
Fundamentals of reproducibility in machine learning

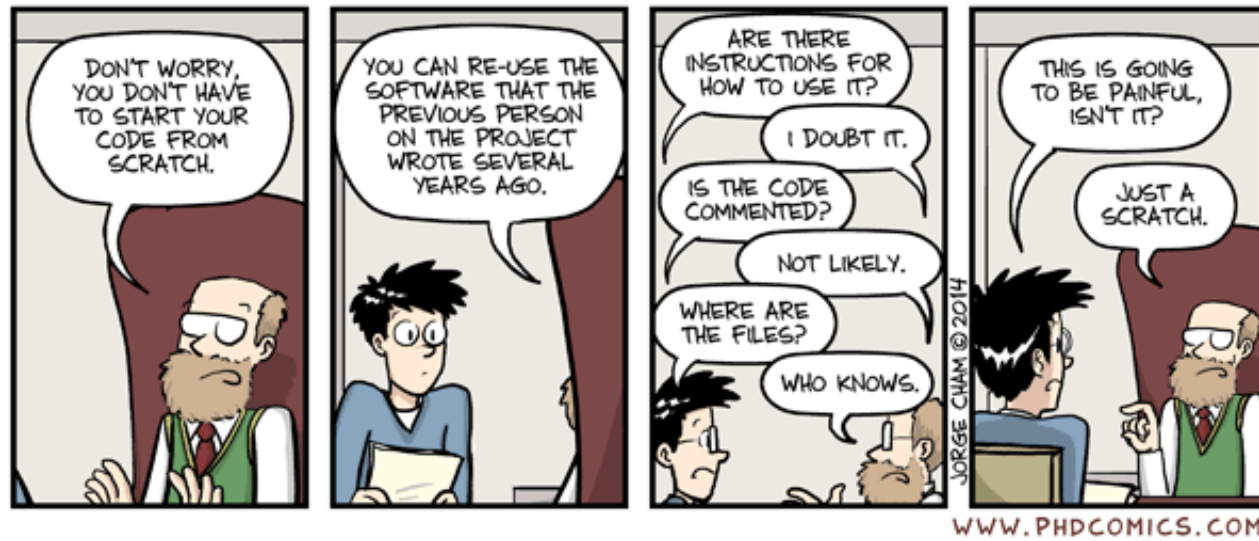
Enrico Glerean DSc., Staff Scientist, Aalto University

Sounds familiar? (Hopefully not!)

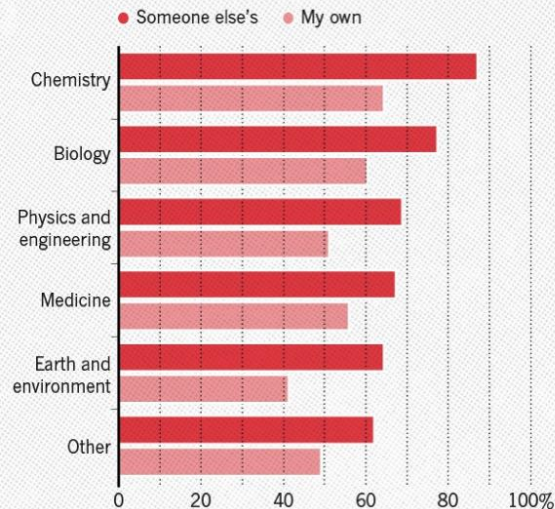
- How did the previous postdoc do that figure???
- What was I doing in this bit of code???
- Did someone modify the data?? I am getting completely different results!
- Reviewer #3 is asking me to re-run everything...

A!

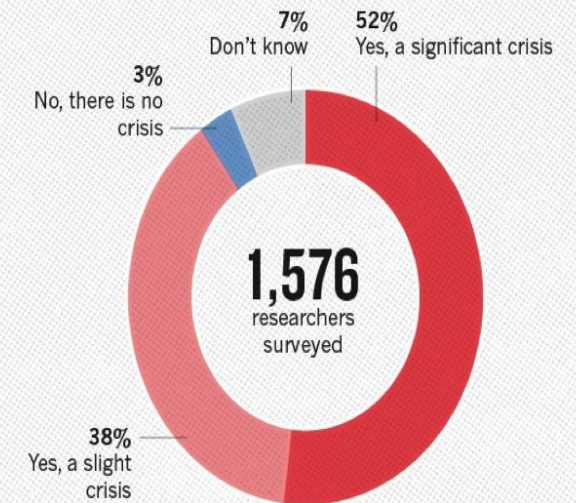
From CodeRefinery "Reproducible Research"



HAVE YOU FAILED TO REPRODUCE AN EXPERIMENT?
Most scientists have experienced failure to reproduce results.



