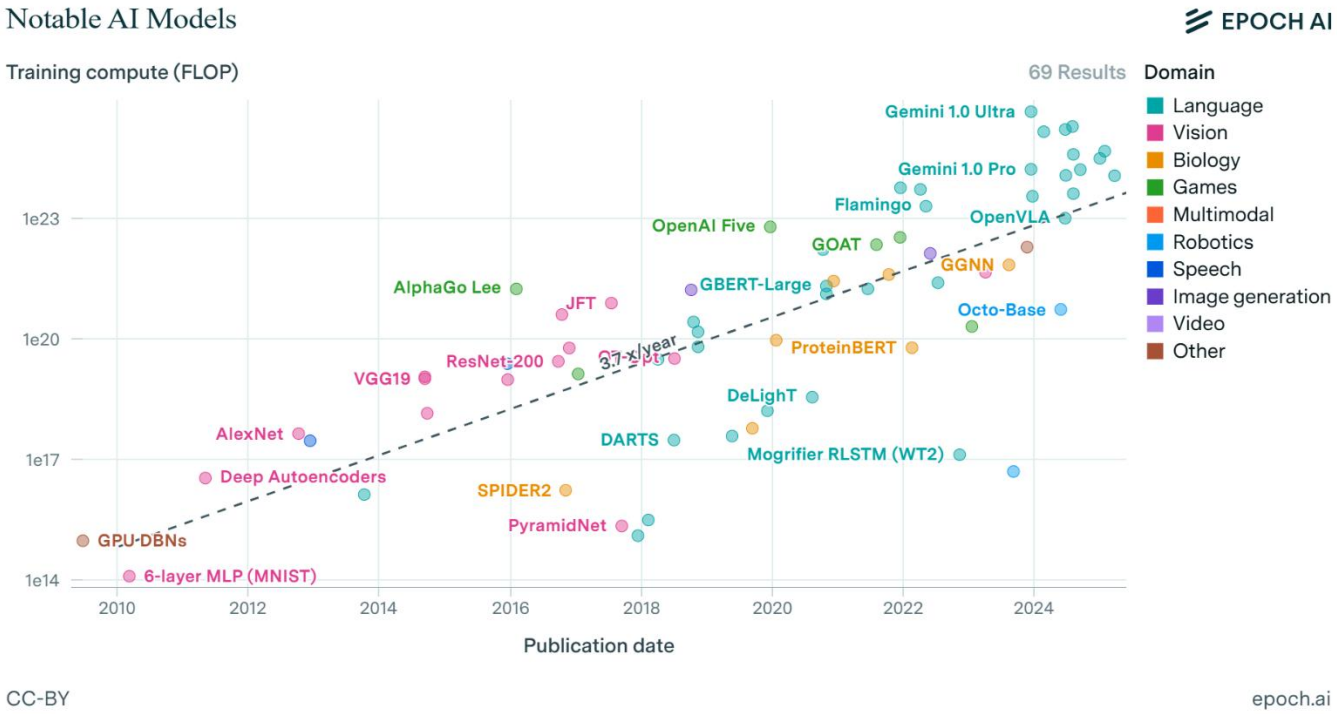
Contributing to reproducible software-based measurement of energy consumption in machine learning

Anthony BERTRAND

Supervisors : David HILL

Engelbert MEPHU NGUIFO

1. Introduction



Graph showing the increase in computing load for machine learning models from 2010 to 2025

- Increasingly large AI models
- increasingly lengthy AI training

- Data on energy consumption?

1. Introduction

Measuring energy consumption

Hardware-based measurement

Physical devices (measuring instruments) for measuring energy consumption.

(Wattmeters, smart plugs)

+ High accuracy
+ OS and software independent
- Lack granularity
- Expansive

Software-based measurement

Software tools to estimate energy consumption based on resources used, or other information related to system activity

+ Granularity
+ Cheap
- Variable accuracy
- OS-dependant

Chen, H., & Shi, W. (2012). Power Measuring and Profiling: State of the Art. In *Handbook of Energy-Aware and Green Computing, Volume 2* (Vol. 2, p. 26). Chapman and Hall/CRC.

1. Introduction

Reproducibility

- In experimental sciences, this criterion distinguishes science from pseudo-science.
- Since May 2025, a “reproducibility checklist” is mandatory to be published in the Journal of Artificial Intelligence Research (JAIR).

