HELMHOLTZ
Open Science

Why Openness and Reproducibility in Machine Learning Matter

Lena Messerschmidt[1], Antonia C. Schrader[1], Peter Steinbach[2], Lea Maria Ferguson[1], Heinz Pampel[1, 3]

[1] Helmholtz Association, Helmholtz Open Science Office

[2] Helmholtz-Zentrum Dresden-Rossendorf

[3] Humboldt-Universität zu Berlin

Hamburg, June 14, 2023

The Helmholtz Open Science Office

- Our Mission: Enabling Open Science practices in Helmholtz!
- The Helmholtz Open Science Office
  - is a service provider for the Association for the cultural change “from closed to open”.
  - promotes dialogue and provides impulses within the Association.
  - offers training and support concerning all aspects of open science.
  - represents Helmholtz positions on open science on a national and international level.

Team

- Roland Bertelmann (Head)
- Christoph Bruch
- Lea Maria Ferguson
- Steffi Genderjahn
- Marcel Meistring
- Lena Messerschmidt
- Heinz Pampel
- Antonia C. Schrader
- Paul Schultze-Motel
- Nina Leonie Weisweiler

HELMHOLTZ
Open Science

open-science@helmholtz.de

Open Science

- “Open Science, the unrestricted access to scientific publications and cultural heritage, is an ongoing and future trend in the scientific landscape worldwide.“
  Helmholtz Open Science Office

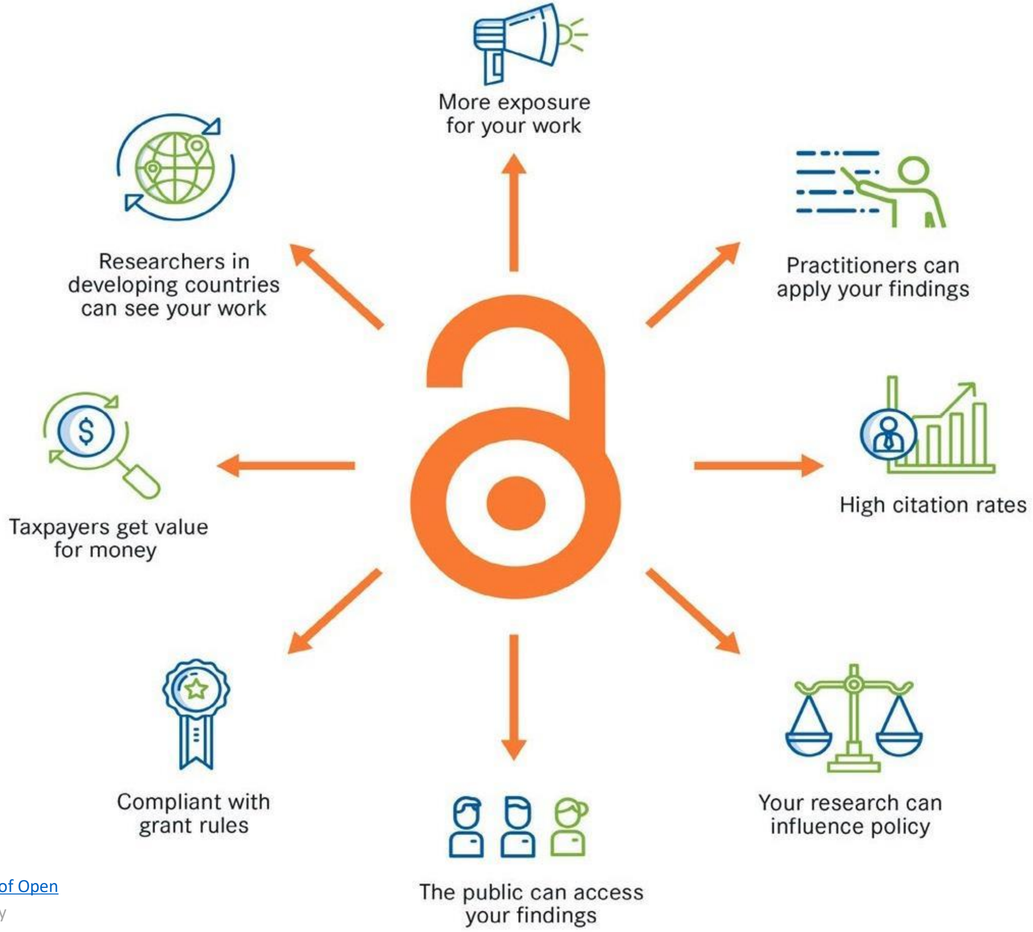
- “Open Science (...) make[s] multilingual scientific knowledge openly available, accessible and reusable for everyone, increase[s] scientific collaborations and sharing of information (...), and open[s] the processes of scientific knowledge creation (...) to societal actors beyond the traditional scientific community.“
  UNESCO Recommendation on Open Science