IS THERE A REPRODUCIBILITY CRISIS?



Number of respondents from each discipline:
Biology 703, Chemistry 106, Earth and environmental 95,
Medicine 203, Physics and engineering 236, Other 233

©nature

©nature

What is reproducibility (of results) in scientific research?

		Data	
		Same	Different
Analysis	Same	Reproducible	Replicable
	Different	Robust	Generalisable

Reproducible: same analysis steps on same dataset produces same answers.

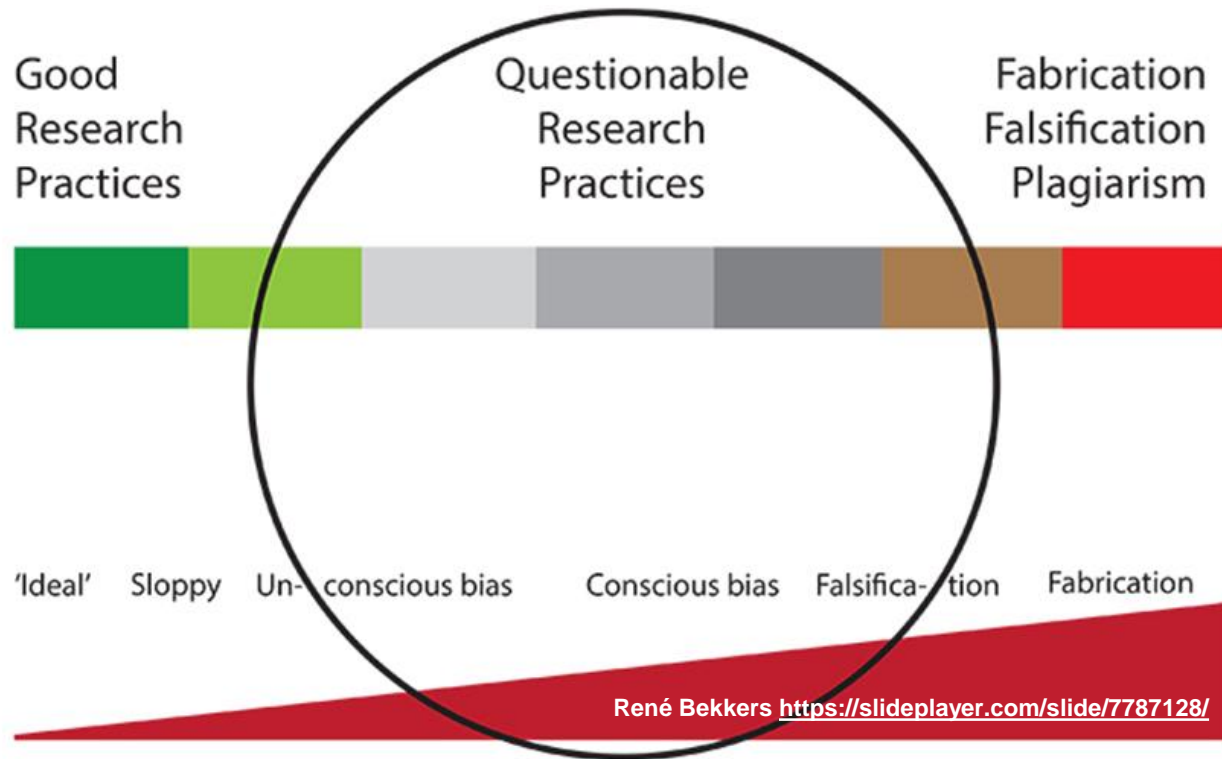
Replicable: same analysis on different datasets produces **similar** answers.

