Reproduzierbare Forschung mit R und Quarto – Einleitung

Workshop für die GSE Wuppertal

Björn S. Siepe

June 27, 2025

Einstieg

Plan für Heute

1. **Einleitung**
    - Was ist Reproduzierbarkeit?
    - Einfache Strategien (“Easy Wins”)
    - Einführung in dynamische Berichte
2. **Praktischer Teil**
    - Erstellung dynamischer Berichte mit Quarto
    - Typische Herausforderungen und Lösungen
3. **Fortgeschrittene Themen**
    - Nächste Schritte für mehr Reproduzierbarkeit
    - Tools und Best Practices
4. **Zusammenfassung & Abschlussdiskussion**
    - Was haben wir gelernt?
    - Offene Fragen & Ausblick

Sehr grober Zeitplan

Uhrzeit	Programmpunkt
10:00 – 10:15	Begrüßung & Einführung
10:15 – 11:15	**Einleitung**
11:15 – 11:20	☕ Kurze Pause
11:20 – 12:30	**Praktischer Teil 1**
12:30 – 13:15	🍽️ Mittagspause
13:15 – 14:20	**Praktischer Teil 2**
14:20 – 14:30	☕ Kurze Pause
14:30 – 15:30	**Fortgeschrittene Themen**
15:30 – 16:00	**Zusammenfassung & Abschlussdiskussion**

Vorabinformation

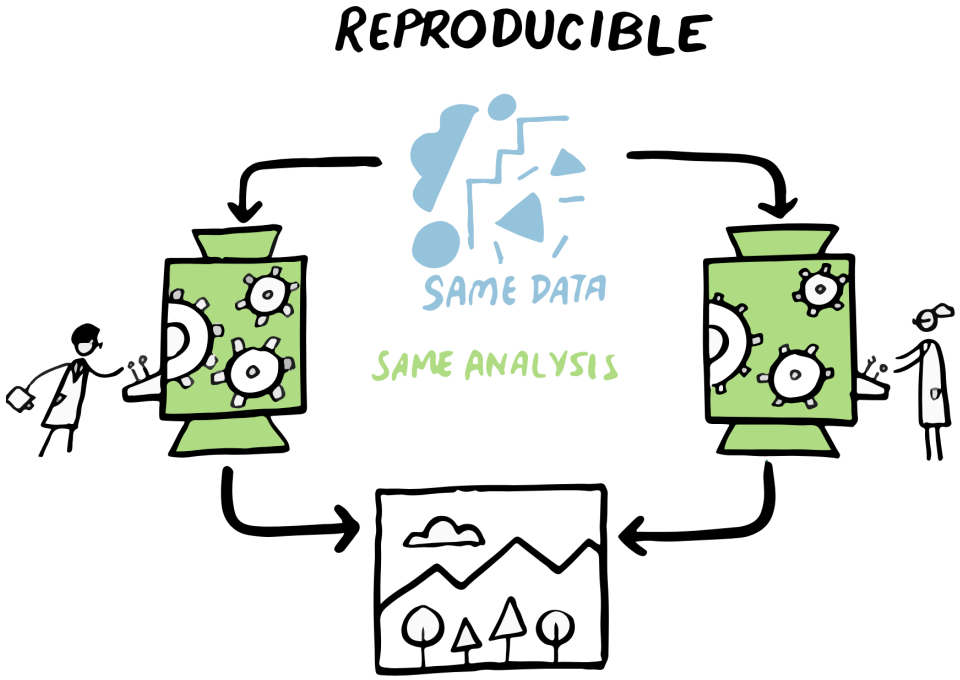
Teile dieses Workshops beruhen teils verbatim oder sehr nah auf folgenden Ressourcen, welche ich mit einer CC-BY Lizenz verwende: - CRS Primer on dynamic reporting - Molo, Fraga González, and Pawel (2025) - Peikert and Diemerling (2025) - Teils recht opinionated

TODO Link zu Materialien nochmals.

Computational reproducibility

Eine sehr einfache Definition:

Die **gleiche Analyse** der **gleichen Daten** führt zu den **gleichen Ergebnissen**



The Turing Way Community and Scriberia (2024)

Example

Computational reproducibility of Jupyter notebooks from biomedical publications

Sheeba Samuel ✉ , Daniel Mietchen ✉ Author Notes

GigaScience, Volume 13, 2024, giad113,
https://doi.org/10.1093/gigascience/giad113
Published: 11 January 2024 **Article history** ▾

- Es wurde versucht, **22.578 Jupyter-Notebooks**, die mit 3.467 Publikationen aus PubMed verknüpft sind, erneut auszuführen.
- Für 10.388 dieser Notebooks konnten alle angegebenen Abhängigkeiten erfolgreich installiert werden
- 1.203 Notebooks liefen ohne Fehler durch
- **879 Notebooks (4 %)** lieferten Ergebnisse, die mit denen des Originals identisch waren

→ **Computational reproducibility kann schwierig sein!**

Terminologie

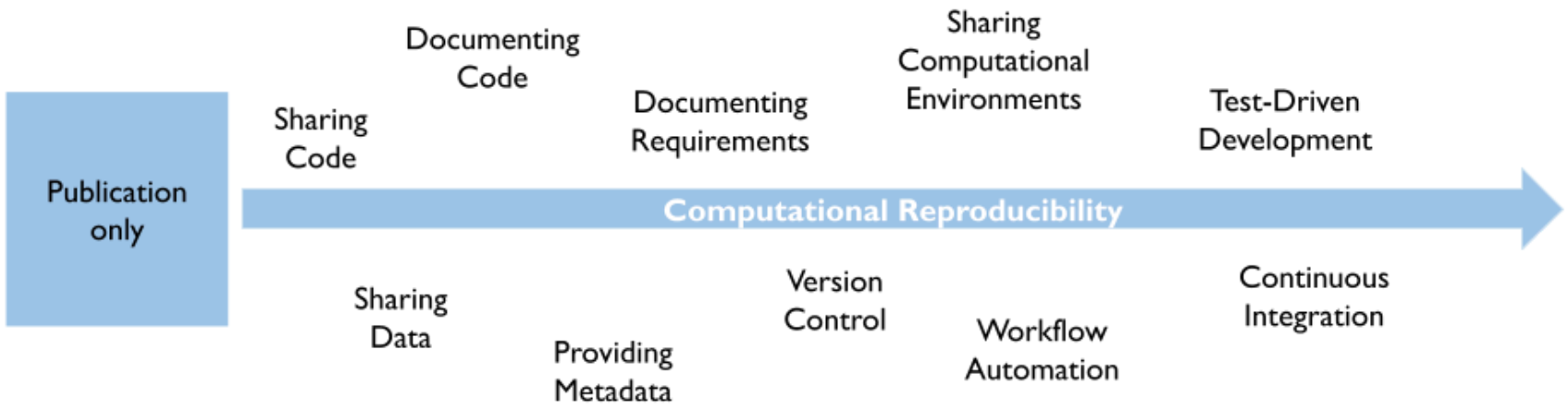
		Data	
		Same	Different
Analysis	Same	Reproducible	Replicable
	Different	Robust	Generalisable

