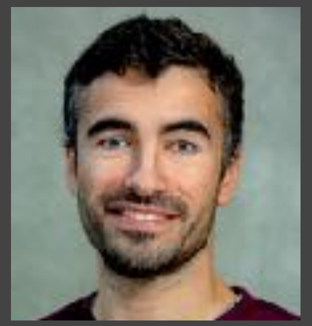
11. Green AI

Sustainable Software Engineering
CS4295



Luís Cruz
L.Cruz@tudelft.nl

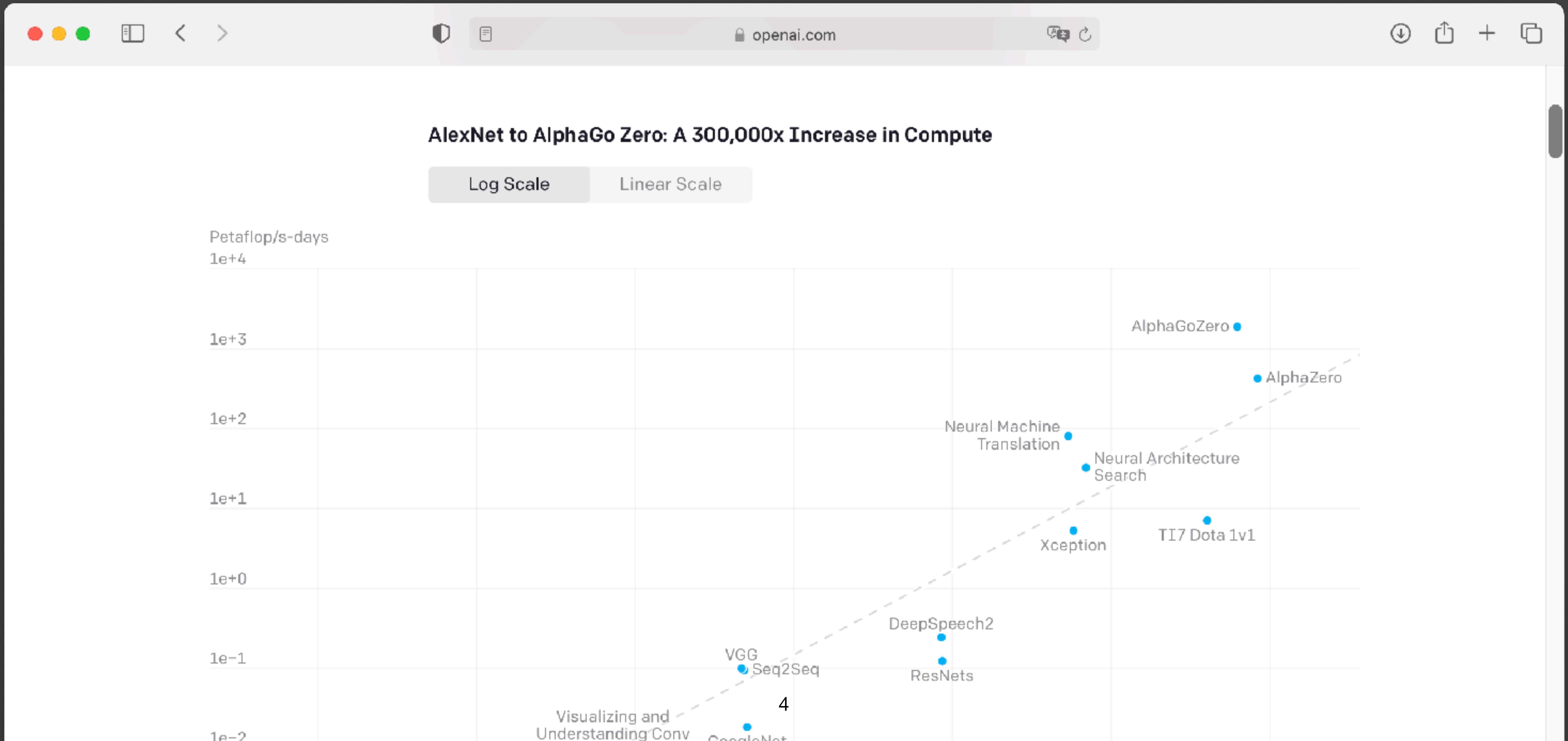
- Overview of Green AI
  - Large language models
- Green data-centric AI
- Model simplification
- Hyper parameter tuning
- Batching for Green AI
- Green AI at Meta

AI

- Artificial Intelligence (AI) is the branch of computer science that deals with **automating** tasks that typically require **human intelligence**.
- In the past years AI has been widely applied across different domains. E.g., health care, transportation, finance.
- To deploy AI systems, we test them against **benchmarks** (or validation sets).
  - The goal is to outperform the previous existing models.
  - E.g., in Machine Learning we usually resort to accuracy metrics. The highest the accuracy, the better the model.

Since 2012, the amount of computing used for AI training has been doubling every 3.4 months

- https://openai.com/blog/ai-and-compute/



- To create better AI systems we are currently adding
  - **More data**
  - **More experiments**
  - **Larger models**

The Equation of Red AI

$$Cost(R) \propto E \cdot D \cdot H$$

Cost of a single (**E**)xample

Number of (**H**)yperparameters

Size of (**D**)ataset

By Schwartz et al. (2020)

Issues of Red AI

- High costs (hardware, electricity, data access, etc.)
- Limited reproducibility.
- Energy consumption.
- Carbon emissions.
- SMEs can hardly be competitive.
- Groundbreaking AI research is mostly done by tech giants.

A few examples of Red AI

- Google’s BERT-large
  - 350 million features
  - Trained for 2.5 days using 512 TPU chips, costing $60K+
- Open-GPT3 (now GPT-3.5)
  - 550 Billion CO2-eq (Patterson, 2021)
  - 175 billion features
  - API is open but no-pretrained model is available
- AlphaGo
  - 1920 CPUs, 280 GPUs, costing $35M

Red AI in Large Language Models (LLMs)

- New Moore's law
- There are some good news:
  - **OPT** by Meta reports **75 tons CO2-eq** (1/7 of OpenGPT's footprint). (Also 175billion params)
    - **Open science**: release includes both the pretrained models and the code needed to train and use them.
  - **Bloom** by Huggingface reports **25 tons**, 51 when considering embodied and operational carbon footprint. (176billion params)

