

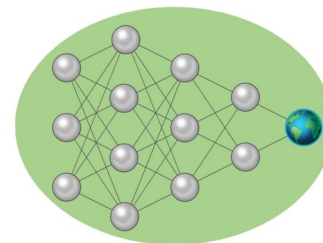
PEARLS Status in Zeuthen

Making Machine Learning Models Efficient

Simran Gurdasani, DESY Zeuthen



Bundesministerium
für Forschung, Technologie
und Raumfahrt



PEARLS

Growth in ML model size

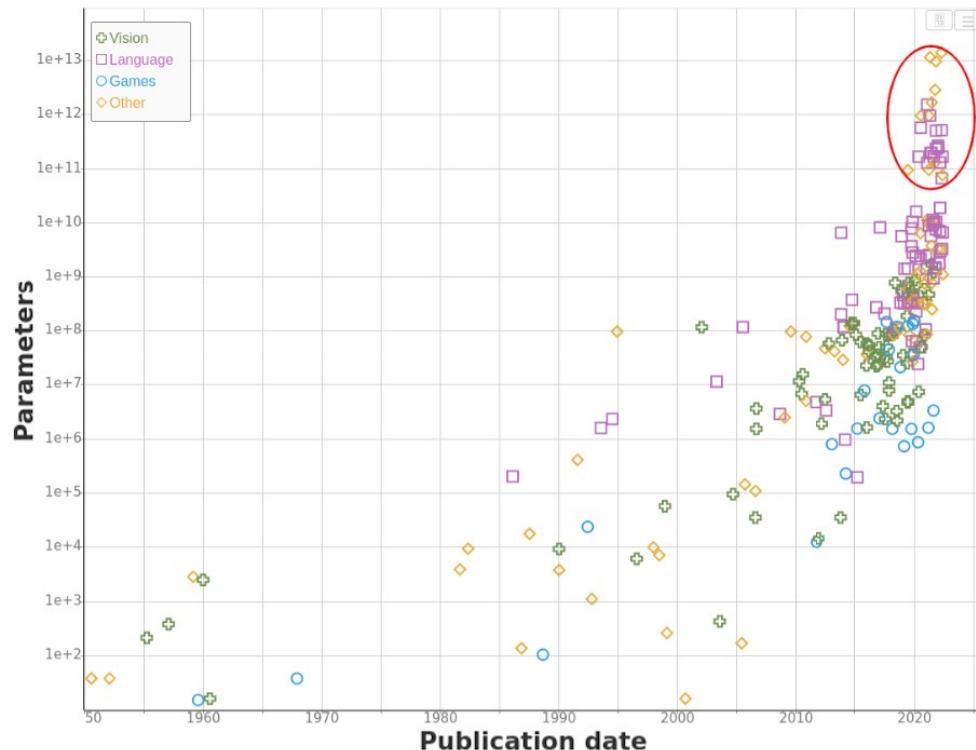
A study of notable models performed in [1] shows that model sizes grew:

- 10^7 between 1952 -> 2018
- 10^4 between 2018 -> 2022

[1] <https://arxiv.org/pdf/2207.02852>

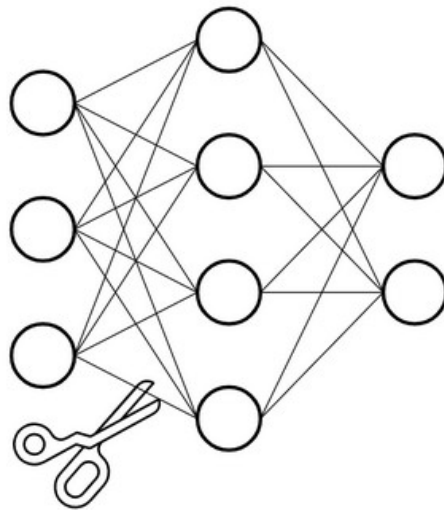
Bigger models cost more: latency, memory, power, CO2

However, often these models are “sparse”/“empty” meaning the parameter values are close to zero and don’t contribute to the output.