The Turing Way Community (2022)

Computational Reproducibility

Molo, Fraga González, and Pawel (2025)

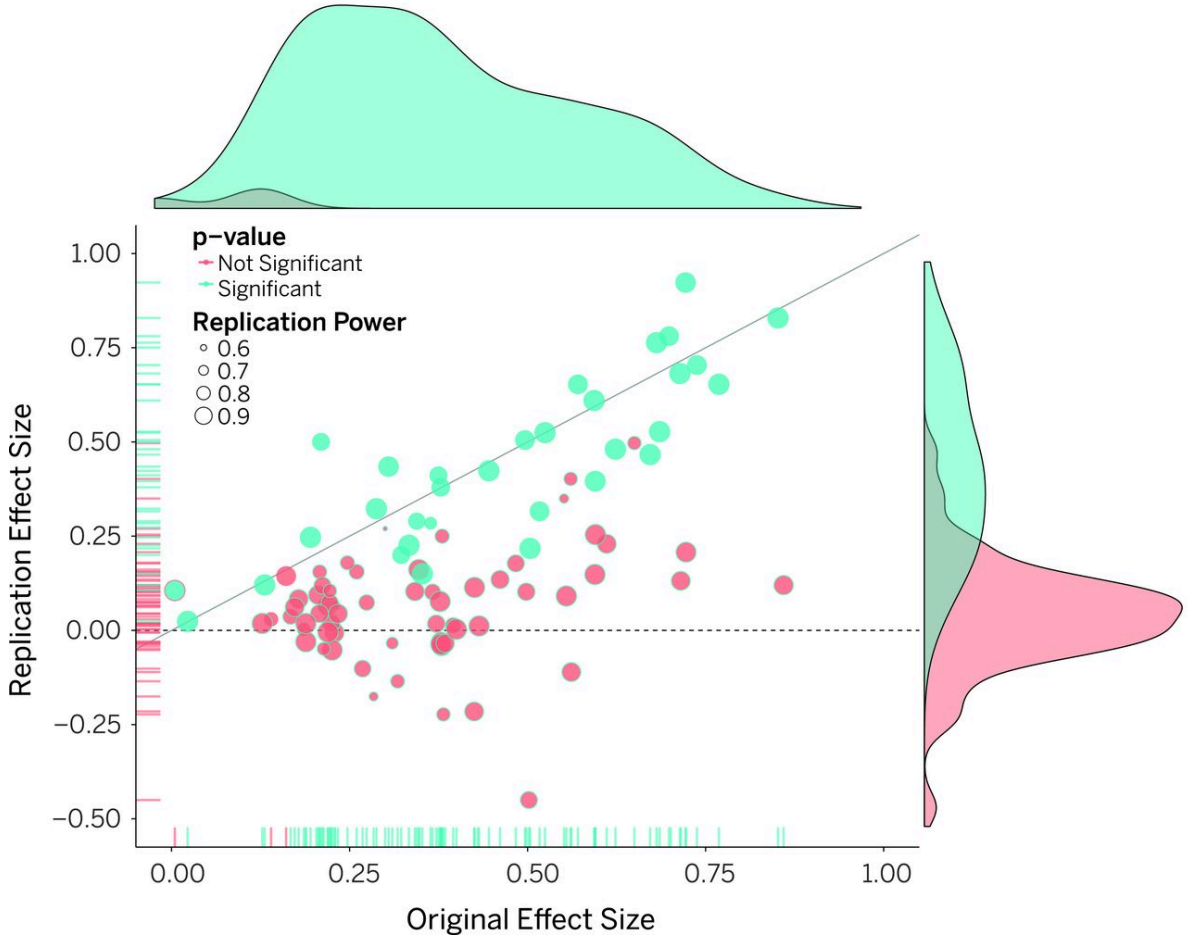
Reproducible	Replicable
Robust	Generalisable



From Molo, Fraga González, and Pawel (2025), inspired by Peng (2011) and Peikert and Brandmaier (2021)

Replicability: Beispiel

Reproducible	Replicable
Robust	Generalisable



Open Science Collaboration (2015)

Robustness: Beispiel

Reproducible	Replicable
Robust	Generalisable

REPRODUCIBILITY GOES ON TRIAL IN ECOLOGY

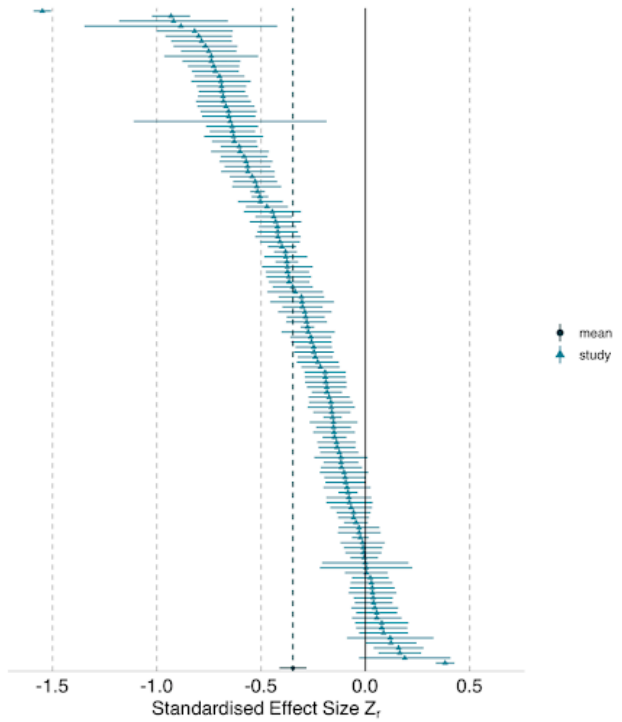
With same data sets, 246 biologists get different results – showing how analytical choices drive conclusions.

