Reproducible analysis workflows

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

ReproducibiliTea

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

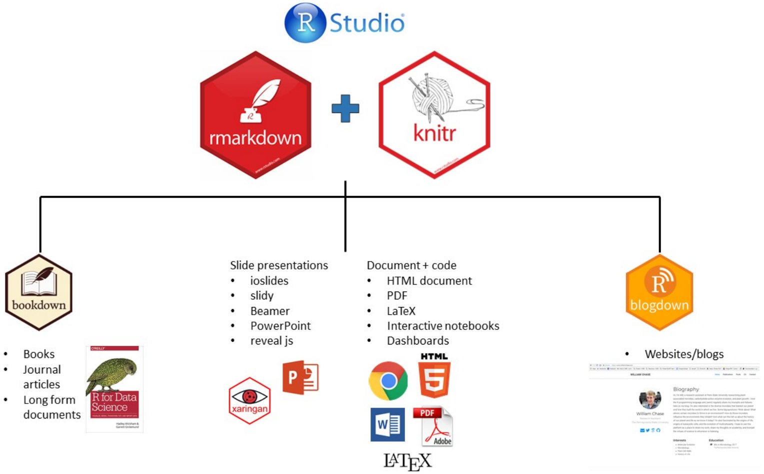
1. Dynamic reports → **R Markdown**

2. Version control → **Git & Github**

3. Dependency management → **Make**

4. Containerization → **Docker**

Highly versatile dynamic documents with R Markdown



- Books
- Journal articles
- Long form documents

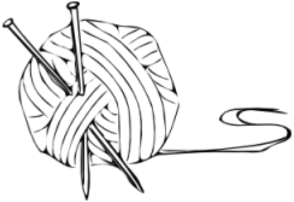
Slide presentations
- ioslides
- slidy
- Beamer
- PowerPoint
- reveal js

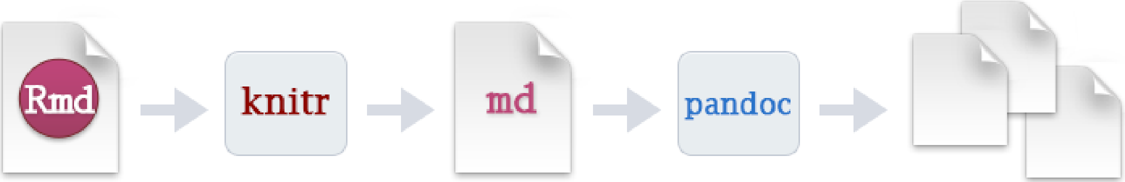
Document + code
- HTML document
- PDF
- LaTeX
- Interactive notebooks
- Dashboards

- Websites/blogs

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

Happy knitting!





Rmd

knitr

md

pandoc

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

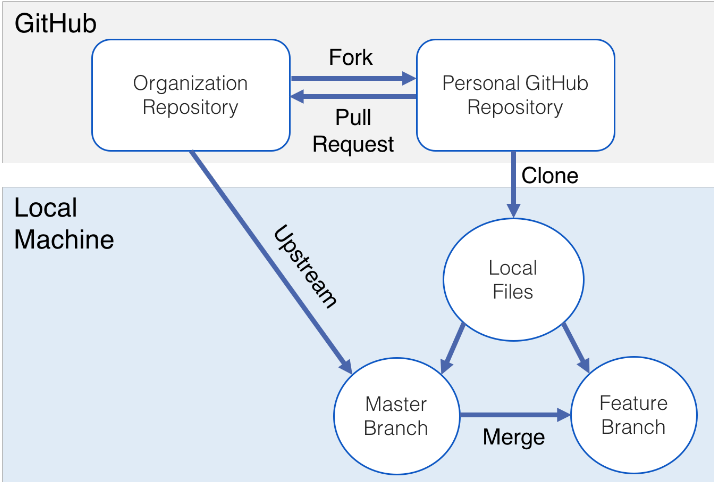
Git

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

Github

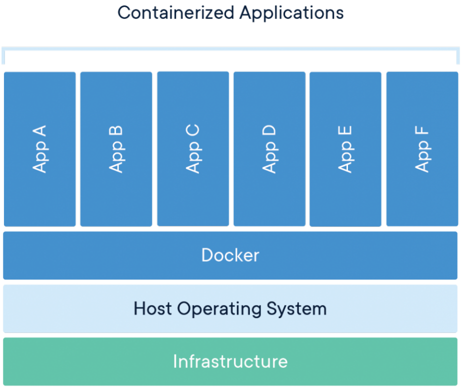
- Git repository hosting service
- **Collaboration**:
  1. Many features for team/project management (scalable!)
  2. Report bugs/issues, get help
  3. Contribute to open-source projects
- Post-publication platform

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

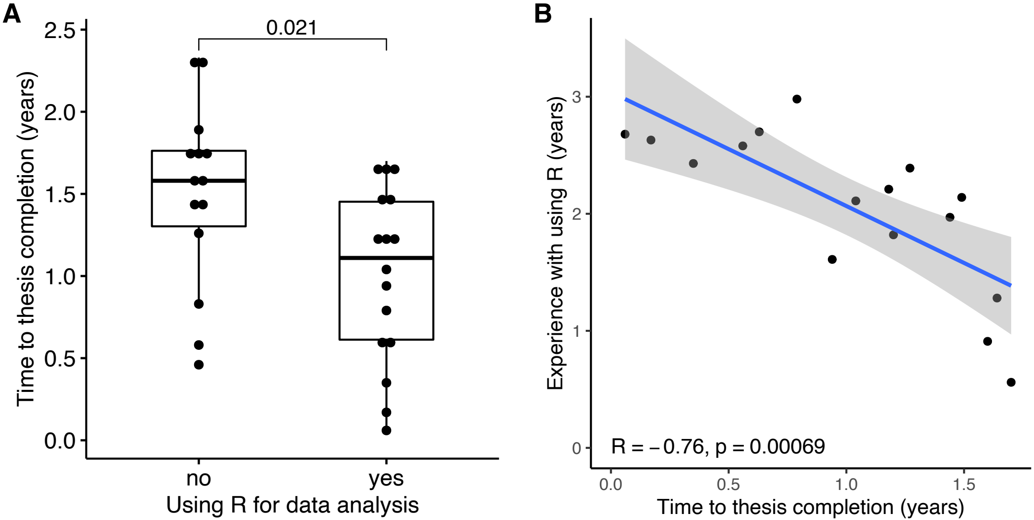
```r
head(data, n = 8) %>%
  knitr::kable()
# glimpse(data)
```

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!

Where to start

Links/ressources for these tools

References

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