Reproducible CCS

Eine Einführung in die Theorie und Praxis reproduzierbarer computational methods

Dr. Tim Schatto-Eckrodt | 17.09.2025 | DGPuK Methoden25 @ Ostfalia

Agenda

- **Was ist Reproducible CCS?**
  - Begriffe & Konzepte
- **Warum ist Reproducible CCS sinnvoll?**
  - Ziele von Reproduzierbarkeit für Forschung und Forschende
- **Wie kann Reproducible CCS umgesetzt werden?**
  - Werkzeuge & Methoden
- **Praktisches Ausprobieren**

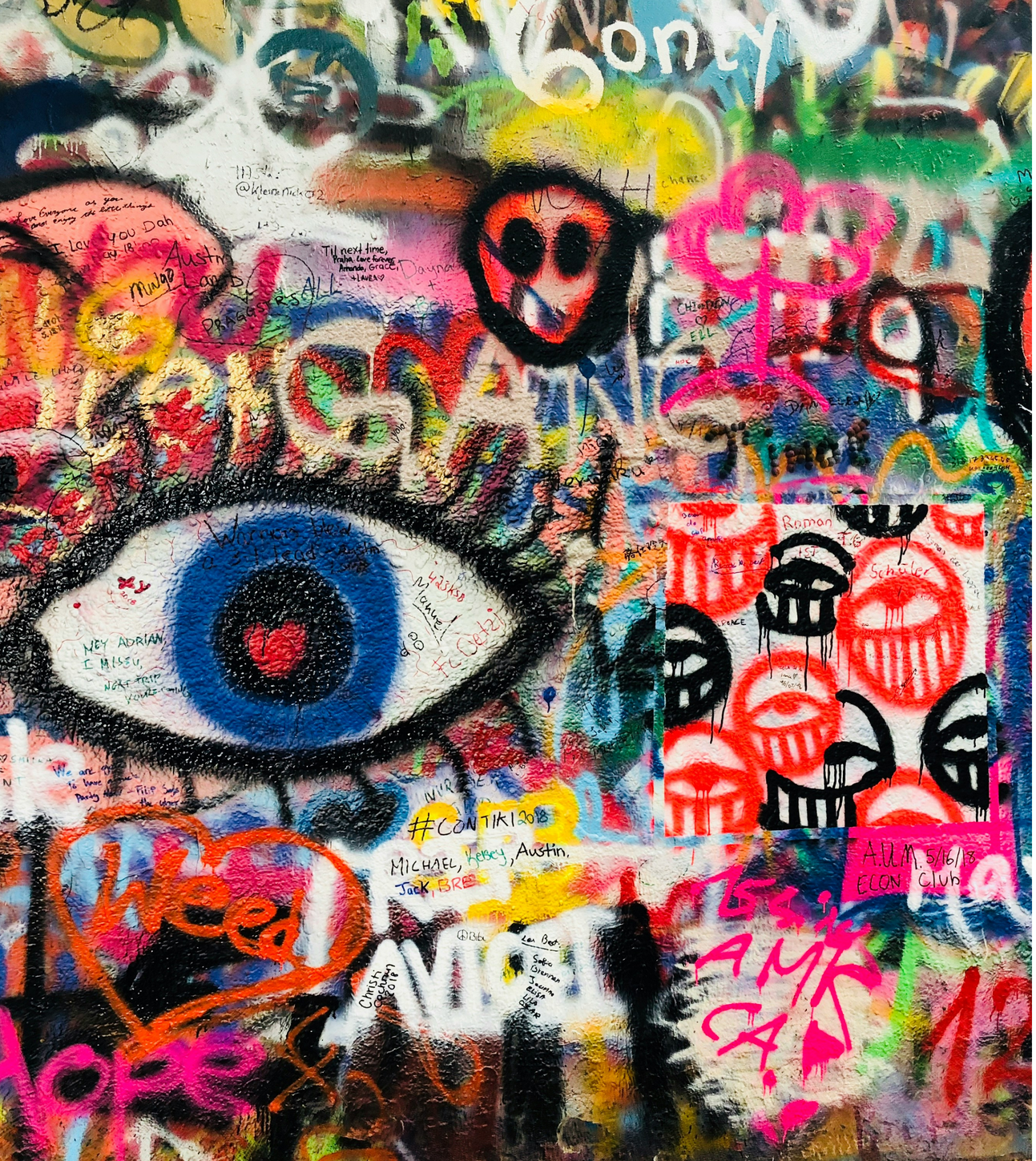
Projekt: Few-Shot (FS) Learning für die Automatische Inhaltsanalyse der Kommunikationswissenschaft (**FLACA**)

Gefördert durch:

Bundesministerium für Forschung, Technologie und Raumfahrt

Finanziert von der Europäischen Union

NextGenerationEU



Vorstellungsrunde

- Wie heißt ihr?
- Was sind eure Forschungsthemen?
- Welche Erfahrungen habt ihr mit CSS & Reproduzierbarkeit?
- Was erwartet ihr vom Workshop?

Was ist Reproducible CCS?

Begriffe & Konzepte

Was ist Reproducible CCS?

- **Computational Communication Science** (CCS) is a variant of Computational Social Science (CSS), an interdisciplinary scientific field in which contributions develop and test theories or provide systematic descriptions of human, organizational, and institutional behavior through the use of computational methods and practices. (Lazer et al., 2009)
  - A **deluge of digitally available data**, ranging from social media messages and other "digital traces" to web archives and newly digitized newspaper and other historical archives
  - **Improved tools** to analyze this data, including network analysis methods and automatic text analysis methods such as supervised text classification, topic modelling, word embeddings, and syntactic methods
  - The emergence of **powerful and cheap processing power**, and easy to use **computing infrastructure** for processing these data, including scientific and commercial cloud computing, sharing platforms such as Github and Dataverse, and crowd coding platforms such as Amazon MTurk and Crowdflower.

Was ist Reproducible CCS?

- "Two of the promises of computational methods for Communication scholars are that it makes research more **transparent** and more **reproducible**." (Chan et al., 2024, p. 2)
  - **Computational reproducibility**: "To what extent results in original studies can be reproduced based on data and code posted or provided by the original authors." (Dreber & Johannesson, 2024)
  - **Recreate reproducibility**: "To what extent results in original studies can be reproduced based on the information in the papers and access to the same raw data or data source, but without having access to the analysis code of the original study and/or the data set it was applied to." (Dreber & Johannesson, 2024)

Was ist Reproducible CCS?

- **Reproducible**: A result is reproducible when the same analysis steps performed on the same dataset consistently produces the same answer.
- **Replicable**: A result is replicable when the same analysis performed on different datasets produces qualitatively similar answers.

		Data	
		Same	Different
Analysis	Same	Reproducible	Replicable
	Different	Robust	Generalisable

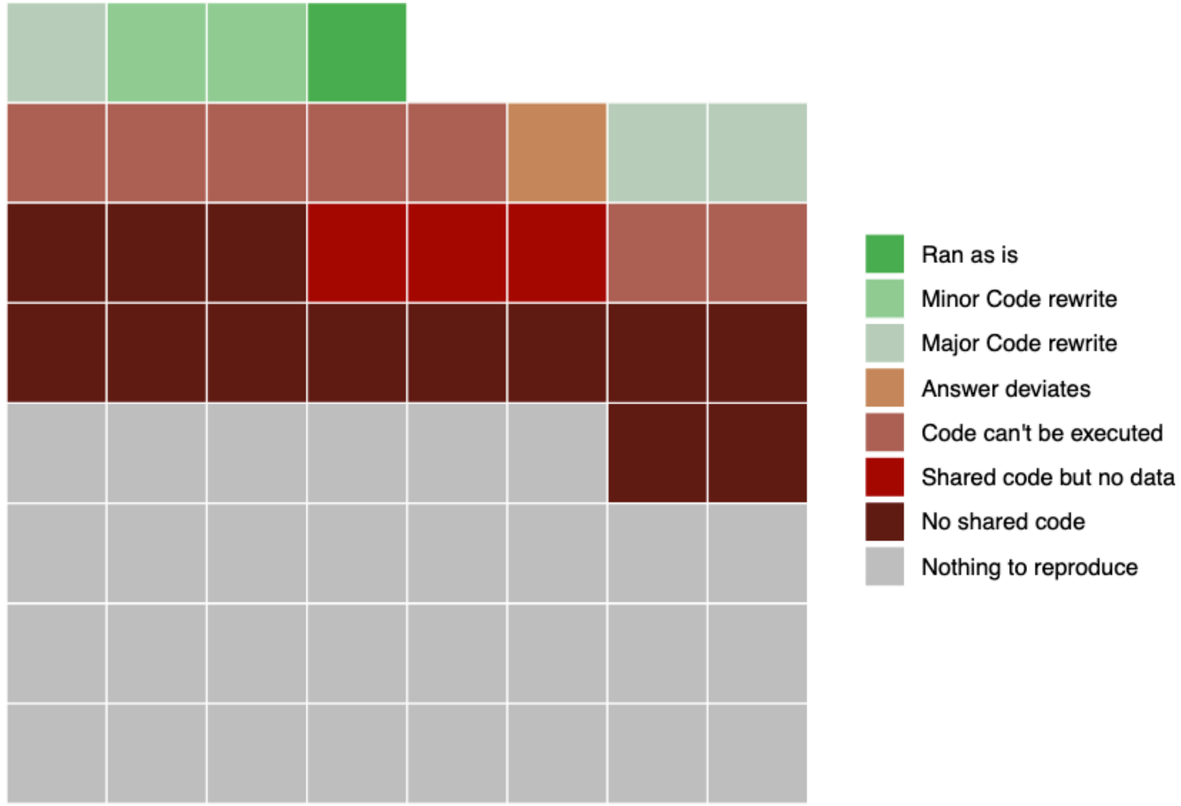
Was ist Reproducible CCS?