Red AI

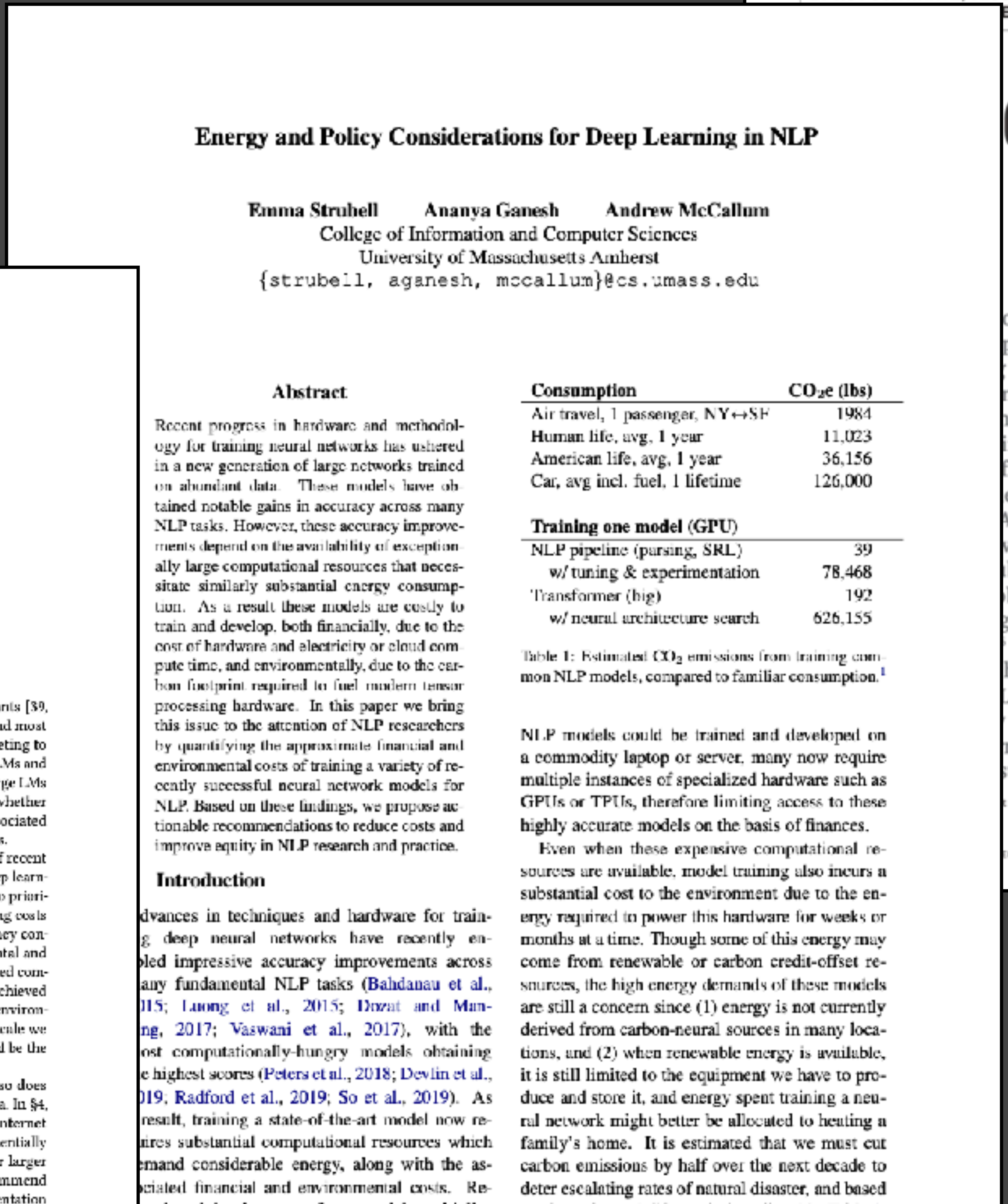
Accuracy: 0.999999999

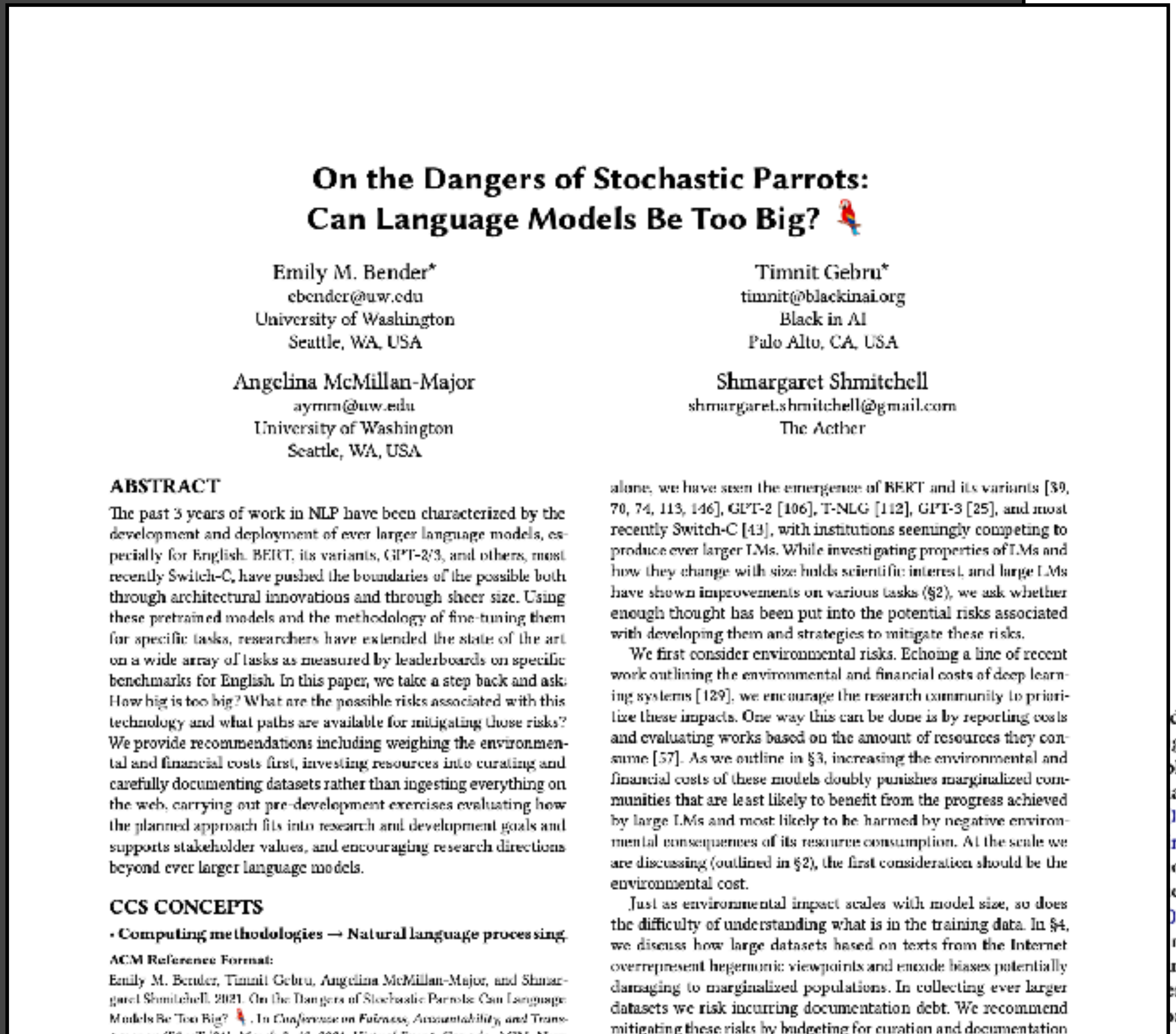
Green AI

- Energy
- Time
- Reproducibility
- Reusage

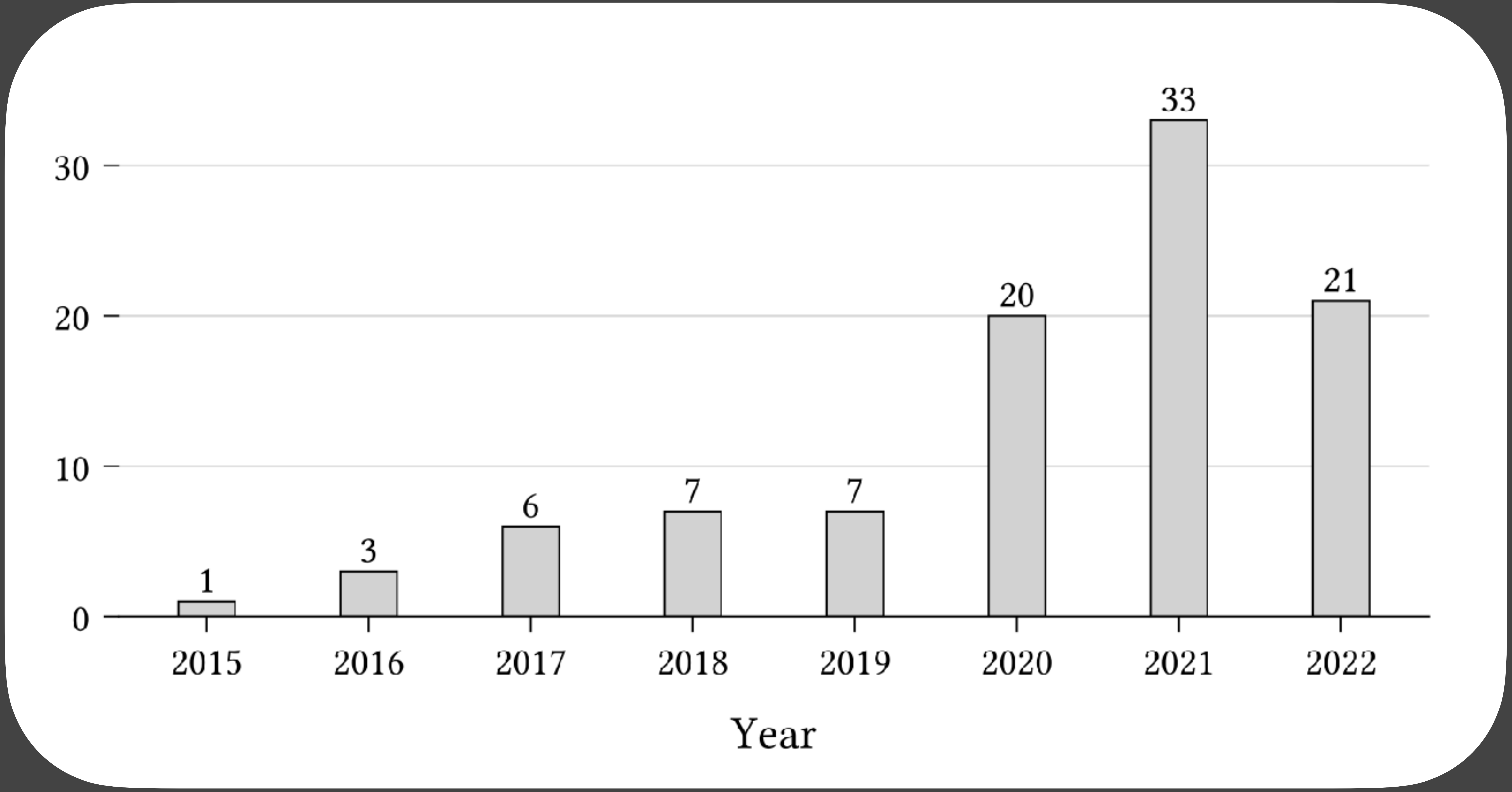
Research on Green AI

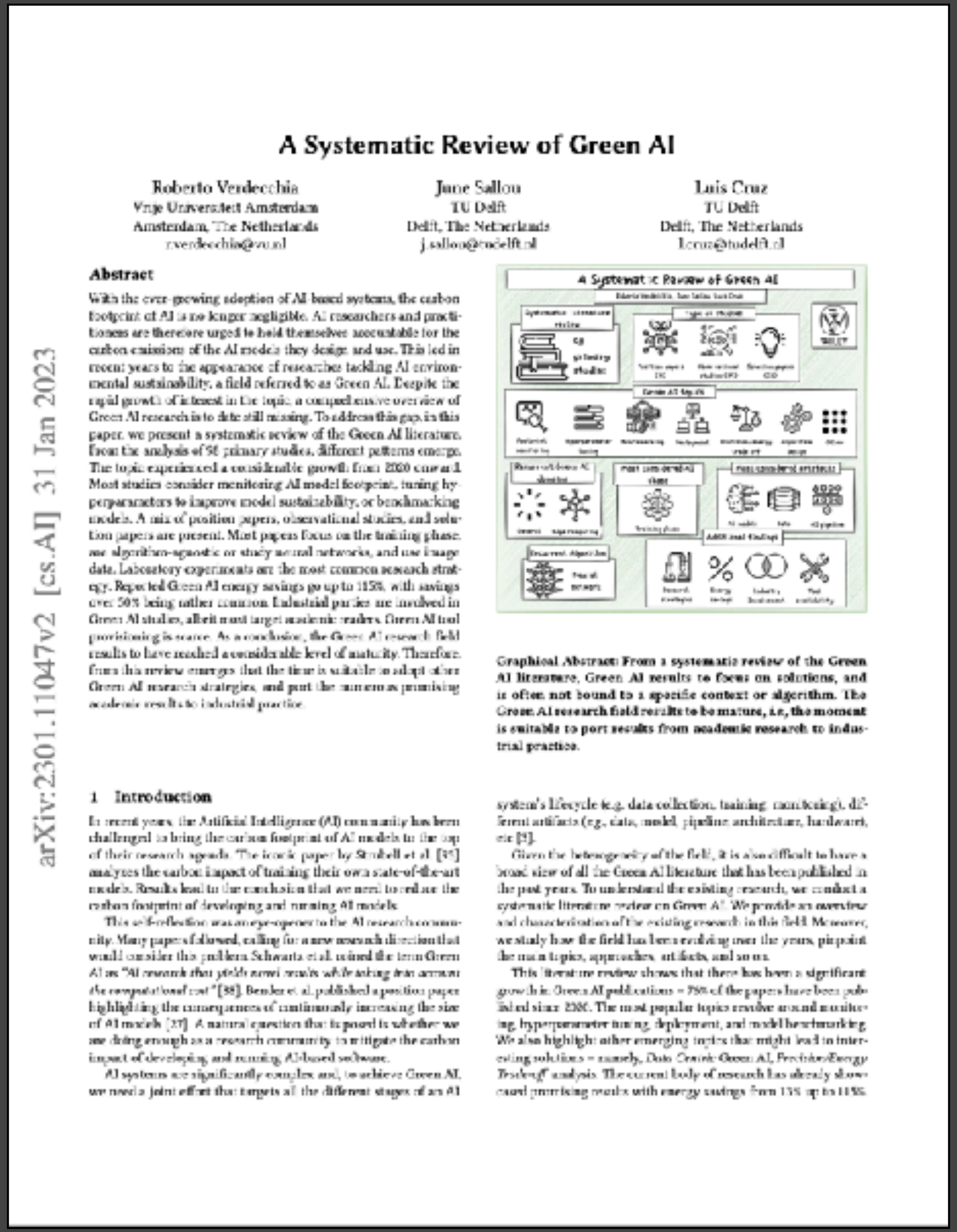
- Most literature revolves around **position papers**. (?)