By Anil Oza

In a massive exercise to examine reproducibility, more than 200 biologists analysed the same sets of ecological data – and got widely divergent results. The first sweeping study of its kind in ecology demonstrates how much results in the field can vary, not because of differences in the environment, but because of scientists' analytical choices (E. Gould *et al*. Preprint at EcoEvoRxiv https://doi.org/gswhwr; 2023).

"There can be a tendency to treat individual papers' findings as definitive," says Hannah Fraser, an ecology meta researcher at the University of Melbourne in Australia and a co-author of the study. But the work shows that "we really can't be relying on any individual result or any individual study to tell us the whole story".

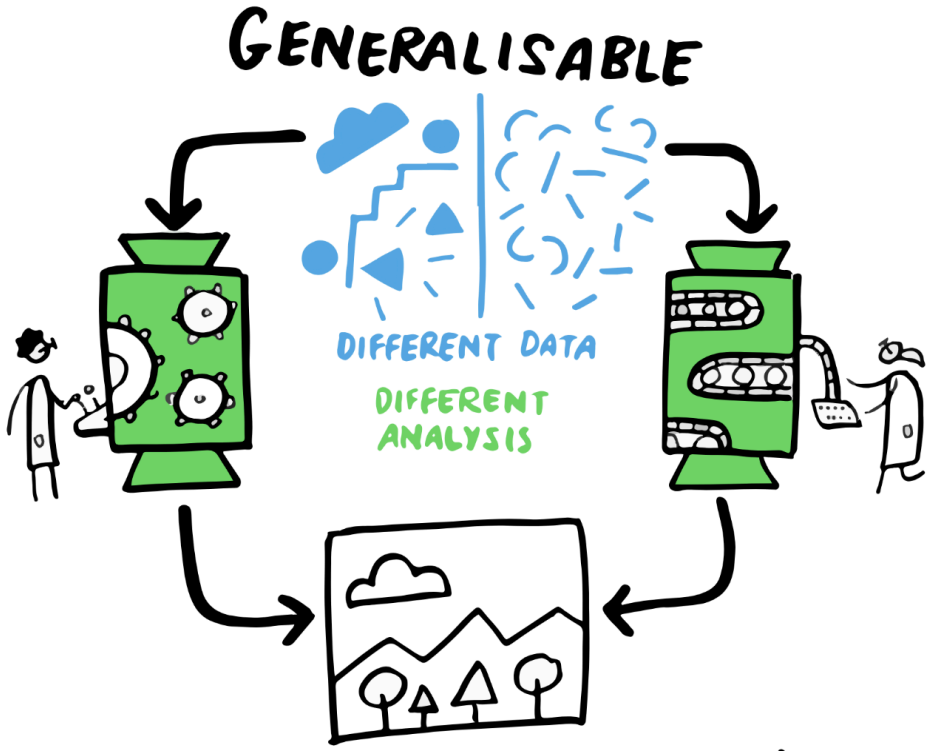
Variation in results might not be surprising, but quantifying that variation in a formal study could catalyse a larger movement to improve reproducibility, says Brian Nosek,



Gould et al. (2023)

Generalizability

Reproducible	Replicable
Robust	Generalisable



The Turing Way Community and Scriberia (2024)

→ **Reproducibility + Robustness + Replicability necessary for Generalizabilty!**

Jeder Schritt zählt

Aber reproduzierbare Forschung ist schwierig und kann sich überwältigend anfühlen.

Kein Grund zur Verzweiflung! Mit jedem Schrit::

- seid Ihr einen Schritt näher an reproduzierbarer Forschung
- verbessert Ihr die Qualität eurer Forschung und erleichtert euer Leben
- lernt Ihr breit anwendbare technische Skills

Easy Wins

Projektstruktur

Vorher: Chaos 📁

- Paper_final_wirklich_final.docx
- daten.xlsx code_alt.R
- code_neu_funktioniert.R
- plot1.png

Nachher: Ordnung ✨

```
projekt/
├── data/
│   ├── raw/
│   └── processed/
├── scripts/
├── output/
└── README.md
```

> **Praxis-Tipp**
>
> Erstellt eine Standard-Ordnerstruktur und kopiert diese für neue Projekte!

Dateinamen

❌ Problematisch

- `Daten.xlsx`
- `neue analyse.R`
- `plot final FINAL.png`
- `Fragebogen - Kopie (2).docx`

✅ Besser

- `2025-06-28_fragebogen-daten-roh.xlsx`
- `01_datenbereinigung.R`
- `02_deskriptive-statistik.R`
- `abb01_altersverteilung.png`

Code-Dokumentation

❌ Unzureichend:

Unklarer Name, keine Dokumentation:

```r
logmlis <- function(x){
  lapply(x, function(x){
   y <- unlist(x)
   y <- y[!is.na(y)]
   y <- y[y != ""]
   return(mean(log(y)))
  })
}
```

✅ Besser:

Klarer Name, konzise Dokumentation:

```r
log_mean_list <- function(x) {
  # Calculates the logarithmic mean of a list of numerical values
  # Removes NA and empty values before the calculation
  y <- unlist(x)
  y <- y[!is.na(y)]
  y <- y[y != ""]
  return(mean(log(y)))
}
```

Code-Dokumentation

- Viele unterschiedliche Meinungen
- In der Wissenschaft: Lieber etwas zu viel als zu wenig

Praxis-Tipp

Nach Peikert and Diemerling (2025):

1. Was standardisiert ist, muss nicht dokumentiert werden.
2. Was einfach ist, braucht nur wenig Dokumentation.
3. Was konsistent ist, muss nur einmal dokumentiert werden.

Dateipfade

❌ Absoluter Pfad (schlecht):

```
daten <- read.csv("C:/Users/Maria/Desktop/Masterarbeit/daten.csv")
```

Problem: Funktioniert nur auf Ihrem Computer!