- Reproduzierbarkeit muss nicht dichotom verstanden werden (reproduziert oder nicht reproduziert), sondern kann ein **Spektrum** darstellen:
  - Wieviele Ergebnisse können reproduziert werden?
  - Wie stark weichen die Ergebnisse der Reproduktion von der ursprünglichen Studie ab?
  - Verändern sich die Schlüsse, die aus den Ergebnissen gezogen werden?

Was ist Reproducible CCS?

Wie steht es um die Reproduzierbarkeit von Ergebnissen der CCS:

- Chan et al. (2024) haben versucht 30 Studien zu reproduzieren, die in Computational Communication Research (CCR) erschienen sind:
  - 16/30 haben Code und Daten geteilt
  - 6/30 konnten (teilweise) reproduziert werden
  - Hindernisse: **Schlechte Dokumentation** & **fehlende (Social Media) Daten**



Was ist Reproducible CCS?

Comment | Published: 02 November 2023

Platform-controlled social media APIs threaten open science

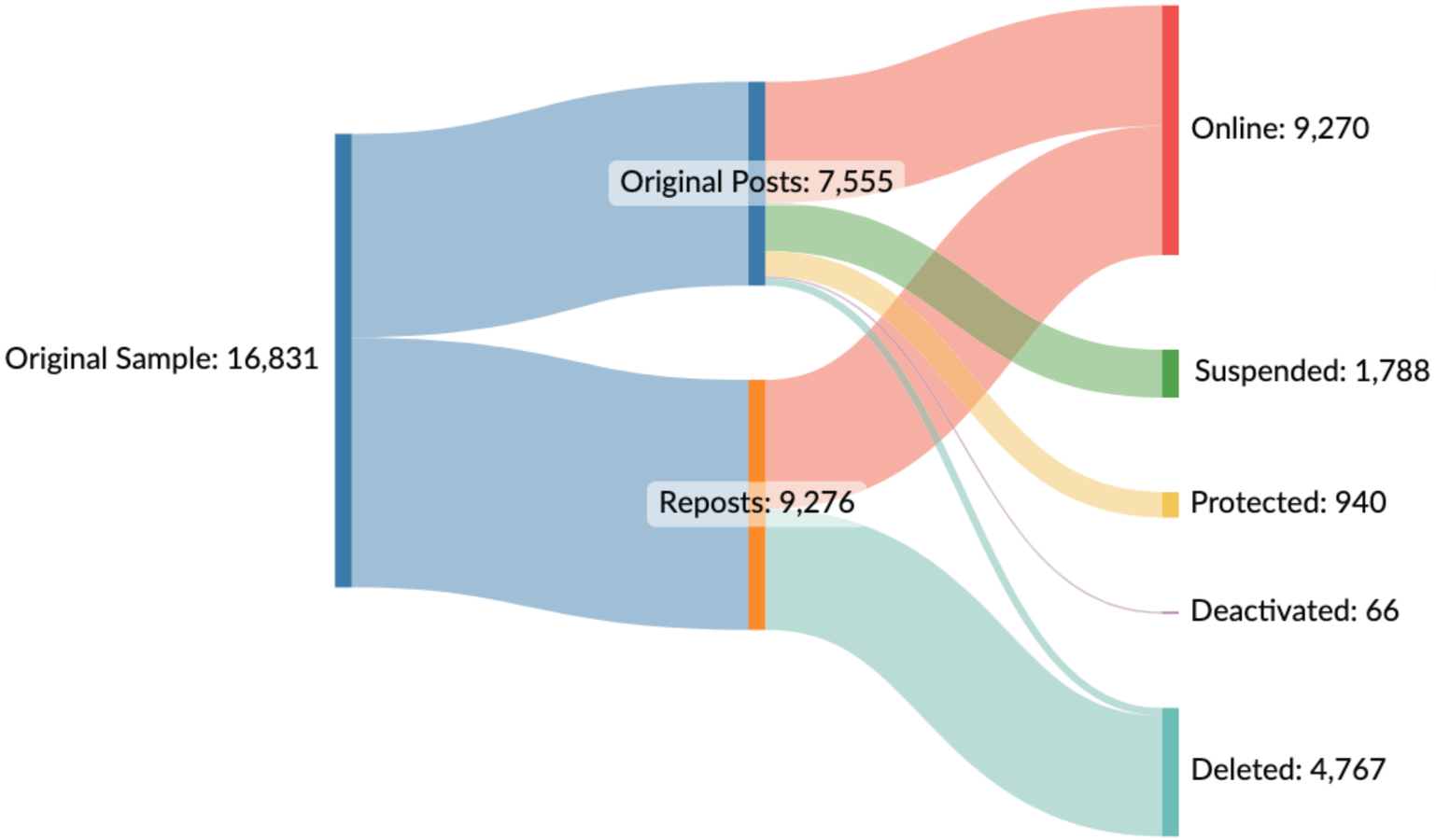
Brittany I. Davidson, Darja Wischerath, Daniel Racek, Douglas A. Parry, Emily Godwin, Joanne Hinds, Dirk van der Linden, Jonathan F. Roscoe, Laura Ayravainen & Alicia G. Cork

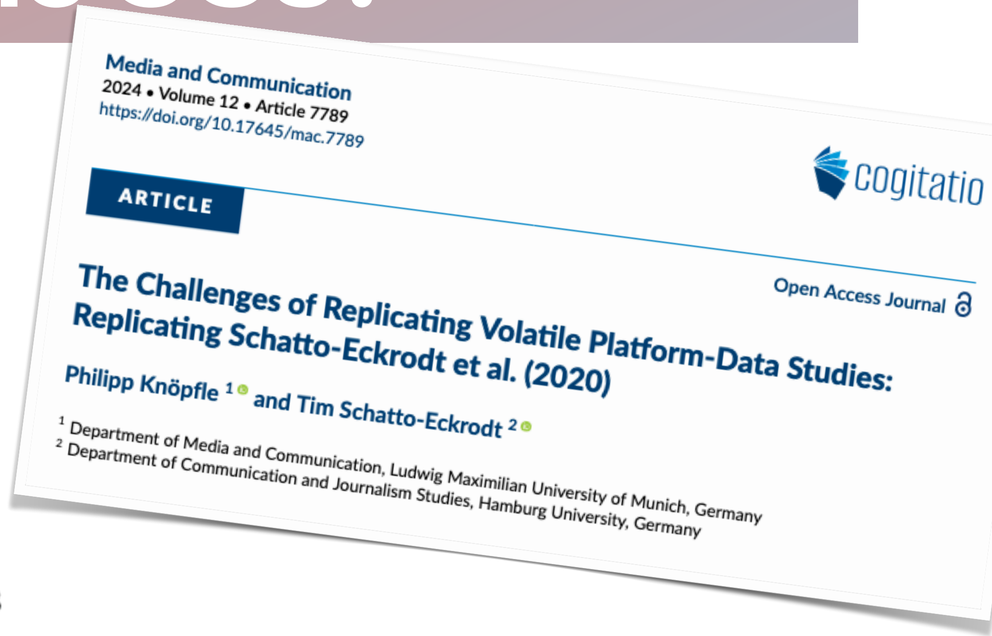
Nature Human Behaviour 7, 2054–2057 (2023) | Cite this article

1756 Accesses | 35 Citations | 103 Altmetric | Metrics

- Verfügbarkeit von Daten als Hindernis für reproduzierbare Analysen:
  - Oft dürften Social Media Daten nicht vollständig veröffentlicht werden (aus juristischen/ethischen Gründen)
  - Teilen von Status IDs nur begrenzt hilfreich, weil Daten schnell nicht mehr verfügbar sind: **10 % Verlust nach den ersten zwei Wochen** (Schatto-Eckrodt, 2022)
  - … und das Fehlen von Posts ist **nicht zufällig**, sondern abhängig von Inhalten und Akteuren (Küpfer, 2024; Schatto-Eckrodt, 2022)

Was ist Reproducible CCS?





Media and Communication
2024 • Volume 12 • Article 7789
https://doi.org/10.17645/mac.7789