- Key aspects of Open Science
  - Open Access
  - Open Research Data
  - Open Research Software

Advantages

- There are manifold advantages resulting from open science, and some of the main benefits include:

- Important: Openness as “intelligent openness”, i.e., “as open as possible and as closed as necessary”.


More exposure for your work
Researchers in developing countries can see your work
Practitioners can apply your findings
Taxpayers get value for money
High citation rates
Compliant with grant rules
Your research can influence policy
The public can access your findings

Open Science in the Helmholtz Association

Core topics

- Open Access Quota
- Open Data Policies and Forum
- Open Software Policy and Forums
- Open Science Policy

General Resources

- Helmholtz Open Science Policy

Resources on Open Research Data

- Position paper "Making information resources more usable"
- Research Data Policies of the Helmholtz Centers

Resources on Open Research Software

- Position paper "Access to and Reuse of Research Software"
- Model policy "Model Policy on Sustainable Software at the Helmholtz Centers "
- Checklist "Checklist to Support the Helmholtz Centers in Implementing Policies on Sustainable Research Software"

Definitions of Reproducibility

Reproducibility is the ability to recalculate a figure from data, parameters and programs

(Schwab M, et al., 2000)

Reproducibility is the ability of independent investigators to draw the same conclusions from an experiment by following the documentation shared by the original investigators.

(Gundersen, O.E., 2021)

Repeatability (same team, same experimental setup)

Replicability (different team, same experimental setup)

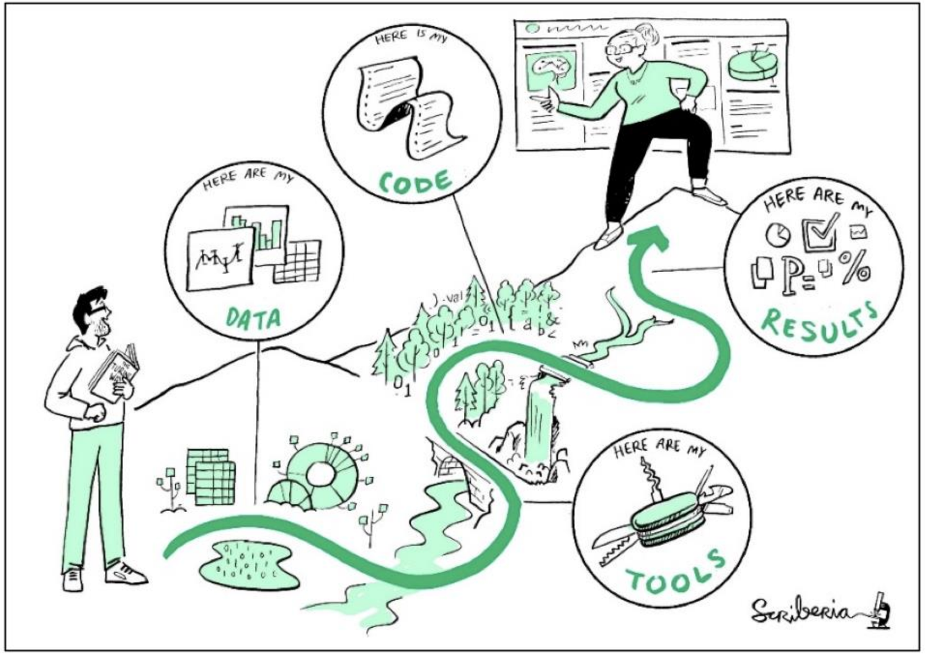
Reproducibility (different team, different experimental setup)