Publications in Green AI over the years




A Systematic Review of Green AI

Roberto Verdecchia
Vrije Universiteit Amsterdam
Amsterdam, The Netherlands

June Sallou
TU Delft
Delft, The Netherlands

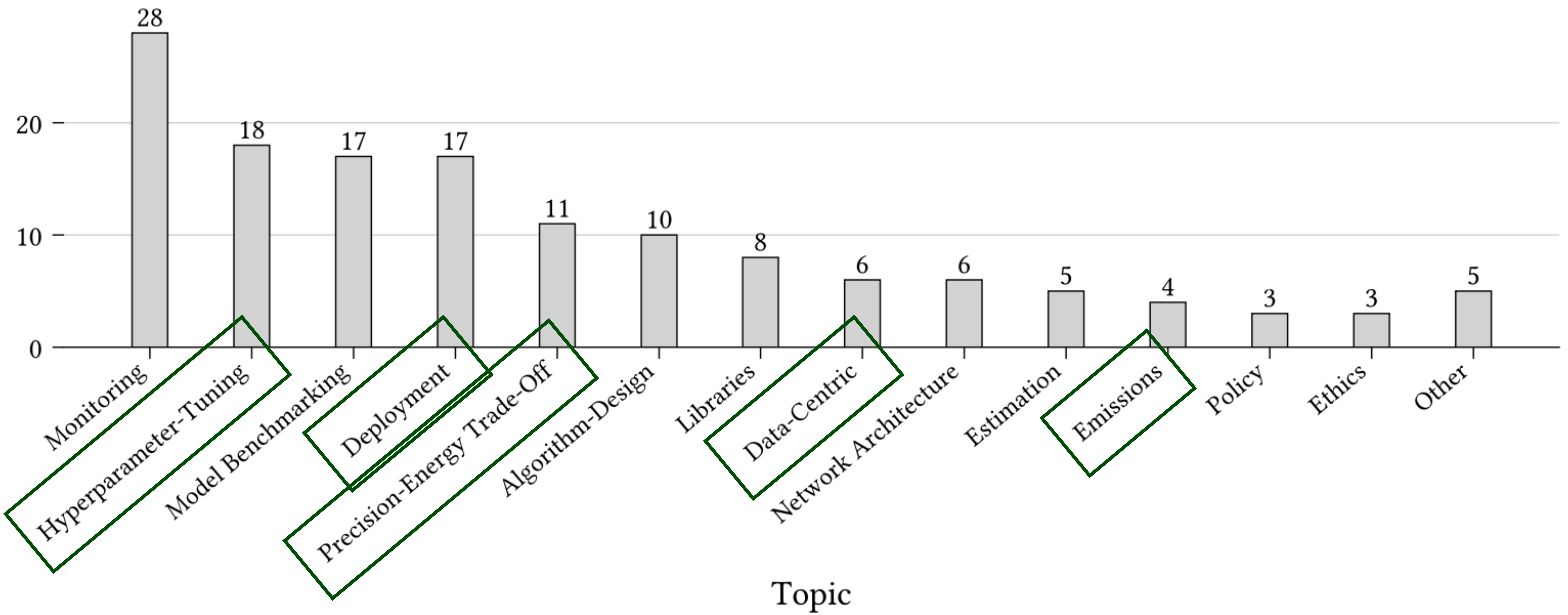
Luís Cruz
TU Delft
Delft, The Netherlands

arXiv:2301.11047v2 [cs.AI] 31 Jan 2023

Abstract

1 Introduction

(Verdecchia, 2022)