ARTICLE

cogitatio

Open Access Journal

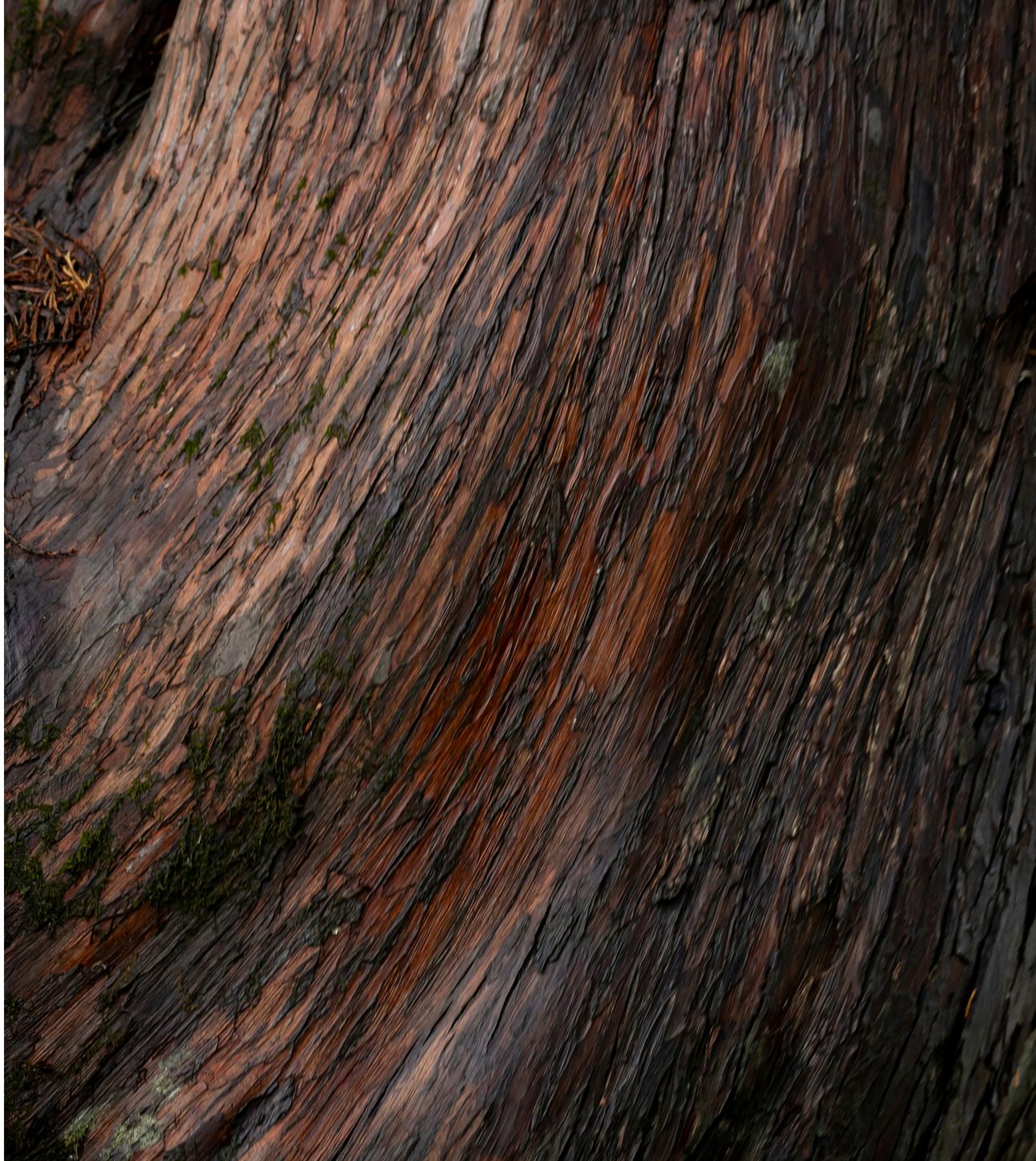
The Challenges of Replicating Volatile Platform-Data Studies: Replicating Schatto-Eckrodt et al. (2020)

Philipp Knöpfle [1] and Tim Schatto-Eckrodt [2]

[1] Department of Media and Communication, Ludwig Maximilian University of Munich, Germany
[2] Department of Communication and Journalism Studies, Hamburg University, Germany

Warum ist Reproducible CCS sinnvoll?

Ziele von Reproduzierbarkeit für Forschung und Forschende



Warum ist Reproducible CCS sinnvoll?

- Reproduzierbarkeit ist ein wichtiges Kriterium für **transparente**, intersubjektiv **nachvollziehbare** und **glaubwürdige** Ergebnisse.
- Sie ist also eine der Voraussetzungen für eine **selbstkorrigierende und kumulative** (Kommunikations-)Wissenschaft.

https://arxiv.org/html/2509.08299v1

e CCS sinnvoll?

m für **transparente**,
vürdige Ergebnisse.

Causal evidence of racial and institutional biases in accessing paywalled articles and scientific data

Hazem Ibrahim[1+], Fengyuan Liu[1+], Khalid Mengal[1], Aaron R. Kaufman[2], Yasir Zaki[1*], and Talal Rahwan[1*]

[1]Computer Science, Science Division, New York University Abu Dhabi, UAE.
[2]Social Science Division, New York University Abu Dhabi, UAE

[+]Joint first author

[*]Corresponding authors. E-m

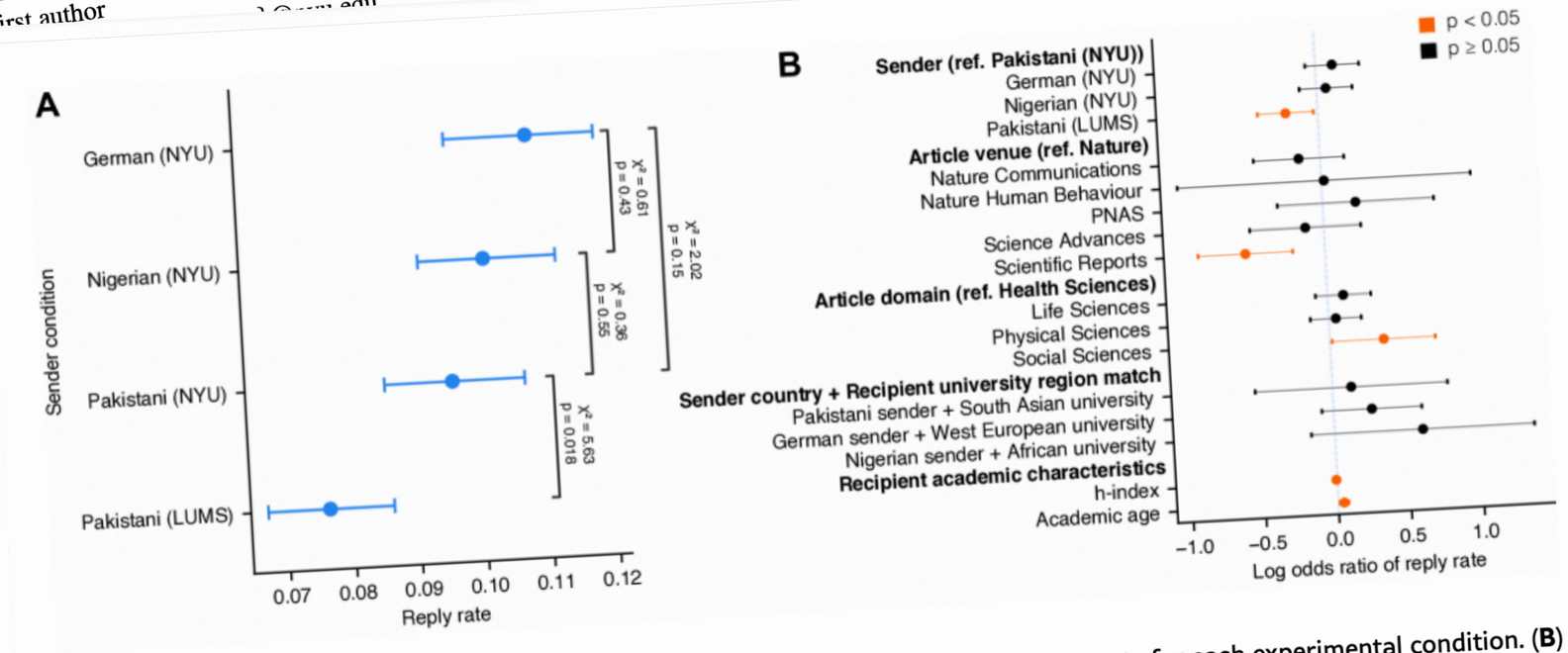


Figure 4: **Summary of dataset audit experiment results.** (**A**) Reply rates for datasets requests for each experimental condition. (**B**) Logistic regression estimated log-odds coefficients of independent variables predicting whether a given email was replied to.

„Our analysis of 250 million articles reveals that researchers in the Global South cite paywalled papers and upon-request datasets at significantly lower rates than their Global North counterparts, and that ***these access gaps are associated with reduced knowledge breadth and scholarly impact****."*

https://arxiv.org/html/2509.08299v1

Warum ist Reproducible CCS sinnvoll?

- Reproduzierbarkeit ist ein wichtiges Kriterium für **transparente**, intersubjektiv **nachvollziehbare** und **glaubwürdige** Ergebnisse.
- Sie ist also eine der Voraussetzungen für eine **selbstkorrigierende und kumulative** (Kommunikations-)Wissenschaft.

Reproduzierbarkeit hat aber auch Vorteile für uns Wissenschaftler*innen!

- Analysen können wiederholt und für neue Forschung angepasst werden
- Gemeinsame Arbeit an Projekten wird vereinfacht
- Fehlersuche wird erleichtert

“Reproducibility is just collaboration with people you don't know, *including yourself next week*”

Prof. Philip B. Stark, University of California

Warum ist Reproducible CCS sinnvoll?

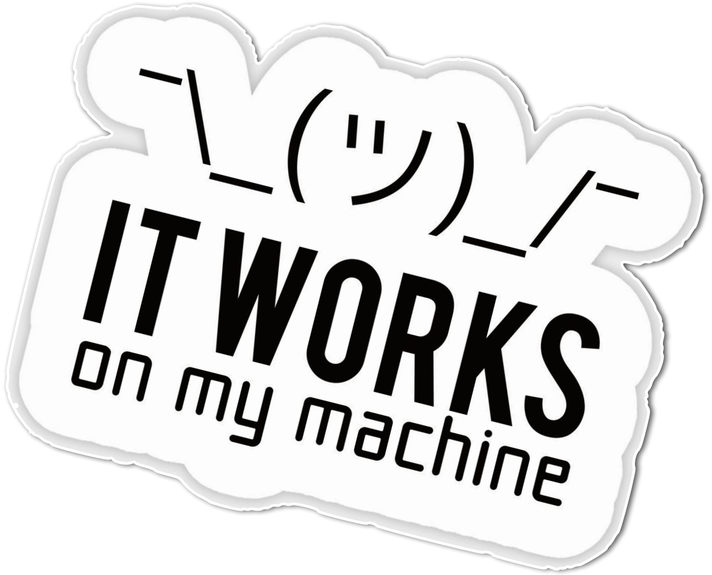
Ziele

- Die Ergebnisse einer empirischen Studie sollten für Dritte wiederholbar sein:
  code + data => same results*

* die Ergebnisse der Analysen sollten die Aussagen aus der veröffentlichten Studie unterstützen

Hürden

- Daten ggf. nicht teilbar
- Analyse-Code läuft nicht überall fehlerfrei
- Relevante Daten und Infos sind verstreut und unvollständig
- ...