(ACM, 2016, https://www.acm.org/publications/policies/artifact-review-badging)

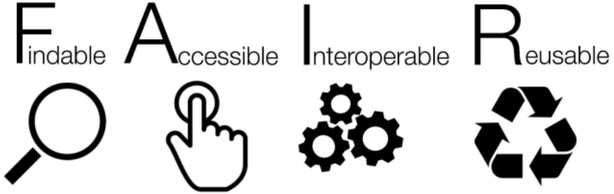
Source: Fritzsch, B. (2022). Reproduzierbarkeit - Viele kleine Schritte zu einem großen Ziel. RDA Deutschland Tagung 2022. https://indico.desy.de/event/31380/contributions/115816/attachments/71905/91885/RDA_DE_2022-Reproduzierbare_Wissenschaft-Reproduzierbarkeit.pdf, Slide 4

Hand in Hand:

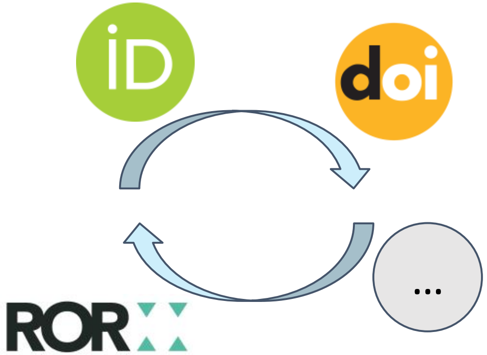
Open Science and Reproducibility



The Turing Way Community, & Scriberia. (2020). Illustrations from the Turing Way book dashes. Zenodo. https://doi.org/10.5281/zenodo.3695300



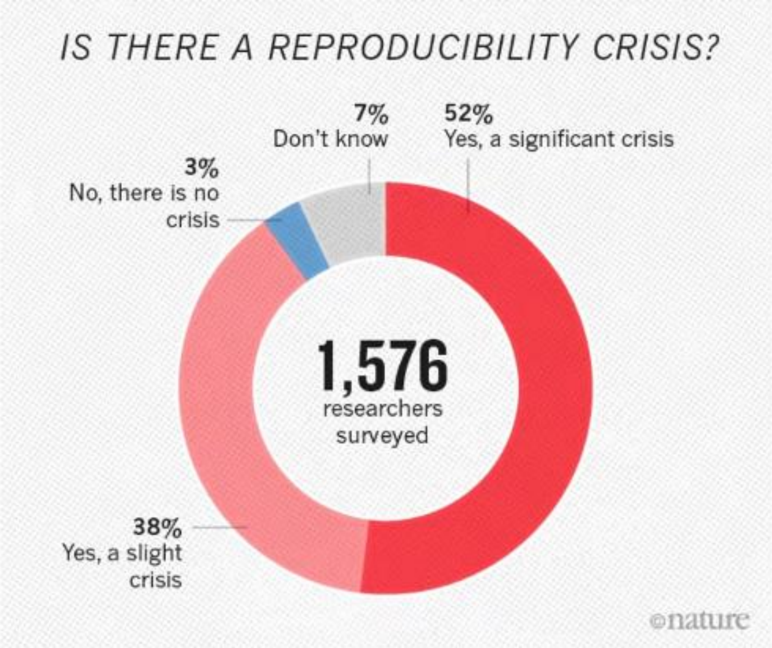
By Sangya Pundir; CC BY SA 4.0, https://commons.wikimedia.org/w/index.php?curid=53414062
More information: www.go-fair.org/fair-principles/