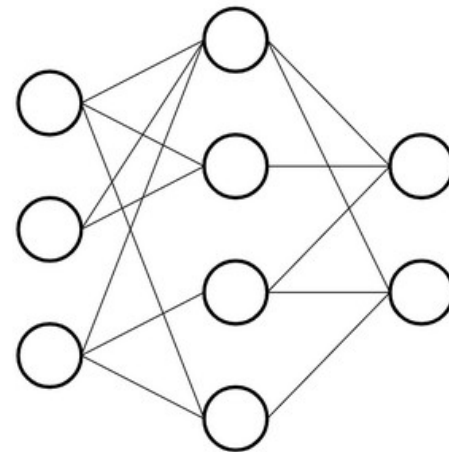


What is Pruning?

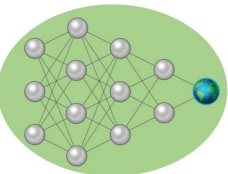
- Pruning identifies and removes the least important parameters — reducing the network size.
- Pruned models cut compute and energy, and hence CO₂.
- Intelligent pruning methods are needed to prune without sacrificing model performance.



Before pruning



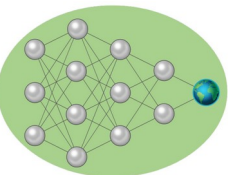
After pruning



When to Prune? – Amortization

- Pruning has an upfront cost – post training calculations and engineering of the model.
- Pays off only with high inference volume relative to that cost.





When to Prune? – Amortization

- Pruning has an upfront cost – post training calculations and engineering of the model.
 - Pays off only with high inference volume relative to that cost.
- 
- Example: Real-time triggering of events in HEP. Models deployed once run continuously for the experiment's data-taking.
 - Contrast: a model used once for a single analysis doesn't benefit the same way.

How does one Prune?

Different ways to
apply pruning:

Magnitude

Prunes smallest
weights

Gradient

Estimates importance
from gradients

Sensitivity

Measures loss
impact directly

Different times to
apply pruning:

Pre-training

Network not yet
trained

During training

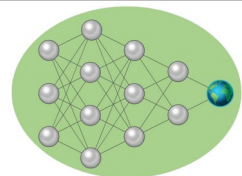
Weights actively
updating

Post-training

Model already
converged

Choose a regime based on our goals:
CO₂ saving means choosing the most efficient in computation

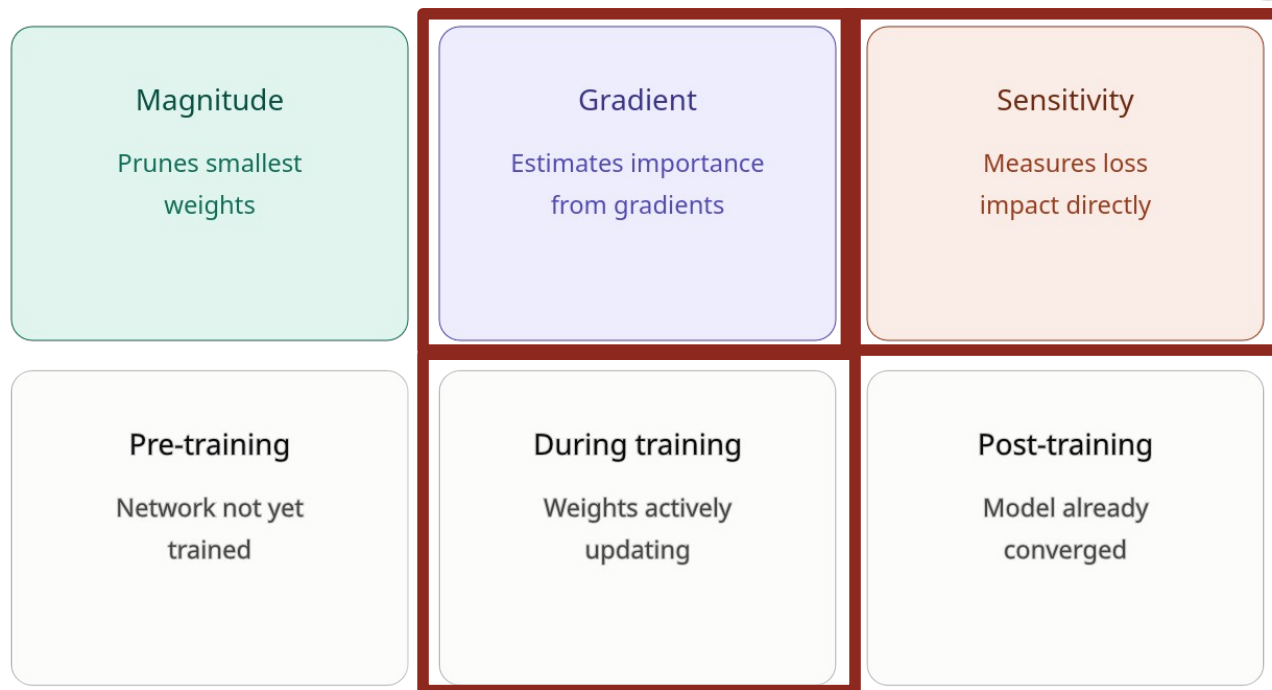
How does one Prune?