Warum ist Reproducible CCS sinnvoll?

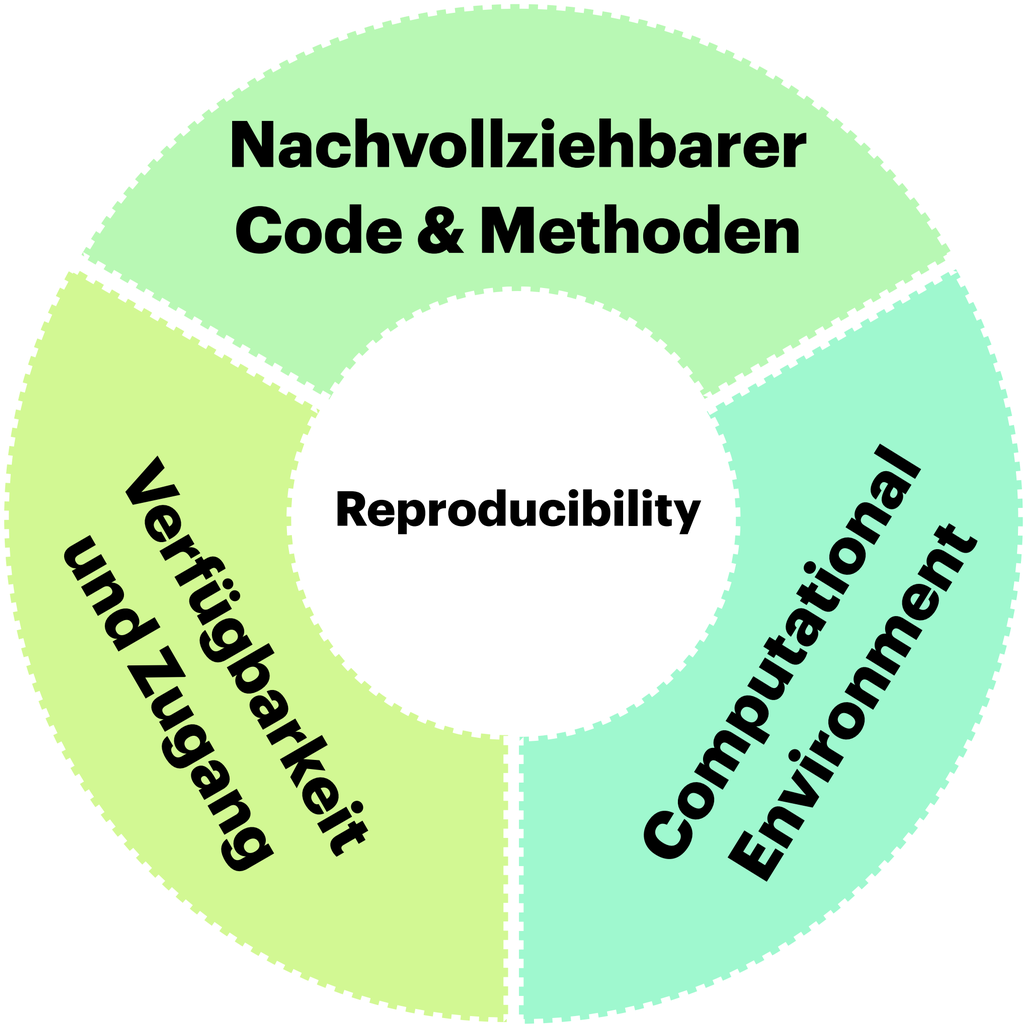
Welche Hürden nehmt ihr wahr?

Wo lagen bisher eure Probleme?

Wie kann Reproducible CCS umgesetzt werden?

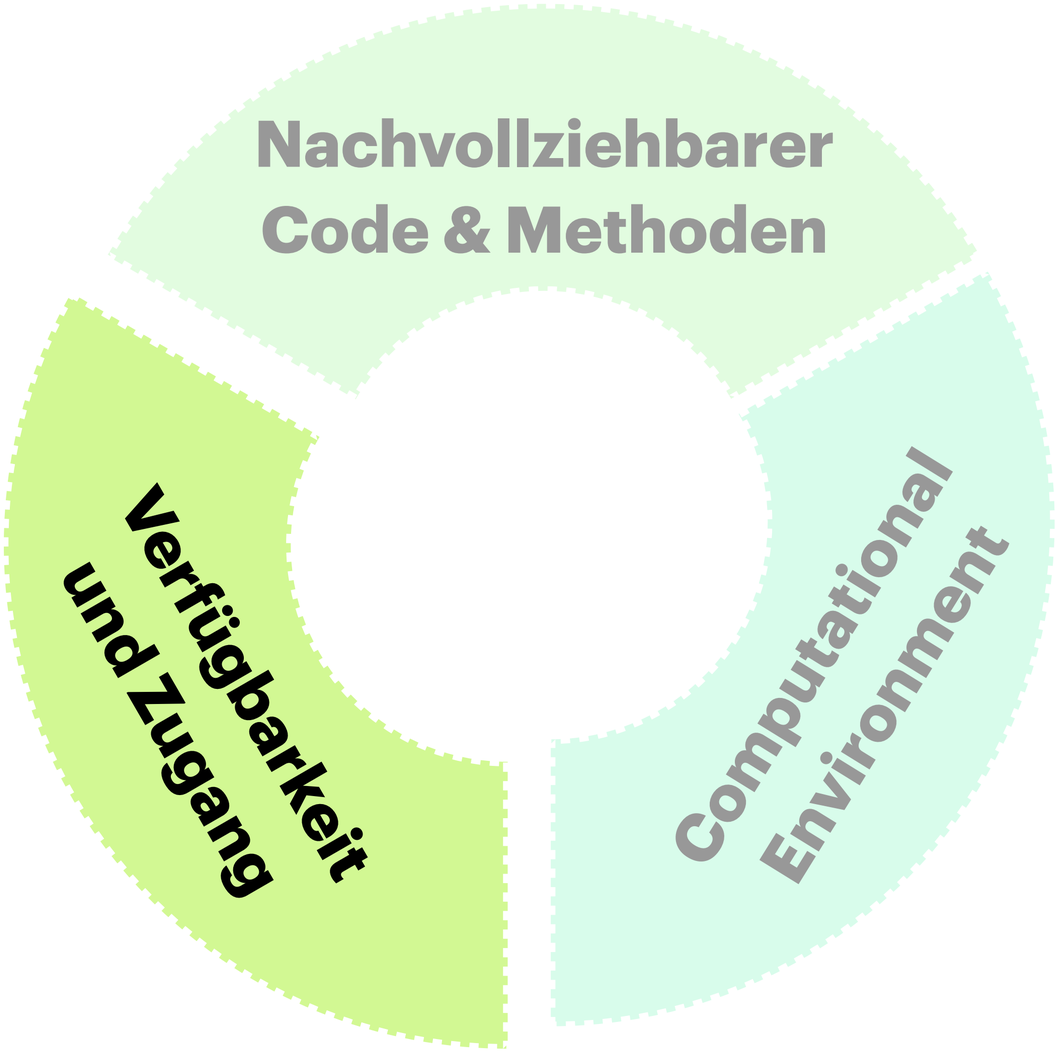
Werkzeuge & Methoden

Wie kann Reproducible CCS umgesetzt werden?



(Chan et al., 2024; Knöpfle & Schatto-Eckrodt, 2024)

Wie kann Reproducible CCS umgesetzt werden?



Verfügbarkeit und Zugang

(wo findet man die Materialien und wie bekommt man Zugang?)

- Wie kann ich Daten und Code teilen?
- Welche Zugangsbeschränkungen sind möglich?
- Wie gestaltet man den Datenzugang nach den FAIR-Prinzipien?

(Chan et al., 2024; Knöpfle & Schatto-Eckrodt, 2024)

Wie kann Reproducible CCS umgesetzt werden?

Hosted git

- *git* ist eine Software, mit der man Code versionieren und managen kann
- Es gibt verschiedene Anbieter, die git als Service hosten:
  - GitHub (for-profit, Microsoft)
  - GitLab (for-profit, GitLab Inc.)
  - Codeberg (non-profit)
  - SourceForge
  - Bitbucket
  - ...