✅ Relativer Pfad (gut):

```
daten <- read.csv("data/raw/fragebogen-daten.csv")
```

Noch besser: Verwendung des `here` packages (funktioniert mit verschiedenen Systemen):

```
library(here)
daten <- read.csv(here("data", "raw", "fragebogen-daten.csv"))
```

Aufteilung in verschiedene Skripte

❌ Ein großes Skript:

```
# complete_analysis.R (500+ lines)
# Data import, cleaning, analysis, plots, export...
```

✅ Modulare Aufteilung:

- `01_data_import.R` - Rohdaten laden
- `02_data_cleaning.R` - Präprozessierung
- `03_descriptive_analysis.R` - Deskriptive Statistiken
- `04_inferential_statistics.R` - Inferenz
- `05_visualization.R` - Abbildungen
- `06_export.R` - Export

Vorteile: Übersichtlicher, leichter zu debuggen, bessere Zusammenarbeit

Funktionsbasiertes Programmieren

❌ Code-Wiederholung:

```r
# Standardization for variable 1
data$var1_std <- (data$var1 - mean(data$var1, na.rm = TRUE)) /
                 sd(data$var1, na.rm = TRUE)

# Standardization for variable 2
data$var2_std <- (data$var2 - mean(data$var2, na.rm = TRUE)) /
                 sd(data$var2, na.rm = TRUE)

# Standardization for variable 3...
```

✅ Funktion erstellen:

```r
standardize_variable <- function(x) {
  # Z-standardization of a numeric variable
  (x - mean(x, na.rm = TRUE)) / sd(x, na.rm = TRUE)
}

# Application
data$var1_std <- standardize_variable(data$var1)
data$var2_std <- standardize_variable(data$var2)
data$var3_std <- standardize_variable(data$var3)

# Or even better
data <- data |>
  mutate(across(starts_with("var"), standardize_variable))
```

Vermeiden von “magic numbers”

❌ Unklare Zahlen im Code:

```r
# What do these numbers mean?
clean_data <- data[data$age >= 18 & data$age <= 65, ]
final_data <- clean_data[clean_data$score > 2.5, ]
significant <- results[results$p_value < 0.05, ]
```

✅ Benannte Konstanten:

```r
# Clear definitions at the beginning of the script
MIN_AGE <- 18
MAX_AGE <- 65
MIN_SCORE <- 2.5
ALPHA_LEVEL <- 0.05

# Usage in code
clean_data <- data[data$age >= MIN_AGE &
                   data$age <= MAX_AGE, ]
final_data <- clean_data[clean_data$score > MIN_SCORE, ]
significant <- results[results$p_value < ALPHA_LEVEL, ]
```

Vorteil: Code wird selbsterklärend und leichter anpassbar

Session Info

Warum wichtig?

- **Reproduzierbarkeit:** Welche R-Version und Pakete wurden verwendet?
- **Fehlerdiagnose:** Unterschiedliche Paketversionen können verschiedene Ergebnisse liefern
- **Dokumentation:** Vollständige Nachvollziehbarkeit der Analyse

✅ Am Ende jedes Skripts:

```
# Session information for reproducibility
sessionInfo()
```

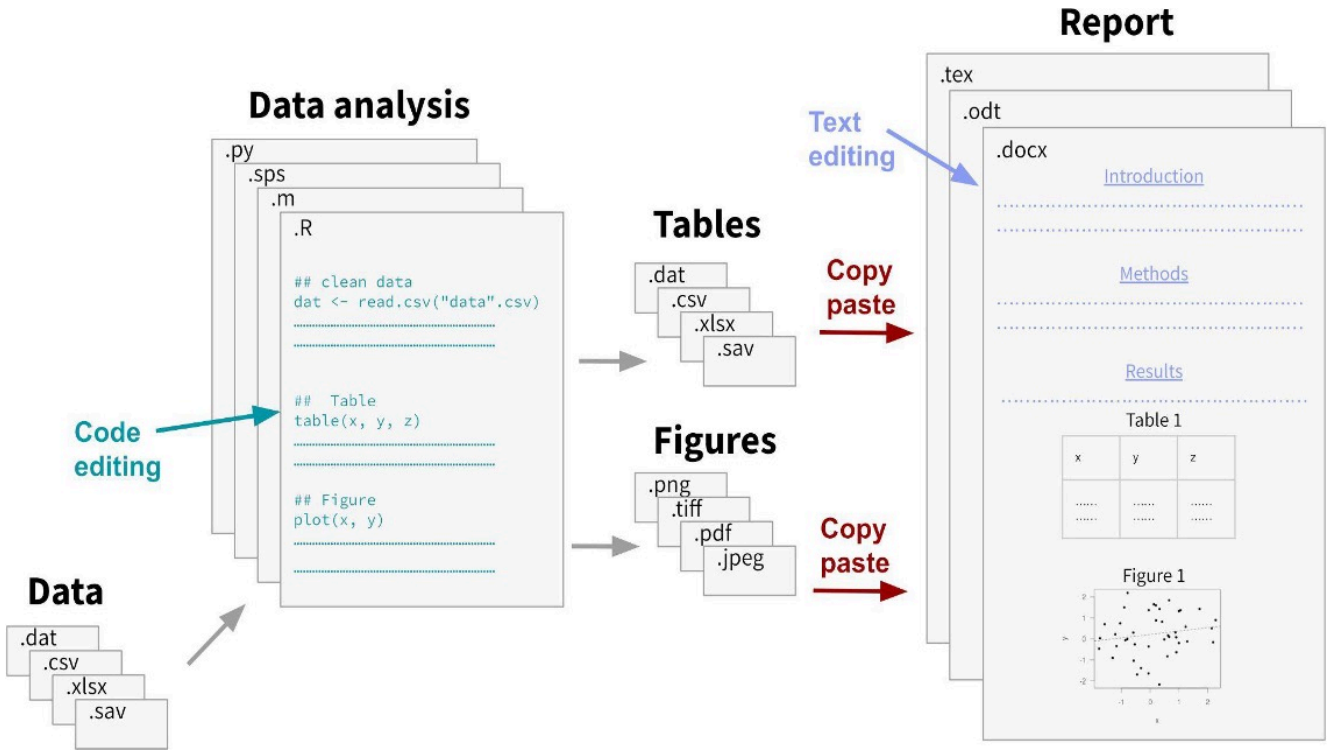
Praxis-Tipp

Entwedern in Quarto direkt rendern oder in separatem File speichern:

```
writeLines(capture.output(sessionInfo()), "session_info.txt")
```

