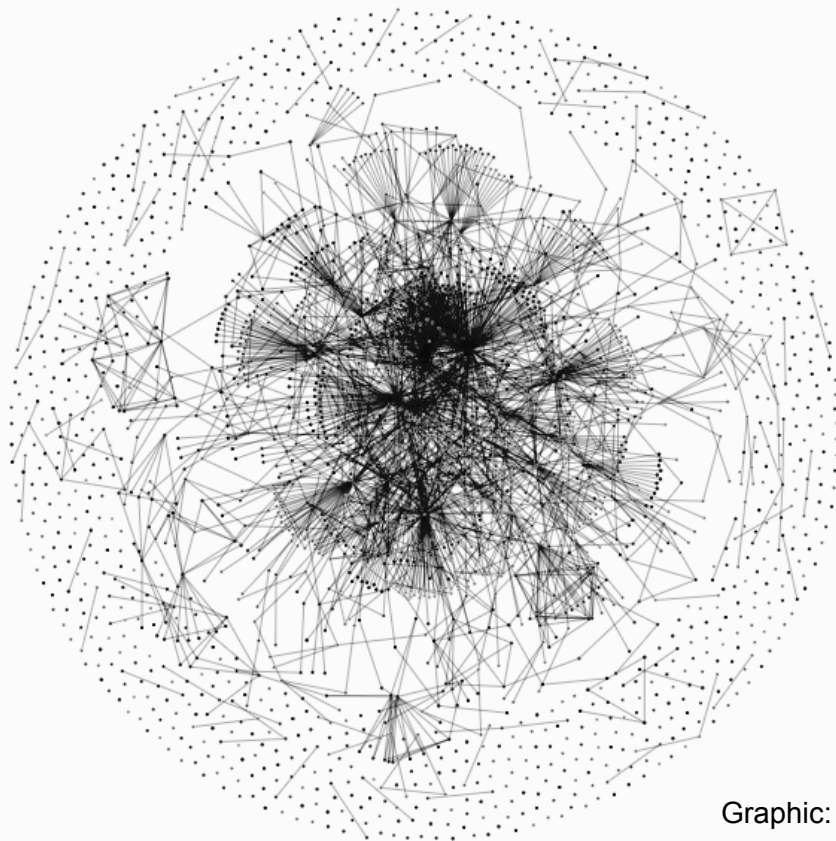
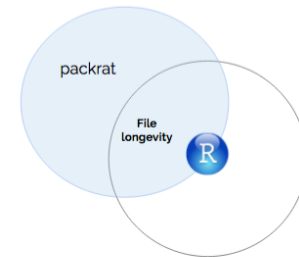
"Grammar of Graphics" plotting

Highly flexible

Easy to customize layouts &
aesthetics

Export options

Packrat - dependency management



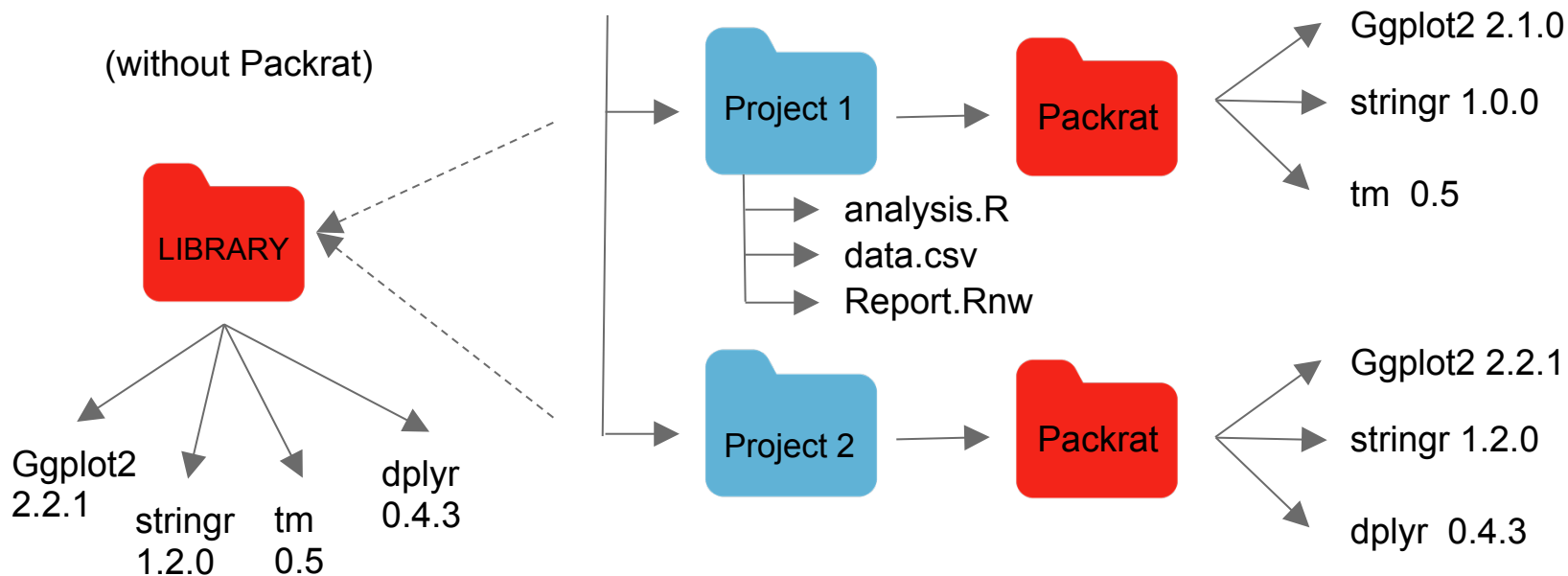
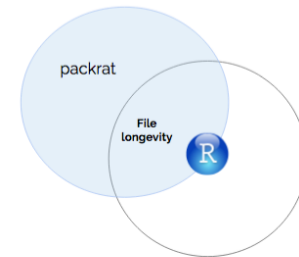
31 versions of ggplot2 in last 10 yrs