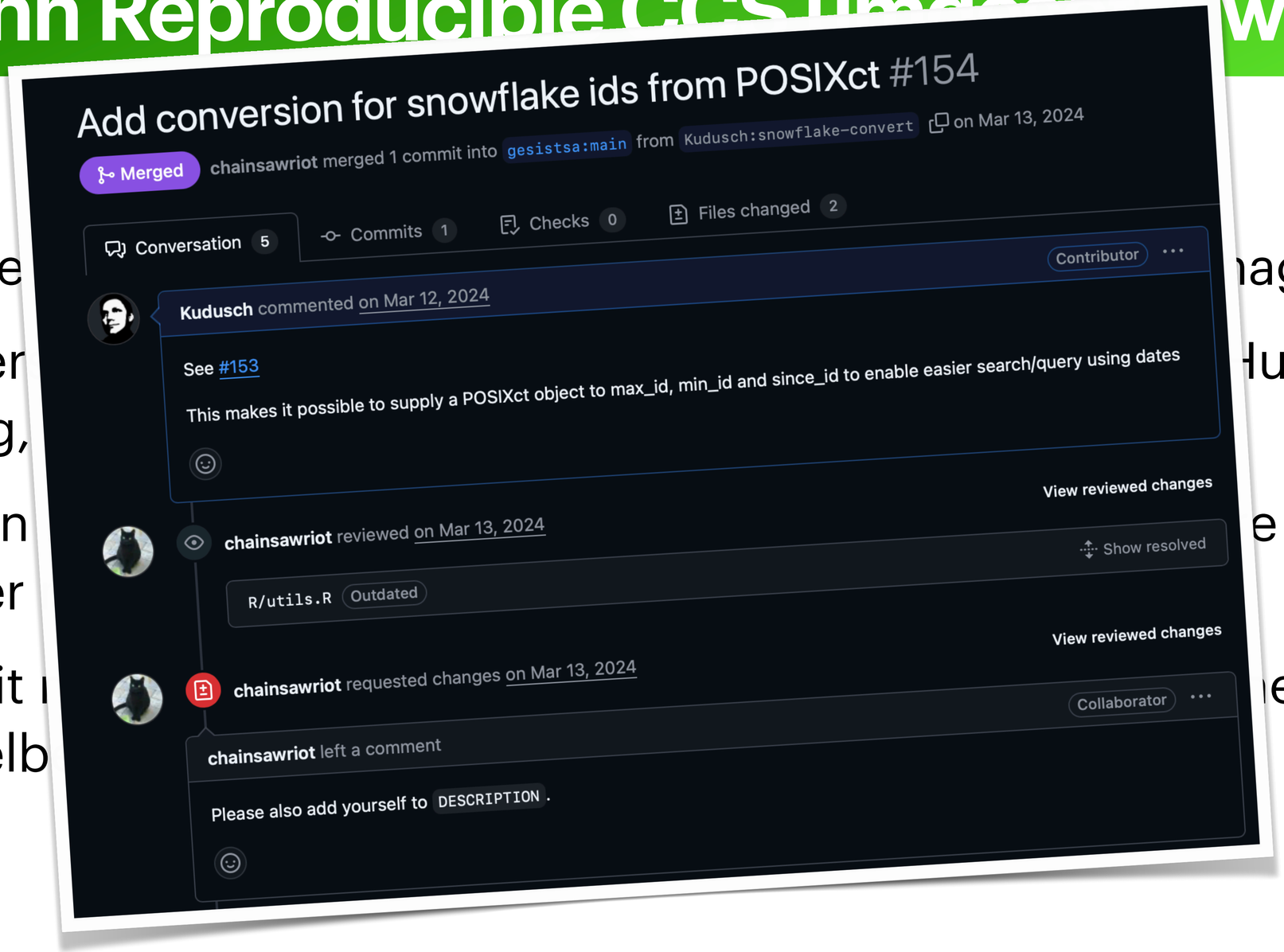
Wie kann Reproducible CCS umgesetzt werden?

Hosted git

- ***git*** ist eine Software, mit der man Code versionieren und managen kann
- Es gibt verschiedene Anbieter, die git als Service hosten (GitHub, GitLab, Codeberg, usw.)
- Auf diesen Plattformen teilt man am besten Code (oder andere plain-text-files), aber nicht unbedingt große Datensätze
- Hosted git macht die Kollaboration zwischen mehreren Personen möglich, die am selben Analysecode arbeiten

Wie kann Reproducible CCS umgesetzt werden?

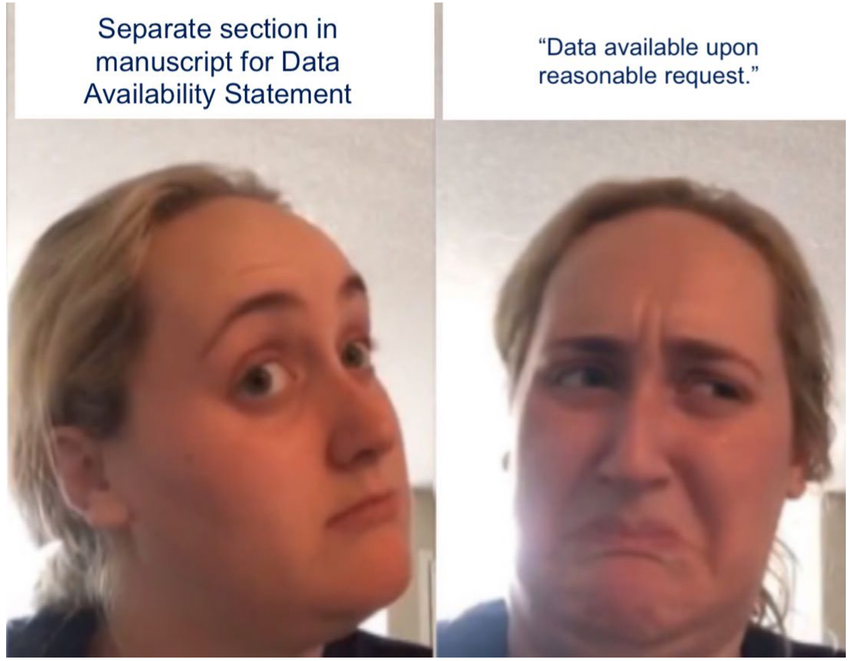
- ***git*** ist eine … nagen kann
- Es gibt ver… Hub, GitLab, Codeberg, …
- Auf diesen … e plain-text-files), aber …
- Hosted git … en möglich, die am selb…



Wie kann Reproducible CCS umgesetzt werden?

Forschungsdatenrepositorien

- Um Datensätze zu teilen und zu archivieren eignen sich spezielle Repositorien zum Forschungsdatenmanagement, z.B.:
  - GESIS Data Services
  - Open Science Framework
  - Zenodo
  - … und viele mehr: re3data.org



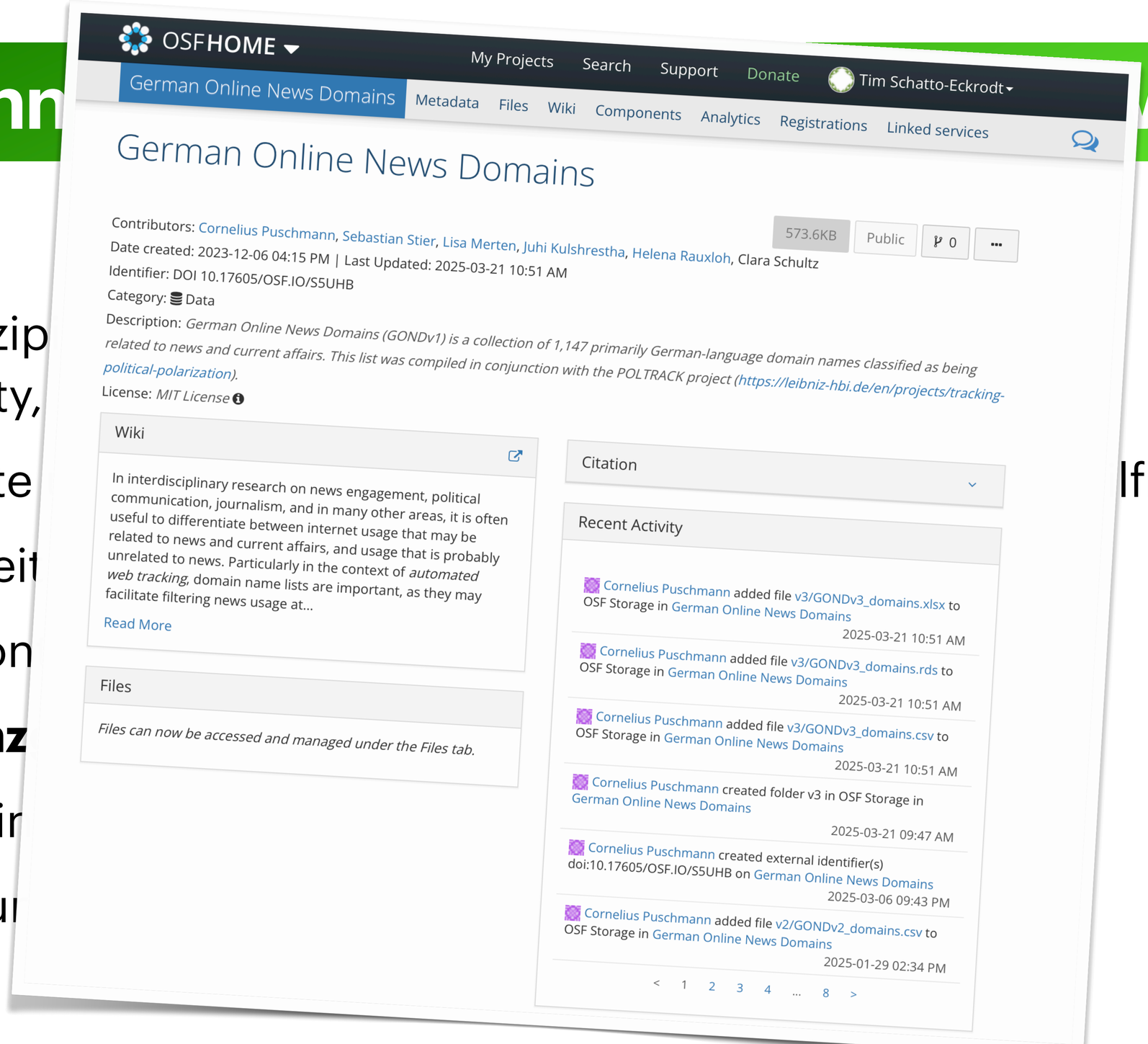
Wie kann Reproducible CCS umgesetzt werden?

