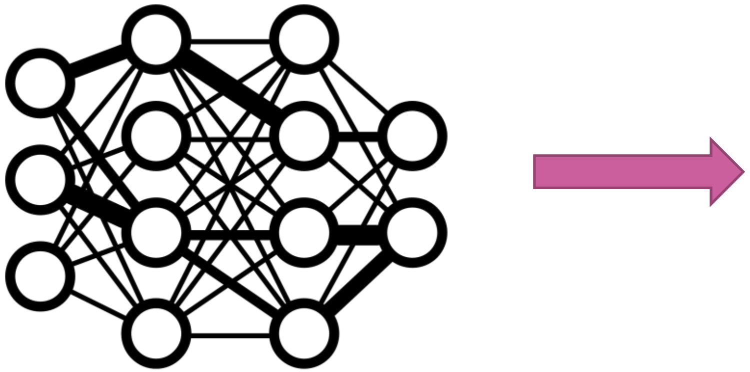
Monitoring Environmental Impact of Machine Listening Systems: Why and How?

Samuele Cornell, **Constances Douwes**, **Francesca Ronchini**,
Romain Serizel, Nicolas Turpault

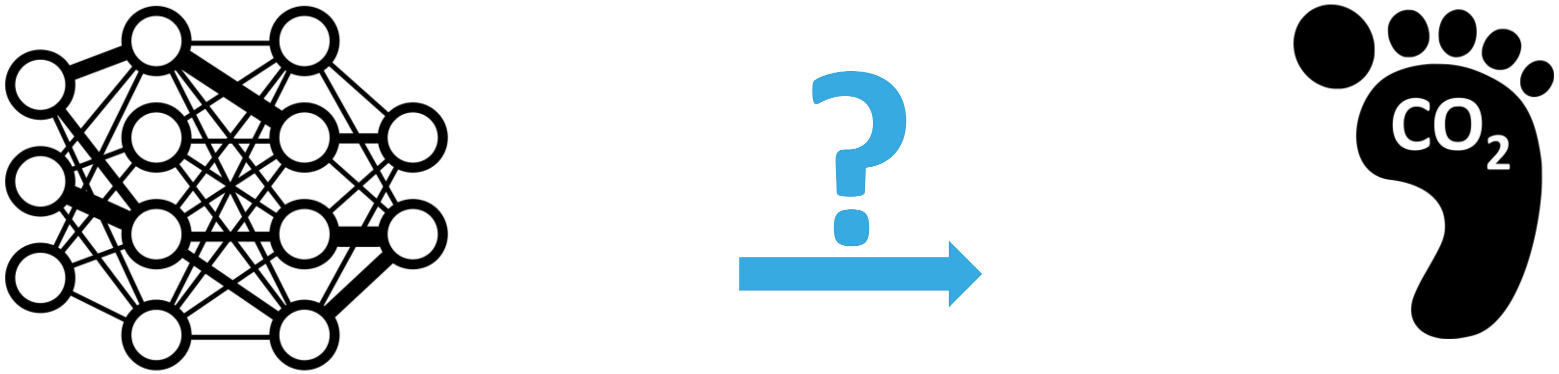
Motivations



What is the footprint of our systems?
What is the cost of performance improvement?

Images : Wikimedia

Monitoring environmental impact: How?



- Which metrics?
- Are they reliable?
- How to relate with performance?

Disclaimer: What we want to do

✓

- Raise awarness
- Compare systems among each other

✗

- Give an absolute estimate of the energy consumption
- Study the footprint at runtime

Outline

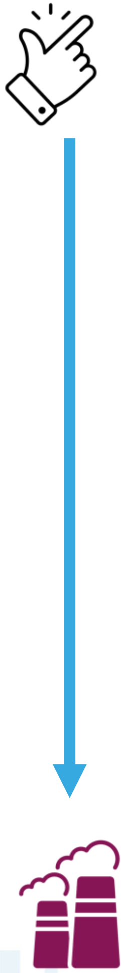
- Comparative study of the metrics

- Towards a fair comparison

- Case study on sound event detection

Comparative study of the metrics

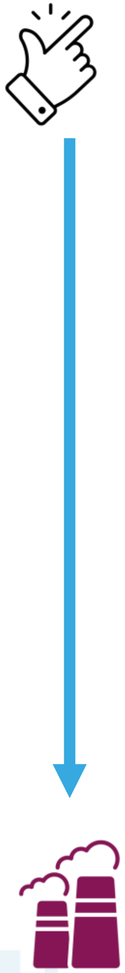
Comparative study of metrics



Runtime

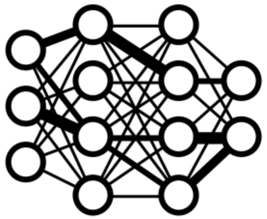
- Straightforward method in every developing environment
- Highly dependent of the model’s implementation
- Number & performance of GPU

Comparative study of metrics



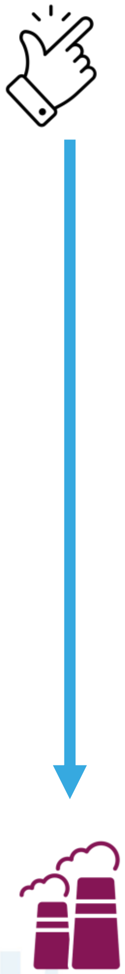
Runtime

Number of parameters



- Correlated with computational complexity
- Support from most DL libraries
- Different operations costs

Comparative study of metrics



Runtime

Number of parameters

Number of operations

+ − ÷ ×

- Hardware independent
- No trivial computation
- Closer to the energy footprint

Comparative study of metrics

Runtime

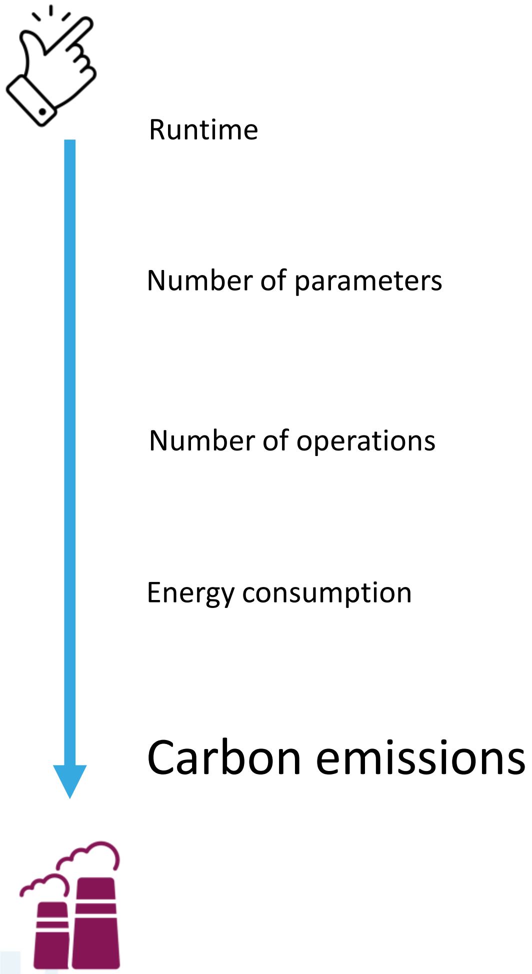
Number of parameters

Number of operations

Energy consumption

- Good indicator of the footprint
- Other jobs running
- Target a particular device

Comparative study of metrics



Runtime

Number of parameters

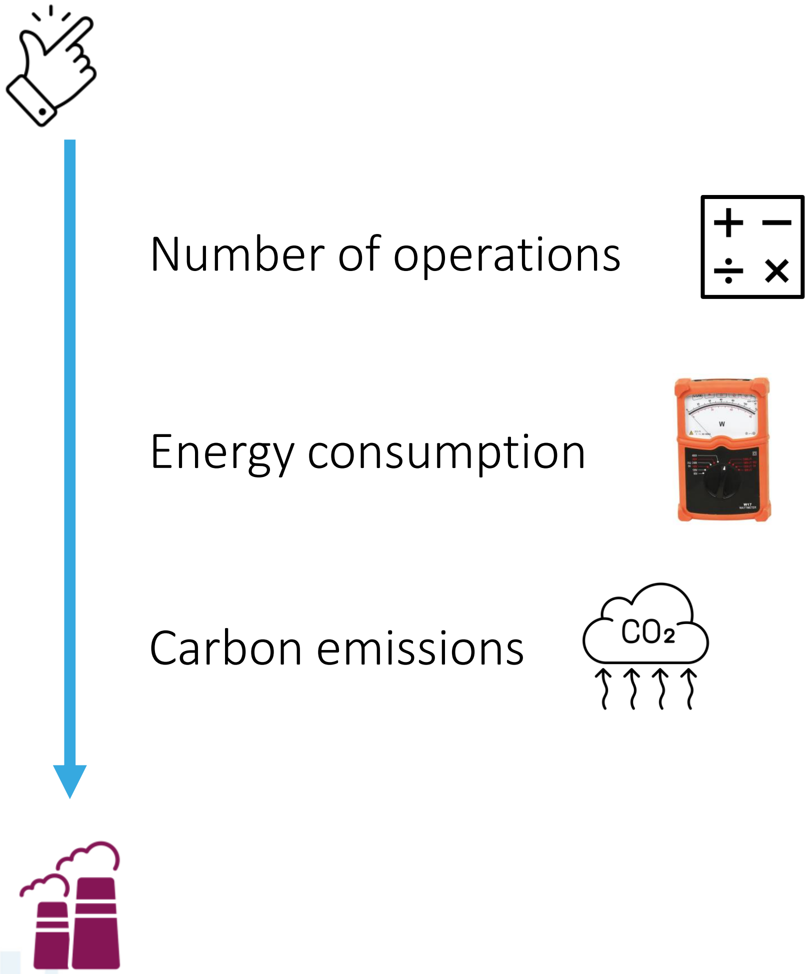
Number of operations

Energy consumption

Carbon emissions

- Direct link with energy consumption
- Real carbon footprint impact
- Depends on local electricity infrastructure