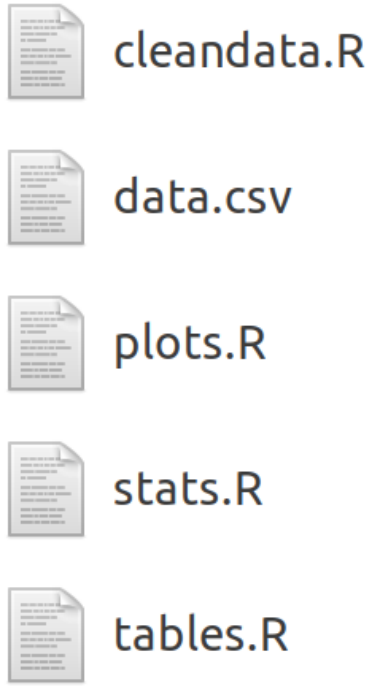
Dynamische Reports

Typischer, nicht-dynamischer Workflow



Molo, Fraga González, and Pawel (2025)

Typische Probleme



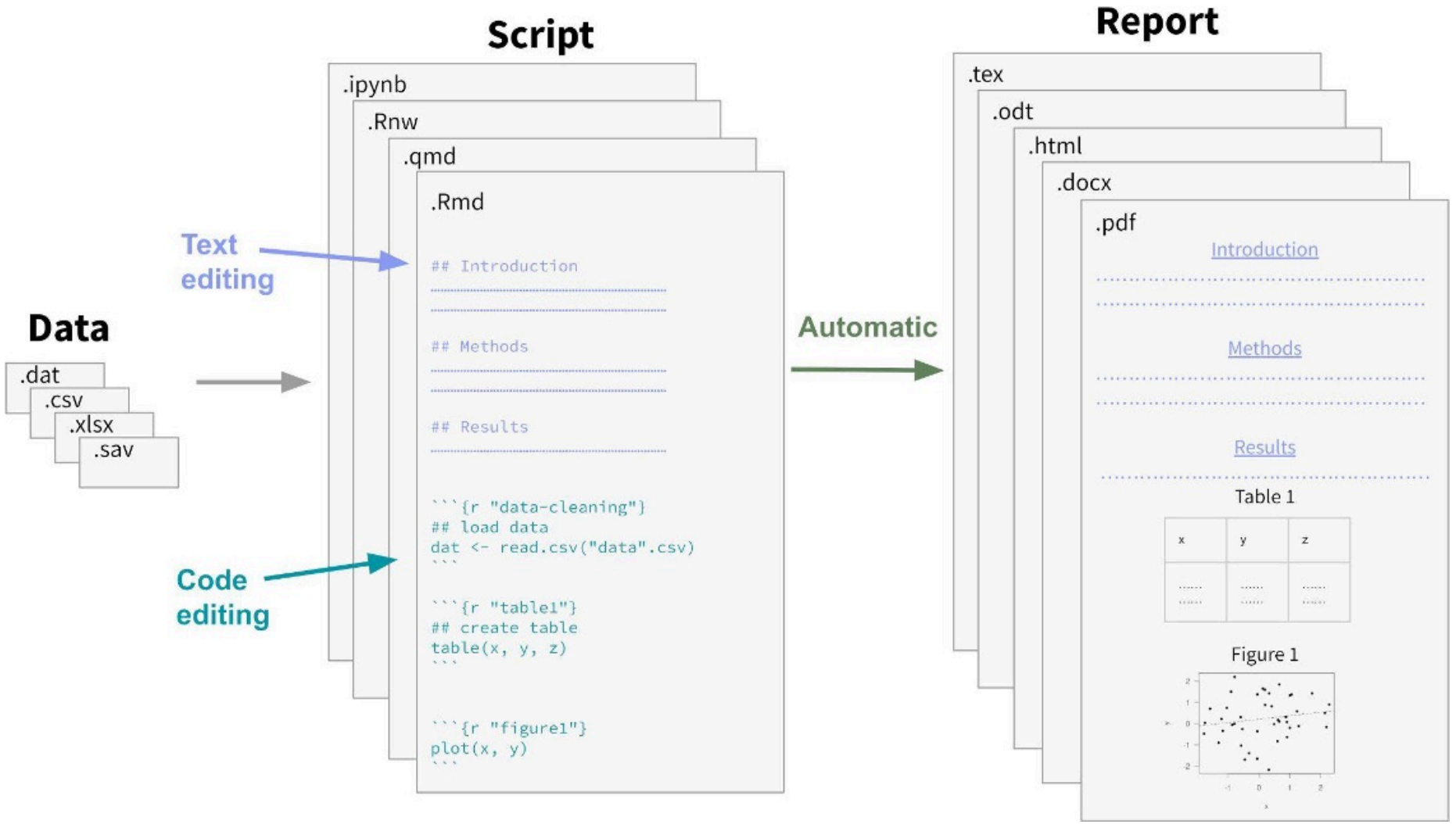
- **unklarer** Ausführungsreihenfolge
- **Zeitaufwendiges** Copy-Pasten
- **Fehleranfälliges** Copy-Pasten

Molo, Fraga González, and Pawel (2025)

