

Reproducible analysis workflows

A short introduction into reproducible analysis tools: Rmarkdown, Github
and others



Cornelius Hennch

27.01.2022

Section 1

Introduction

Why do we need reproducible data analysis?

“Reproducibility is the ability to obtain identical results from the same statistical analysis and the same data”

= **long-term** and **cross-platform** reproducibility of data analyses

– Peikert and Brandmeier [1]




Reproducibility \neq Replicability

same analysis, **same data** / same analysis, **new data**

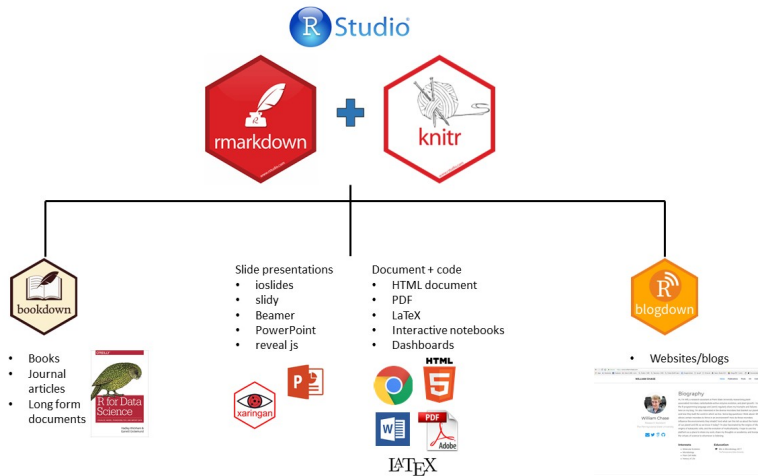
Goals of reproducible workflows

- ① **Reported** results are consistent with the **actual** results
- ② Computational reproducibility (= hardware and software change over time)
- ③ Version control (= keep track of any changes at any time)

Four essential tools for reproducible workflows

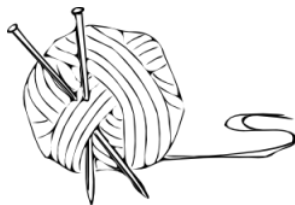
- ① Dynamic reports → **R Markdown** 
- ② Version control → **Git & Github** 
- ③ Dependency management → **Make**
- ④ Containerization → **Docker** 

Highly versatile dynamic documents with R Markdown



https://timotheenivalis.github.io/workshops/RforRSB/rmarkdown_notes.html

Happy knitting!



https://rmarkdown.rstudio.com/authoring_quick_tour.html

Git & Github



Git

- “Distributed version control system”
- Track and document changes (“commits”)
- Retrieve older versions of code
- Enables collaboration on any kind of programming projects (scalable!)

Git & Github



Git

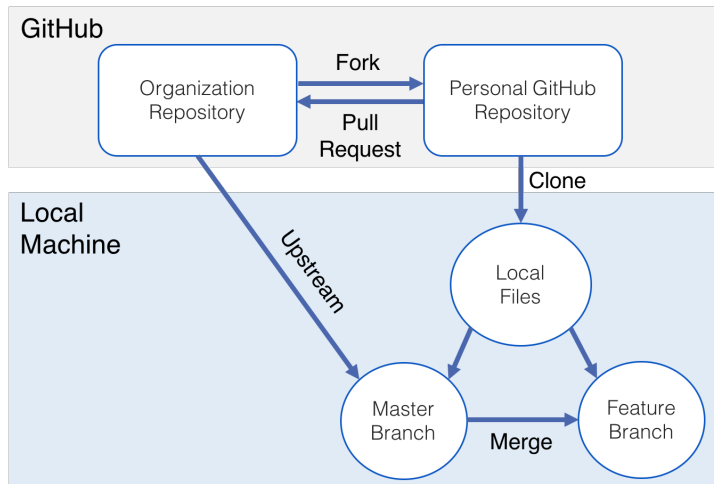
- “Distributed version control system”
- Track and document changes (“commits”)
- Retrieve older versions of code
- Enables collaboration on any kind of programming projects (scalable!)