Different ways to
apply pruning:

Different times to
apply pruning:

Computationally
Expensive



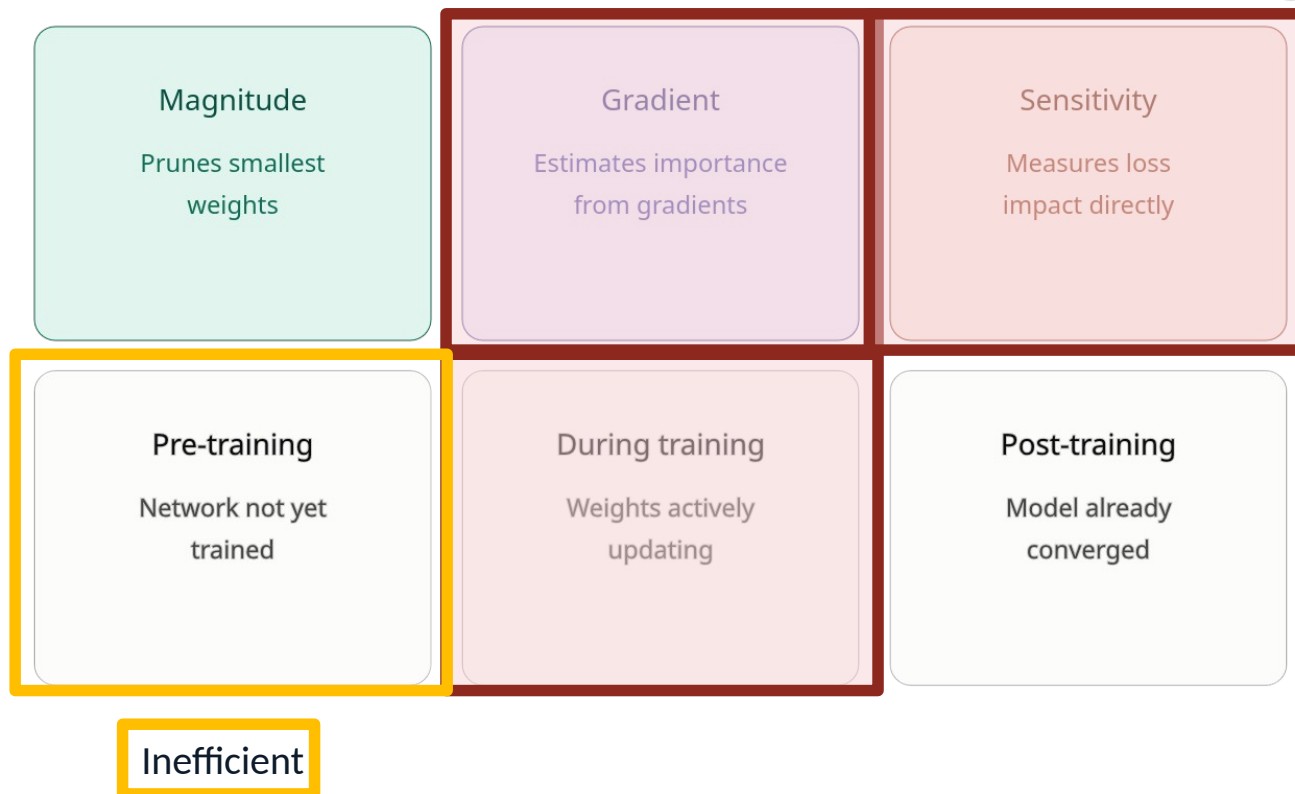
How does one Prune?