Forschungsdatenrepositorien

- FAIR-Prinzipien sollten eingehalten werden:
  (**F**indability, **A**ccessibility, **I**nteroperability, and **R**eusability)
- Dauerhafte und eindeutige **Identifikatoren** (wie DOIs) sind hilfreich
- Möglichkeit **Lizenzierungen** vorzunehmen
- Zugriffskontrollen, bzw. **Zugriffsmanagement**
- **Persistenzgarantien** für langfristige Archivierung
- Standort in der **EU**
- DSGVO- und **Datenschutz**konformität

Wie kann … werden?

- FAIR-Prinzip
  (**F**indability,
- Dauerhafte … lfreich
- Möglichkeit
- Zugriffskon
- **Persistenz**
- Standort in
- DSGVO- u



OSFHOME

My Projects | Search | Support | Donate | Tim Schatto-Eckrodt

German Online News Domains | Metadata | Files | Wiki | Components | Analytics | Registrations | Linked services

German Online News Domains

573.6KB | Public | 0

Contributors: Cornelius Puschmann, Sebastian Stier, Lisa Merten, Juhi Kulshrestha, Helena Rauxloh, Clara Schultz

Date created: 2023-12-06 04:15 PM | Last Updated: 2025-03-21 10:51 AM

Identifier: DOI 10.17605/OSF.IO/S5UHB

Category: Data

Description: *German Online News Domains (GONDv1) is a collection of 1,147 primarily German-language domain names classified as being related to news and current affairs. This list was compiled in conjunction with the POLTRACK project (https://leibniz-hbi.de/en/projects/tracking-political-polarization).*

License: *MIT License*

Wiki

In interdisciplinary research on news engagement, political communication, journalism, and in many other areas, it is often useful to differentiate between internet usage that may be related to news and current affairs, and usage that is probably unrelated to news. Particularly in the context of *automated web tracking*, domain name lists are important, as they may facilitate filtering news usage at...

Read More

Files

Files can now be accessed and managed under the Files tab.

Citation

Recent Activity

Cornelius Puschmann added file v3/GONDv3_domains.xlsx to OSF Storage in German Online News Domains
2025-03-21 10:51 AM

Cornelius Puschmann added file v3/GONDv3_domains.rds to OSF Storage in German Online News Domains
2025-03-21 10:51 AM

Cornelius Puschmann added file v3/GONDv3_domains.csv to OSF Storage in German Online News Domains
2025-03-21 10:51 AM

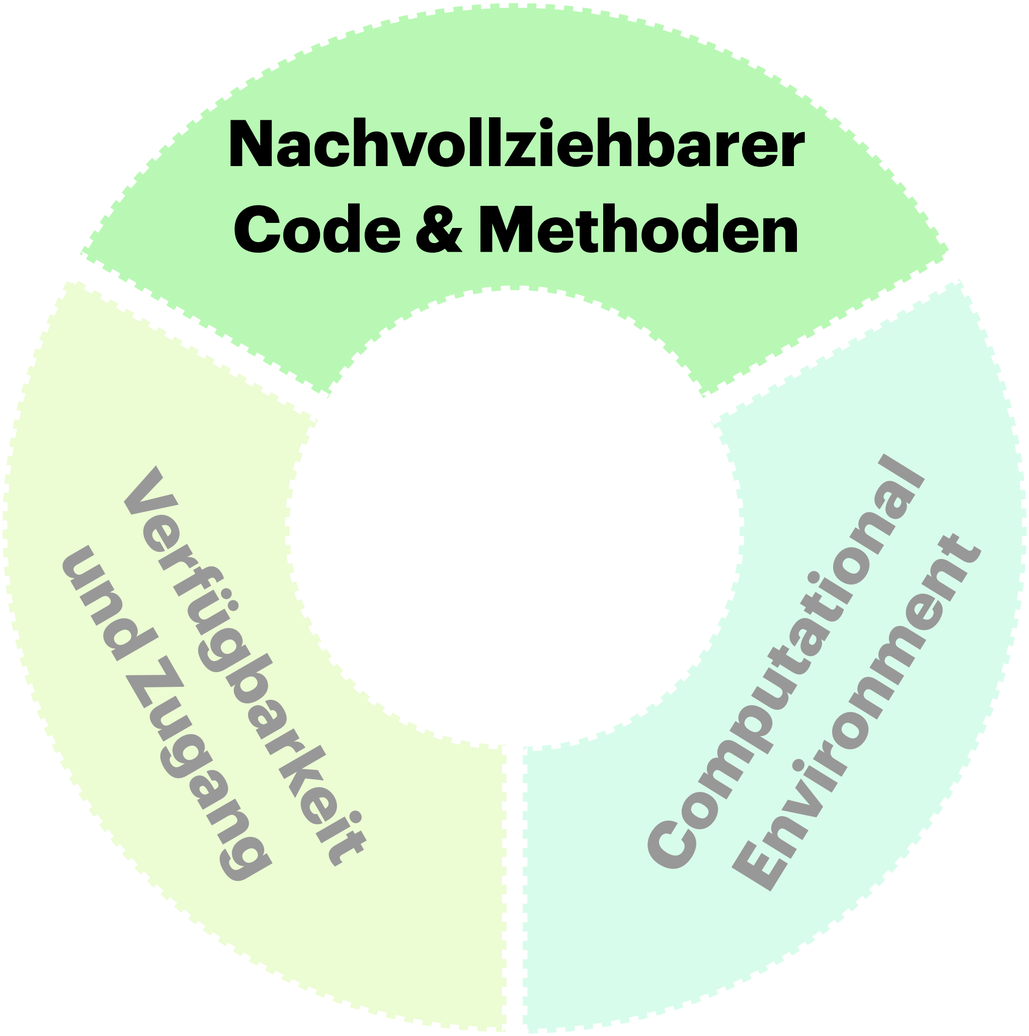
Cornelius Puschmann created folder v3 in OSF Storage in German Online News Domains
2025-03-21 09:47 AM

Cornelius Puschmann created external identifier(s) doi:10.17605/OSF.IO/S5UHB on German Online News Domains
2025-03-06 09:43 PM

Cornelius Puschmann added file v2/GONDv2_domains.csv to OSF Storage in German Online News Domains
2025-01-29 02:34 PM

< 1 2 3 4 ... 8 >

Wie kann Reproducible CCS umgesetzt werden?



Nachvollziehbarer Code & Methoden

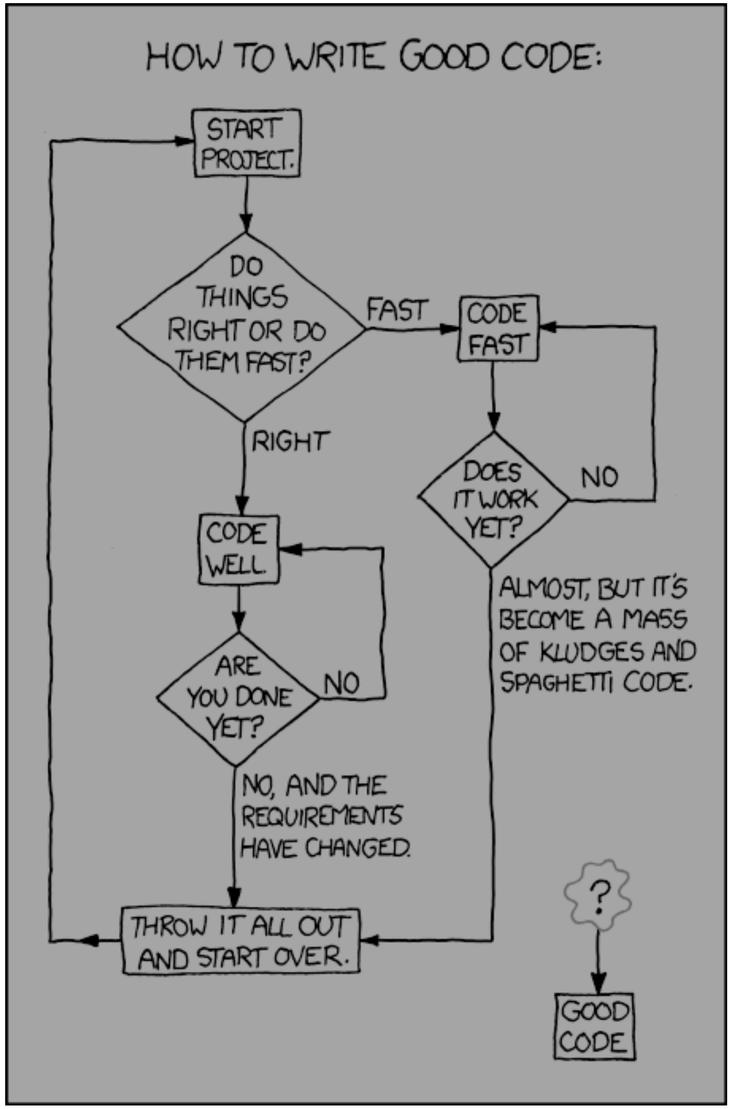
(Verständlichkeit und Dokumentation der Analyse)

- Welche Methoden wurden wie angewendet?
- Welche Skripte/Teile des Codes gehören zu welchen Ergebnissen?
- Wie soll der Code ausgeführt werden?