Comparative study of metrics

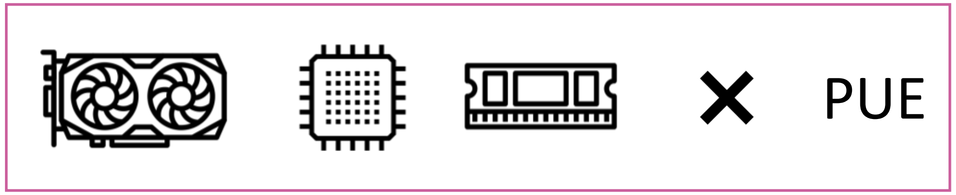


Number of operations

Energy consumption

Carbon emissions

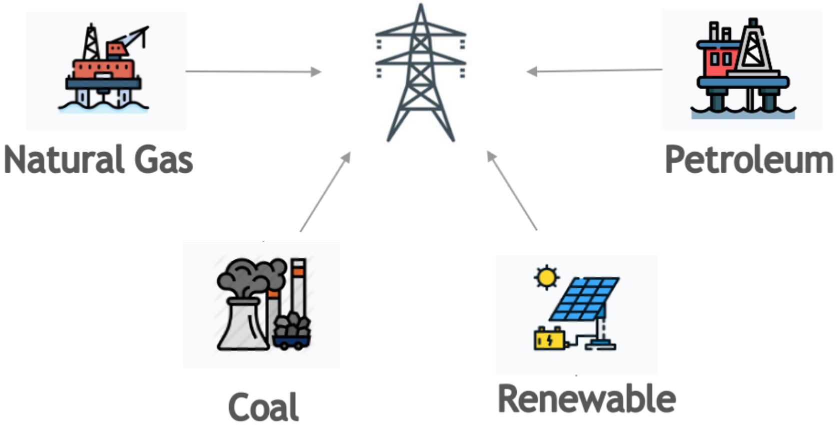
Carbon emissions



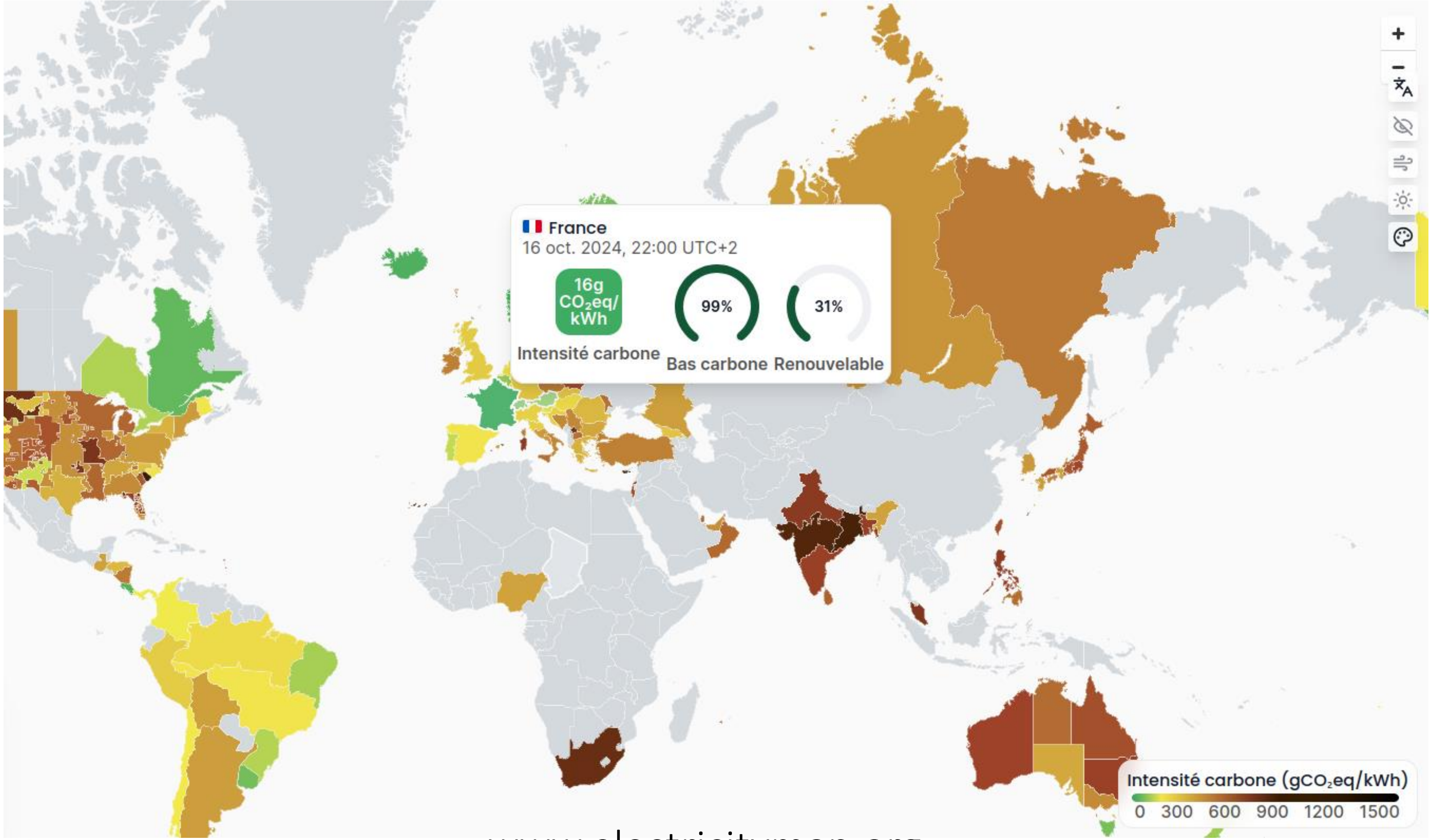
Energy consumption

×

Carbon intensity factor



☹ Leads to unfair comparisons



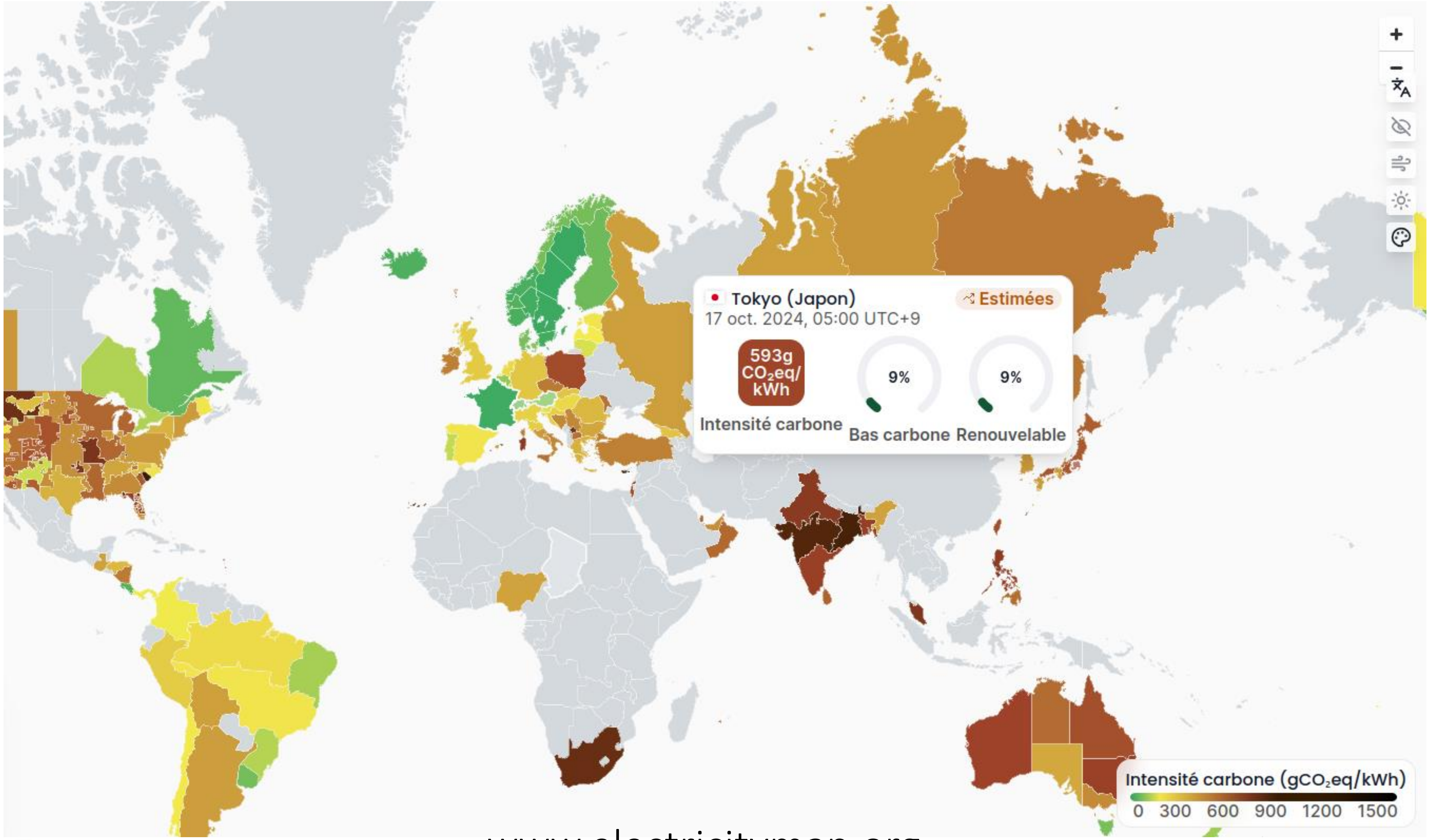
France
16 oct. 2024, 22:00 UTC+2

16g CO_2eq/kWh
Intensité carbone

99%
Bas carbone

31%
Renouvelable

www.electricitymap.org



Tokyo (Japon) Estimées

17 oct. 2024, 05:00 UTC+9

593g CO₂eq/kWh — Intensité carbone

9% — Bas carbone

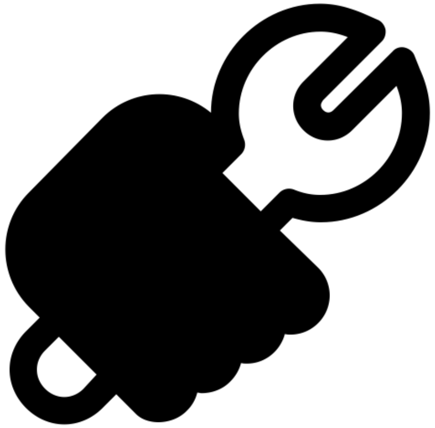
9% — Renouvelable

www.electricitymap.org