Different ways to
apply pruning:

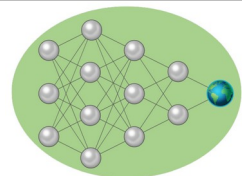
Different times to
apply pruning:

Computationally
Expensive

Inefficient



How does one Prune?



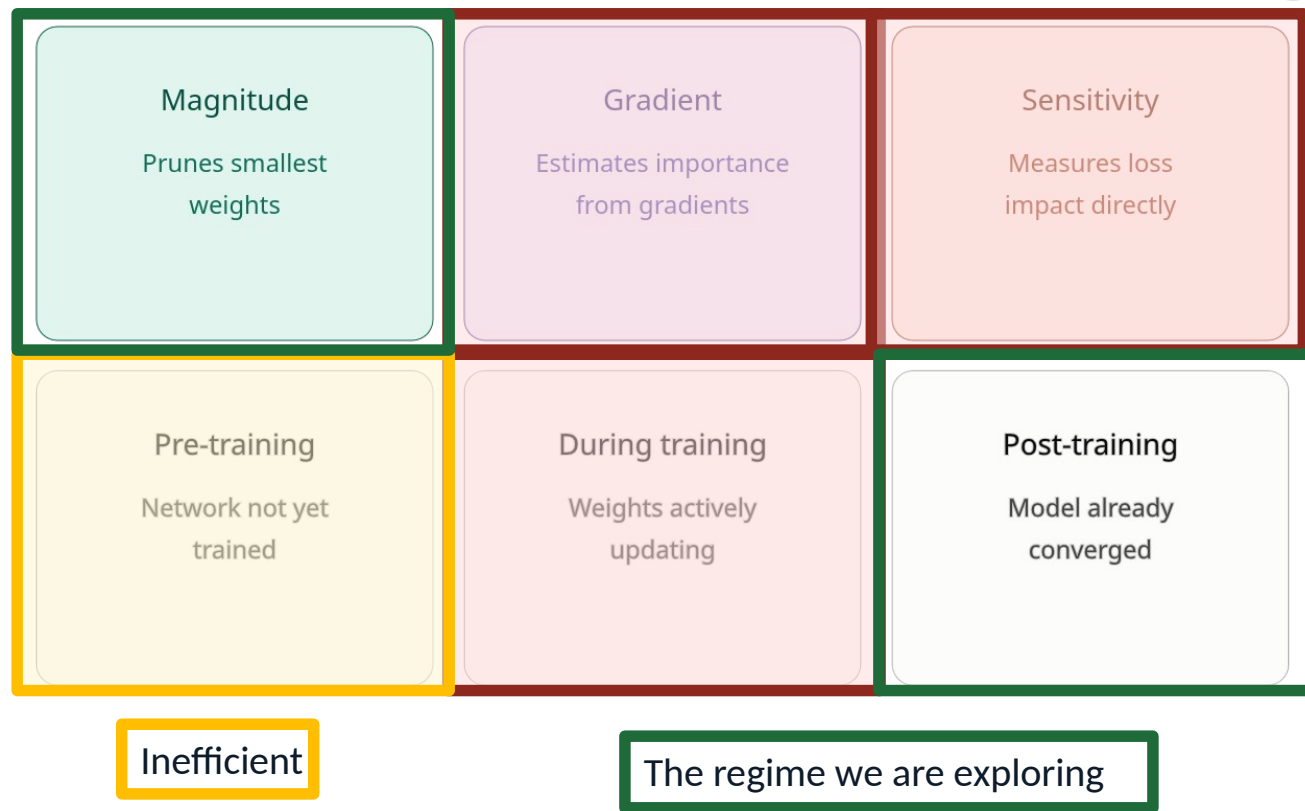
Different ways to
apply pruning:

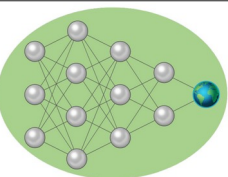
Different times to
apply pruning:

Computationally
Expensive

Inefficient

The regime we are exploring





And who is “We”???