(Chan et al., 2024; Knöpfle & Schatto-Eckrodt, 2024)

Wie kann Reproducible CCS umgesetzt werden?

- Im Kern einer reproduzierbaren Analyse steht ***guter Code***.
- Um Code zu produzieren, der besonders gut *lesbar*, *verständlich*, *nicht-fehleranfällig* ist und uns bei der *Reproduktion* nicht im Weg steht, können wir uns an ein paar Regeln halten.
- Wir können hier von den Regeln der Software-Entwicklung lernen, aber: *Nicht alles, was für Software Engineers sinnvoll ist, ist auch für uns (Computational) Social Scientist wichtig*.



Wie kann Reproducible CCS umgesetzt werden?

Ein ***Research Compendium*** ist eine Sammlung aller digitalen Teile eines Forschungsprojekts:

Daten, Code, Texte (Protokolle, Fragebögen, Stimuli, etc.)

Die Sammlung wird so angelegt, dass alle Ergebnisse problemlos reproduziert werden können (Nuest et al., 2017; Gentleman & Temple Lang, 2007).

Wie kann Reproducible CCS umgesetzt werden?

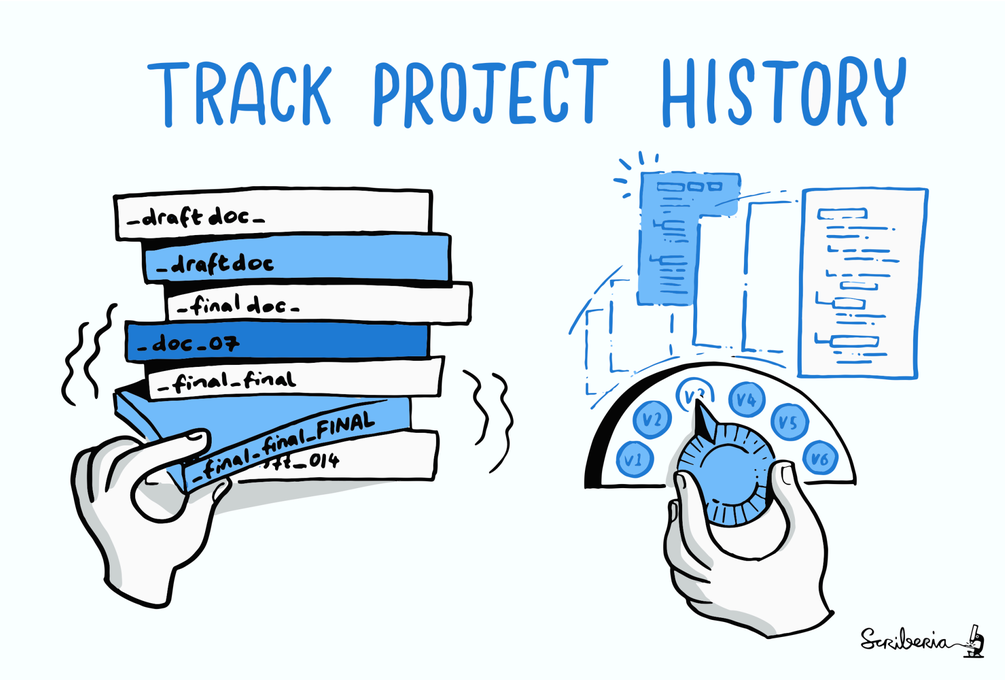
Organisation von Code

- Ordnerstrukturen können helfen, Code und Daten zu organisieren

```
.
├── _quarto.yml
├── analysis.Rproj
├── data
├── docs
├── output
├── README.md
├── renv
├── renv.lock
└── scripts
```

- analysis.Rproj → RStudio-Projekt
- data → Daten (optional: schreibgeschützt)
- docs → Dokumente (Quarto, RMarkdown, etc.)
- output → Output (Tabellen, Grafiken, Modelle, etc.)
- README.md → Info/Anleitung zum Projekt
- scripts → Skripte (Preprocessing, einzelne Analysen, etc.)

Wie kann Reproducible CCS umgesetzt werden?



git (oder andere Versionskontrollsysteme) helfen uns, auch bei Änderungen am Code den Überblick zu behalten.

Änderungen, die im Laufe der Zeit an einer Datei vorgenommen wurden, werden aufgezeichnet, sodass deren Verlauf verfolgt, alle Änderungen überprüft und zu früheren Versionen zurückgekehrt werden kann.

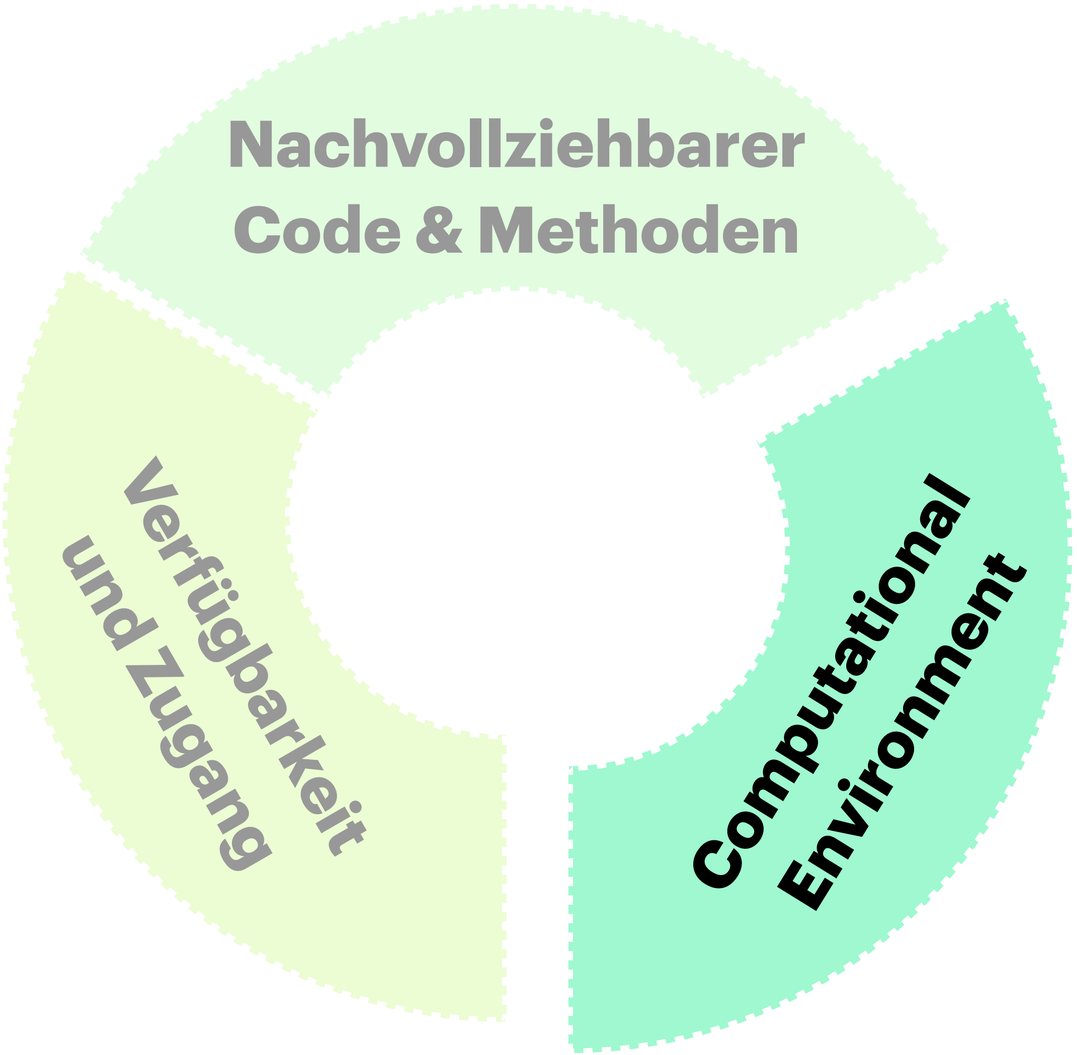
Besonders hilfreich, wenn man nicht alleine an etwas arbeitet!

Wie kann Reproducible CCS umgesetzt werden?

Allgemeine Tipps:

- Nutzt immer **relative Pfade** (`/Users/max/work/reproducible_ccs/reproducible_template/data.csv` vs `data.csv`)
- **Kommentiert** euren Code für euch und andere!
  (`# remove all participants under 18 and with less than 60s time in survey`)
- Setzt **Seeds** in euren Skripten, um den Zufall einzufrieren: `set.seed(1234)`
- **Proaktive Reproduzierbarkeit:**
  - Bestehende Projekte reproduzierbar zu machen ist aufwändiger, als von vornherein Reproduzierbarkeit mitzudenken!
  - *interaktiv entwickeln, standalone testen*

Wie kann Reproducible CCS umgesetzt werden?