Robust: Robust results show that the work is not dependent on the specificities of the implementation of the analysis.

Generalisable: Results that are meaningful beyond the specific dataset or analysis pipeline used.

Figure from The Turing Way

Why do we care? Research Integrity



ALLEA European Code of Conduct for Research Integrity – Principles:

Reliability, Honesty, Respect, Accountability

Violations of research integrity:

Fabrication: Making up data or results and recording them as if they were real.

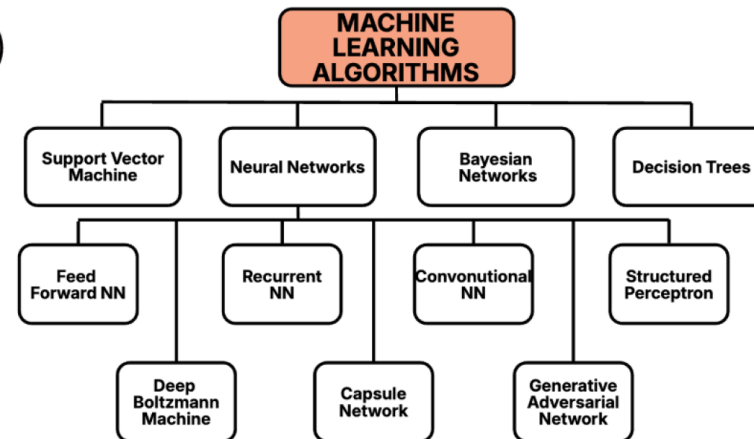
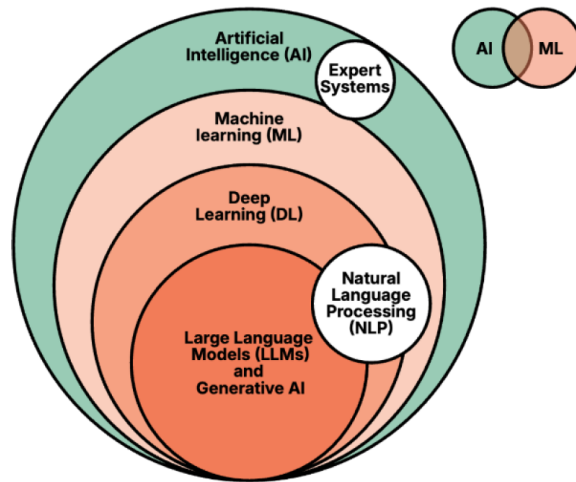
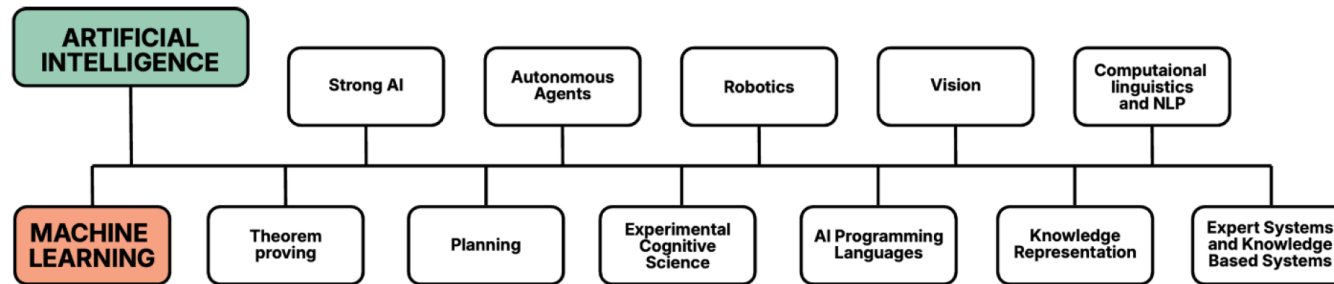
Falsification: Manipulating research materials, equipment, images, or processes, or changing, omitting, or suppressing data or results without justification

If your findings are not reproducible, the integrity of your research can be questioned.

Figure by René Bekkers <https://slideplayer.com/slide/7787128/>

What about artificial intelligence and machine learning?

What is Artificial Intelligence (AI)?