Towards a fair comparison

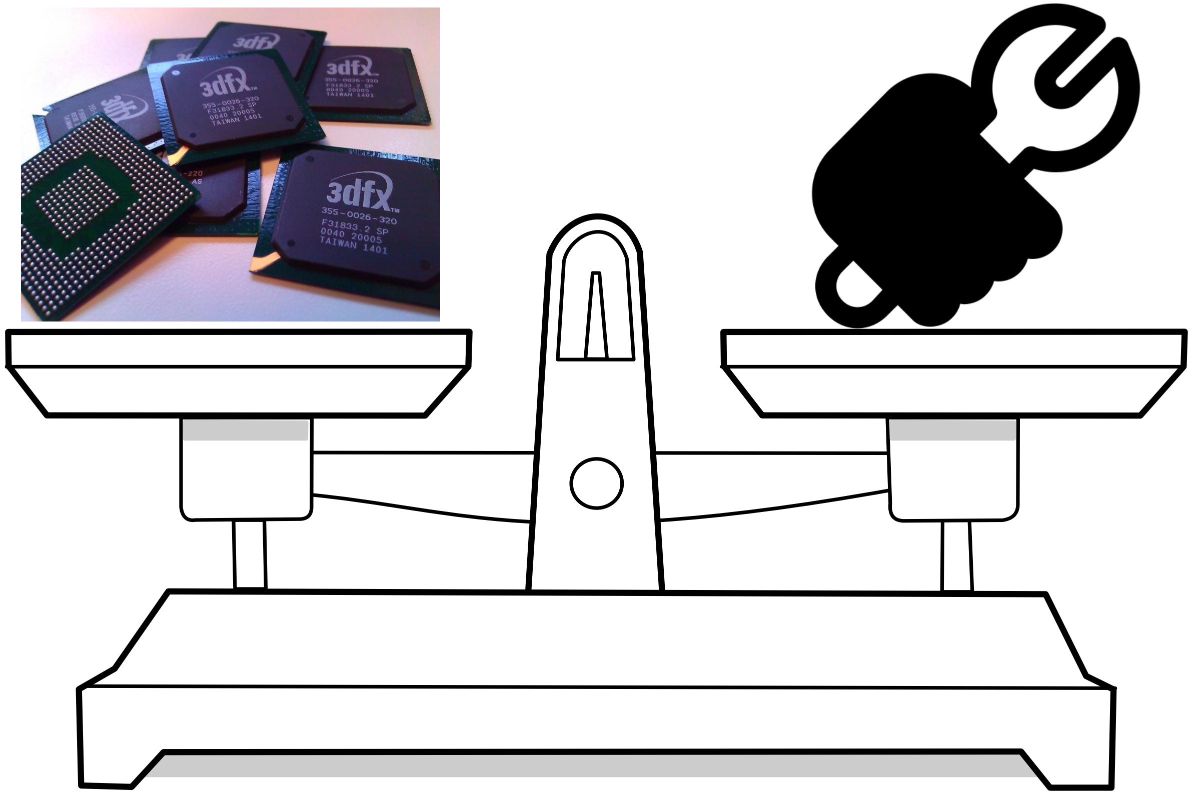
Of systems energy consumption

Motivations



Easy way to compare energy consumption across sites?

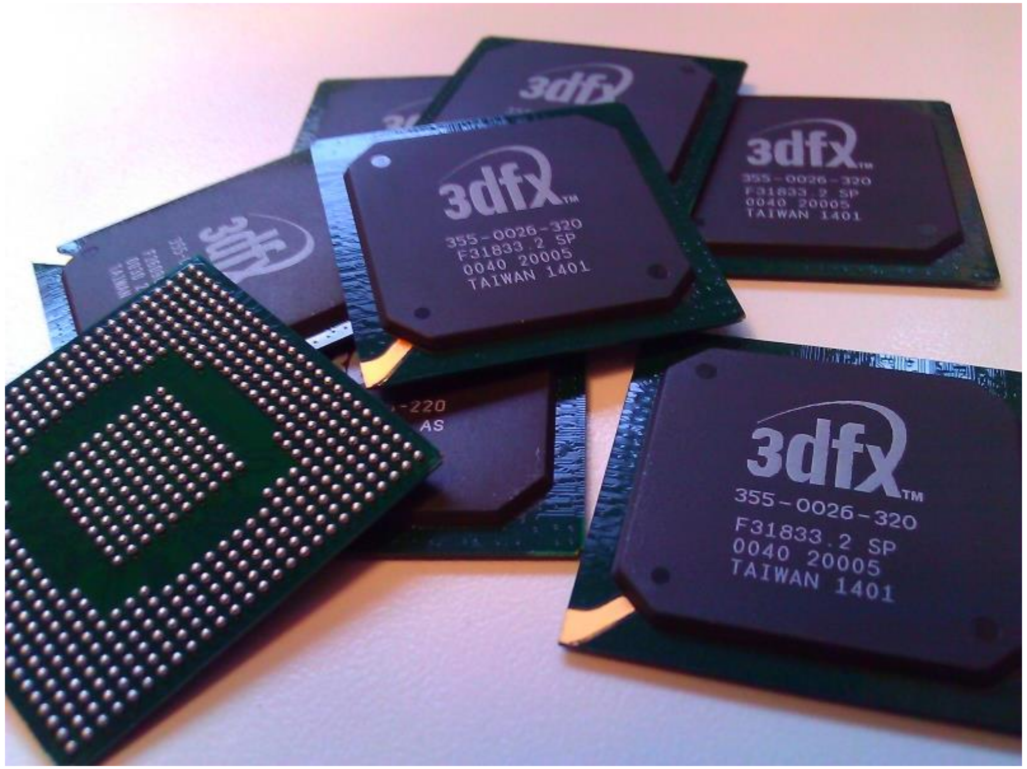
Images : Wikimedia



Which aspect impact the energy consumption most?
(for a same system)

Method

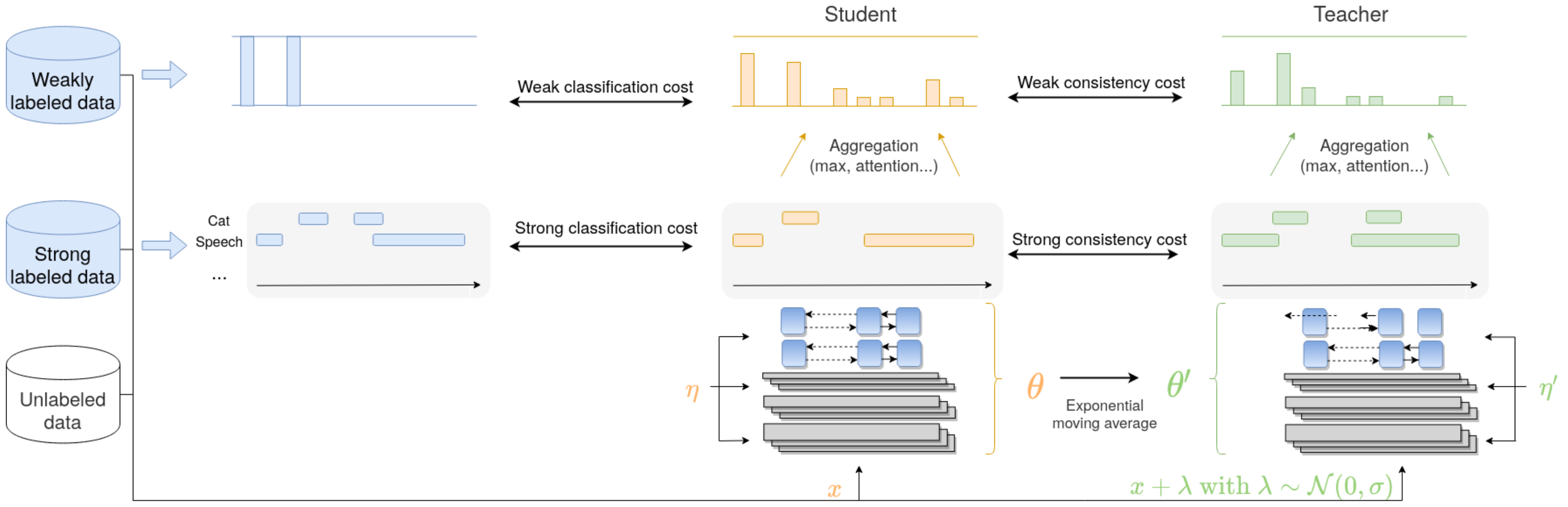
Benchmark training on several GPUs



Images : Wikimedia

Initial experiment

- Mean-teacher system
  - ~ 1M parameters

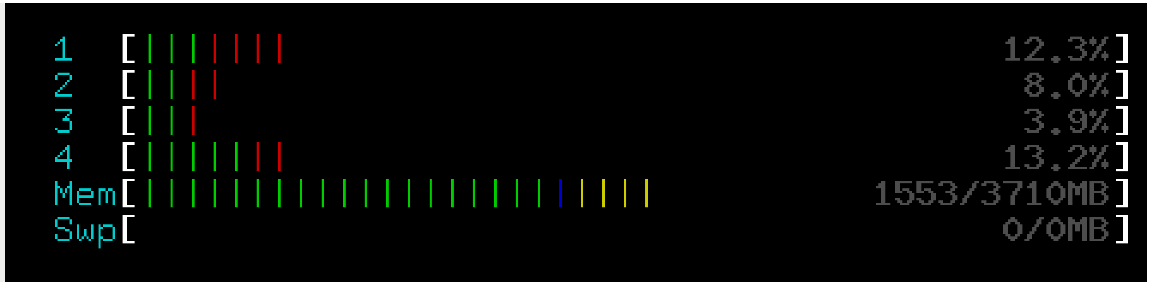


- **6 types of GPU** (from GTX 980 to A100)
- Train the baseline until convergence
  - **3 runs** for each hardware
  - Several batch sizes