Dynamische Reports: Das Ziel

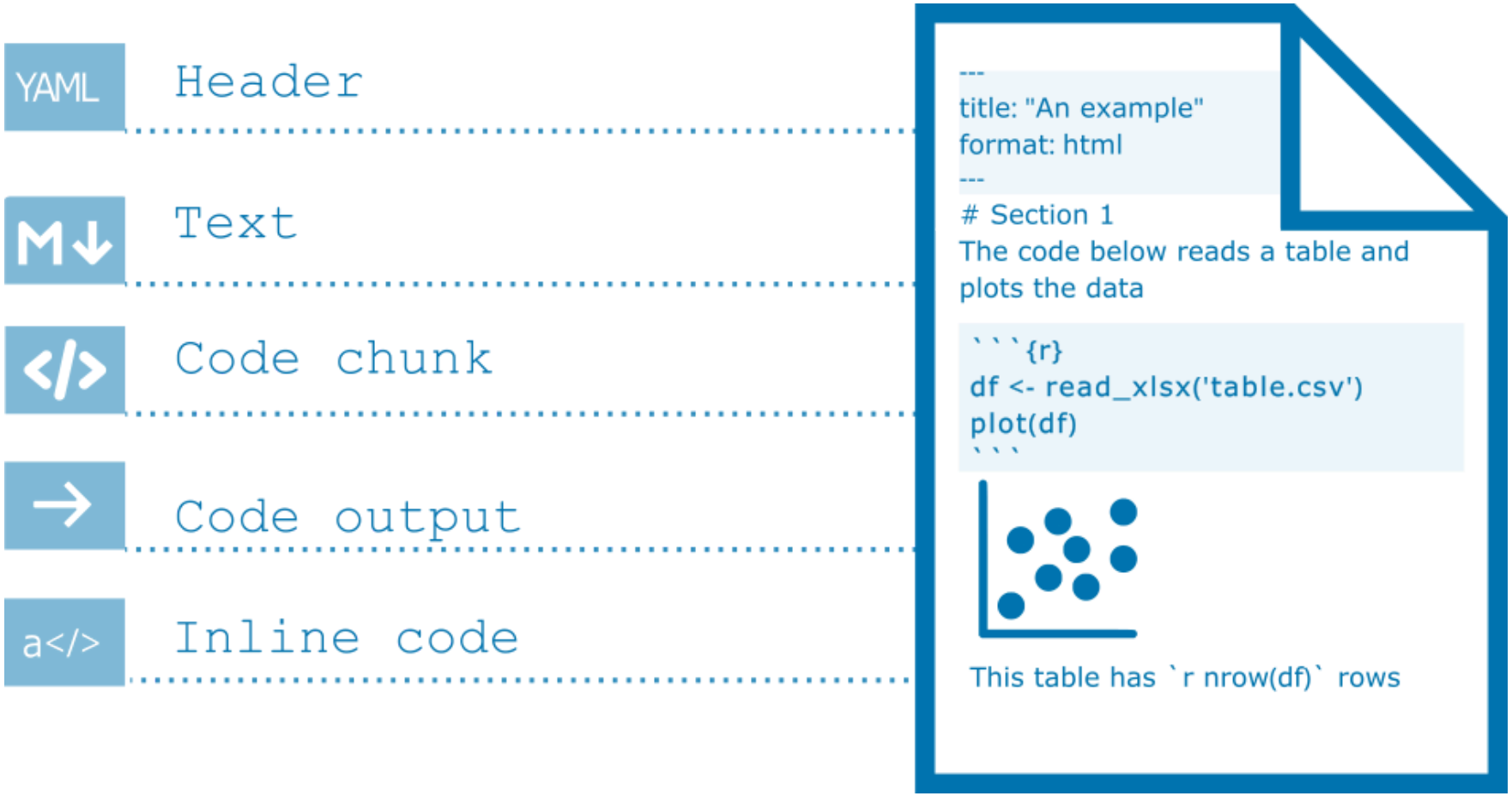
`Daten` → dynamischer Report → `Manuskript`

Ein Workflow für dynamische Reports



Molo, Fraga González, and Pawel (2025)

Komponenten eines dynamischen Reports



Molo, Fraga González, and Pawel (2025)

Text: Typische Markup-Sprachen

HTML

```
<!DOCTYPE html>
<html>
<head>
<title>Document</title>
</head>
<body>
 <h1>Heading Level 1</h1>
 <h2>Heading Level 2</h2>
 <p><strong>Bold text</strong>
and <em>italic text</em> in a
paragraph.</p>

<ol> <li>First item in an ordered
list</li> <li>Second item in an ordered
list</li> </ol>

<ul> <li>Unordered list item</li>
<li>Another unordered list item</li>
</ul> <p><code>Inline code</code>
and</p>

<pre>Block code</pre>

</body>
</html>
```

LaTeX

```
\documentclass{article}
\begin{document}
\section{Heading Level 1}
\subsection{Heading Level 2}

\textbf{Bold text} and \textit{italic
text} in a paragraph.
\begin{enumerate}
  \item First item in an ordered list
  \item Second item in an ordered list
\end{enumerate}

\begin{itemize}
  \item Unordered list item
  \item Another unordered list item
\end{itemize}

\texttt{Inline code} and
\begin{verbatim} Block code \end{verbatim}
\end{document}
```

Markdown

```
# Heading Level 1
## Heading Level 2

**Bold text** and *italic text*
in a paragraph.

1. First item in an ordered list
2. Second item in an ordered list

- Unordered list item
- Another unordered list item

`Inline code` and

```
Block code
```
```

Molo, Fraga González, and Pawel (2025)

Manche Implementierungen

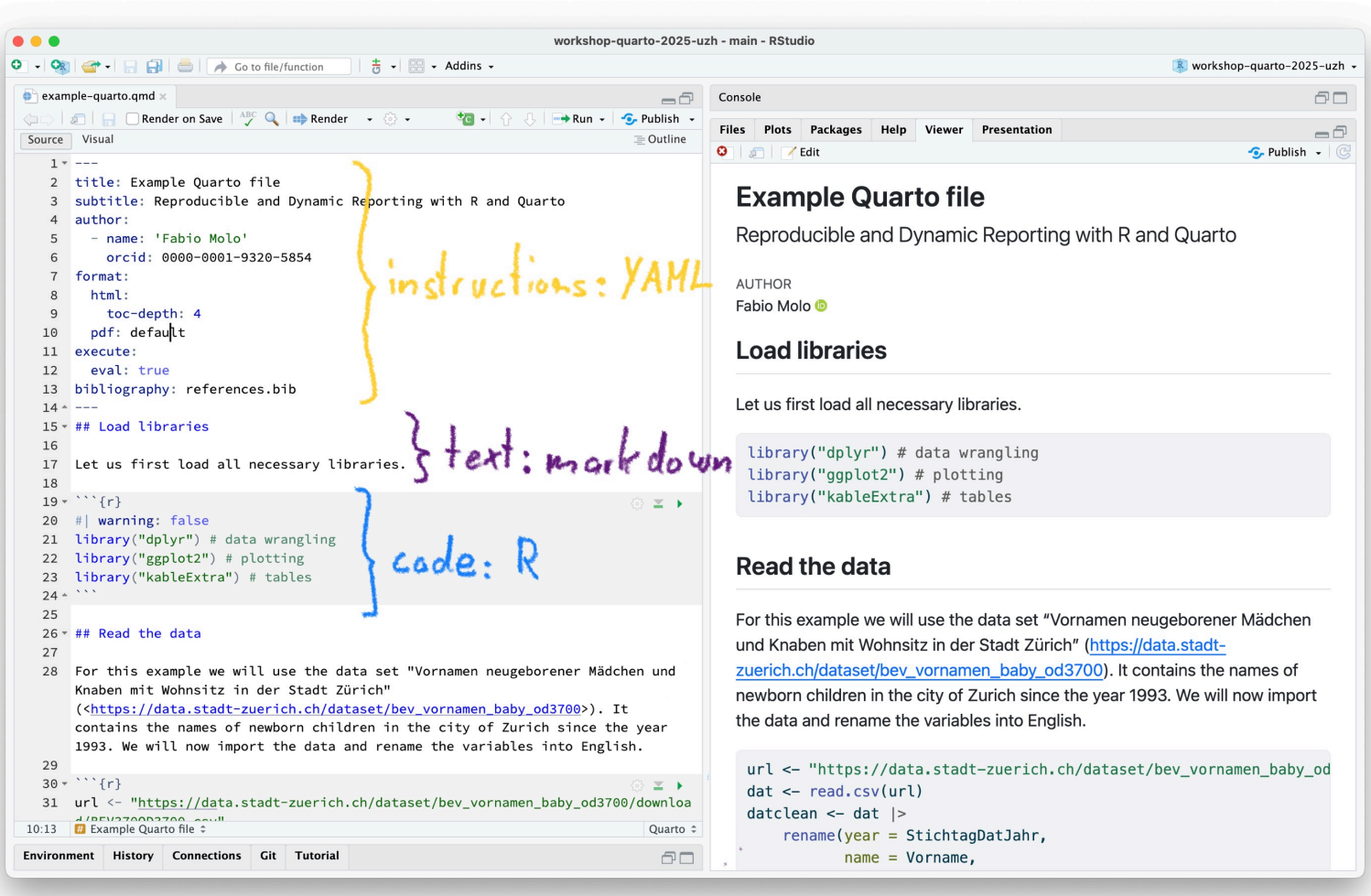
	Datei	Code	Markup	Output Formate
Sweave	.Rnw	R	LaTeX	pdf, tex
R Markdown	.Rmd	R	Markdown	pdf, html, tex, docx, pptx
Jupyter Notebook	.ipynb	Julia, Python, R	Markdown	pdf, html, tex, py, …
Quarto	.qmd	Julia, Python, R	Markdown	pdf, html, tex, docx, pptx, …