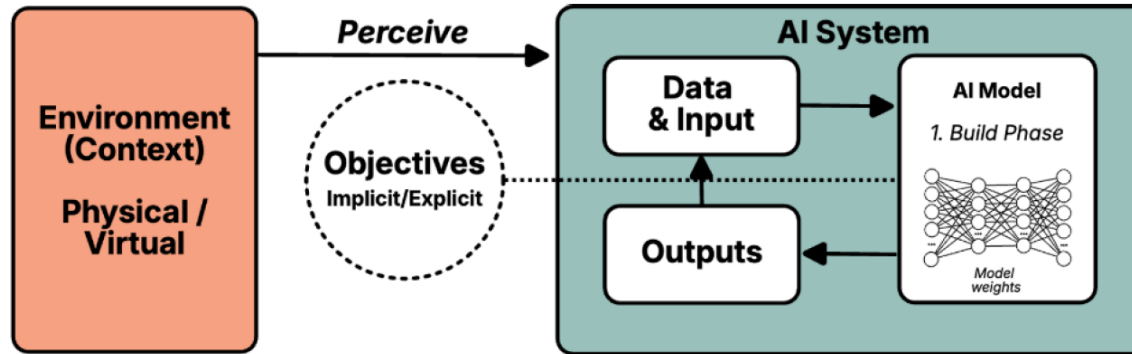


‘AI system’ means a **machine-based system** that is designed to operate with varying levels of **autonomy** and that may exhibit **adaptiveness** after deployment, and that, for explicit or implicit objectives, **infers, from the input it receives, how to generate outputs** such as **predictions, content, recommendations, or decisions** that can influence physical or virtual environments;

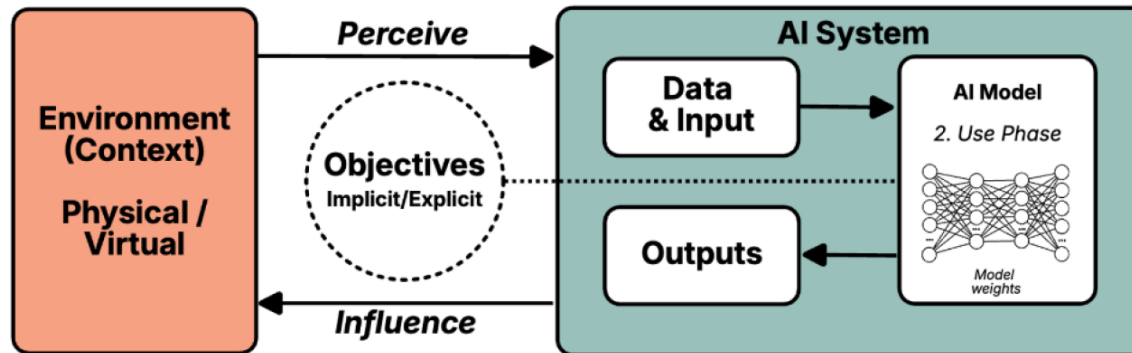
(Article 3(1), Artificial Intelligence Act)

Other AI legal definitions collected by IAPP.