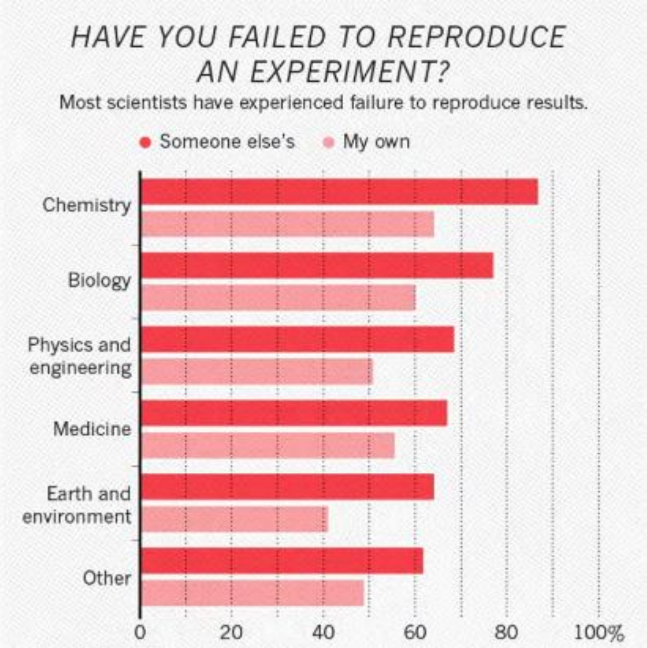


More information: https://pidforum.org/

Why does this matter?

Reproducibility in Machine Learning





Baker, M. 1,500 scientists lift the lid on reproducibility. *Nature* **533**, 452–454 (2016). https://doi.org/10.1038/533452a

Challenges

There are many factors impacting the reproducibility of Machine Learning research:

- Lack of records
- Changes in data
- Hyperparameter inconsistency
- Randomness
- Experimentation
- Framework changes
- Computational discrepancy
- Nondeterminism

What can be done to handle these factors?