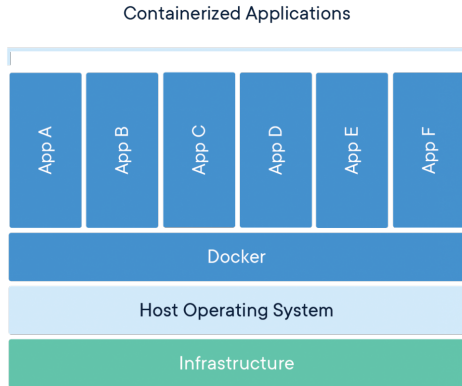
Github

- Git repository hosting service
- **Collaboration:**
 - 1 Many features for team/project management (scalable!)
 - 2 Report bugs/issues, get help
 - 3 Contribute to open-source projects
- Post-publication platform

Collaboration with Git & Github



How to Update a Fork in Git



Section 2

Reproducible data analysis in action

Example analysis: How do R skills influence time to thesis completion.

Hypothesis: Years of experience with R are inversely correlated with the estimated time to thesis completion.

Simulate data

Examine data structure

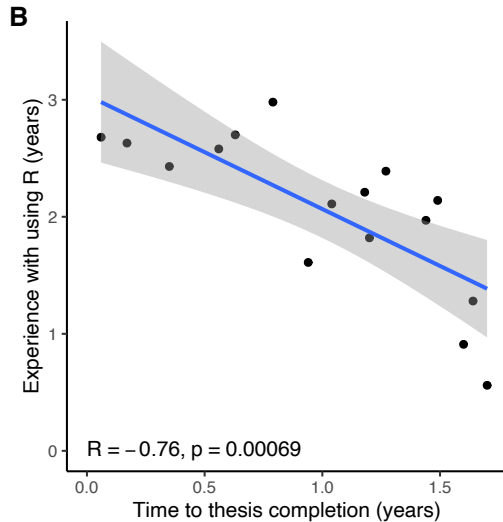
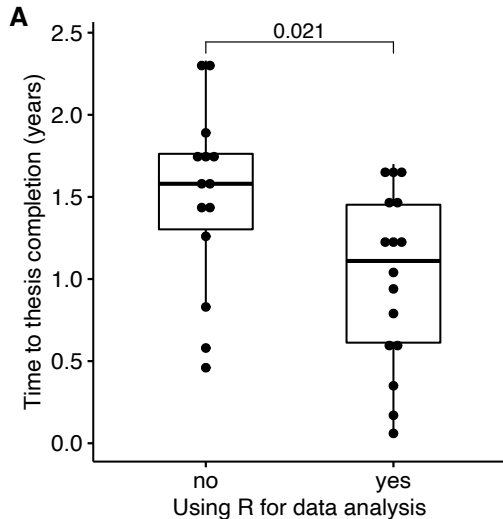
```
head(sim_data, n = 8) %>%  
  knitr::kable()  
# glimpse(data)
```

using_r	r_exp	thesis_compl
no	0.00	1.74
yes	1.97	1.44
no	0.00	2.27
no	0.00	1.26
no	0.00	1.61
no	0.00	1.43
yes	0.91	1.60
yes	2.11	1.04

Data summary

Dependent: all		all
Experience with R (years)	Mean (SD)	1.1 (1.2)
Est. time to thesis completion	Mean (SD)	1.2 (0.6)
Using R for analysis	no	14 (46.7)
	yes	16 (53.3)