The PEARLS Pruning Team



Simran Gurdasani



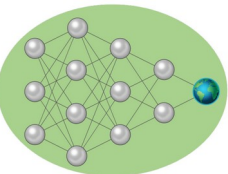
Josephine Larissa Prochnow

Simran:

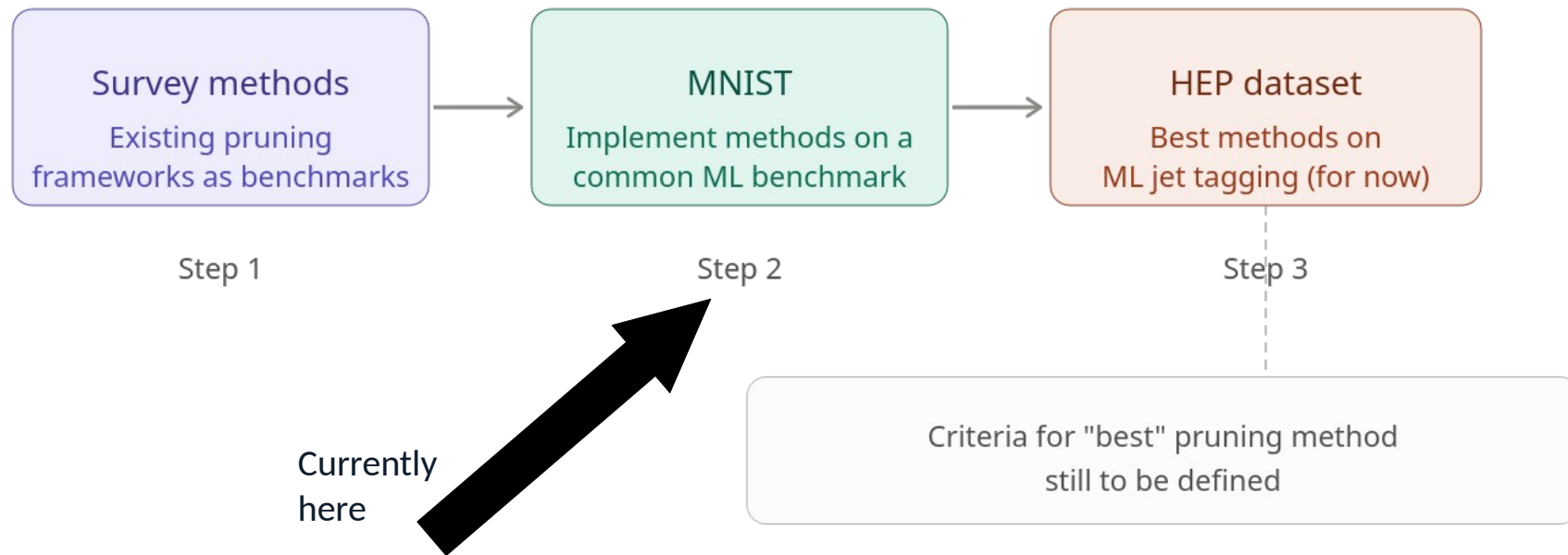
- PostDoc @ 10% on Pruning

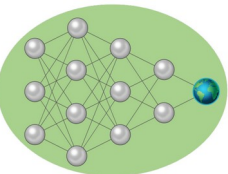
Larissa:

- March - Sep 2026:
HiWi @ 50% on Pruning
- Sep 2026 - Sep 2027:
WiMi @ 100% Pruning