Reproducibility in Machine Learning

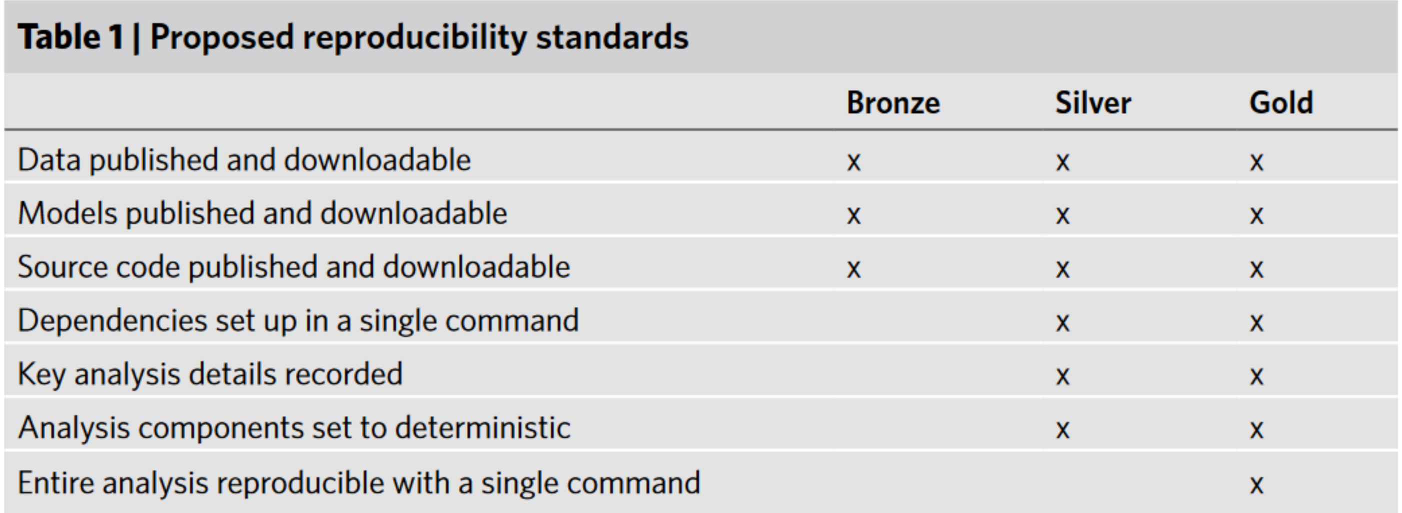


Table 1 | Proposed reproducibility standards

	Bronze	Silver	Gold
Data published and downloadable	x	x	x
Models published and downloadable	x	x	x
Source code published and downloadable	x	x	x
Dependencies set up in a single command		x	x
Key analysis details recorded		x	x
Analysis components set to deterministic		x	x
Entire analysis reproducible with a single command			x

Heil, B.J., Hoffman, M.M., Markowetz, F. *et al.* Reproducibility standards for machine learning in the life sciences. *Nat Methods* **18**, 1132–1135 (2021). https://doi.org/10.1038/s41592-021-01256-7

Helpful resources

Sustainable Research Software in Helmholtz

Open Science Office

- Task Group Research Software
- Open Science Fora on Research Software
- Workshops and Seminars on Reproducibility
- Task Group Helmholtz Quality Indicators for Data and Software Products

More Initiatives in Helmholtz

- Helmholtz Research Software Directory, The HERMES Project, HIDA Course Catalog

Publications

- Ferguson, L. M., Schrader, A. C., Seibold, H., Weisweiler, N. L. (2021): Open Science Factsheet No. 2: [ ] Practical Steps Towards Open and Reproducible Research, Potsdam: Helmholtz Open Science Office. https://doi.org/10.48440/os.helmholtz.025
- Messerschmidt, R., Schrader, A. C. (2021): Offene Forschungssoftware als integraler Bestandteil von Open Science und guter wissenschaftlicher Praxis, CampusSource Tagung 2021 am 26. Mai 2021. Online.

Helpful resources

Reproducibility in practice

ReproHack Hub

Host a hackathon doing live peer reviews trying to reproduce your publication

Turing Way Project

The place-to-go for all things reproducibility

German Reproducibility Network

Join a network which promotes reproducible and robust research on a national level

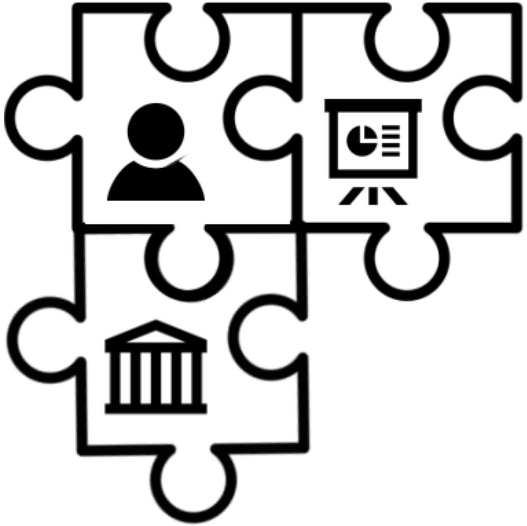
The Carpentries

Managing Open and Reproducible Computational Projects training material

Synergies and Collaboration

Open Science and Reproducibility

- Open Science and Reproducibility go hand in hand.



- Reproducibility is essential for sustainable data science and ML research
- Open Science – Cultural change in the scientific way of working, organization, and communication
- Achieve reproducibility – make all scientific outputs (data, code, environment) open, comprehensible, reusable, and easily accessible
- Adjustments of incentives and rewards system needed – Change process

Keep in touch

- Email – open-science@helmholtz.de

- Mastodon – @HelmholtzOpenScienceOffice@helmholtz.social
- Twitter – @helmholtz_os

- Website – www.os.helmholtz.de

- Mailing list for members of Helmholtz – Helmholtz Open Science Professionals

- Open Science Newsletter

HELMHOLTZ
Open Science

Thank you for your attention!

Lena Messerschmidt

lena.messerschmidt@os.helmholtz.de

https://orcid.org/0000-0002-3406-9933

@lenamesserschmidt@openbiblio.social

@lenes_messer