Reproducible analysis workflows

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

Cornelius Hennch

24.01.2022

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 (2021)

Reproducibility ≠ Replicability

(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)

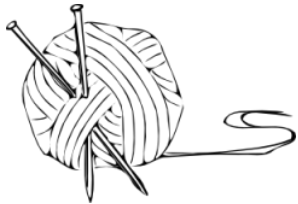
Four essential tools for reproducible workflows

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

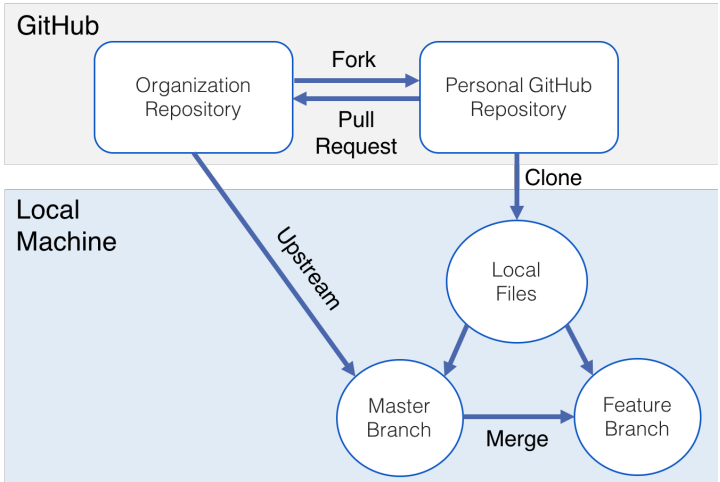
Git

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

Github

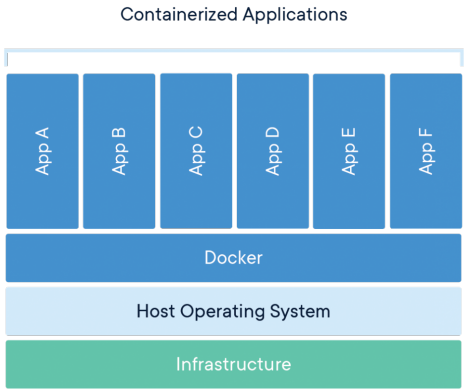
- Git repository hosting service
- **Collaboration**:
  1. Many features for team/project management (scalable!)
  2. Report bugs/issues, get help
  3. Contribute to open-source projects
- Post-publication platform

Collaboration with Git & Github



How to Update a Fork in Git

Docker


Containerized Applications
App A
App B
App C
App D
App E
App F
Docker
Host Operating System
Infrastructure

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

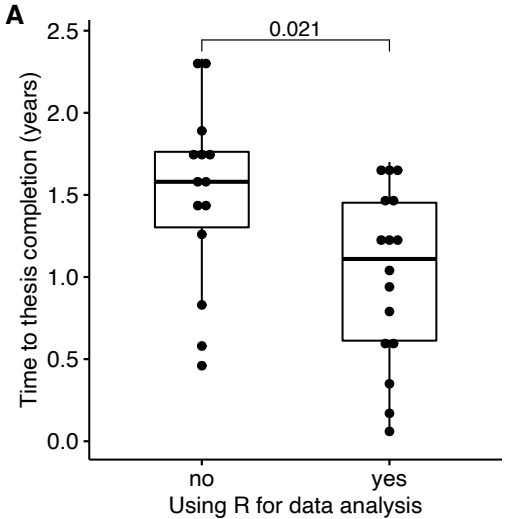
```
head(data, n = 8) %>%
  knitr::kable()
# glimpse(data)
```

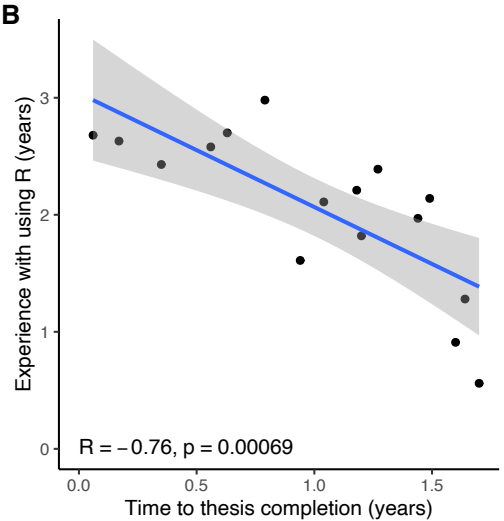
r_exp	using_r	thesis_compl
0.00	no	1.74
1.97	yes	1.44
0.00	no	2.27
0.00	no	1.26
0.00	no	1.61
0.00	no	1.43
0.91	yes	1.60
2.11	yes	1.04

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





Get the data

link/QR code to google forms

Run the code!

Section 3

How do I learn these tools?

Where to start

- Reproducible research with R
- Books
- Datacamp
- Website

Tidy data!

Links/ressources for these tools

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   httr_1.4.2
##  [5] tools_4.0.2       backports_1.3.0   utf8_1.2.2        R6_2.5.1
##  [9] DBI_1.1.0         mgcv_1.8-33       colorspace_2.0-2  withr_2.4.2
## [13] tidyselect_1.1.1  curl_4.3          compiler_4.0.2    cli_3.1.0
## [17] rvest_1.0.2       mice_3.13.0       xml2_1.3.2        labeling_0.4.2
## [21] scales_1.1.1      digest_0.6.28     foreign_0.8-80    rmarkdown_2.11
## [25] rio_0.5.16        jpeg_0.1-8.1      pkgconfig_2.0.3   htmltools_0.5.2
## [29] dbplyr_2.1.1      fastmap_1.1.0     highr_0.9         rlang_0.4.12
## [33] readxl_1.3.1      rstudioapi_0.13   generics_0.1.1    farver_2.1.0
## [37] jsonlite_1.7.2    zip_2.1.1         car_3.0-10        magrittr_2.0.1
## [41] Matrix_1.2-18     Rcpp_1.0.7        munsell_0.5.0     fansi_0.5.0
```

References

Peikert, Aaron, and Andreas Brandmeier. 2021. “A Reproducible Data Analysis Workflow with R Markdown, Git, Make, and Docker Aaron.” *Preprint*, 1–47.