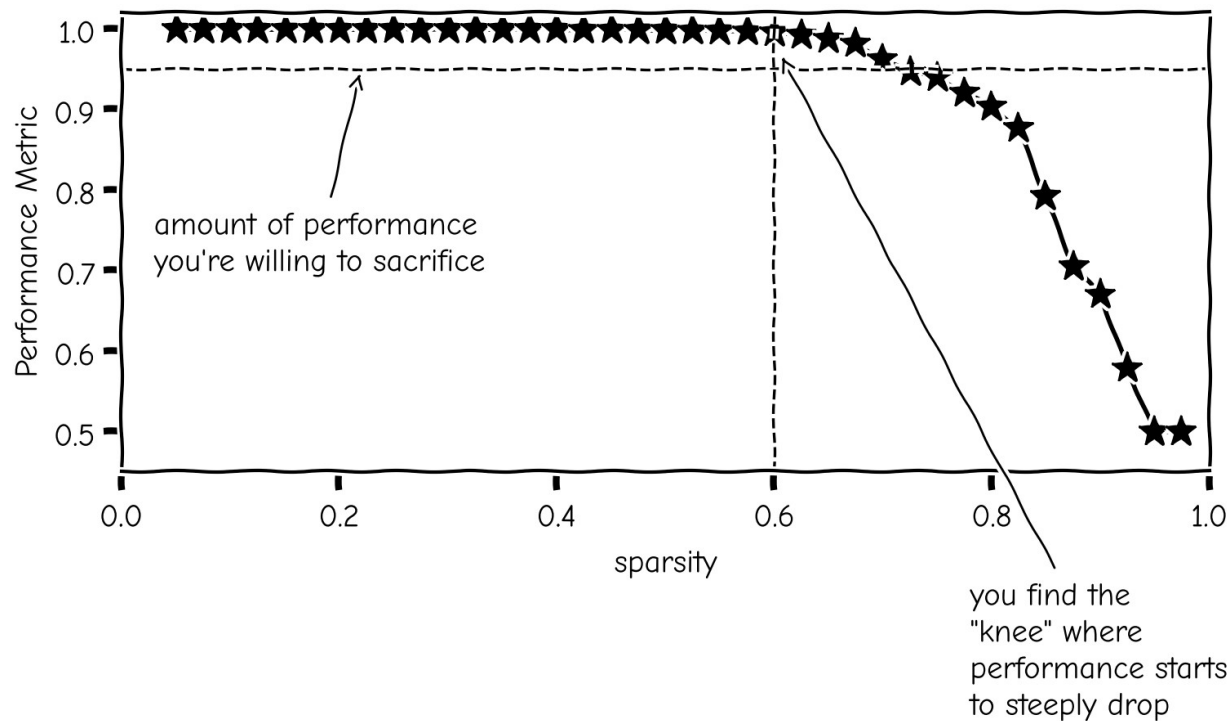


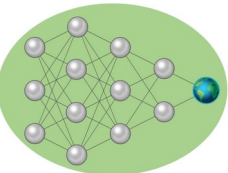
Research Plan





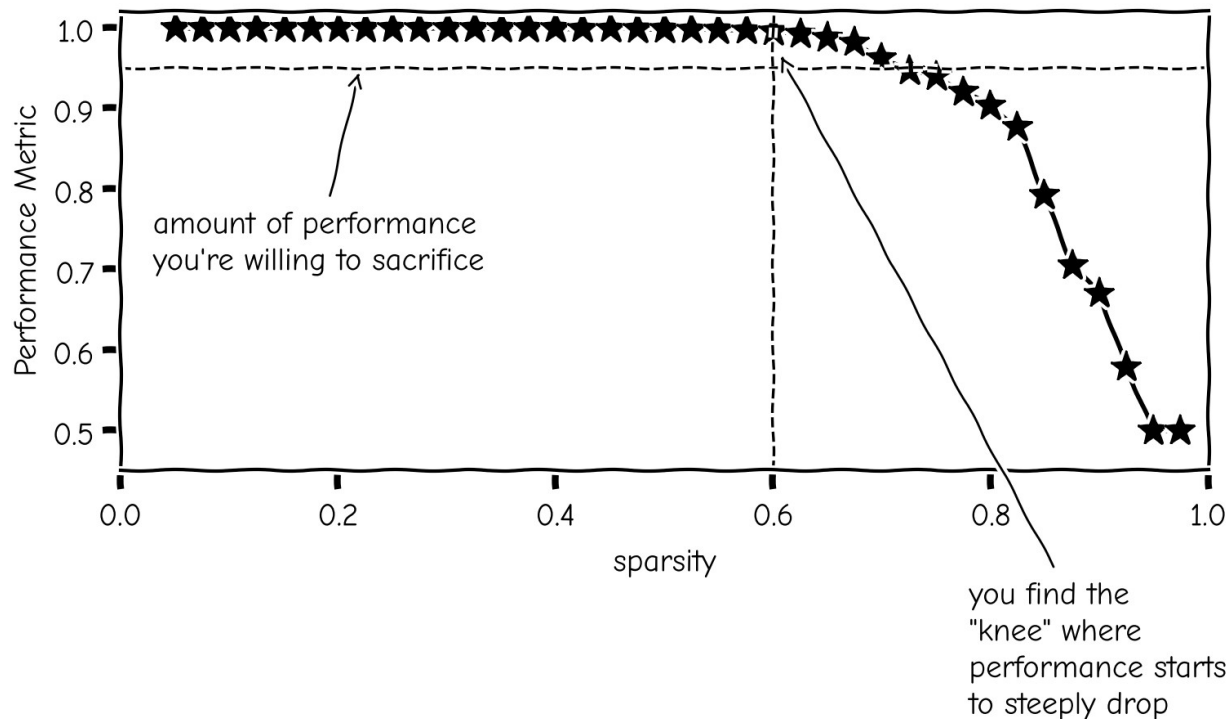
“Best” Pruning Method

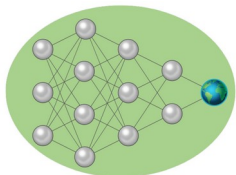




“Best” Pruning Method

- We currently track AUC vs. sparsity as our core metric.
- Sparsity is expected to scale with CPU time and memory, (when pruning is applied in a certain way).
- AUC makes sure we take into account false rate and not on accuracy.

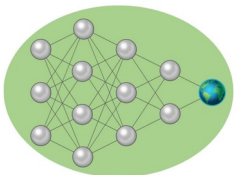




Open Questions



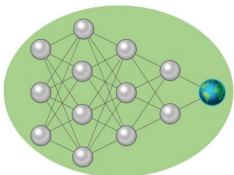
- Does sparsity reduce time and memory consumption for GPUs as well?
Or do we need completely different pruning for different computing architecture?



Open Questions



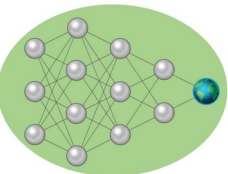
- Does sparsity reduce time and memory consumption for GPUs as well?
Or do we need completely different pruning for different computing architecture?
- Should we consider something besides CPU/GPU Memory and Time?
Example: FLOPs or bit operations as a hardware-independent proxy?



Open Questions



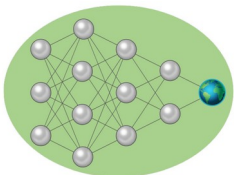
- Does sparsity reduce time and memory consumption for GPUs as well?
Or do we need completely different pruning for different computing architecture?
- Should we consider something besides CPU/GPU Memory and Time?