Energy consumption

Large impact

Minor impact



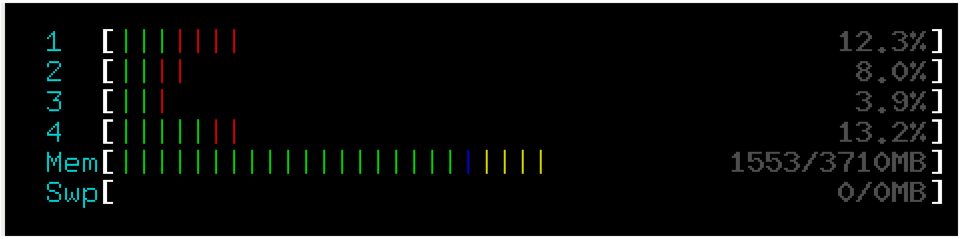
```
1   [||||||                                            12.3%]
2   [|||                                                8.0%]
3   [|||                                                3.9%]
4   [||||||                                            13.2%]
Mem[|||||||||||||||||||||||                       1553/3710MB]
Swp[                                                   0/0MB]
```

Images : Wikimedia

Energy consumption

Minor impact

Batch size



```
 1   [|||||||                               12.3%]
 2   [||||                                   8.0%]
 3   [|||                                    3.9%]
 4   [|||||||                               13.2%]
 Mem[||||||||||||||||||||||||          1553/3710MB]
 Swp[                                       0/0MB]
```

- No impact on the energy consumption per minute
- Impact on the training time…

Images : Wikimedia

Energy consumption

Large impact

GPU models

- Large impact on the energy consumption per minute
- Impact on the training time

Images : Wikimedia

First proposal

Normalize consumption depending on the hardware

- Everybody runs a reference training on local hardware
- Weight the reported energy consumption

We did try this. More about it later…

First proposal

Assumptions that **<u>need being checked</u>**

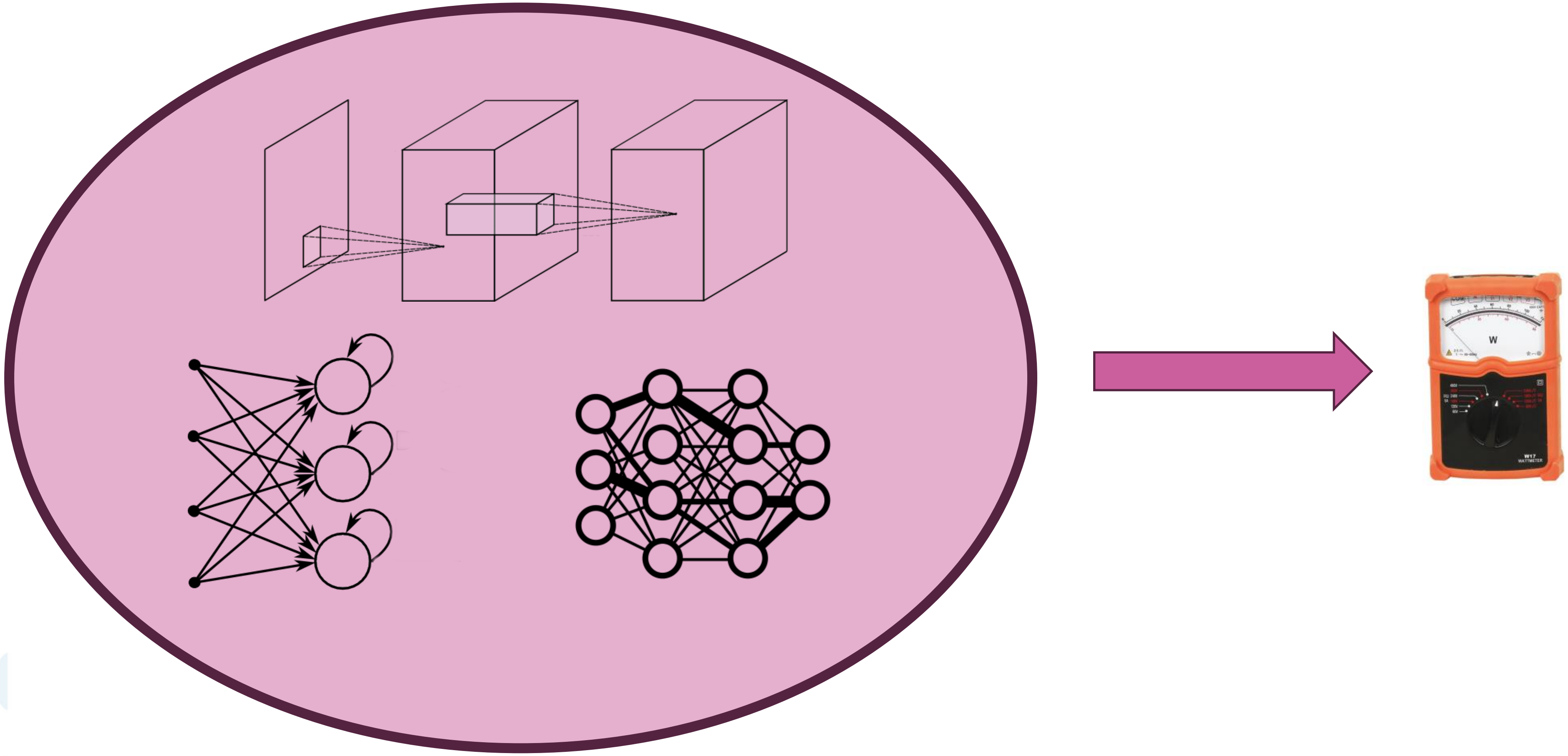
- Is the batch size a good proxy to measure GPU use?
- Is one point sufficient to normalize?

Is the batch size a good proxy to measure GPU use?

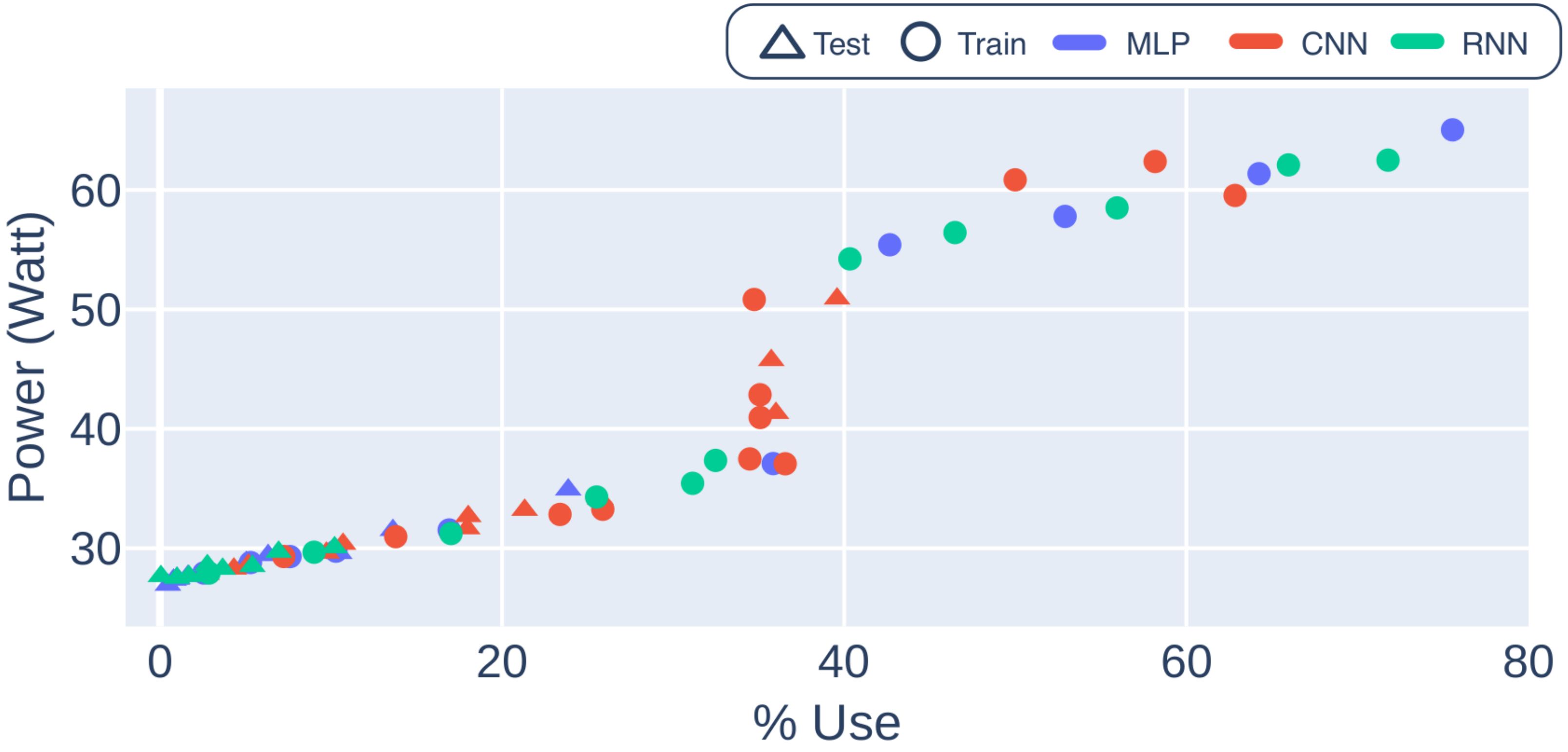
Is energy consumption **independent from GPU use**?