13 versions of dplyr in last 3 yrs

12 versions of tidyr in last 3 yrs

3 versions of tidyverse in last yr

Single library vs project library

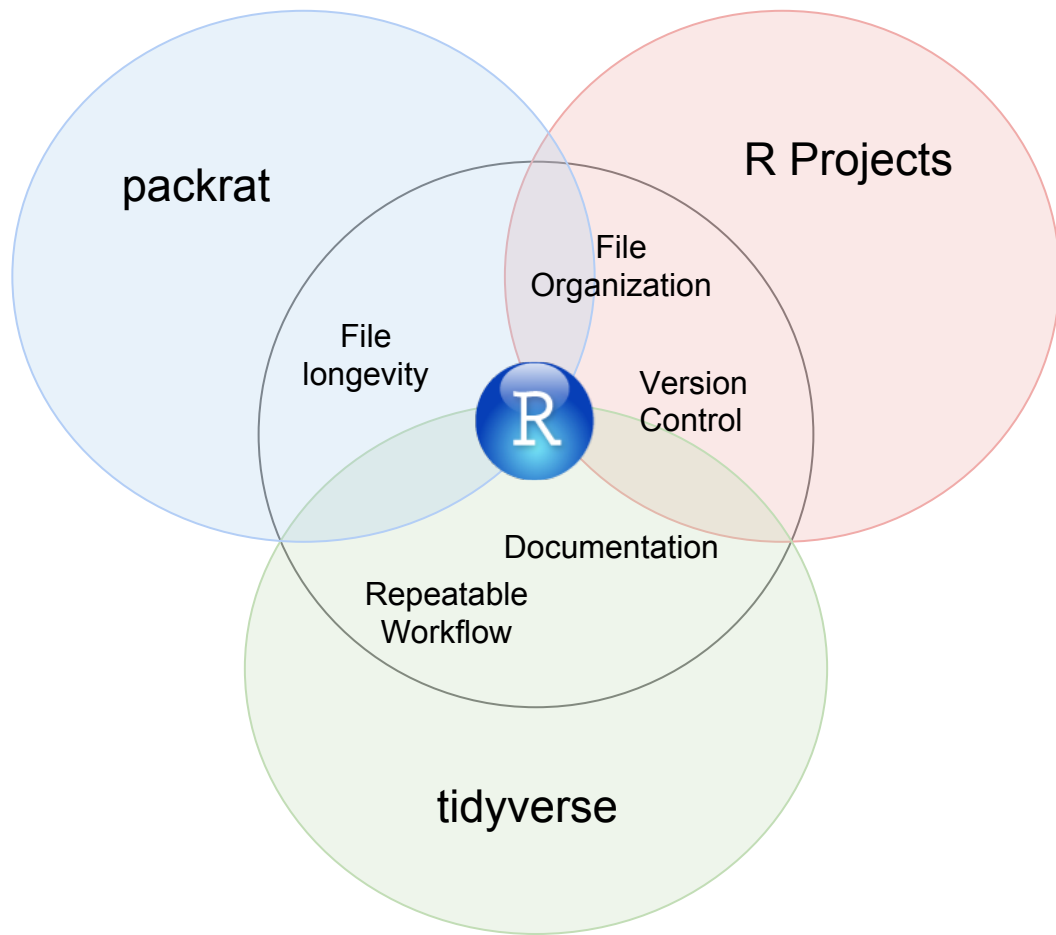


Summary

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/>