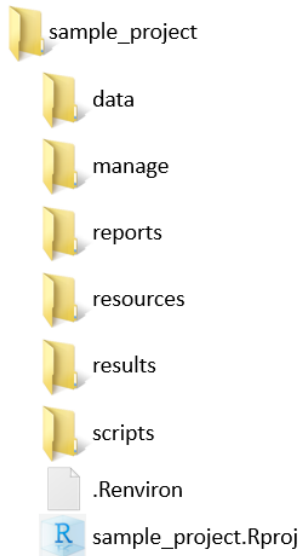


# Principles for a reproducible workflow

497 / 597 Reproducible Research

Richard Layton

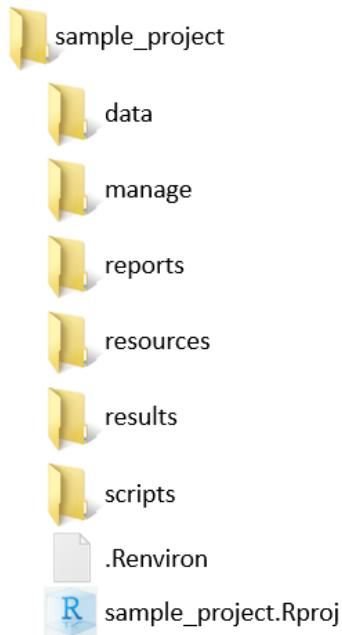
Rose-Hulman Institute of Technology  
Fall 2018




# Organize for reproducibility from the beginning

- ▶ Plan your directory structure
- ▶ Script everything — point/click/copy/paste is not reproducible
- ▶ Strive for simplicity & readability
- ▶ Link files explicitly
- ▶ Adopt a file naming scheme
- ▶ DRY (don't repeat yourself)
- ▶ Use version control

# From the beginning — plan your directory structure



# From the beginning — plan your directory structure

 sample\_project

▶ working directory (relative file paths start here)

 data


 manage


 reports

 resources


 results

 scripts

 .Renviron

 sample\_project.Rproj

# From the beginning — plan your directory structure

 sample\_project

▶ working directory (relative file paths start here)

 data

▶ unaltered raw data


 manage


 reports

 resources


 results

 scripts

 .Renviron

 sample\_project.Rproj

# From the beginning — plan your directory structure

 sample\_project

▶ working directory (relative file paths start here)

 data

▶ unaltered raw data

 manage


▶ administrative files, not version controlled


 reports

 resources


 results

 scripts

 .Renviron

 sample\_project.Rproj

# From the beginning — plan your directory structure

 sample\_project

▶ working directory (relative file paths start here)

 data


▶ unaltered raw data

 manage

▶ administrative files, not version controlled


 reports


▶ Rmd file(s) of the project report(s)

 resources


 results

 scripts

 .Renviron

 sample\_project.Rproj

# From the beginning — plan your directory structure

 sample\_project

▶ working directory (relative file paths start here)

 data

▶ unaltered raw data

 manage

▶ administrative files, not version controlled

 reports


▶ Rmd file(s) of the project report(s)


 resources

▶ images and pdfs from other sources

 results

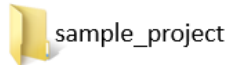
 scripts

 .Renviron

 sample\_project.Rproj



# From the beginning — plan your directory structure



sample\_project

- ▶ working directory (relative file paths start here)



data

- ▶ unaltered raw data



manage

- ▶ administrative files, not version controlled



reports

- ▶ Rmd file(s) of the project report(s)



resources

- ▶ images and pdfs from other sources



results

- ▶ save script output (tidy data and graphs) here



scripts

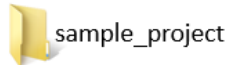


.Renvirom



sample\_project.Rproj

# From the beginning — plan your directory structure



- ▶ working directory (relative file paths start here)



- ▶ unaltered raw data



- ▶ administrative files, not version controlled



- ▶ Rmd file(s) of the project report(s)



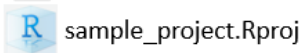
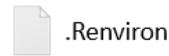
- ▶ images and pdfs from other sources



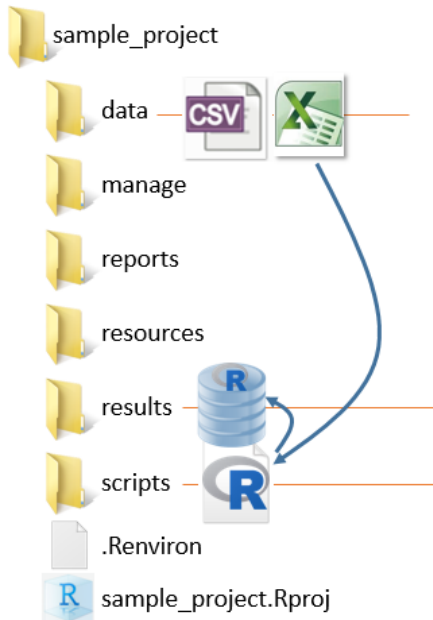
- ▶ save script output (tidy data and graphs) here