Energy vs GPU use (C. Douwes)

Experiment setup



Energy vs GPU use



Energy consumption <u>depends on GPU use</u>

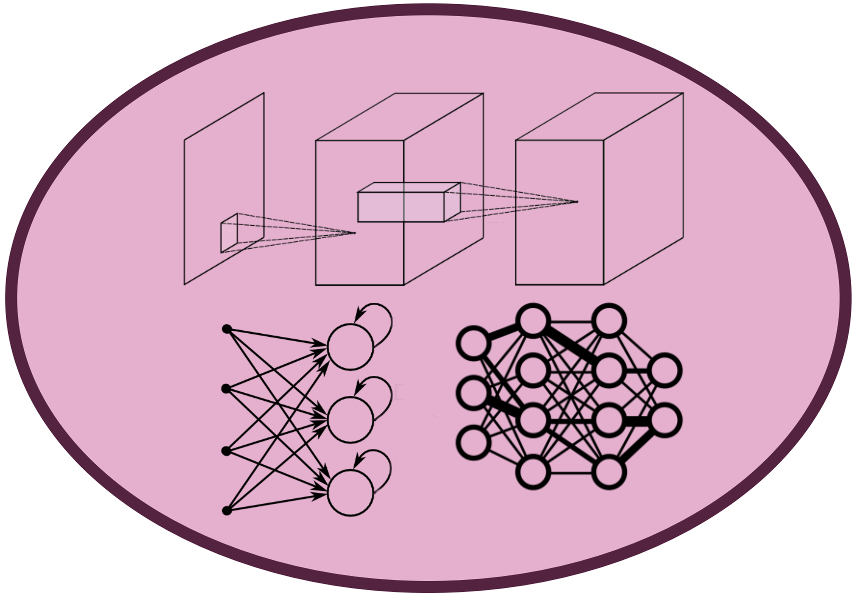
And the relationship is not linear ☹

How can we normalize the energy consumption?

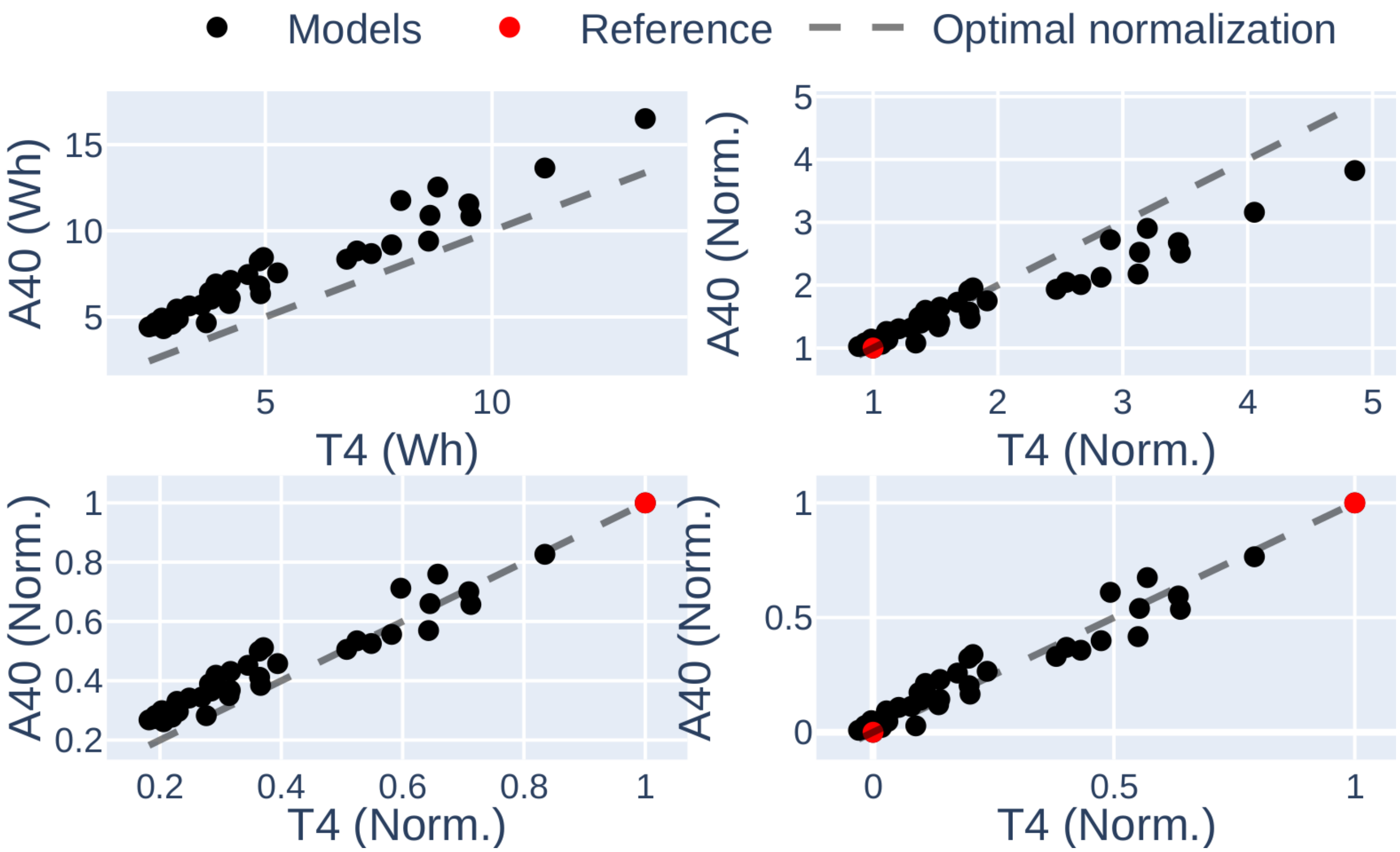
A single reference point is probably not the most appropriate…

Energy normalization (C. Douwes)

