SFB 1294
Data Assimilation

Woran Reproduzierbarkeit scheitert – Erkenntnisse und Lösungen aus dem SFB 1294 „Data Assimilation“

Hendrik Geßner, Anja Seegebrecht

Hintergrund

SFB 1294 „Data Assimilation“

- 11 Forschungsprojekte (13 Projekte ab 2021)
- Drei Bereich
  - Theorie, Anwendung, Verwaltung und Infrastruktur

DFG legt Wert auf gutes Forschungsdatenmanagement

INF-Projekt Z03

- Infrastruktur bereitstellen
- Workshops anbieten
- Forschende unterstützen
- Qualitäts- und Bewertungs-modell entwickeln

→ Zustand und Bedarf ermitteln

Reproduzierbarkeit im Detail

SFB 1294
Data Assimilation

Wiederholen (Repeat)	Reproduzieren (Reproduce)	Replizieren (Replicate)
• Gleiche Methode	• Gleiche Methode	• Gleiche Methode
• Gleiche Ausrüstung	• Gleiche Ausrüstung	• Andere Ausrüstung
• Gleiche Daten u. Code	• Gleiche Daten u. Code	• Andere Daten u. Code
• Gleiches Labor	• Gleiches oder anderes Labor	• Anderes Labor
• Gleiche Wissenschaftl.	• Andere Wissenschaftl.	• Andere Wissenschaftl.
• Anderer Durchlauf	• Anderer Durchlauf	• Anderer Durchlauf

Reproduzierbarkeitsrate ist lächerlich niedrig

1,7%

Stagge, James H.; Rosenberg, David E.; Abdallah, Adel M.; Akbar, Hadia; Attallah, Nour A.; James, Ryan (2019): Assessing data availability and research reproducibility in hydrology and water resources. In: *Scientific data* 6, S. 190030. DOI: 10.1038/sdata.2019.30.

Reproduzierbarkeitsrate ist lächerlich niedrig

Wissenschaftliche Gemeinschaft

1,7%

Stagge, James H.; Rosenberg, David E.; Abdallah, Adel M.; Akbar, Hadia; Attallah, Nour A.; James, Ryan (2019): Assessing data availability and research reproducibility in hydrology and water resources. In: *Scientific data* 6, S. 190030. DOI: 10.1038/sdata.2019.30.

SFB 1294 „Data Assimilation“

0%

Publikation ausstehend

Studiendesign

Stagge et al. (2019)

- 360 Publikationen aus sechs Hydrology-Journals von 2017
- Fragebogen mit 15 Fragen (Metadaten, Verfügbarkeit, Reproduzierbarkeit, Dauer)
- 5 Stopp-Kriterien bei fehlenden Informationen, Daten oder Code

Geßner, Seegebrecht (2020)

- Alle SFB-Publikationen (90)
- 26 zusätzliche Fragen (u.a. SFB-Metadaten, Lizenz, Komplexität, Formate, Fehler)
- Gleicher Ablauf

→ **Kompatibel zu Stagge et al.**

Fragebogen Stagge et al. (2019)

SFB 1294 Data Assimilation

Paper Metadata

Q1. Assessor's name
Q2. Journal name
Q3. Article DOI
Q4. Full paper citation

Availability

Q5. How accessible to users?

Some or all applicable | Not specified where | Not applicable

Q6. Where available?

All online | Third party | Author | In article

Q7. What is present?

Required: Input Data | Directions | Code / Software

Optional: License | Metadata | Identiers | Hardware / software requirements | File format

Q8. Comments on availability [open response].

Q9. Do you estimate you and readers could use the available artifacts to generate results?

Yes | Not sure | Not familiar with resources | No

Q10. Continue to reproduce results?

Yes | No

Reproducibility

Q11. Do the outputs verify published results (in text, figures, and tables)?

Yes (explain in Q12) | No (explain in Q13 and Q14)

Q12. If yes, explain what made the work reproducible and other comments [open response].