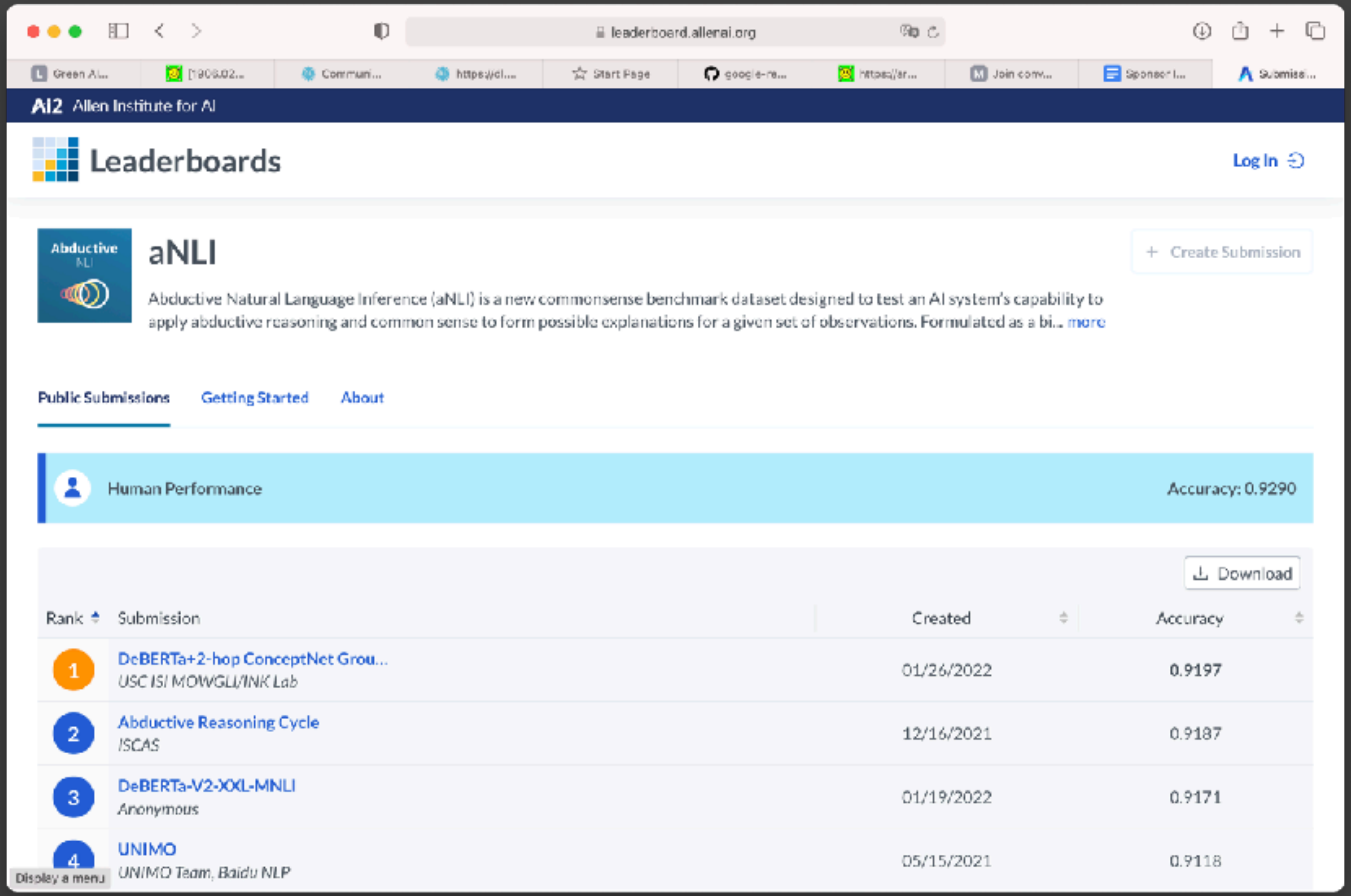
28
18
17
17
11
10
8
6
6
5
4
3
3
5
20
10
0
Monitoring
Hyperparameter-Tuning
Model Benchmarking
Deployment
Precision-Energy Trade-Off
Algorithm-Design
Libraries
Data-Centric
Network Architecture
Estimation
Emissions
Policy
Ethics
Other
Topic

How can we adopt Green AI

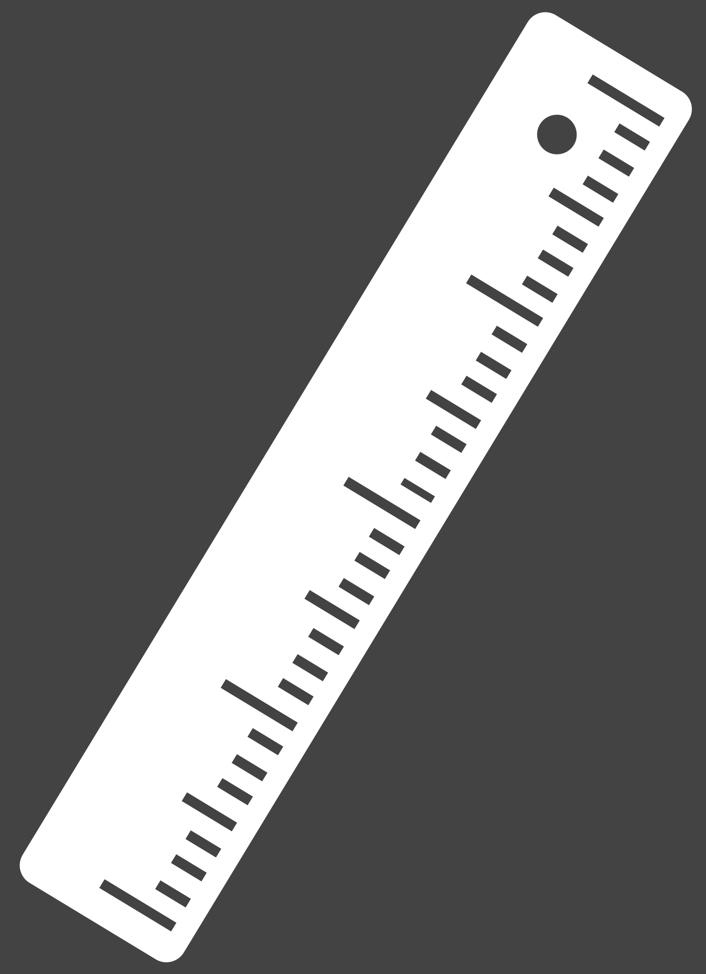
- **Check whether AI is needed.**
- Select green datacenters.
- Run on **low carbon intensity** hours.
- Opt for **GPU-optimised** solutions
- Opt for **low-power hardware** (e.g., Nvidia Jetson boards)
    - Or GPUs that provide energy metrics (e.g., NVIDIA GPUs via the **nvidia-smi** tool)
- **Report** energy/carbon metrics (e.g., embed in MLFlow?)
- Use pre-trained models (Transfer Learning)
- Preprocess dataset to reduce size.
- Improve parameter-tuning strategy.