Experiment setup

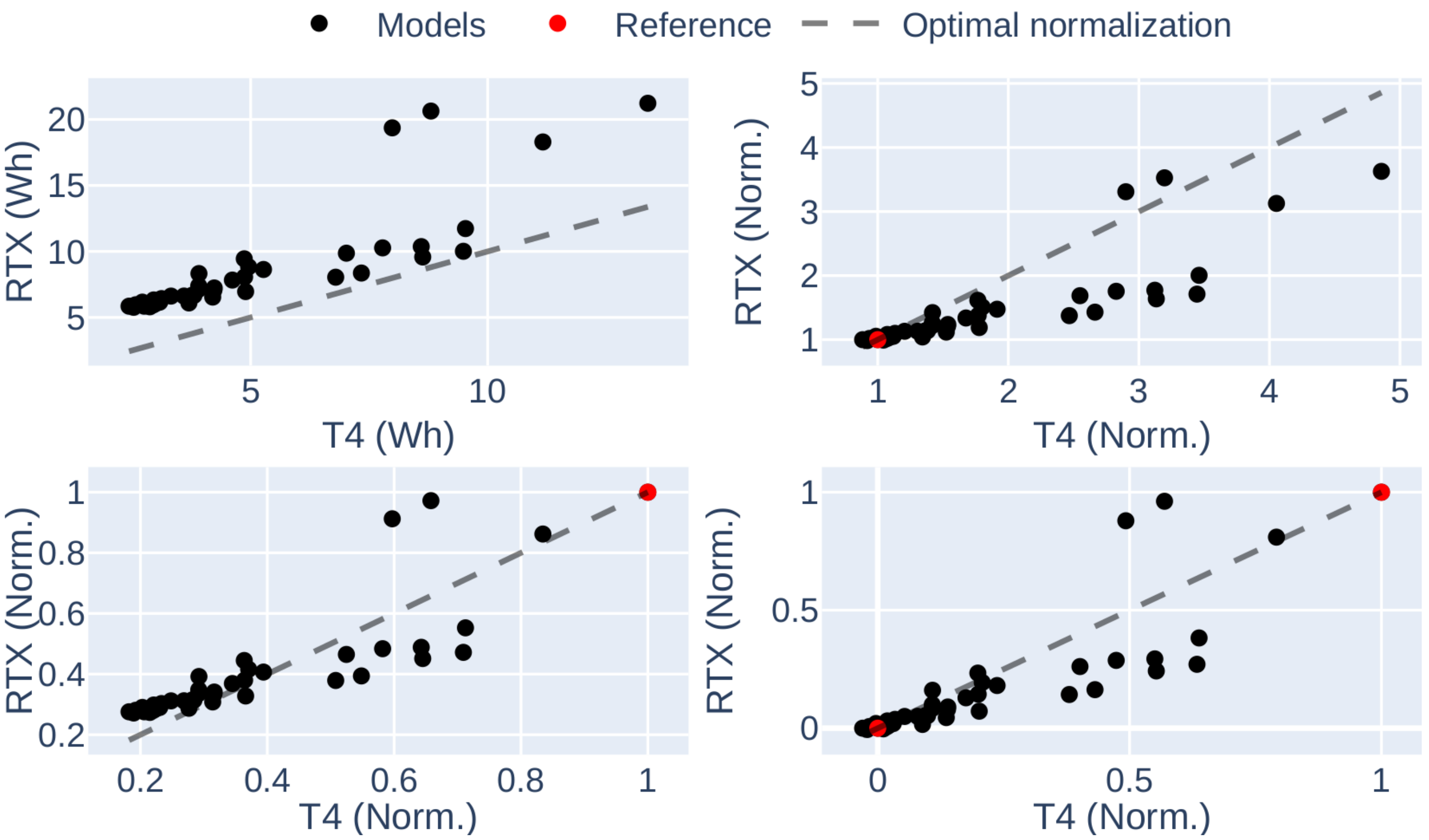


Linear regression



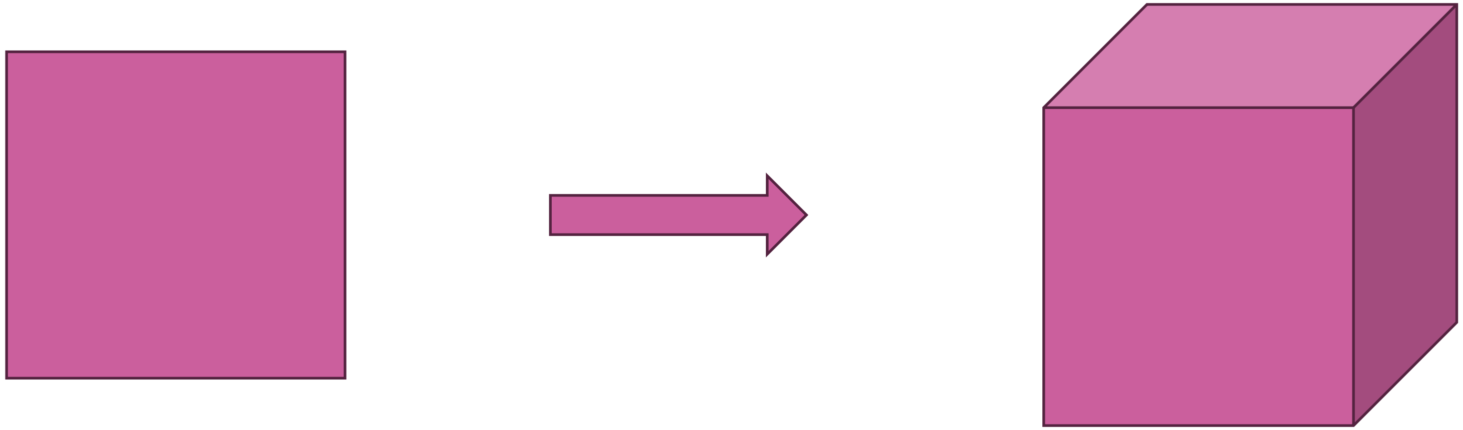
Two points seems to work well

Linear regression



Well, not always… ☹

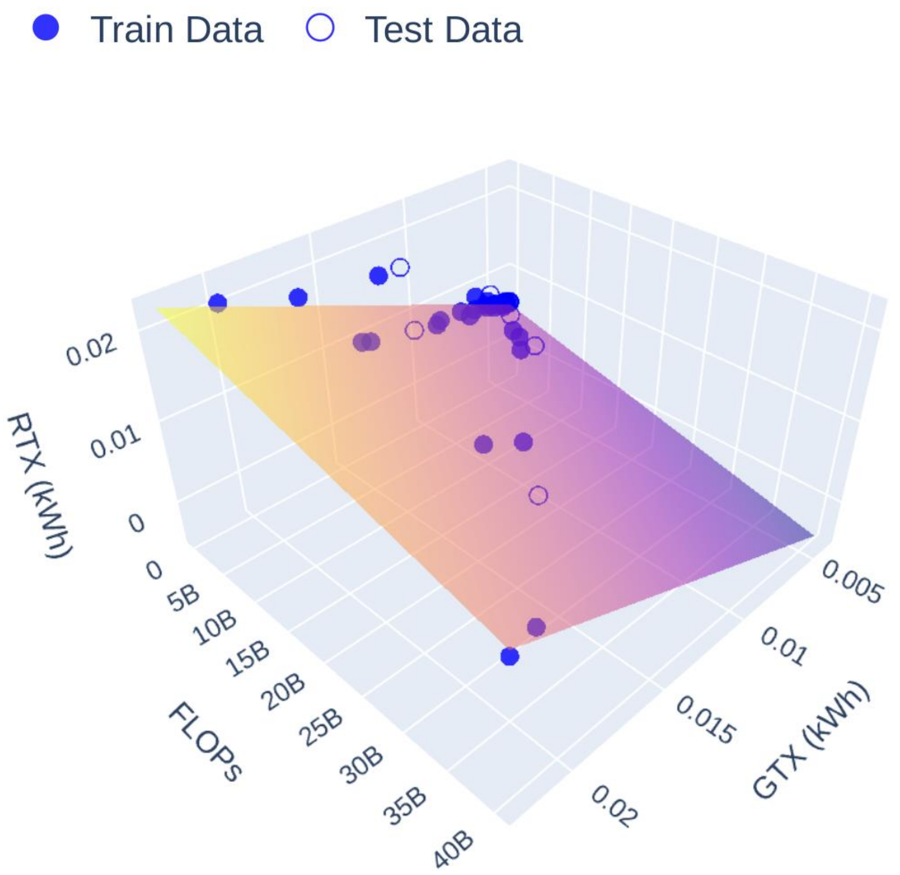
Potential workaround



Add other axis to the regression

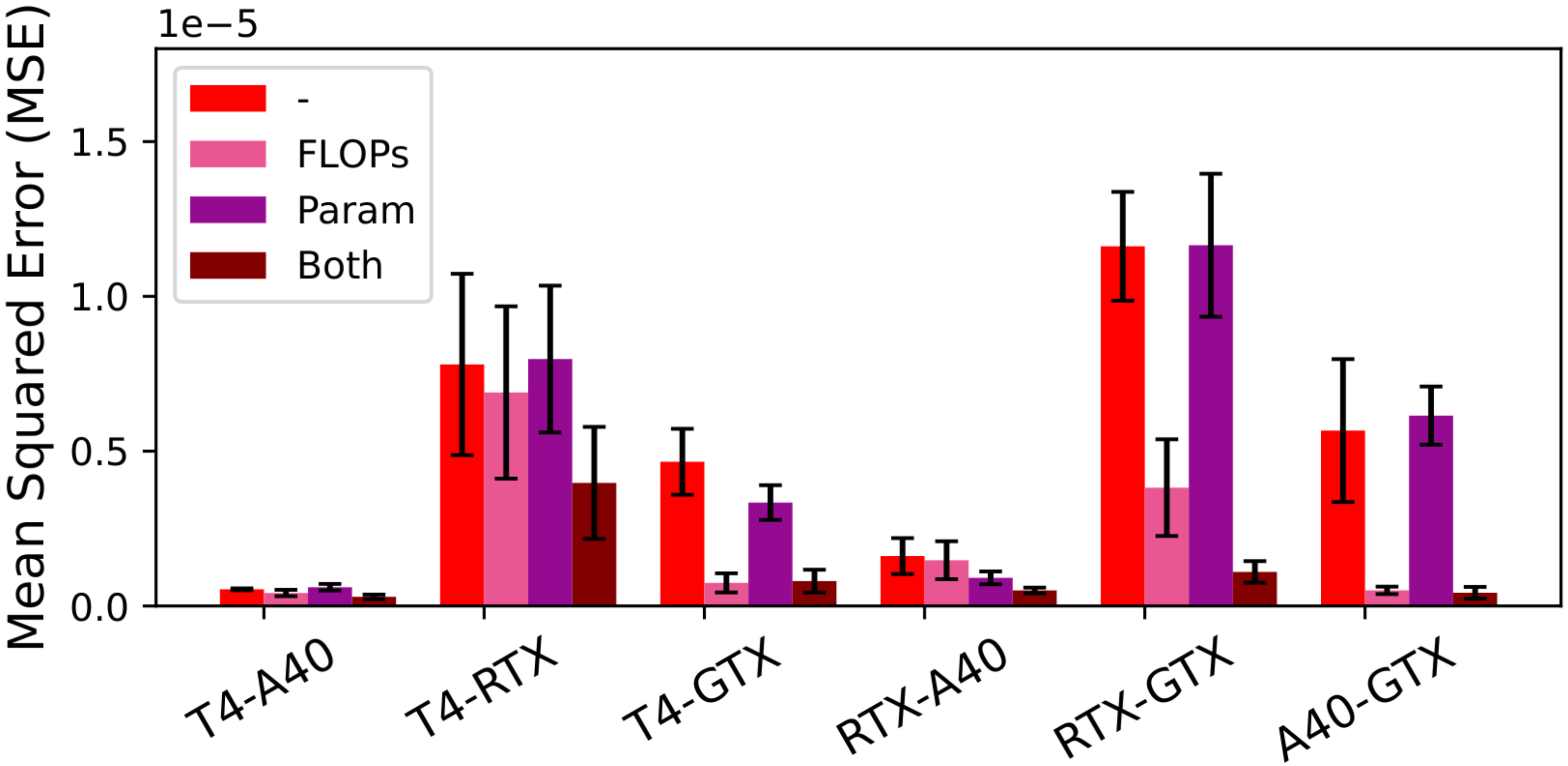
We did try other regressions and more reference points too.

Linear regression (with FLOPS)



Seems to be improving

Linear regression (with FLOPS and number of parameters)



Improving!

But we need more reference points…

Case study
on sound event detection

Efficiency / Accuracy

How?