Q13. If no, why did reproducing the work fail?

Hardware / software errors | Did not generate results | Results differed | Unclear directions | Other

Q14. Other comments on why reproducing the work failed [open response].

Time to Complete

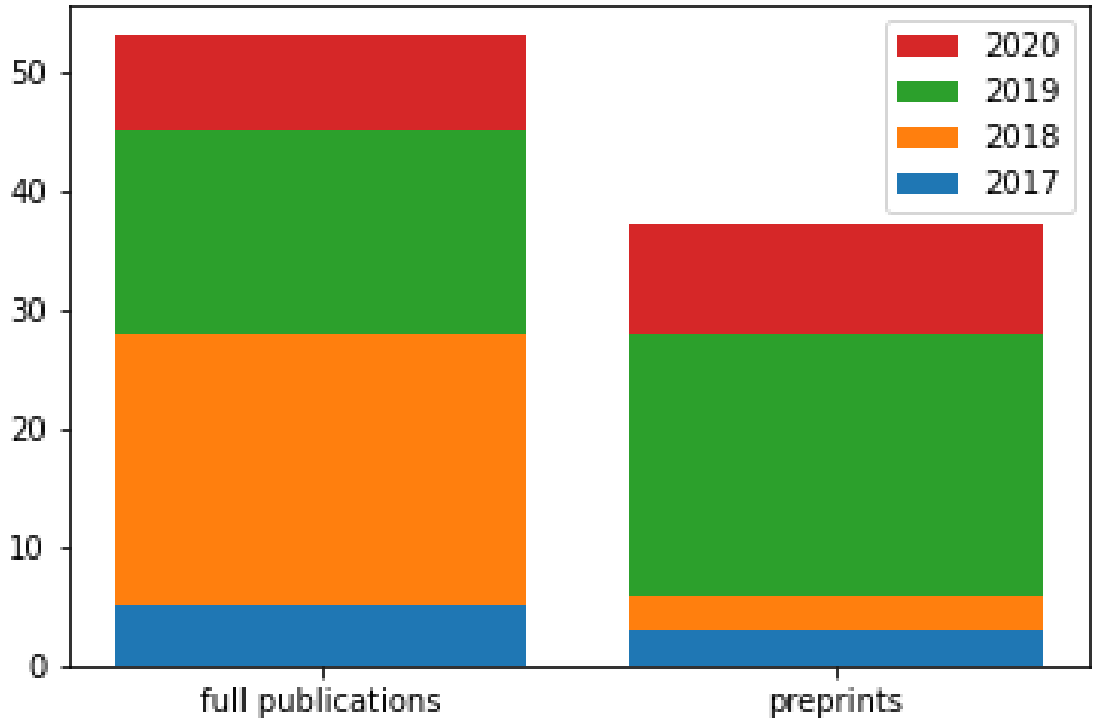
Q15: How many minutes did the survey take?

Stagge, James H.; Rosenberg, David E.; Abdallah, Adel M.; Akbar, Hadia; Attallah, Nour A.; James, Ryan (2019): Assessing data availability and research reproducibility in hydrology and water resources. In: *Scientific data* 6, S. 190030. DOI: 10.1038/sdata.2019.30.