Visualize simulated data



 Please provide some “real” data



<https://forms.gle/Z3RVbscYMYp3aThr5>

Let's get the real data!

```
url <- "https://docs.google.com/spreadsheets/d/17UDIyzhZknffptP0FQTGC0409QDDgfG9juZ39

# auth
googledrive::drive_auth(email = "cornelius.hennch@gmail.com")

# get the data from the google sheet
real_data <- googlesheets4::read_sheet(url) %>%
  select(-Zeitstempel)

# rename columns
colnames(real_data) <- colnames(sim_data)

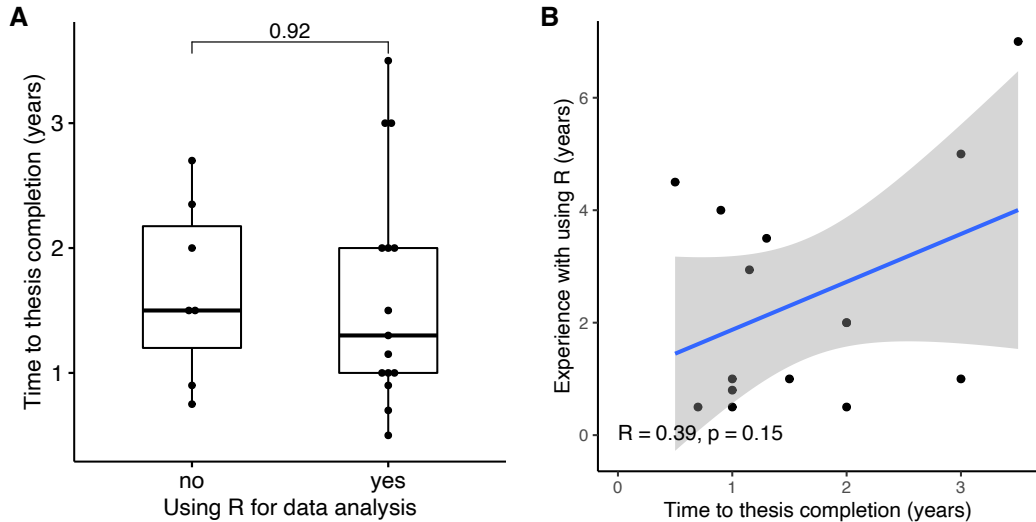
# wrange (convert everything to correct variable type)
real_data <- real_data %>%
  mutate(using_r = factor(using_r),
         r_exp = map_chr(r_exp, as.double) %>% as.double(),
         thesis_compl = map_chr(thesis_compl, as.double) %>% as.double())
```

Real data overview

Table 2: Survey summary, n = 22

Dependent: all		all
Experience with R (years)	Mean (SD)	1.6 (2.0)
Est. time to thesis completion	Mean (SD)	1.6 (0.8)
Using R for analysis	no	7 (31.8)
	yes	15 (68.2)

Vizualization of the “real” data with the same script



Section 3

How do I learn these tools?

Where to start

- Reproducible research with R:\

<https://www.bihealth.org/de/translation/innovationstreiber/quest-center/mission-ansaetze/ausbildung-und-training/reproducible-research-with-r>

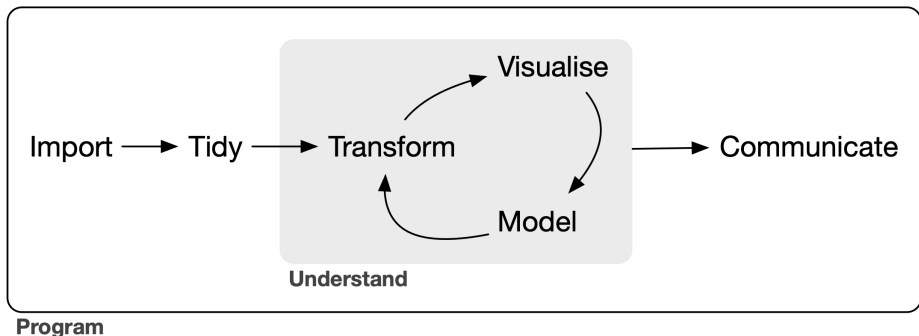
-  datacamp.com

-  Books e.g. R for Data Science by Hadley Wickham

- More resources on my website:

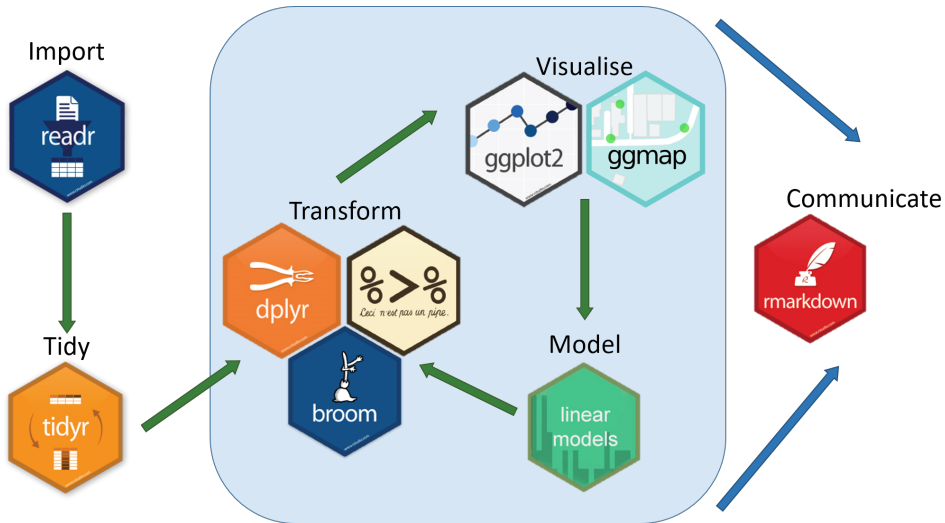
<https://www.hennch.co/post/free-r-learning-resources/>

Tidy data and analyses are essential for reproducibility



– Wickham and Grolemund [2]

Tidyverse tools



<https://medium.com/@kadek/how-to-install-the-tidyverse-r-via-homebrew-macos-10-14-d749d2136cf1>

Session Info

```
## R version 4.0.2 (2020-06-22)
## Platform: x86_64-apple-darwin17.0 (64-bit)
## Running under: macOS 10.16
##
## Matrix products: default
## BLAS: /Library/Frameworks/R.framework/Versions/4.0/Resources/lib/libRblas.dylib
## LAPACK: /Library/Frameworks/R.framework/Versions/4.0/Resources/lib/libRlapack.dylib
##
## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
##
## attached base packages:
## [1] stats      graphics  grDevices  utils      datasets  methods   base
##
## other attached packages:
## [1] finalfit_1.0.3  forcats_0.5.1  stringr_1.4.0  dplyr_1.0.7
## [5] purrr_0.3.4    readr_2.1.0    tidyr_1.1.4    tibble_3.1.6
## [9] ggplot2_3.3.5  tidyverse_1.3.1 knitr_1.36
##
## loaded via a namespace (and not attached):
## [1] nlme_3.1-150      fs_1.5.0          lubridate_1.8.0
## [4] httr_1.4.2        tools_4.0.2       backports_1.3.0
## [7] utf8_1.2.2        R6_2.5.1          DBI_1.1.0
## [10] mgcv_1.8-33       colorspace_2.0-2  withr_2.4.2
## [13] tidyselect_1.1.1  curl_4.3          compiler_4.0.2
## [16] cli_3.1.0         rvest_1.0.2       mice_3.13.0
## [19] xml2_1.3.2        labeling_0.4.2    scales_1.1.1
## [22] askpass_1.1       rappdirs_0.3.3    digest_0.6.28
## [25] foreign_0.8-80    rmarkdown_2.11    rio_0.5.16
## [28] jpeg_0.1-8.1      pkgconfig_2.0.3   htmltools_0.5.2
## [31] labelled_2.9.0    dbplyr_2.1.1      fastmap_1.1.0
```

References

Github repository of this talk: https://github.com/corneliushennch/repro_workflow

- [1] Aaron Peikert and Andreas Brandmeier. “A Reproducible Data Analysis Workflow with R Markdown, Git, Make, and Docker”. In: *Preprint* (2021), pp. 1–47.

- [2] Hadley Wickham and Garrett Golemund. *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*. 1st ed. O’Reilly Media, Jan. 2017. ISBN: 1491910399. URL: <http://r4ds.had.co.nz/>.