- ▶ R files to tidy data, do analysis, & create graphs



# Script everything

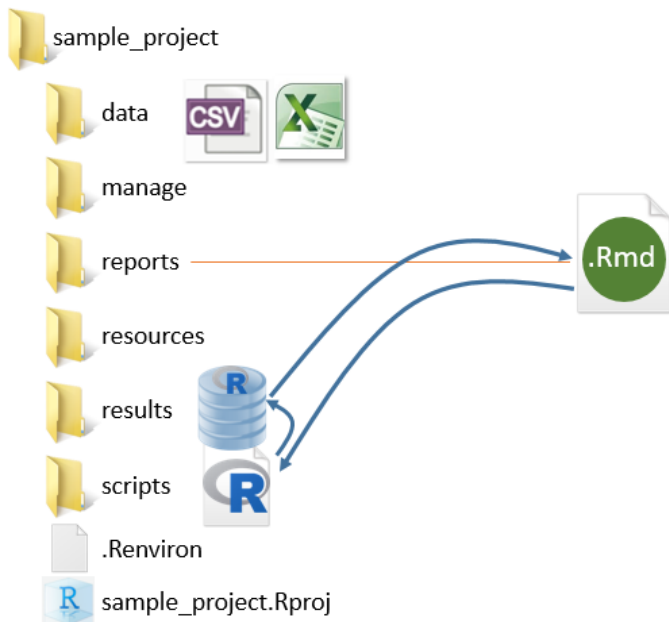


Use an R script to

- ▶ read a raw data file
- ▶ produce tidy data saved to results

Raw data files are stored unaltered.

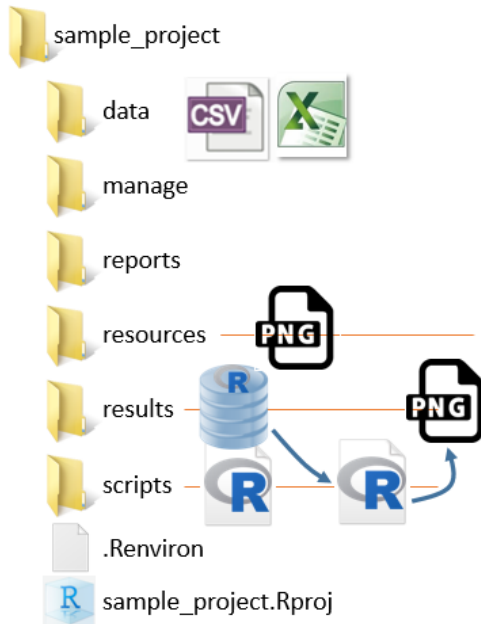
# Link files explicitly



Start the Rmd script

- ▶ write prose to explain the work
- ▶ write R code chunks to execute the scripts
- ▶ import data from results to create data tables