Reporting energy/carbon footprint

- We need **benchmarks**.
- AllenAI leaderboard https://leaderboard.allenai.org
  - **No carbon metrics**, yet
- Report comparable proxies for energy consumption.
  - ⚠️ Learning algorithms behave in a non-deterministic
  - ⚠️ Different data-points lead to different energy consumption



Reporting energy/carbon footprint



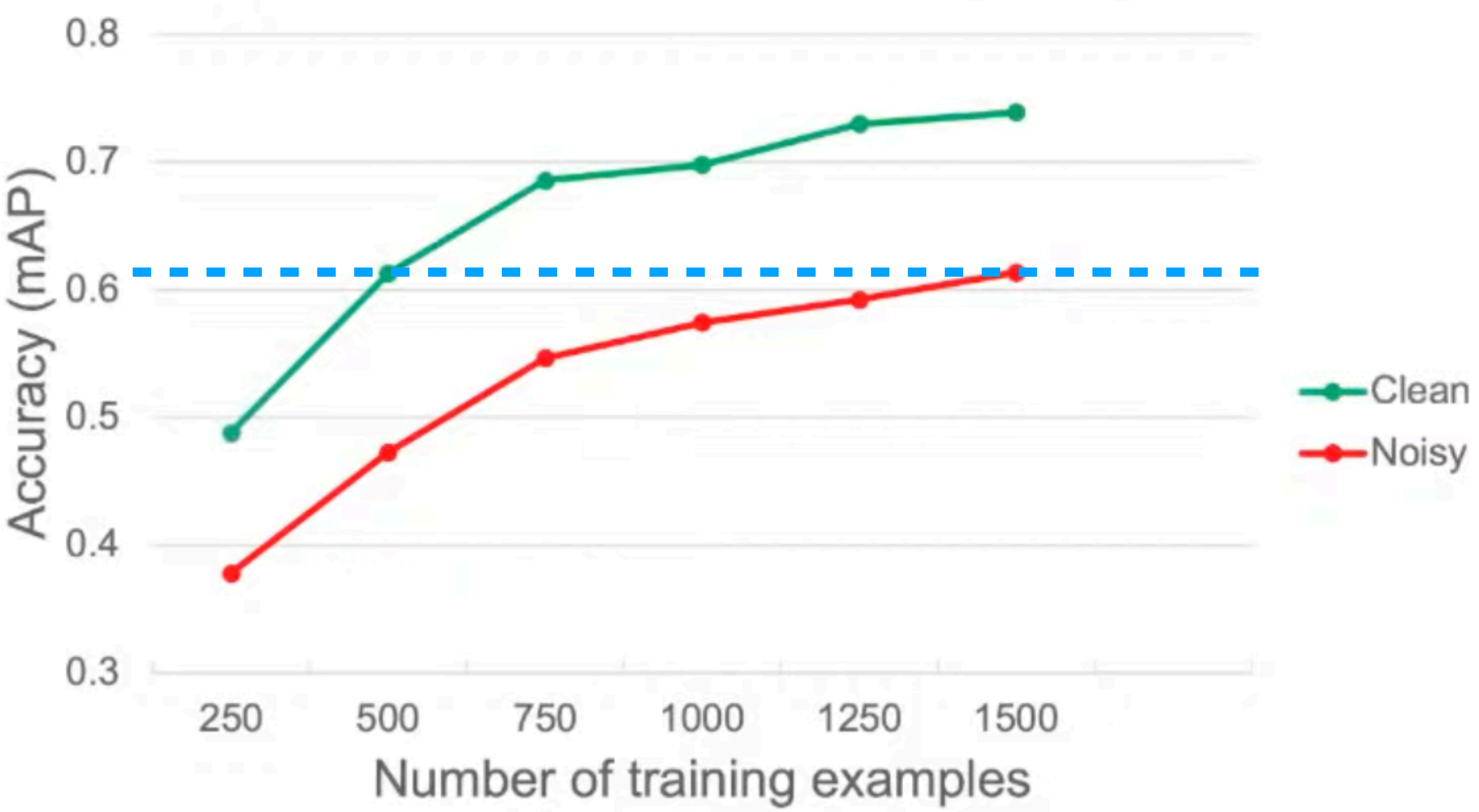
- Reporting **measured energy consumption**
  - + Accurate
  - + Easy to map to carbon emissions
  - - Hard to measure
  - - Low replicability
- Reporting **time** / estimation based on **time & hardware**
  - + Easy to measure
  - + Correlates with energy consumption in most cases.
  - - Difficult to compare with measurements from other setups
- E.g., **floating point operations** (FPOs) (?)
  - + comparable across different setups
  - + cheap
  - - does not factor in memory energy consumption
  - - does not reflect carbon emissions

Data-centric AI

Data-centric AI

- Emerging discipline that deals with systematically engineering data to build AI systems.
  - Shift from **improving the training strategy** to **improving the data**.
  - It is better to have **small but reliable** datasets than **large but noisy** datasets.
    - => Improve **data collection**, **data labelling**, and **data preprocessing**.
  - More about data-centric AI by Andrew Ng: https://www.youtube.com/watch?v=06-AZXmwHjo

Example: Clean vs. noisy data



Green Data-centric AI

- How do different ML algorithms compare in terms of energy consumption?
- How does **number of rows** relate to the energy consumption of ML models?
- How does **number of features** relate to the energy consumption of ML models?
- What is the impact of reducing data in the **performance** of the model?
- Method -> results -> discussion

Data-Centric Green AI
An Exploratory Empirical Study

Authors Blinded for Review*
*name of organization (of Aff.)
City, Country
{Author1, Author2,..., AuthorN}@....com

***Abstract*—With the growing availability of large-scale datasets, and the popularization of affordable storage and computational capabilities, the energy consumed by AI is becoming a growing concern. To address this issue, in recent years, studies have focused on demonstrating how AI energy efficiency can be improved by tuning the model training strategy. Nevertheless, how modifications applied to datasets can impact the energy consumption of AI is still an open question.**

To fill this gap, in this exploratory research, we evaluate if data-centric approaches can be utilized to improve AI energy efficiency. To achieve our goal, we conduct an empirical experiment, executed by considering 6 different AI algorithms, a dataset comprising 5,574 data points, and two dataset modifications (number of data points and number of features).

Our results show evidence that, by exclusively conducting modifications on datasets, energy consumption can be drastically reduced (up to 92.16%), often at the cost of a negligible or even absent accuracy decline. As additional introductory results, we demonstrate how, by exclusively changing the algorithm used, energy savings up to two orders of magnitude can be achieved.

In conclusion, this exploratory investigation empirically demonstrates the importance of applying data-centric techniques to improve AI energy efficiency. Our results call for a research agenda that focuses on data-centric techniques, to further enable and democratize Green-AI.

***Index Terms*—Energy Efficiency, Artificial Intelligence, Green AI, Data-centric, Empirical Experiment**

I. Introduction

We live in the era of artificial intelligence (AI): new intelligent technologies are emerging every day to change people's lives. Many organizations identified the massive potential of using intelligent solutions to create business value. Hence, in the past years, the *modus operandi* is collecting as much data as possible so that no opportunity is missed. Data science teams are constantly looking for problems where AI can be applied to existing data to train models that can provide more personalized and optimized solutions to their operations customers and operations [1].

Nevertheless, the energy consumption of developing AI applications is starting to be a concern. Previous studies observed that AI-related tasks are particularly energy-greedy [2], [3]. In fact, since 2012, the amount of computing used for AI training has been doubling every 3.4 months [4]. Hence, a new subfield is emerging to make the development and application of AI technologies environmentally sustainable: *Green AI* [5].

On a related note, AI practitioners have realised that the current trend of collecting massive amounts of data is not necessarily yielding better models. Being able to collect high-quality data is more important than collecting big data – a trend coined as *Data-centric AI*[1]. Instead of creating learning techniques that squeeze every bit of performance, data-centric AI focuses on leveraging systematic, reliable, and efficient practices to collect high-quality data.

Therefore, in this study, we conduct an exploratory empirical study on the intersection of Green AI and Data-centric AI. We investigate the potential impact of modifying datasets to improve the energy consumption of training AI models. In particular, we focus on machine learning, the branch of AI that deals with the automatic generation of models based on sample data – machine learning and AI are used interchangeably throughout this paper. In addition to investigate the energy impact of dataset modifications, we also analyze the inherent trade-offs between energy consumption and performance when reducing the size of the dataset – either in the number of data points or features. Moreover, the analysis is performed in six state-of-the-art machine learning model applied in the detection of Spam messages.