Warum Quarto?

- **open source**
- **modern**, wird kontinuierlich weiterentwickelt
- **einfache Handhabung**
- **Mehrere Programmiersprachen** (z.B. R, Python, Julia)
- **vielfältige Editoren/IDEs** (z. B. RStudio, VS Code)
- **allgemeines** wissenschaftliches Publikationssystem (z. B. Artikel, Folien, Websites)

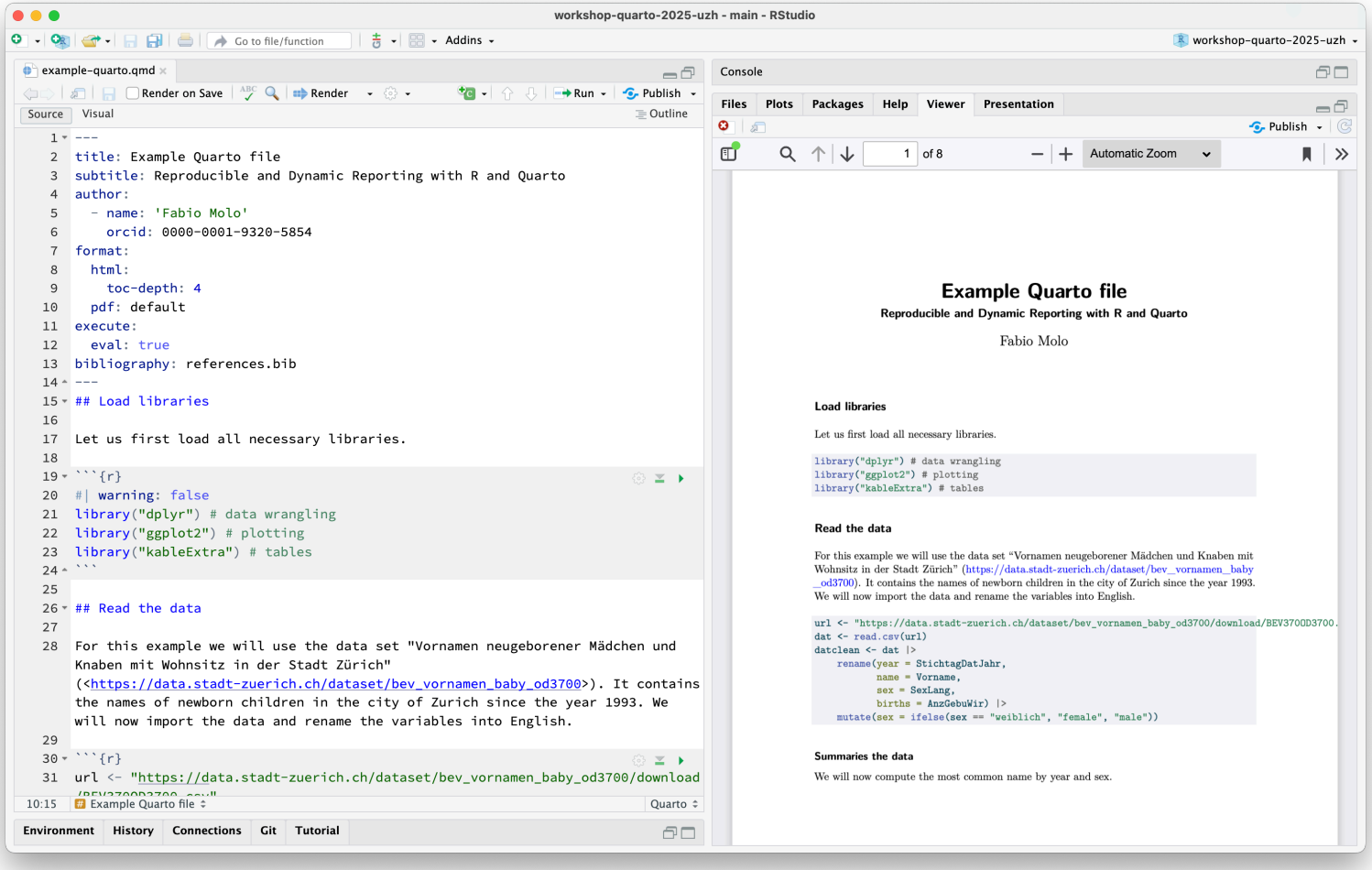
https://quarto.org/quarto.png

Quarto in RStudio



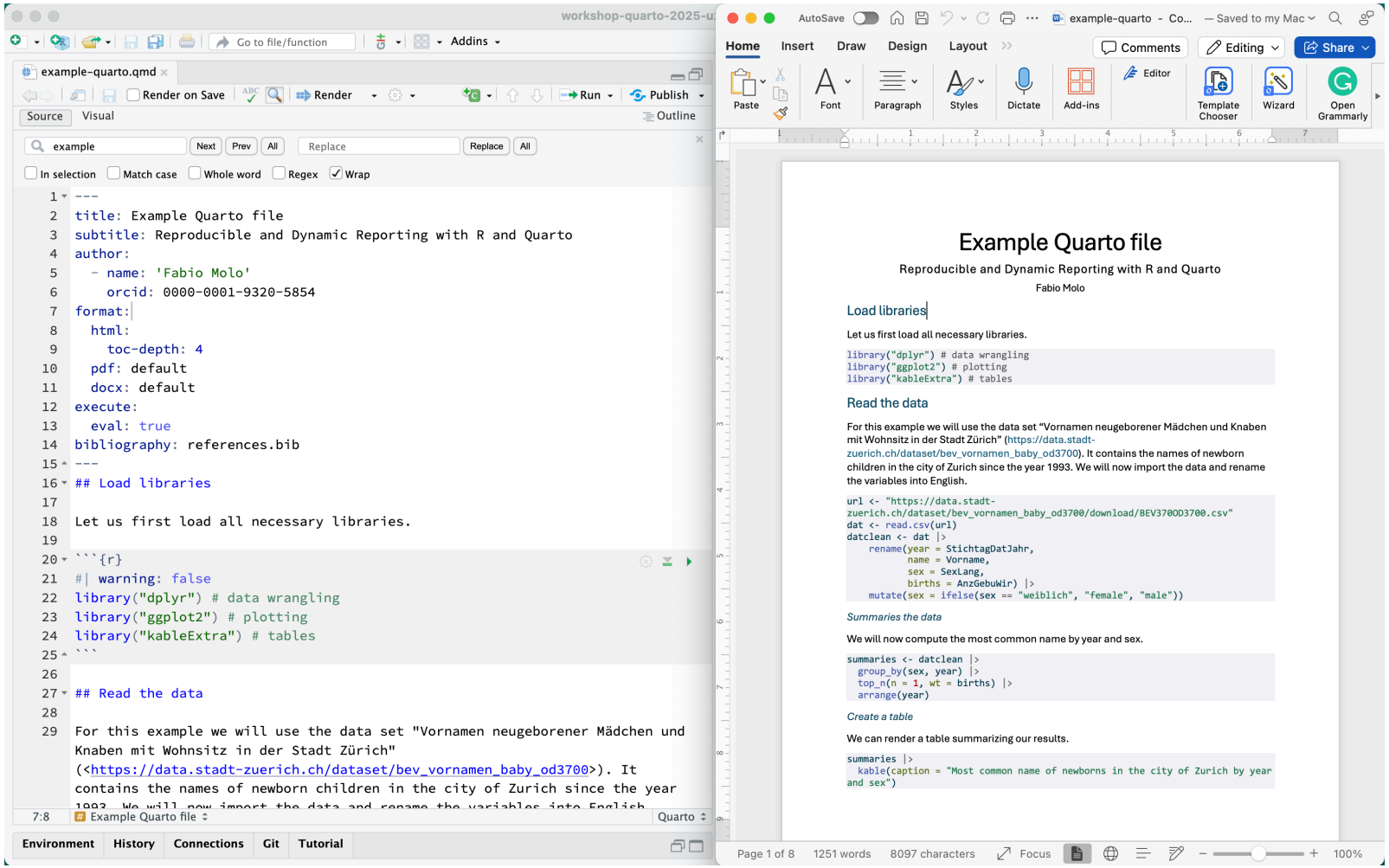
Molo, Fraga González, and Pawel (2025)

Quarto in RStudio: PDF Output



Molo, Fraga González, and Pawel (2025)

Quarto in RStudio: MS Word output



Molo, Fraga González, and Pawel (2025)

Weitere Literatur

- CRS Primer zu dynamischen Reports: https://doi.org/10.5281/zenodo.7565735
- Quarto Intro Tutorial: https://quarto.org/docs/get-started/hello/rstudio.html
- Quarto Authoring Tutorial: https://quarto.org/docs/get-started/authoring/
- Quarto Layout: https://quarto.org/docs/authoring/article-layout.html
- R4DS Kapitel zu Quarto: https://r4ds.hadley.nz/quarto
- Yihui Xie's Blog: With Quarto coming, is R Markdown going away? No.
- Reproducible manuscripts with Quarto: Folien von Mine Cetinkaya-Rundel
- Quarto/RMarkdown – What's different?: Folien von Ted Laderas