# Script everything



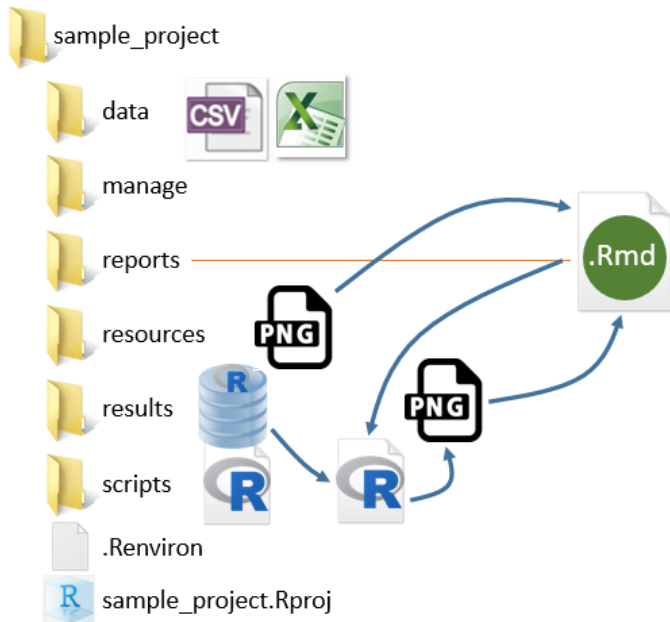
Use an R script to

- ▶ read tidy data from results
- ▶ produce a graph saved to results



Non-reproducible images stored in resources

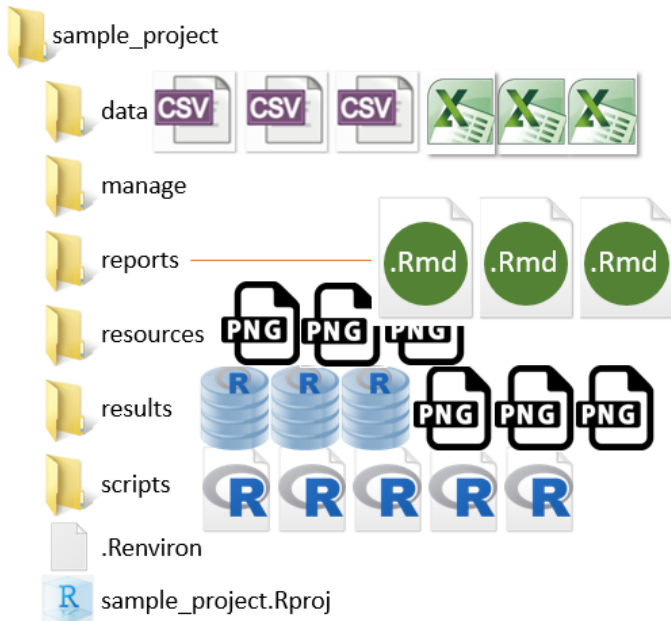
# Link files explicitly



## Continue the report

- ▶ write prose to explain the work
- ▶ write R code chunks to execute the scripts
- ▶ import images

# Strive for simplicity & readability



One Rmd script for each project milestone

- ▶ proposal
- ▶ progress report
- ▶ final report

# Strive for simplicity & readability

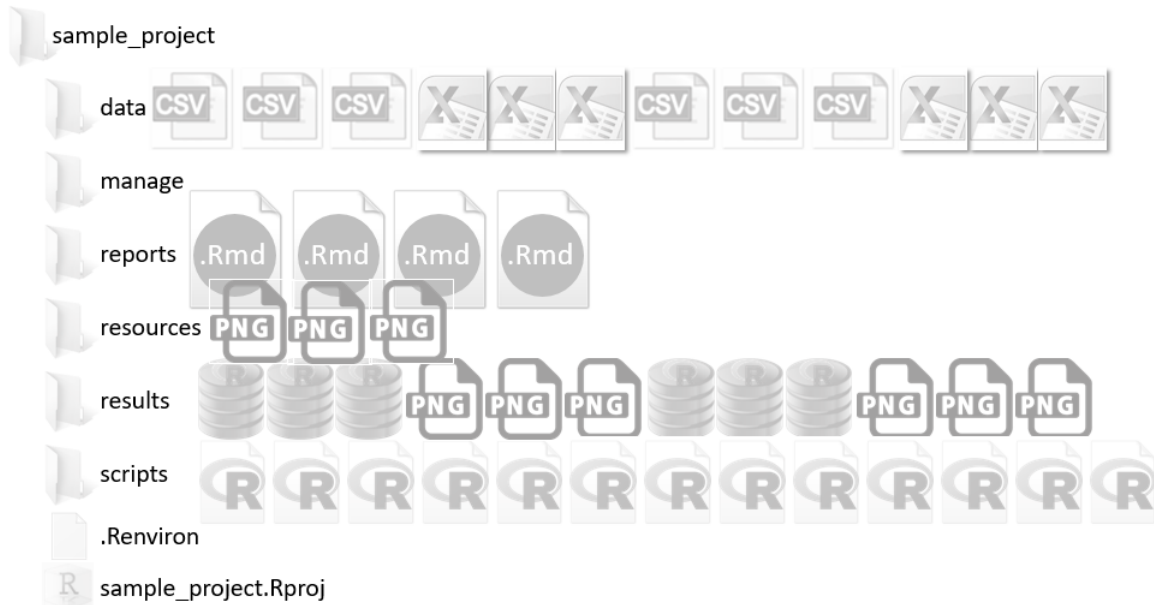
R scripts are generally short, between 60–100 lines, to

- ▶ produce one object written to file, e.g., CSV, PNG
- ▶ simplify editing, testing, & debugging
- ▶ improve readability





# From the beginning — adopt a file naming scheme



# From the beginning — adopt a file naming scheme

## *000-series* manage

001\_RFP\_2018-05-25.pdf  
002\_contract\_2018-06-05.pdf  
invoice\_201801.pdf  
invoice\_201802.pdf

## *100-series* data

101\_raw-data\_2018-07-25.xlsx  
102\_raw-data\_2018-08-01.xlsx

## *200-series* resources

201\_apparatus\_2018-08-12.png  
202\_load-cell\_2018-08-12.png

## *300-series* reports

301\_proposal\_2018-07-05.Rmd  
302\_progress\_2018-08-12.Rmd  
303\_report\_2018-09-03.Rmd

## *400-series* scripts

401\_data-wide.R  
402\_data-tidy.R  
403\_calibr-graph.R

## *400-series* also used for results

401\_data-wide.rda  
402\_data-tidy.rda  
403\_calibr-graph.png

# Use version control

See the website for instructions

**GitHub**



obtain a free account for remote, asynchronous collaboration



create an online repository for each project



link each repo to a local RStudio Project