Our results show that feature selection can reduce energy consumption up to 76% while preserving the performance of the model. The improvement in energy efficiency is more impressive when reducing the number of data points: up to 92% in the case of Random Forrest. However, in this case, it is not cost-free: the trade-off between energy and performance needs to be considered. Finally, we also show that KNN tends to be the most energy-efficient algorithm while ensemble classifiers tend to be the most energy greedy.

This paper provides insights to define the most relevant and energy-efficient modifications of datasets used during the elaboration of the AI models while ensuring minimal accuracy loss. We argue that more research in Data-centric AI will help more practitioners in developing green AI models. To the best of our knowledge, this is the first study to explore the potential of preprocessing data to reduce the energy consumption of AI.

The entirety of our experimental scripts and results are made available with an open-source license, to enable the independent verification and replication of the results presented in this study: https://github.com/GreenAIproject/ICT4S22.

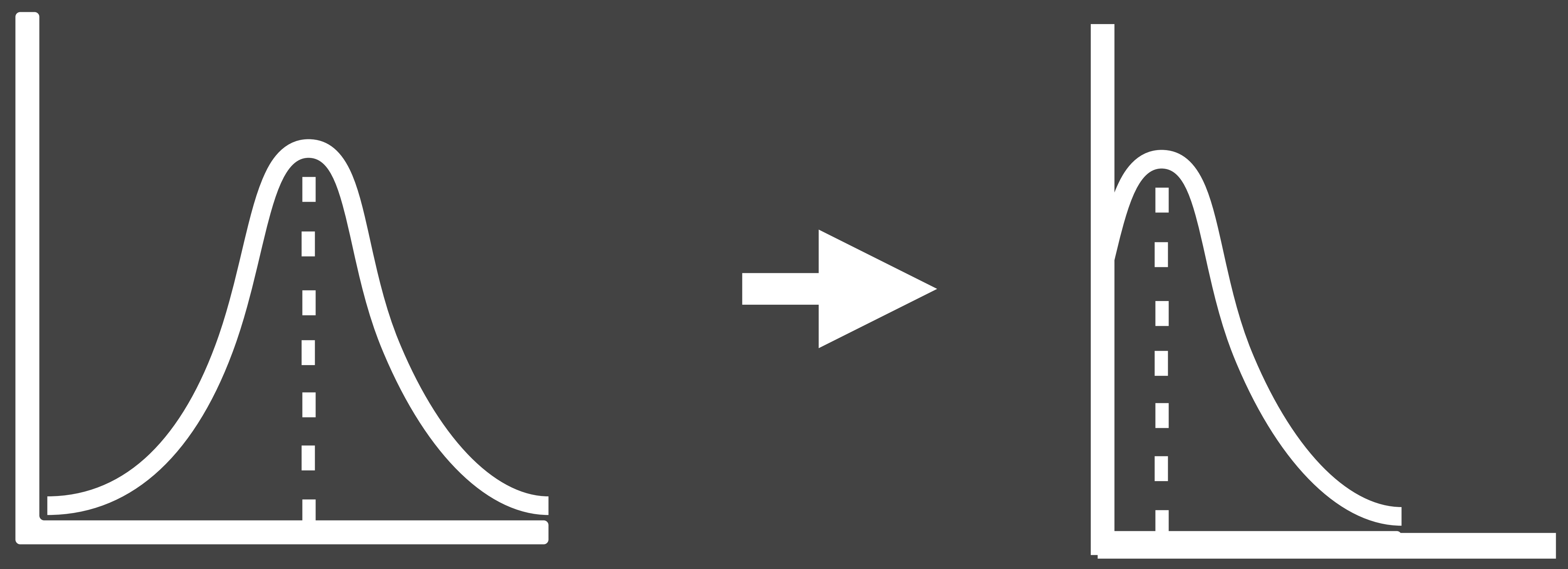
The remainder of this paper is structured as follows. Section II presents the related work on the energy consumption

[1]Understanding Data-Centric AI: https://landing.ai/data-centric-ai/. Accessed 24th January 2022.

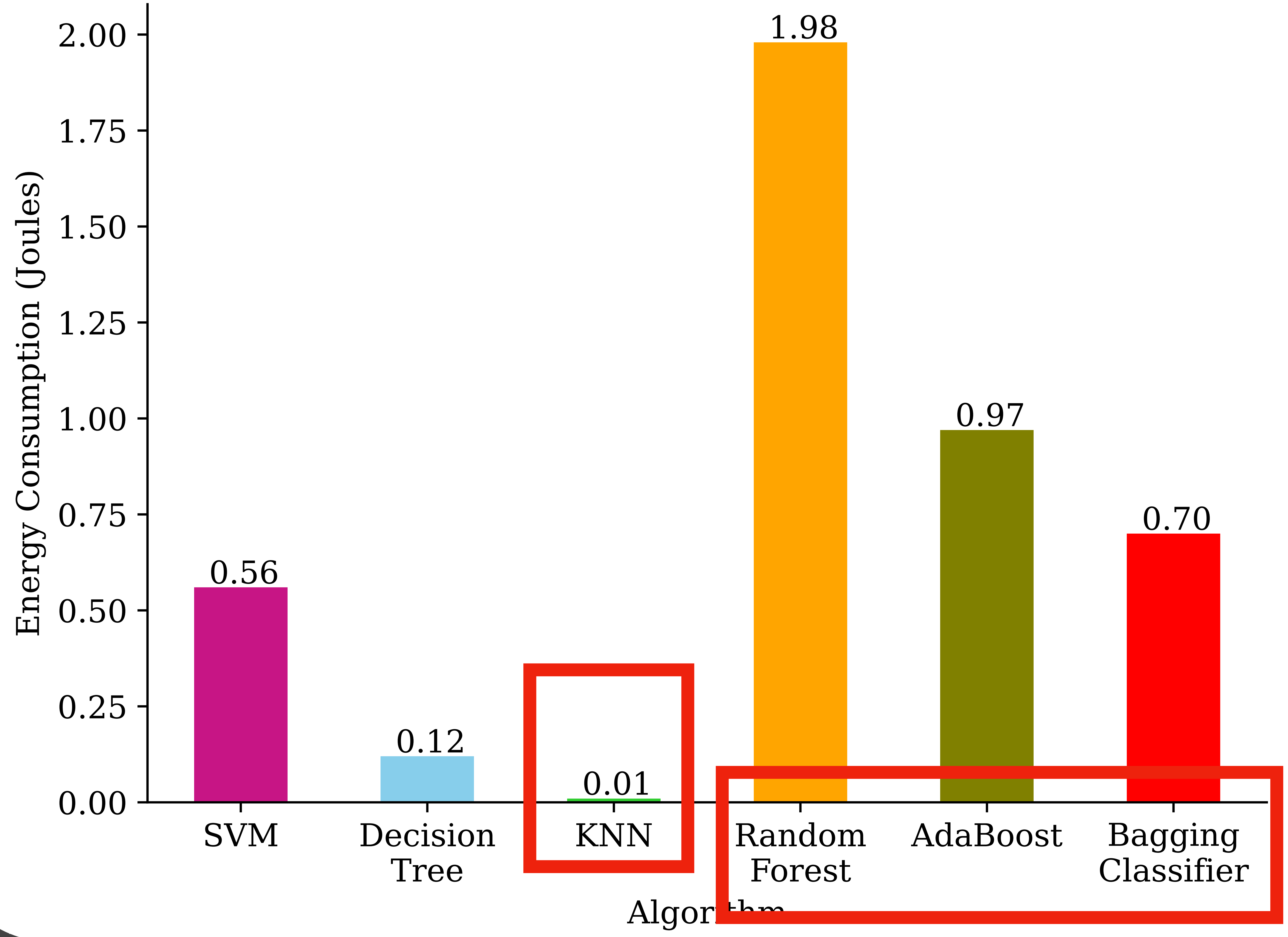
Method

- Single object of study: natural language model to **detect spam messages**.
- 6 machine learning algorithms: **SVM, Decision Tree, KNN, Random Forrest, AdaBoost, Bagging Classifier**.
- Reduce the number of rows. 10%, 20%, .., 100%
  - **Stratified random sampling** (?)
- Reduce the number of features. 10%, 20%, .., 100%
  - **Feature importance** metric based on the Chi-Square Test (Chi2)
- Estimate energy consumption using a RAPL-based tool. (?)