Efficiency

/

Accuracy

Efficiency

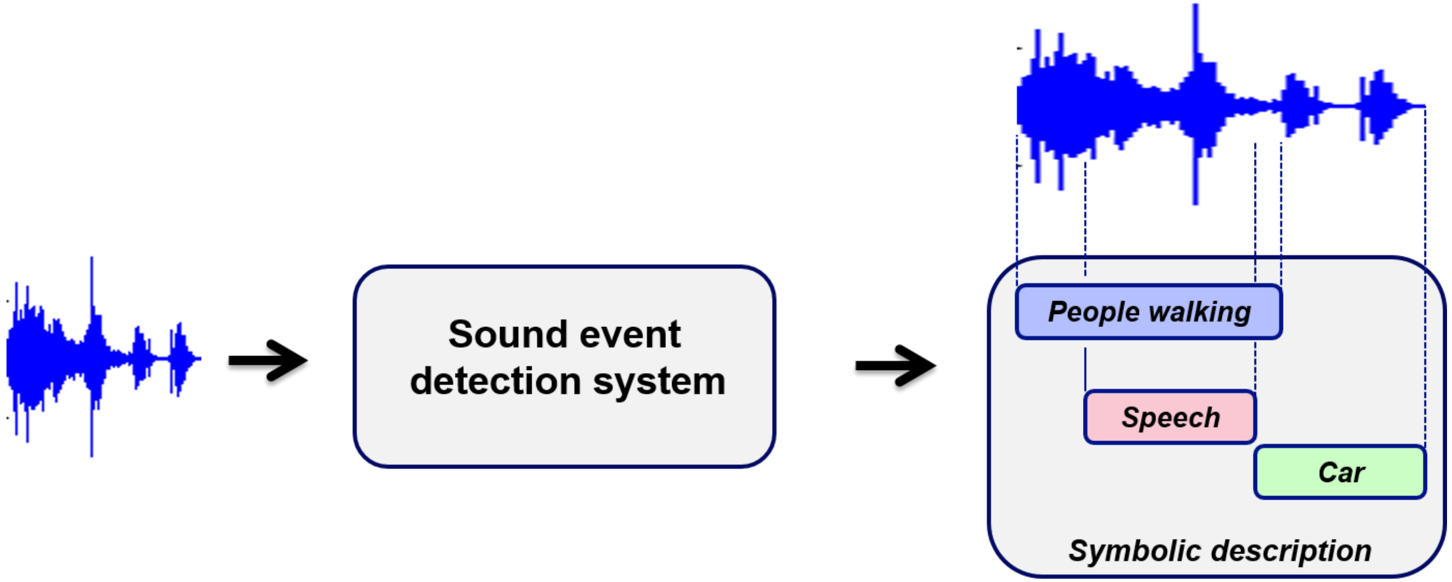
Accuracy

DCASE Challenge Task 4

DCASE CHALLENGE

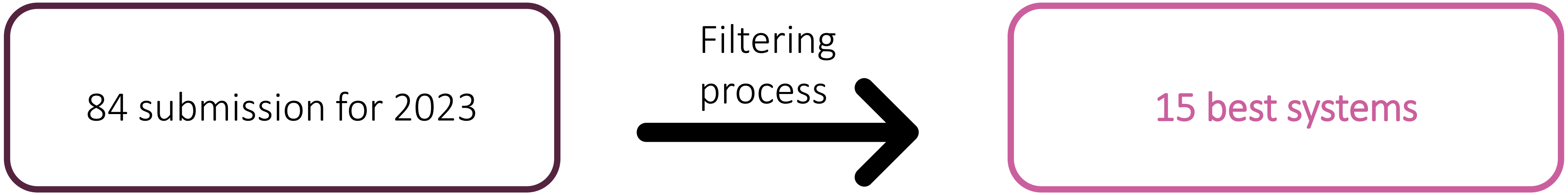
Domestic

Task 4



Loria

Analysis setup (F. Ronchini)



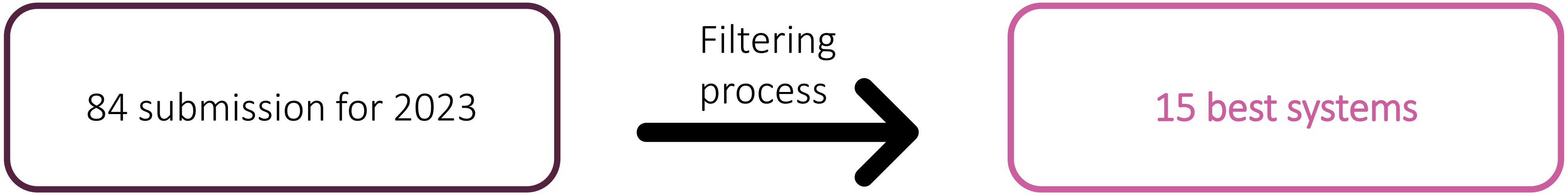
84 submission for 2023

Filtering process

15 best systems

Relation between energy consumption and SED metrics

Analysis setup (F. Ronchini)



84 submission for 2023

Filtering process

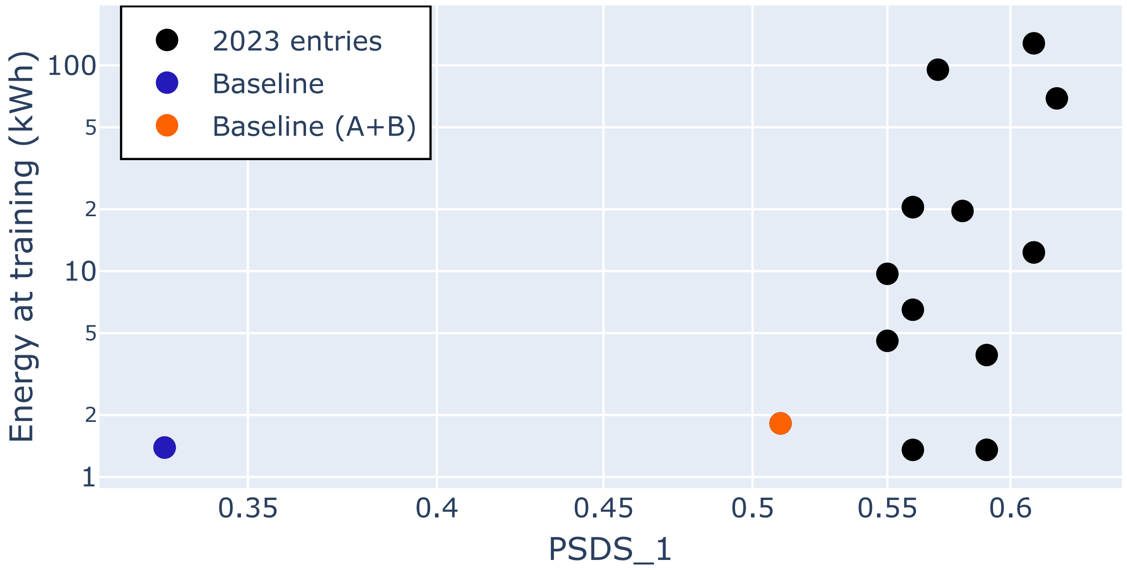
15 best systems

⚠ We normalize with a single point!

Relation between energy consumption and SED metrics

Performance and energy balance: a comprehensive study of state-of-the-art sound event detection systems - https://arxiv.org/abs/2310.03455
Ronchini Francesca, Serizel Romain

Performance vs. energy consumption

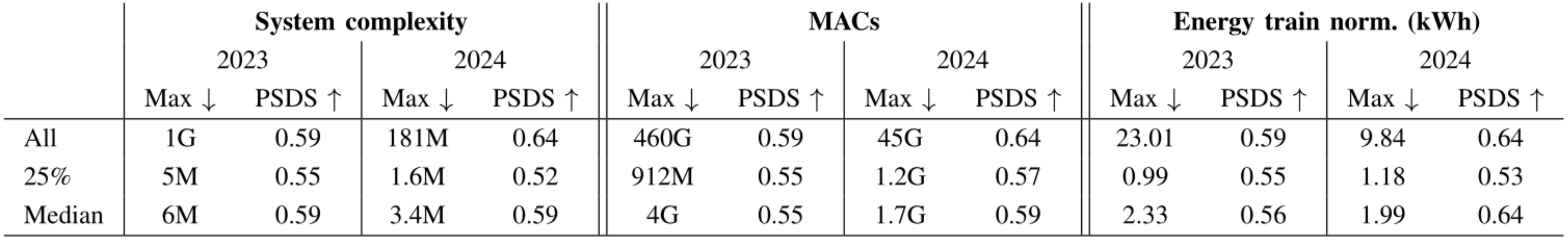


Top-performing systems are not always the systems that consume the most energy!!! ☺

PSDS: High is good!

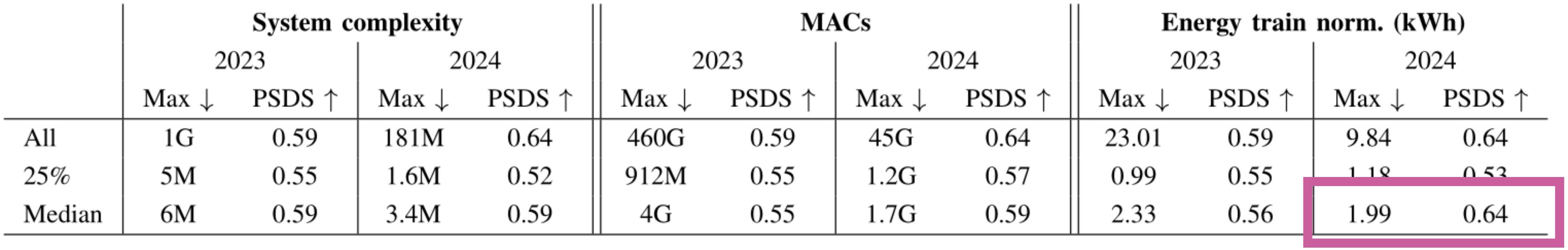
Threshold on energy consumption

How much does performance degrade with a footprint cap?