AI systems vs AI models



Build phase, pre-deployment

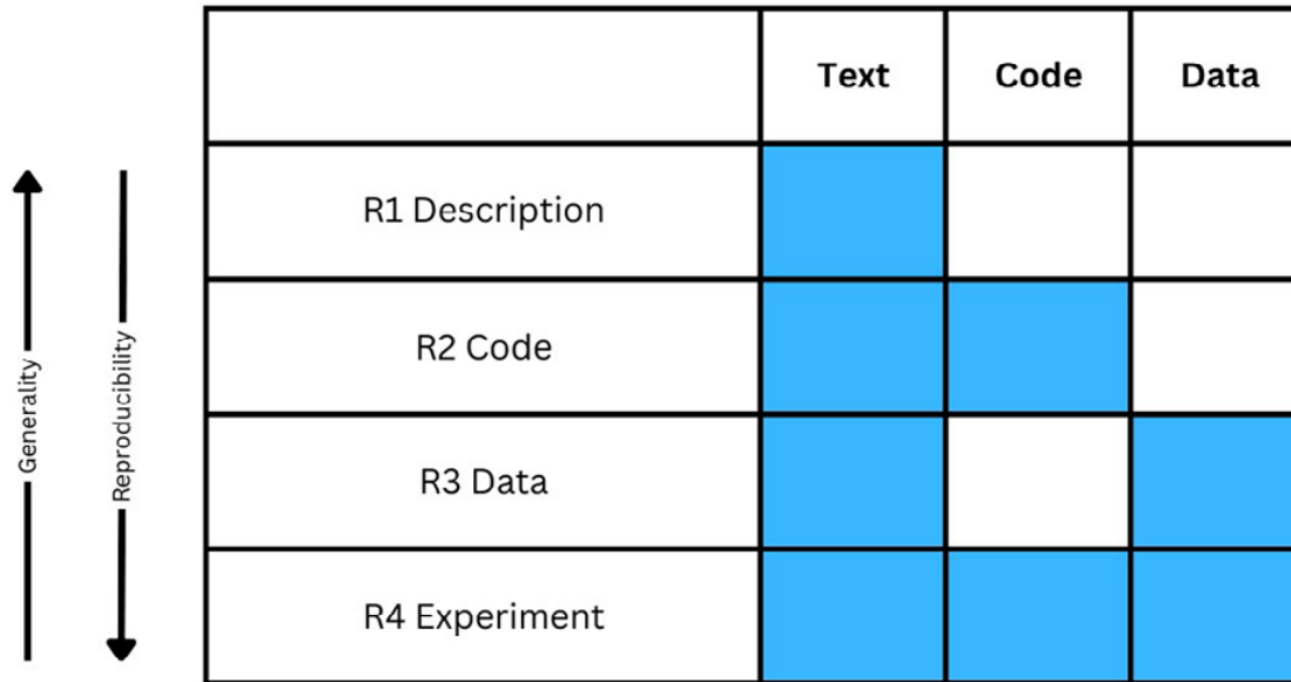


Use phase, post deployment

The **AI model** is like the **engine of the car**: it can be very powerful and even dangerous, but if it is not put into a system (the rest of the car) it cannot do anything.

The **AI system** enables the use of the AI model: by feeding input to the model, it can produce recommendations, classifications, translations, synthetic text, synthetic images, code, ... and **influence** the external physical/virtual environment

What is reproducibility in machine learning research?



	Text	Code	Data
R1 Description			
R2 Code			
R3 Data			
R4 Experiment			

“...the general community continues to take this issue too lightly.”

Data: not only availability but also issues specific to ML: **data leakage** (poor split between training/test datasets), **bias** (various types of imbalances in the data)

Experiment: indeterminism (randomness of certain methods), environment (GPU vs CPU vs TPU architecture, version of libraries), computational resources (not everyone has access)

So is data and code availability enough?