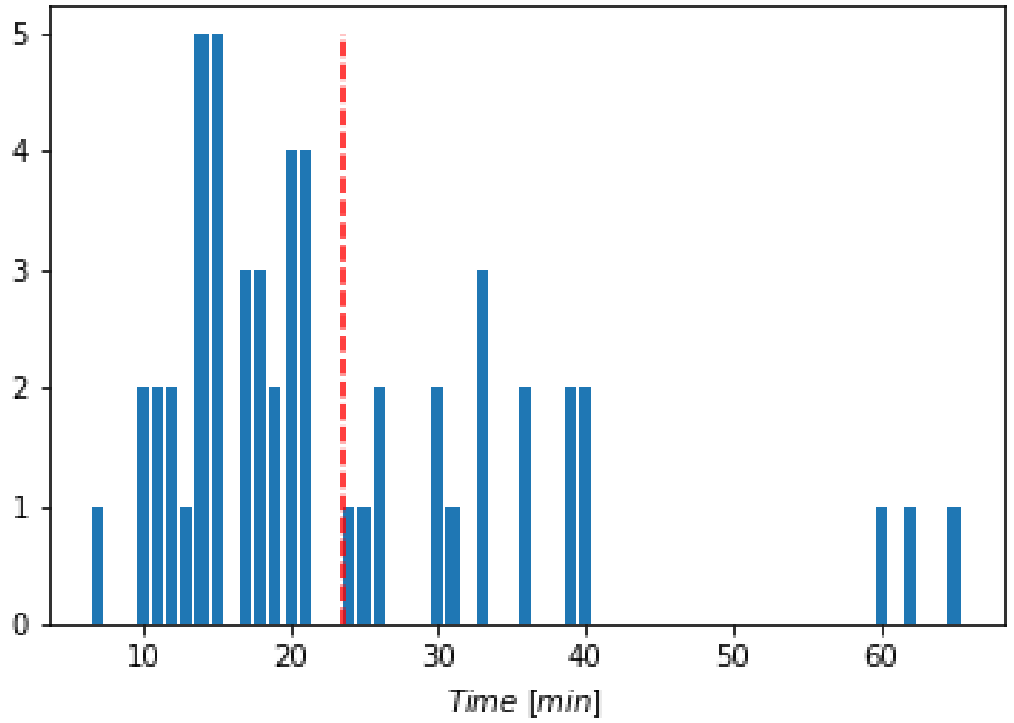
Ergebnisse im SFB 1294

SFB 1294
Data Assimilation

Vollpublikationen vs. Preprints



Zeitaufwand pro Vollpublikation

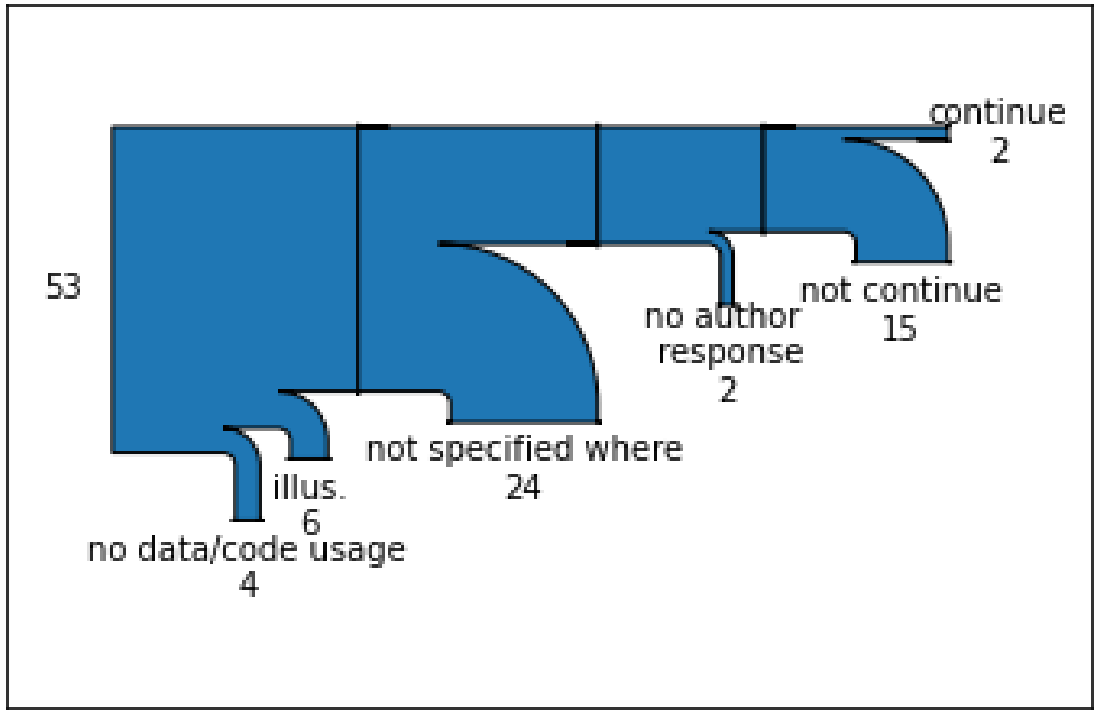