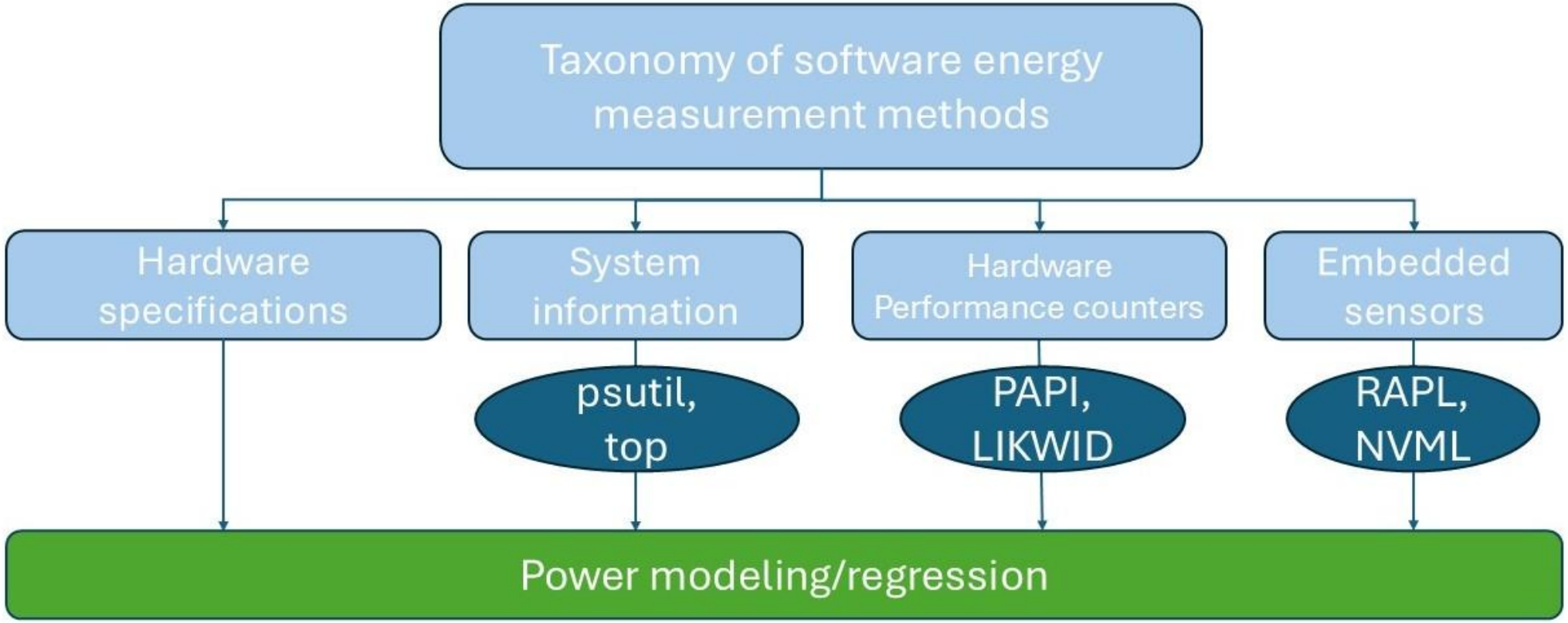
My job:

- Tackle reproducibility challenges in ML.
  - Spread good practices.
  - Raise awareness.
- Make my own research reproducible.

Gundersen, O. E., Helmert, M., & Hoos, H. (2024). Improving Reproducibility in AI Research: Four Mechanisms Adopted by JAIR. *Journal of Artificial Intelligence Research*, *81*, 1019–1041.

2. Energy measurement

Taxonomy



2. Energy measurement

Tools

Date	Tools
2019	Energy Usage
2020	Experiment Impact Tracker Carbon Tracker
2021	Cumulator CodeCarbon
2022	Power Measurement Toolkit PowerJoular perun eco2AI
2023	Rjoules
2024	EA2P CPPJoules
2025	Alumet