Full reproducibility from full transparency

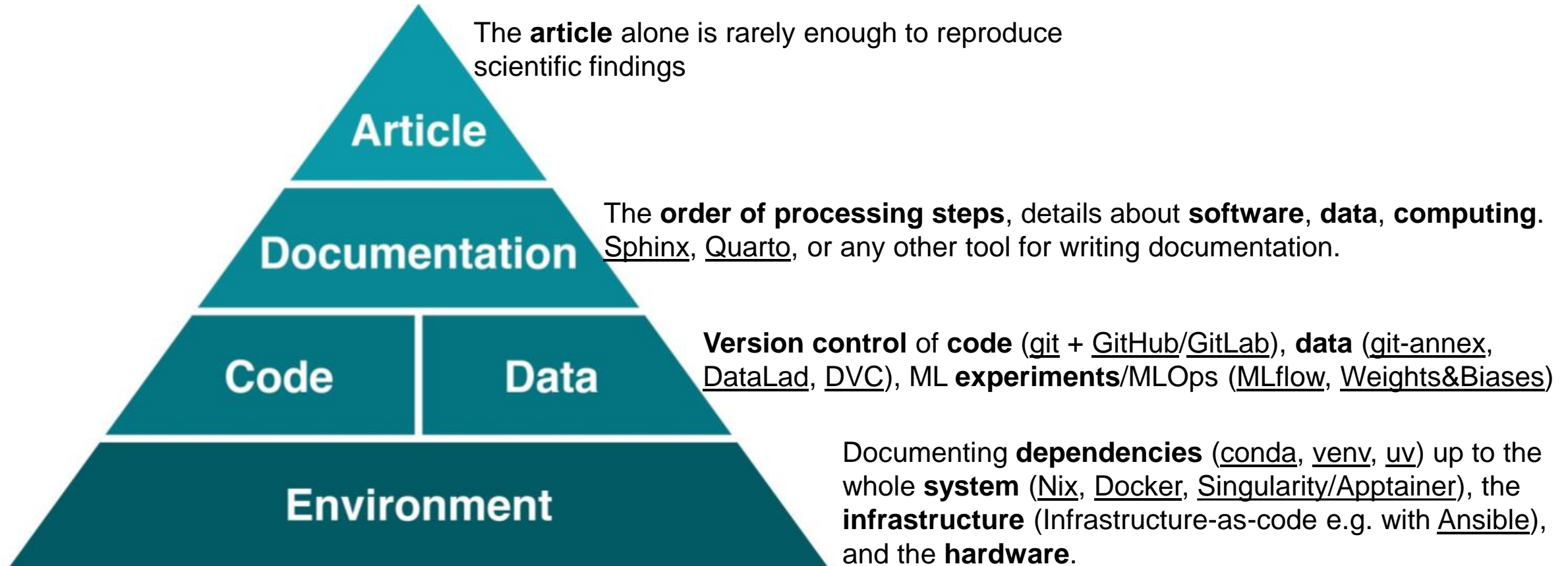


Figure from Steeves, Vicky (2017) in “Reproducibility Librarianship”
Collaborative Librarianship: Vol. 9: Iss. 2, Article 4.

Reproducibility from full transparency in machine learning

BARRIERS		Technology-driven					Procedural			Awareness
		Hosting services	Virtualization	Managing sources of randomness	Privacy-preserving technologies	Tools, platforms	Standardized datasets, evaluation	Guidelines, checklists	Model info sheets, model cards	Training, policies, initiatives
R1 Description	Completeness, quality of reporting									
	Spin practices and publication bias									
R2 Code	Limited access to code									
R3 Data	Limited access to data									
	Data leakage									
	Bias									
R4 Experiment	Inherent nondeterminism									
	Environmental differences									
	Limited resources									

Mapping solutions to barriers

(selection)

- Data and Code (and Models) version control
- Training pipelines
- Standardised datasets (benchmarks)
- Model sharing, model cards
- Checklists

Practical example:
reproducibility in the
NeurIPS paper checklist

A! Figure from Semmelrock et al. (2025) Reproducibility in machine-learning-based research: Overview, barriers, and drivers

Conclusion and importance in the real world

- **Reproducibility** is a fundamental component of research integrity
- Beyond research, ML reproducibility is at the core of **AI system robustness**: important in **products in the real world** (ISO/IEC 5259-4:2024, ISO/IEC 42005:2025)
- Not covered: **concept drift** (when real world data changes, but the system should still work), **explainability of AI** (the better we can explain how it works, the better we can ensure robustness, reproducibility, and replicability)

References

Belz, Anja, Shubham Agarwal, Anastasia Shimorina, and Ehud Reiter. "A Systematic Review of Reproducibility Research in Natural Language Processing." *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 2021, 381–93.

Herrmann, Moritz, F Julian D Lange, Katharina Eggensperger, et al. "Position: Why We Must Rethink Empirical Research in Machine Learning." *arXiv Preprint arXiv:2405.02200*, 2024.

Hill, David, Benjamin A Antunes, Anthony Bertrand, et al. "Machine Learning and Reproducibility Impact of Random Numbers." *38th European Simulation and Modelling Conference (ESM)*, 2024, 65–70.

Kapoor, Sayash, and Arvind Narayanan. "Leakage and the reproducibility crisis in machine-learning-based science." *Patterns* 4, no. 9, 2023.

Semmelrock, Harald, Tony Ross-Hellauer, Simone Kopeinik, et al. "Reproducibility in Machine-Learning-Based Research: Overview, Barriers, and Drivers." *AI Magazine* 46, no. 2 (2025): e70002.