- Repeat 30 times
- Fix random seeds
- …
- Data was **not Normal** => **tailed Normal distribution.** (?)



Results: energy consumption of algorithms



Results: energy vs data shape

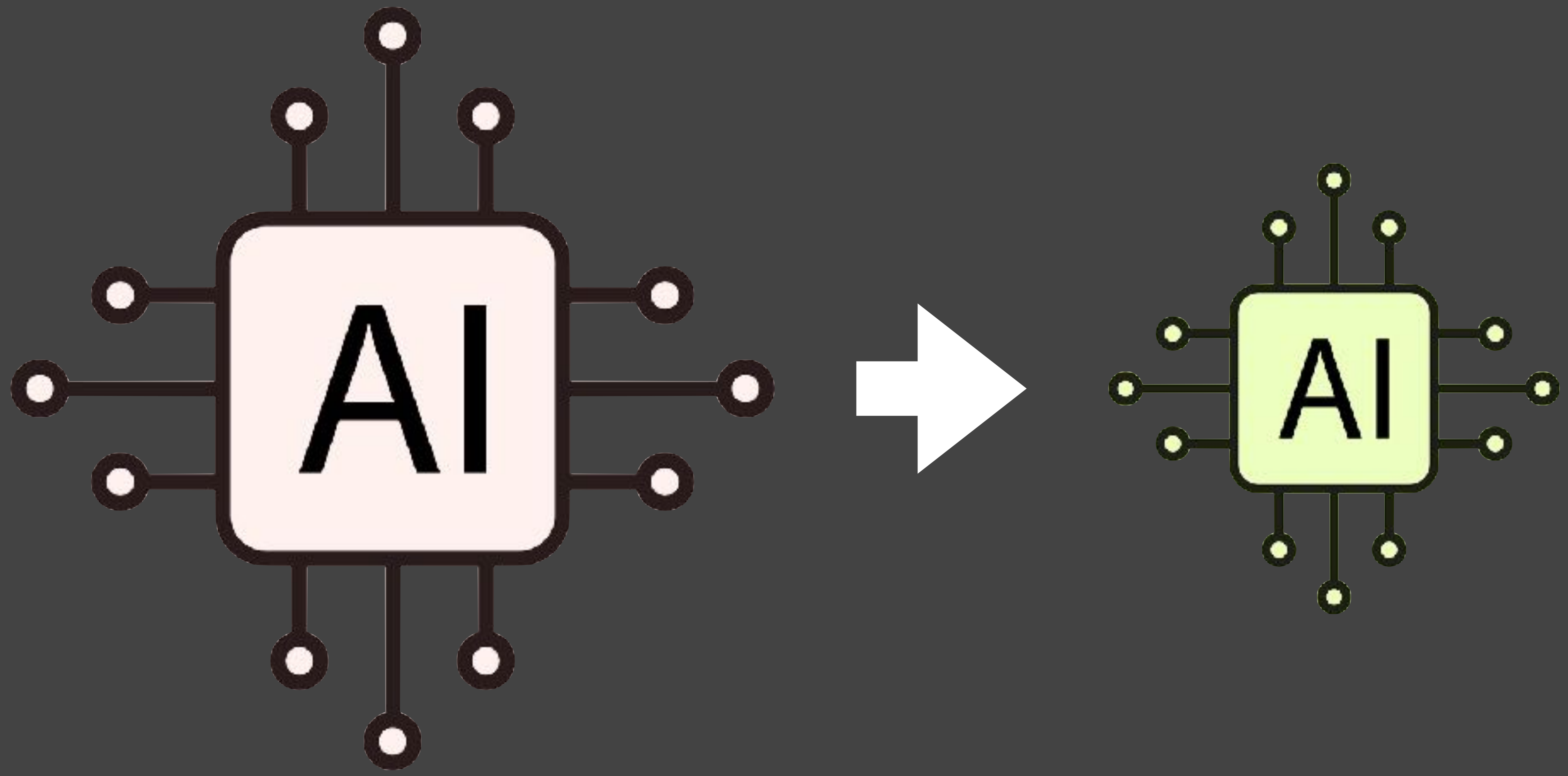
Results: performance vs data shape

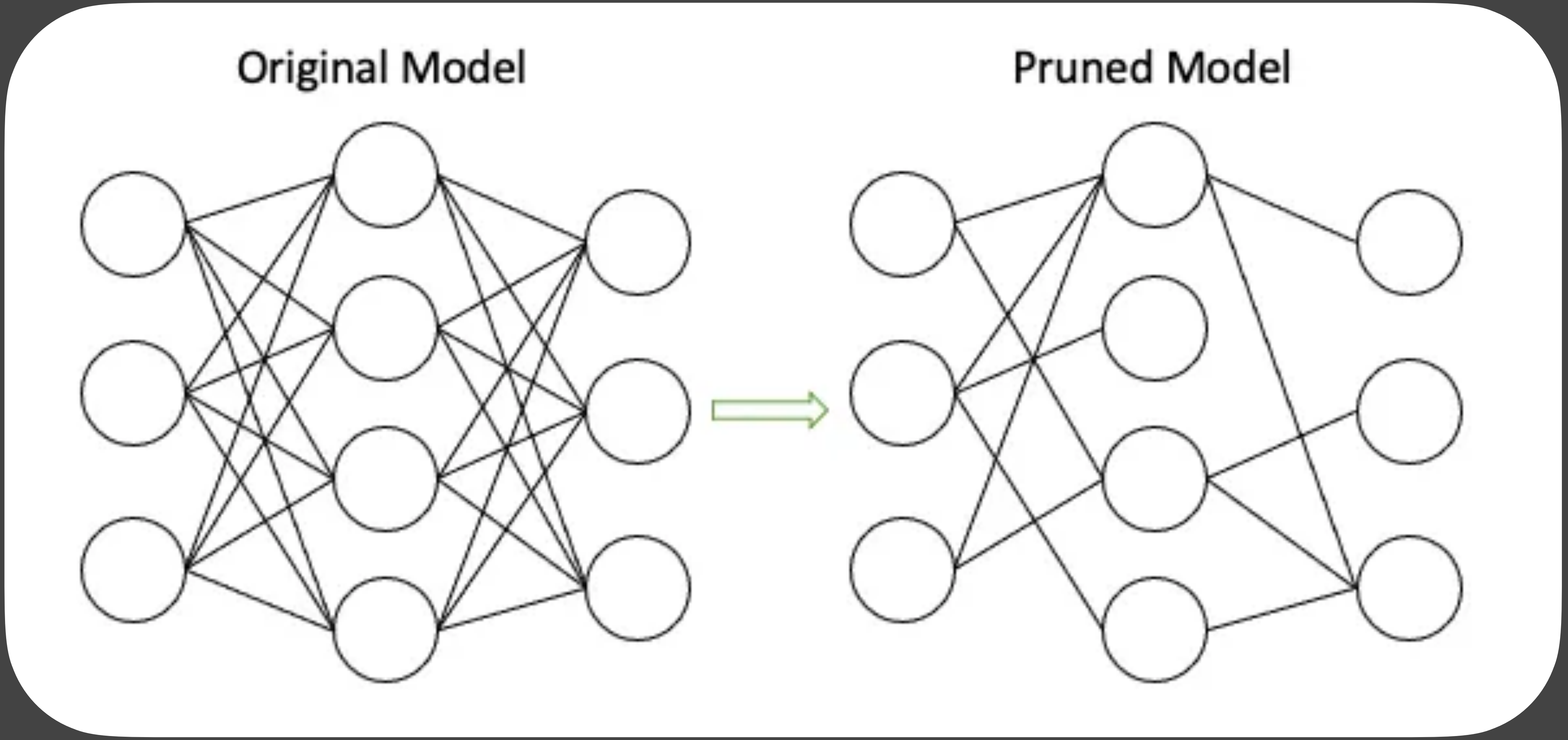
Discussion

- Other data properties should be investigated.
  - E.g., data types
- **Reporting energy data** is essential. It can lead to different model selection without hindering model performance.
- There is a big opportunity in **Model and Data Simplification.**