	System complexity				MACs				Energy train norm. (kWh)			
	2023		2024		2023		2024		2023		2024	
	Max ↓	PSDS ↑	Max ↓	PSDS ↑	Max ↓	PSDS ↑	Max ↓	PSDS ↑	Max ↓	PSDS ↑	Max ↓	PSDS ↑
All	1G	0.59	181M	0.64	460G	0.59	45G	0.64	23.01	0.59	9.84	0.64
25%	5M	0.55	1.6M	0.52	912M	0.55	1.2G	0.57	0.99	0.55	1.18	0.53
Median	6M	0.59	3.4M	0.59	4G	0.55	1.7G	0.59	2.33	0.56	1.99	0.64

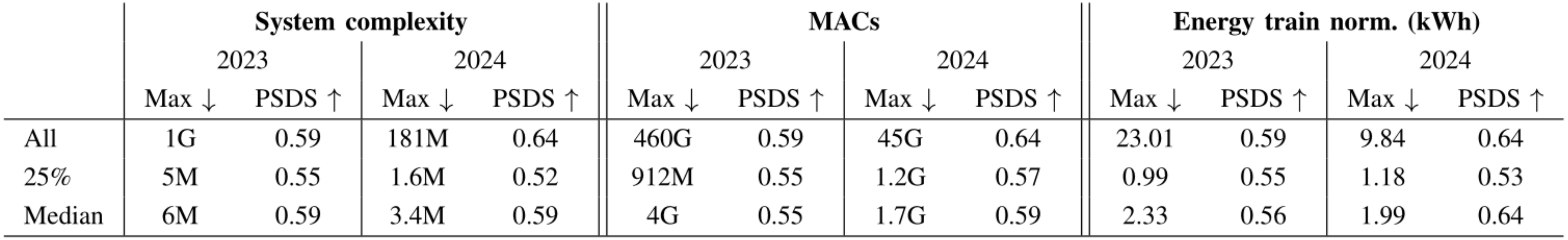
Threshold on energy consumption



	System complexity				MACs				Energy train norm. (kWh)			
	2023		2024		2023		2024		2023		2024	
	Max ↓	PSDS ↑	Max ↓	PSDS ↑	Max ↓	PSDS ↑	Max ↓	PSDS ↑	Max ↓	PSDS ↑	Max ↓	PSDS ↑
All	1G	0.59	181M	0.64	460G	0.59	45G	0.64	23.01	0.59	9.84	0.64
25%	5M	0.55	1.6M	0.52	912M	0.55	1.2G	0.57	0.99	0.55	1.18	0.53
Median	6M	0.59	3.4M	0.59	4G	0.55	1.7G	0.59	2.33	0.56	1.99	0.64

- Performance remains rather stable regardless of the threshold cap
- For 2024 the best system is below the median energy!

Threshold on energy consumption



	System complexity				MACs				Energy train norm. (kWh)			
	2023		2024		2023		2024		2023		2024	
	Max ↓	PSDS ↑	Max ↓	PSDS ↑	Max ↓	PSDS ↑	Max ↓	PSDS ↑	Max ↓	PSDS ↑	Max ↓	PSDS ↑
All	1G	0.59	181M	0.64	460G	0.59	45G	0.64	23.01	0.59	9.84	0.64
25%	5M	0.55	1.6M	0.52	912M	0.55	1.2G	0.57	0.99	0.55	1.18	0.53
Median	6M	0.59	3.4M	0.59	4G	0.55	1.7G	0.59	2.33	0.56	1.99	0.64

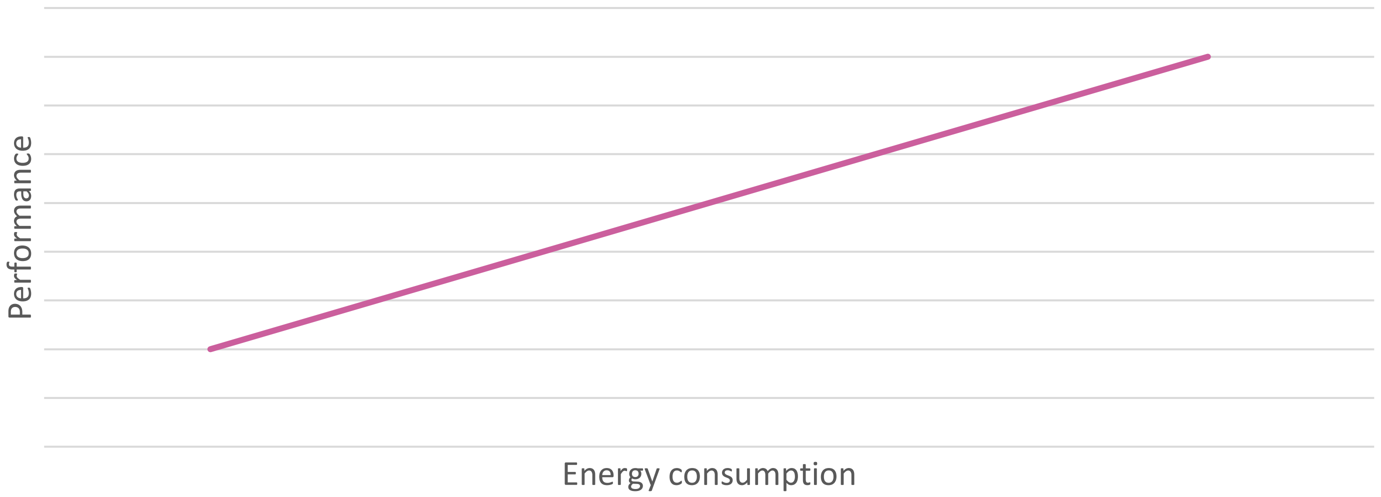
- Performance remains rather stable regardless of the threshold cap
- For 2024 the best system is below the median energy!

We are spending a large amount of energy and computation to increase the performance only marginally. ☹ ☹ ☹

Bonus question

Initial study (see Section 2)

Is performance vs energy consumption linear?

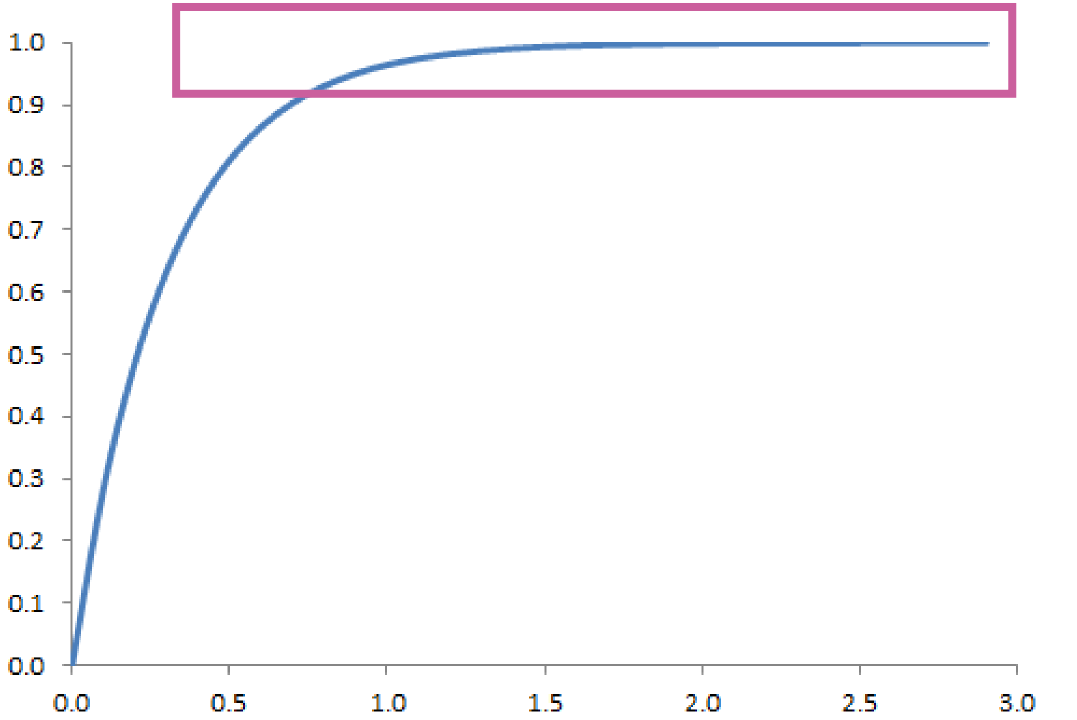


When should I stop training?

Energy vs performance

50%+ energy

5 % increase in performance...



© 2013 Alan Fletcher – permission is given to copy for non-commercial use.

Images : Wikimedia

Conclusions

Take home

- Many (complementary) metrics
  - A single one is not sufficient
- Many potential shortcomings when comparing systems
  - Across site, Hardware, Configuration

➡ Need for standardize procedures

- Combining footprint/performance metric is not obvious
  - Balance between the criterion
  - Fit actual application needs

- How can we can this attractive at community level?

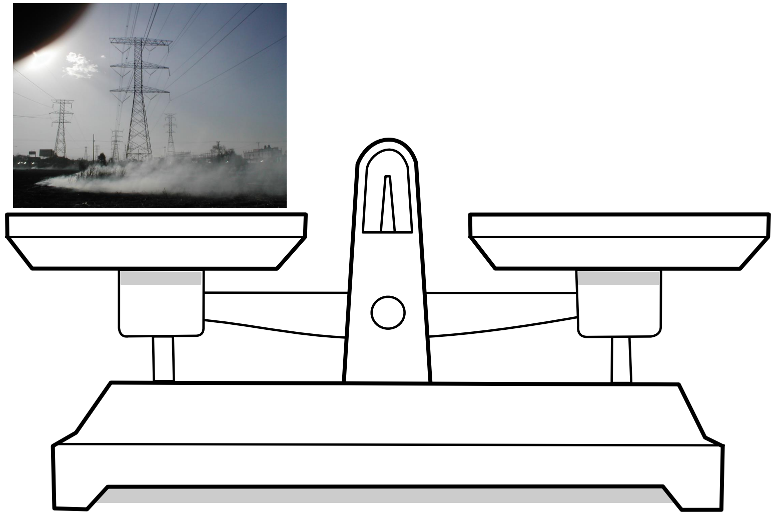
Proposal

- Steer away from simple performance comparison

What is a worthy improvement?

✕

- Define the cost we are ready to pay for this improvement



Questions? Remarks?

Images : Wikimedia