Computational Environment

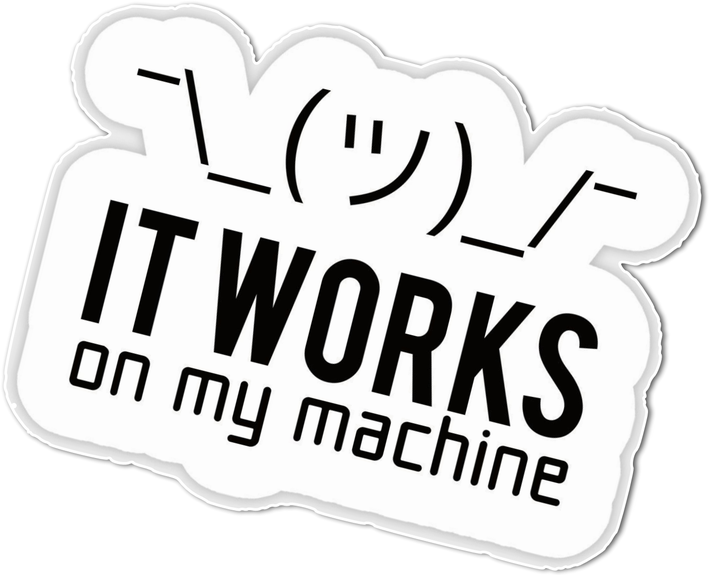
(die Umgebung in der die Analyse durchgeführt wurde)

- Mit welcher Software wurde die Analyse durchgeführt?
- Welche Pakete wurden verwendet?
- Welche zusätzlichen Infos/Daten standen zur Verfügung?

(Chan et al., 2024; Knöpfle & Schatto-Eckrodt, 2024)

Wie kann Reproducible CCS umgesetzt werden?

- Es gibt keine Garantie dafür, dass Analysen auch auf anderen Computern laufen, auf denen sie nicht entwickelt wurden.
- Jeder Computer verfügt über eine spezifische Rechenumgebung (***environment***), die aus seinem *Betriebssystem*, der *installierten Software*, den *Versionen* der installierten Softwarepakete und anderen *Funktionen* besteht.
- Damit Analysen reproduzierbar werden, muss die Rechenumgebung, in der sie durchgeführt wurde, so erfasst werden, dass andere sie **nachbilden können**.



Wie kann Reproducible CCS umgesetzt werden?

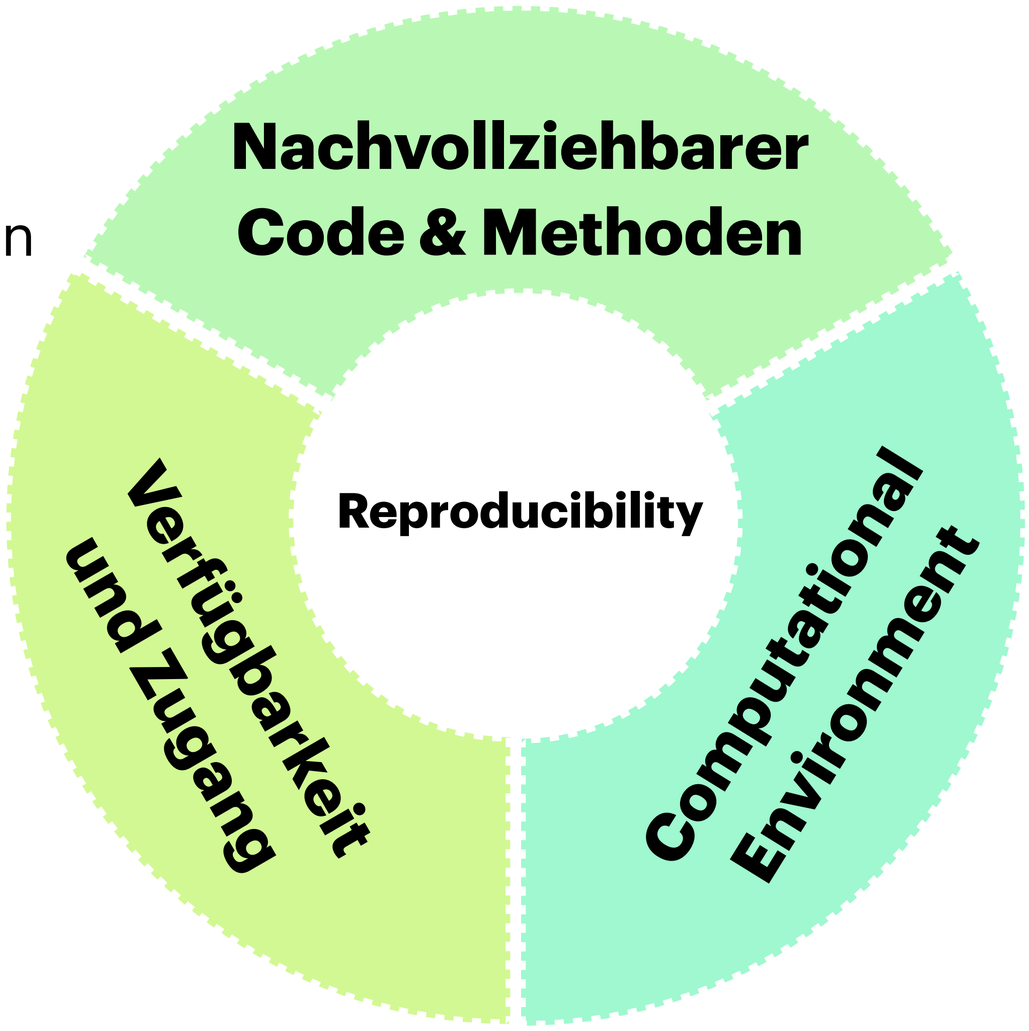
Sparsamkeit bei Imports!

- **renv/pacman:** R-Packete, mit dem man eine Liste der installierten Software (und den jeweiligen Versionen) erstellen kann. Mit dieser Liste (renv.lock) können wir den beschrieben Zustand auch wiederherstellen!
- **docker/nix:** Software-Frameworks mit denen vollständige Systeme als sog. Container verwaltet werden. Mit docker und nix notieren wir nicht nur welche Software wir genutzt haben, sondern erstellen sozusagen einen Snapshot des gesamten Systems.
- **sessionInfo():** Die einfachste Variante, die ohne weitere Software auskommt. Speichert den „Fingerabdruck“ der aktuellen Session, kann aber nichts wiederherstellen.

```
writeLines(capture.output(sessionInfo()), ''sessionInfo.txt'')
```

Wie kann Reproducible CCS umgesetzt werden?

- ***Code*** schreiben, den wir selbst auch in zwei Wochen noch verstehen (z.B. mit Kommentaren & Struktur)
- Die ***Analysen*** verfüg- und nachvollziehbar machen (z.B. mit hosted git und Research Compendia)
- Das ***computational environment*** beschreiben und Wiederherstellung erleichtern (z.B. mit `renv`)
- Datensätze und Materialen ***FAIR veröffentlichen*** (z.B. in OSF)
- **Reproduzierbarkeit direkt von Anfang an mitdenken!** (z.B. mit Kolleg*innen testen)



Wie kann Reproducible CCS umgesetzt werden?

- Weitere Links und Tipps:
  - https://www.gesis.org/fileadmin/admin/Dateikatalog/pdf/guides/02_Bleier_computational_reproducibility.pdf
  - https://book.the-turing-way.org/reproducible-research/compendia

Wie kann Reproducible CCS umgesetzt werden?

“Le mieux est le mortel ennemi du bien”

(The best is the mortal enemy of the good)

Montesquieu, 1726

Praktisches Ausprobieren

Ein Template für reproduzierbare Analysen

Praktisches Ausprobieren

Template für reproduzierbare Analysen

README

A template for reproducible css

This repository holds the research compendium for the reproducible_css workshop. Data is stored on OSF.

To reproduce the three reports located in the `docs` directory, follow the following steps:

- Install the `renv` package

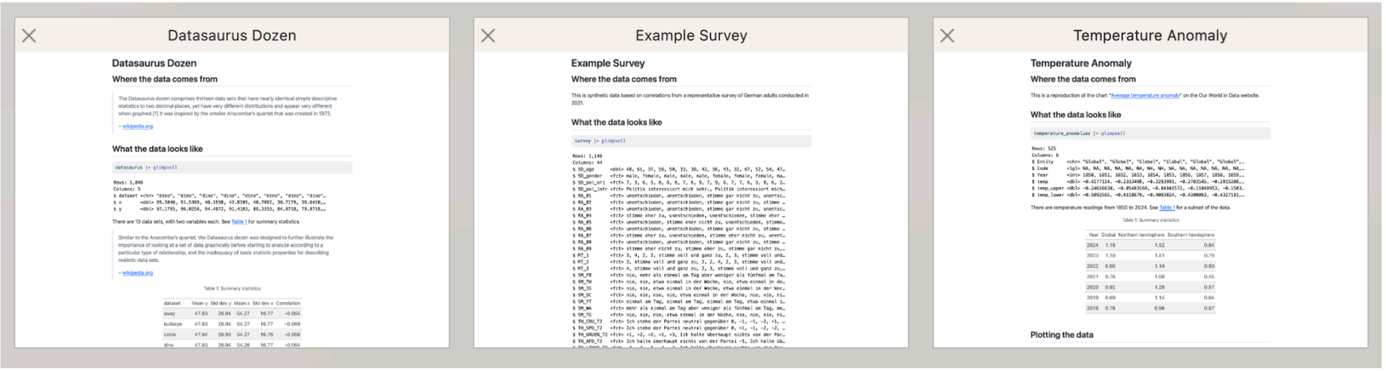
```
install.packages("renv")
```

- Install RStudio
- Run these commands in the terminal to setup the project:

```
git clone https://github.com/Kudusch/reproducible_template
cd reproducible_template
R --no-echo -e 'renv::init()'
R --no-echo -e 'renv::restore()'
```

- Open `analysis.Rproj` and render the three reports (`datasaurus.qmd` , `temperature_anomaly.qmd`)

https://github.com/Kudusch/reproducible_template



https://kudusch.github.io/reproducible_template/temperature_anomaly.html

https://kudusch.github.io/reproducible_template/datasaurus.html

https://kudusch.github.io/reproducible_template/example_survey.html

git + OSF + renv + quarto + GitHub Pages