Literatur

Gould, Elliot, Hannah S. Fraser, Timothy H. Parker, Shinichi Nakagawa, Simon C. Griffith, Peter A. Vesk, Fiona Fidler, et al. 2023. “Same Data, Different Analysts: Variation in Effect Sizes Due to Analytical Decisions in Ecology and Evolutionary Biology,” October. https://ecoevorxiv.org/repository/view/6000/.

Molo, Fabio, Gorka Fraga González, and Samuel Pawel. 2025. “Workshop ‘Reproducible and Dynamic Reporting with r and Quarto – Getting Started’ at UZH.” https://doi.org/doi.org/10.5281/zenodo.14169002.

Open Science Collaboration. 2015. “Estimating the Reproducibility of Psychological Science.” *Science* 349 (6251): aac4716. https://doi.org/10.1126/science.aac4716.

Peikert, Aaron, and Andreas M. Brandmaier. 2021. “A Reproducible Data Analysis Workflow with r Markdown, Git, Make, and Docker.” *Quantitative and Computational Methods in Behavioral Sciences* 1. https://doi.org/10.5964/qcmb.3763.

Peikert, Aaron, and Hannes Diemerling. 2025. “Reproducible Research in r: A Workshop on How to Do the Same Thing More Than Once.” Zenodo. https://doi.org/10.5281/zenodo.15038481.

Peng, Roger D. 2011. “Reproducible Research in Computational Science.” *Science* 334