Reproducible analysis workflows

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

ReproducibiliTea

Cornelius Hennch

23.01.2022

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

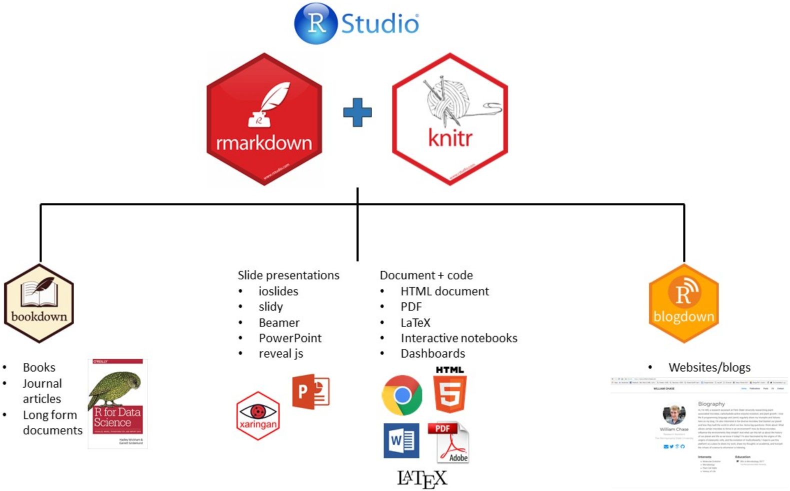
1. Dynamic reports → **R Markdown**

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

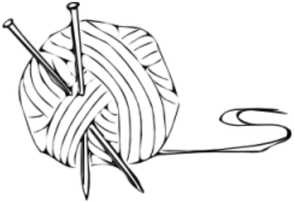
Document + code

- HTML document
- PDF
- LaTeX
- Interactive notebooks
- Dashboards

- Websites/blogs

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

Happy knitting!



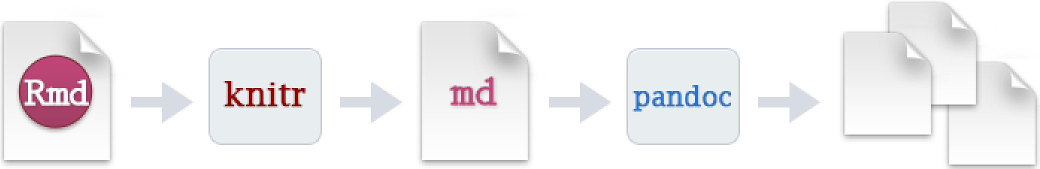


Figure 1: R Markdown under the hood

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!)

Github

Git & Github

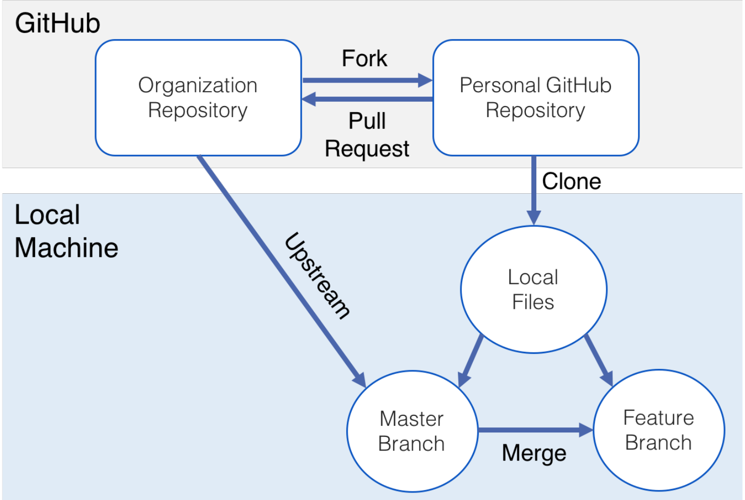
Git

- “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**:
  - Many features for team/project management (scalable!)
  - Report bugs/issues, get help
  - Contribute to open-source projects
- Post-publication platform

Collaboration with Git & Github



How to Update a Fork in Git

Docker

Docker is a tool that allows encapsulation, sharing, and re-creation of a computational environment on most operating systems (Windows, macOS, & Linux).

Reproducible data analysis in action

Hypothesis: R skills predict early PhD completion

Simulate and vizualize data

Get the data

link/QR code to google forms

Run the code!

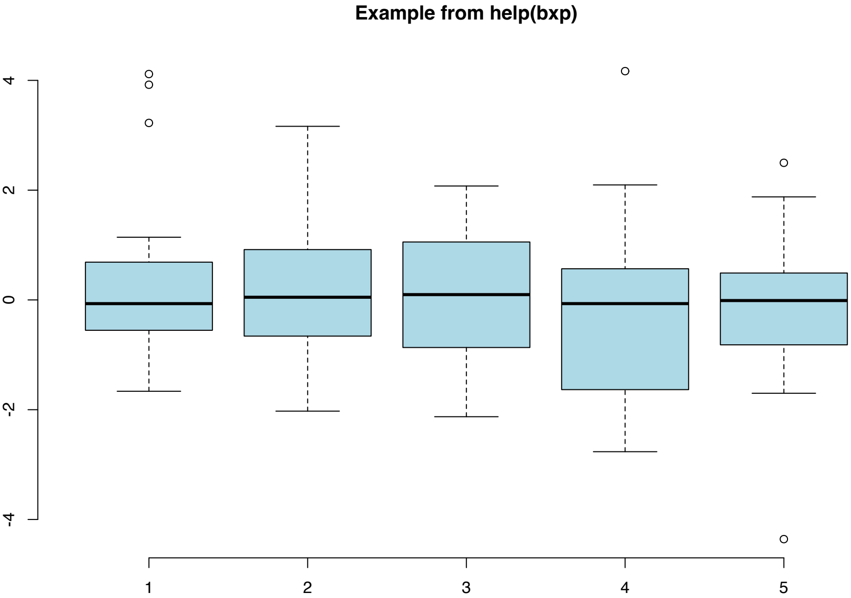
Where to start

Links/ressources for these tools

R Appendix: R Figure Example

The following code generates the plot on the next slide (taken from `help(bxp)` and modified slightly):

R Appendix: R Figure Example



R Appendix: R Table Example

A simple `knitr::kable` example:

Table 1: (Parts of) the mtcars dataset

	mpg	cyl	disp	hp	drat	wt	qsec	vs
Mazda RX4	21.0	6	160	110	3.90	2.620	16.46	0
Mazda RX4 Wag	21.0	6	160	110	3.90	2.875	17.02	0
Datsun 710	22.8	4	108	93	3.85	2.320	18.61	1
Hornet 4 Drive	21.4	6	258	110	3.08	3.215	19.44	1
Hornet Sportabout	18.7	8	360	175	3.15	3.440	17.02	0

References

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