Kernprobleme im SFB 1294

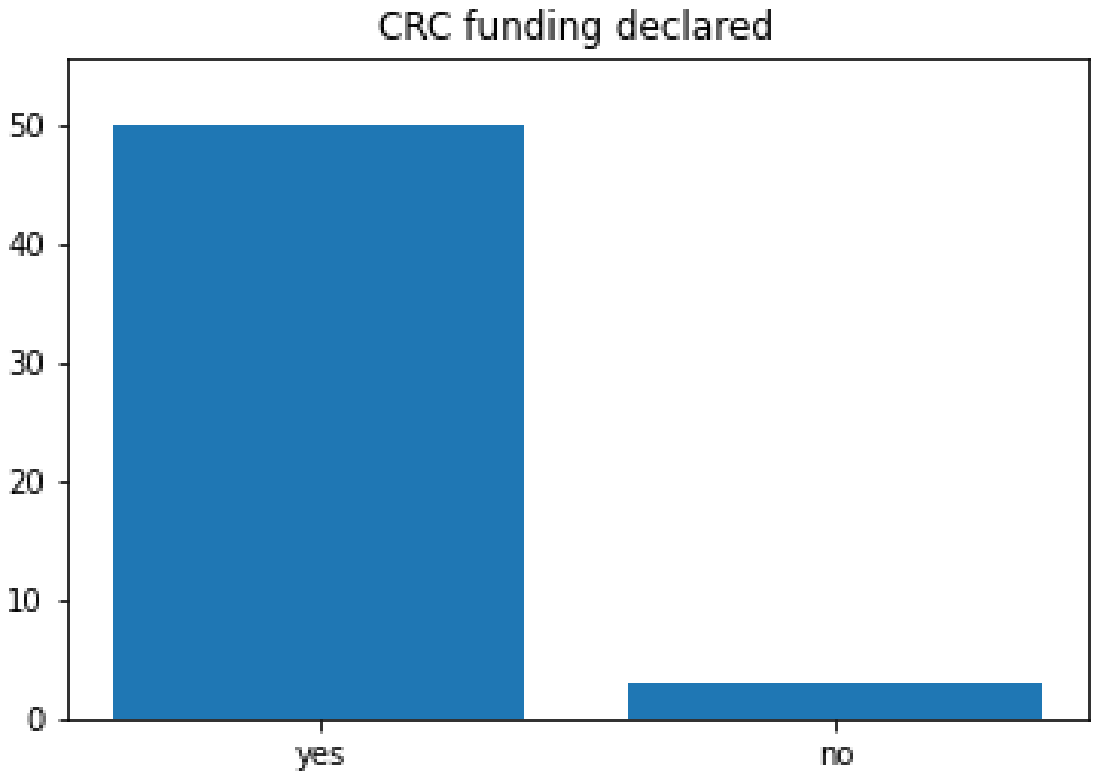
- 45% nennen keinen Ort für Daten und Code
- 28% verfügbarer Daten und Codes sind unvollständig
- Alle verfügbaren Daten und Codes (4%) hatten Laufzeitfehler