- All these tools use RAPL
- RAPL uses proprietary power models

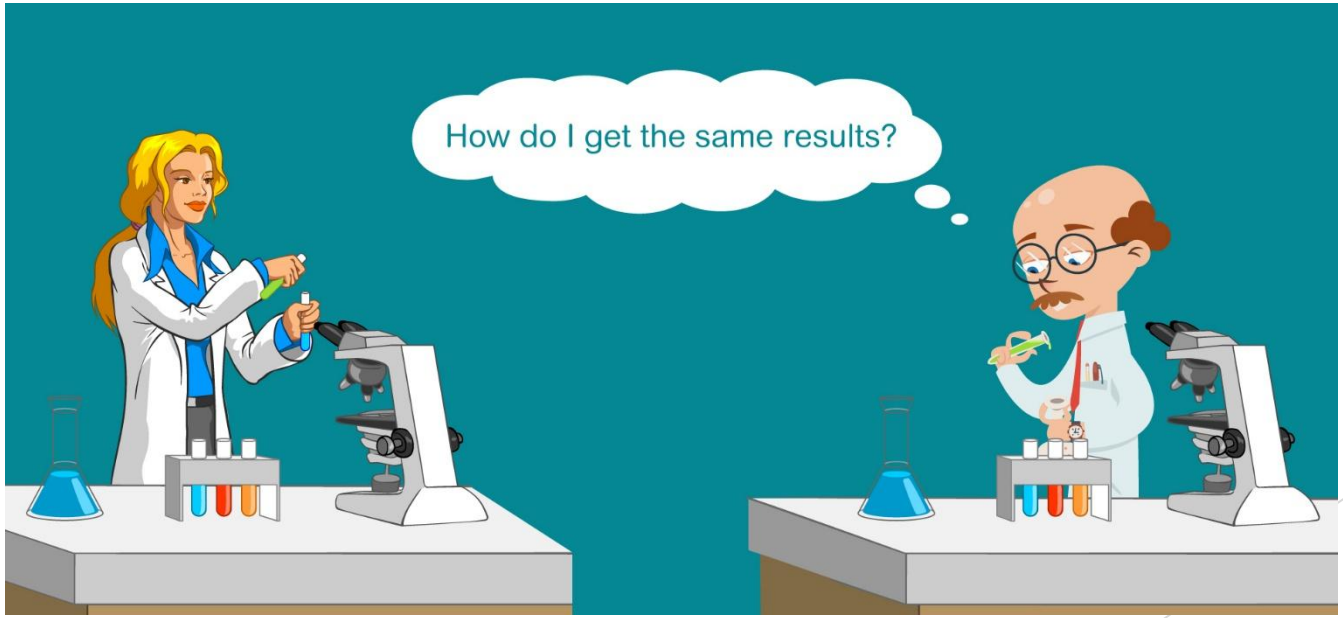
- What about Storage? Network? Fans?

3. Reproducibility

Definition of reproducibility

The measurement can be obtained with **stated precision** by a **different team** using the **same measurement procedure**, the **same measuring system**, under the **same operating conditions**, in the same or a different location on multiple trials.

For computational experiments, this means that an independent group can obtain the same result using the author's own artifacts.



Antunes, B., & Hill, D. R. C. (2024). Reproducibility, Replicability and Repeatability: A survey of reproducible research with a focus on high performance computing. Computer Science Review, 53, 100655.

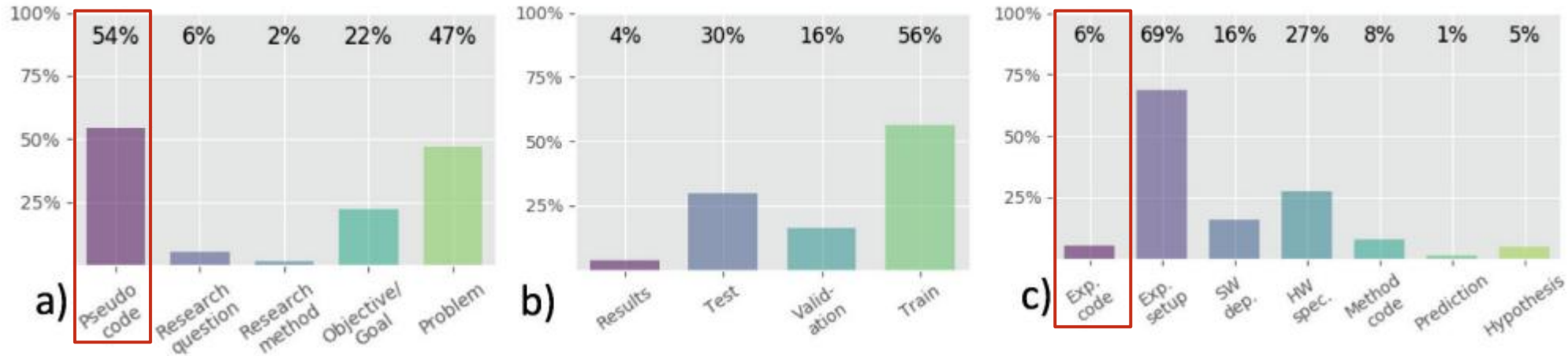
3. Reproducibility

Reproducibility in AI

AAAI : Association for the Advancement of Artificial Intelligence

IJCAI : International Joint Conference on Artificial Intelligence

400 articles published between 2013 and 2016:



Percentage of items meeting the criteria for each variable used in the experiment

Gundersen, O. E., & Kjensmo, S. (2018). State of the Art: Reproducibility in Artificial Intelligence. *Proceedings of the AAAI Conference on Artificial Intelligence*, *32*(1).