Example: FLOPs or bit operations as a hardware-independent proxy?
- Should we measure CO₂ cost directly?
Or infer it from sparsity?



Open Questions



- Does sparsity reduce time and memory consumption for GPUs as well?
Or do we need completely different pruning for different computing architecture?
- Should we consider something besides CPU/GPU Memory and Time?
Example: FLOPs or bit operations as a hardware-independent proxy?
- Should we measure CO₂ cost directly?
Or infer it from sparsity?
- Shall we include other target hardware: GPU, FPGA, ASIC?



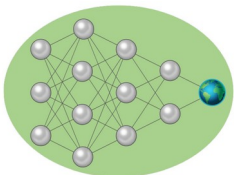
Status Overview



- Larissa is now hired full-time since last week!
- Currently: In the process of concluding our studies on the MNIST Dataset.
- Next steps:
 - Expand to HEP datasets
 - Measure exactly how sparsity $\leftrightarrow$ CPU Time / CPU Memory
 - Explore other parameters: FLOPs / bits / IO operations



With Freiburg



Status Overview



- Larissa is now hired full-time since last week!
- Currently: In the process of concluding our studies on the MNIST Dataset.
- Next steps:
 - Expand to HEP datasets
 - Measure exactly how sparsity $\leftrightarrow$ CPU Time / CPU Memory
 - Explore other parameters: FLOPs / bits / IO operations



With Freiburg

Thank You