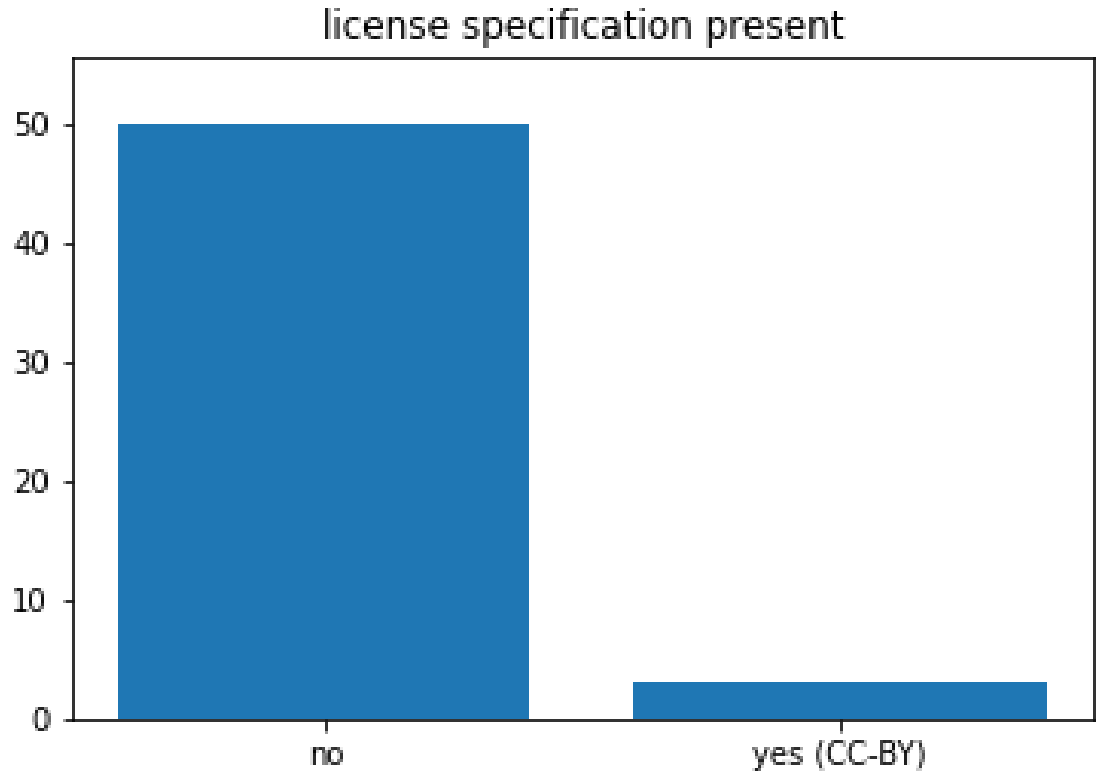
Stagge et al. (2019):
20% ohne Ort, 43% unvollständig

Dropout für Vollpublikationen



Weitere Probleme im SFB 1294?





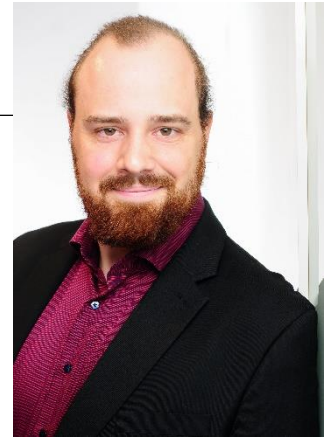
Reproduzierbarkeit verbessern

SFB 1294
Data Assimilation

Problem	Lösung
Daten und Code nicht verfügbar	Deklarieren, wo Daten und Code gefunden werden können (z.B. „request code and data from first author“ oder „code and data are available under the following OSF link“)
Reproduktion schlägt fehl wegen Software-Fehlern	Arbeitskollege lässt Daten und Code laufen
Daten und Code können nicht legal genutzt werden	Lizenz deklarieren

SFB 1294
Data Assimilation

Hendrik Geßner, M.Sc.



Universität Potsdam

Institut für Informatik und Computational Science

SFB 1294 „Data Assimilation“ / Forschungsgruppe Maschinelles Lernen

Campus Griebnitzsee

Haus 4, Büro 0.21

hgessner@uni-potsdam.de