3. Reproducibility

Definition of repeatability

The measurement can be obtained with **stated precision** by the **same team** using the **same measurement procedure**, the **same measuring system**, under the **same operating conditions**, in the **same location** on multiple trials.

For computational experiments, this means that a researcher can reliably repeat their own computation

For scientific programs, the stated precision must be 0. We need bitwise identical results to debug!

3. Reproducibility

Work on clustering methods

- Study of K-Means, DBSCAN and Ward algorithms
  - Divide them into steps
  - Explain what must be define to have a deterministic algorithm
- Study of the Scikit-learn implementation
  - K-Means is not bitwise repeatable!
  - May come from OpenMP mismanagement of floating-point operation order
  - Try to compare it with other implementations

3. Reproducibility

Work on clustering methods

8 Datasets:

Datasets \ Parameters	# Instances	# Features	# Classes
Iris	150	4	3
Toxicity	171	1 203	2
Wine	178	13	2
Breast cancer	569	30	2
Taiwan	6 819	95	2
Letter recognition	20 000	16	26
Default of credit card clients	30 000	23	2
Blob dataset (generated)	60 000	2	10

5 implementations:

With Scikit-learn
- DBSCAN
- Ward
- K-Means

With SciPy
- K-Means

In addition
- Custom K-Means with OpenMP

Everything is bitwise-repeatable, except K-Means of scikit-learn

3. Reproducibility

Work on clustering methods

		Bitwise-repeatability check for Scikit-Learn's K-Means																													
datasets	nb threads	1				2				3				4				16				64				128				192			
	Results	C	L	I	M	C	L	I	M	C	L	I	M	C	L	I	M	C	L	I	M	C	L	I	M	C	L	I	M	C	L	I	M
Iris												X				X				X				X				X				X	
Toxicity																				X				X				X				X	
Wine																X				X				X				X				X	
Breast cancer										X		X		X		X		X		X		X		X		X		X		X		X	
Taiwan										X		X		X		X		X		X		X		X		X		X		X		X	
Letter										X		X		X		X		X		X		X		X		X		X		X		X	
Credit card																X		X		X		X		X		X		X		X		X	
Generated										X		X		X		X		X		X		X		X		X		X		X		X	

K-Means results repeatability for each dataset depending on the number of OpenMP threads used (up to 192 threads) over 30 replications.
C=final centers, L=final labels, I=inertia score, M=best iteration.
A cross indicates that different results were found in at least two runs out of the 30 replications.

Bertrand, A., Mephu Nguifo, E., Antoine, V., & Hill, D. (2025). *A K-MEANS, WARD AND DBSCAN REPEATABILITY STUDY*. ⟨hal-05426697v2⟩

Conclusion

Main goals :

- Measure the energy consumption of ML models (train/inference) with software-based tools.
- Tackle reproducibility issues in ML field.

Sub Goals :

- Have a deterministic and repeatable ML program. This will reduce measurement inaccuracies.
- Spread good practices and raise awareness of reproducibility pitfalls in ML.

Thank you for your attention

References

- Antunes, B., & Hill, D. R. C. (2024). Reproducibility, Replicability and Repeatability: A survey of reproducible research with a focus on high performance computing. *Computer Science Review*, *53*, 100655.
- Bertrand, A., Mephu Nguifo, E., Antoine, V., & Hill, D. (2025). *A K-MEANS, WARD AND DBSCAN REPEATABILITY STUDY*. ⟨hal-05426697v2⟩
- Chen, H., & Shi, W. (2012). Power Measuring and Profiling: State of the Art. In *Handbook of Energy-Aware and Green Computing, Volume 2* (Vol. 2, p. 26). Chapman and Hall/CRC.
- Gundersen, O. E., & Kjensmo, S. (2018). State of the Art: Reproducibility in Artificial Intelligence. *Proceedings of the AAAI Conference on Artificial Intelligence*, *32*(1).
- Gundersen, O. E., Helmert, M., & Hoos, H. (2024). Improving Reproducibility in AI Research: Four Mechanisms Adopted by JAIR. *Journal of Artificial Intelligence Research*, *81*, 1019–1041.