Data/Model Simplification

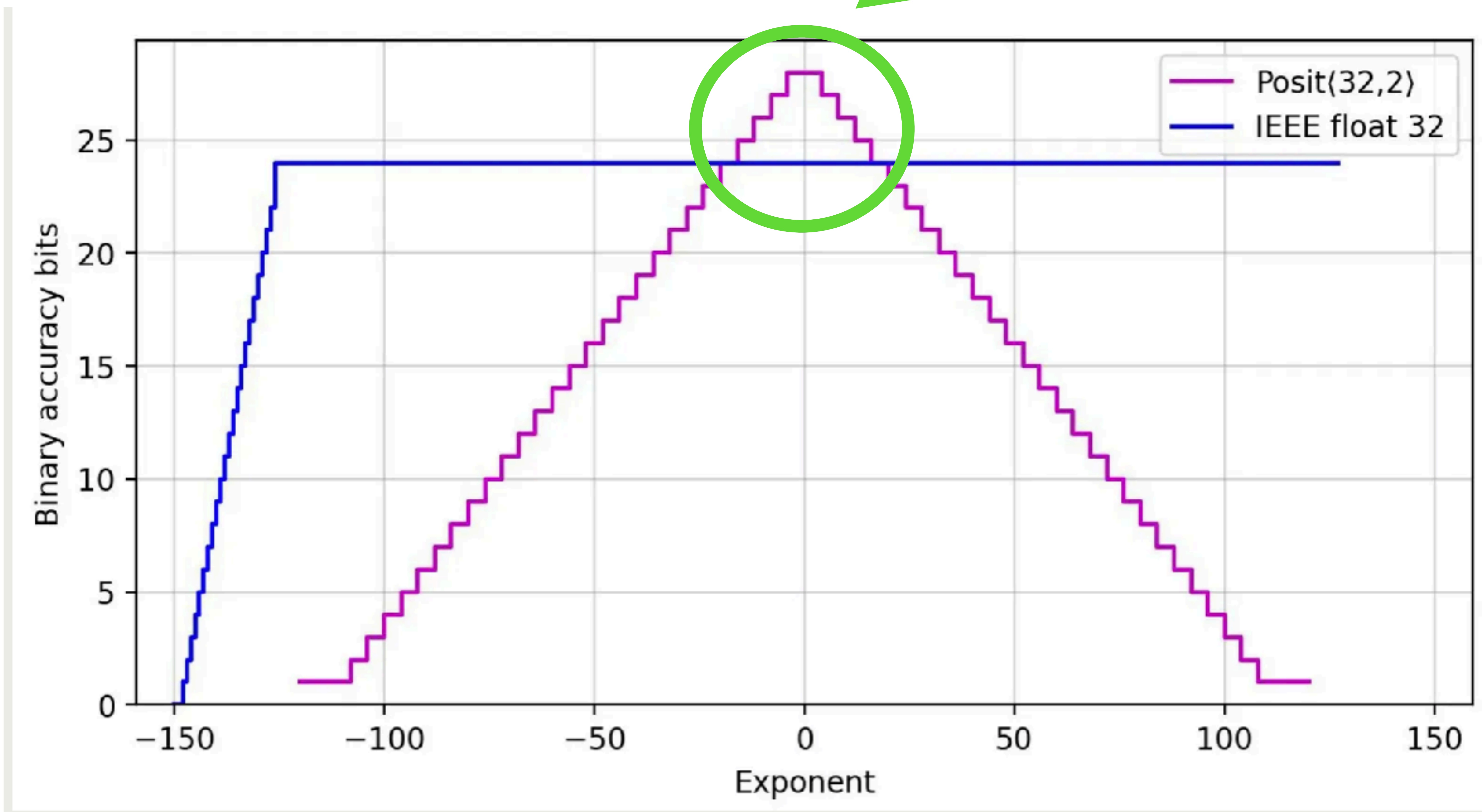
- (?)
- Data selection
- Data quantisation. **Posit?**
- Data distillation
- Coreset extraction (?)
- Model distillation
- Model quantisation
- Model pruning
- …





Posit vs Float

Better for DL use cases



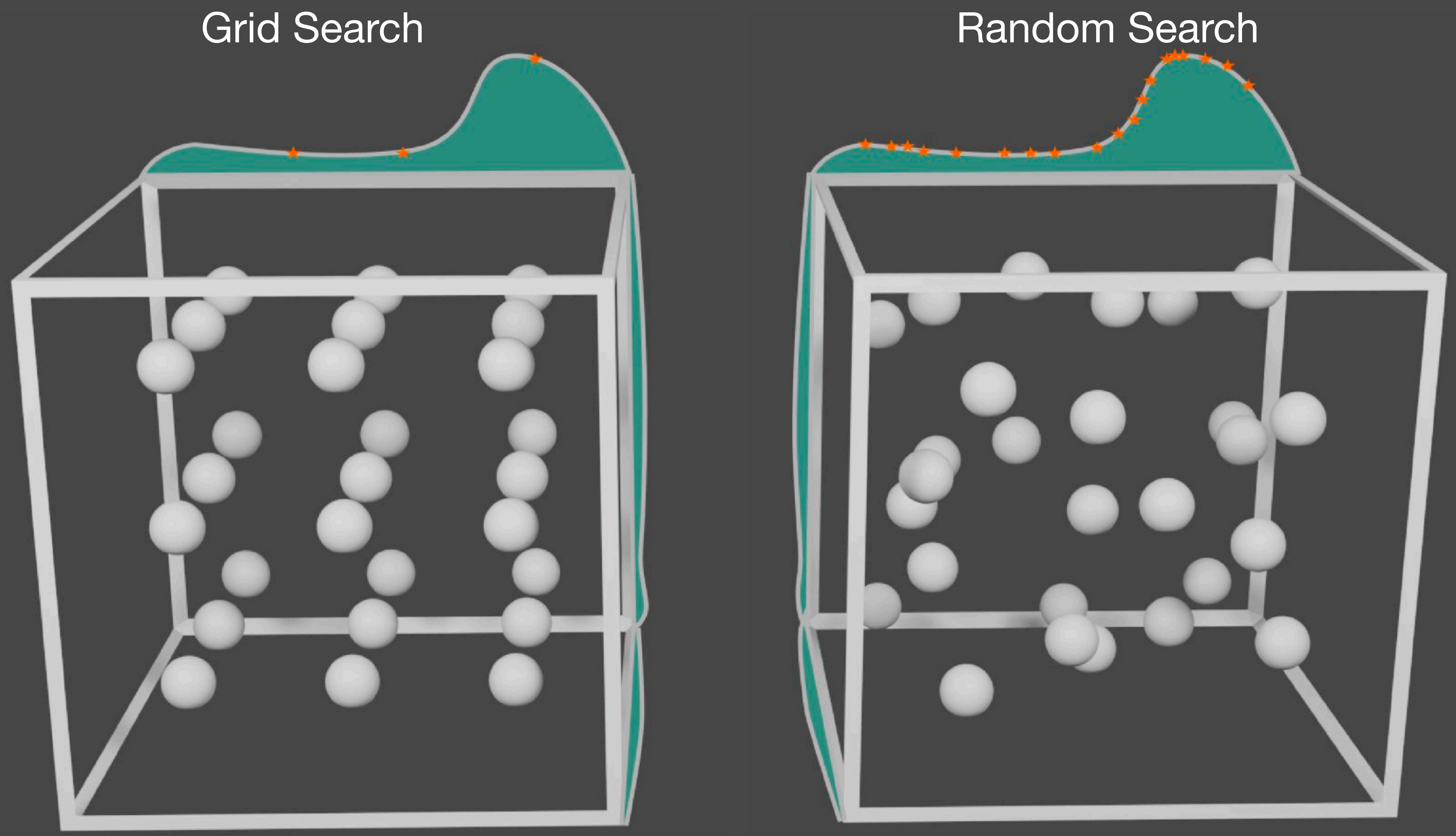
More about this: https://spectrum.ieee.org/floating-point-numbers-posits-processor

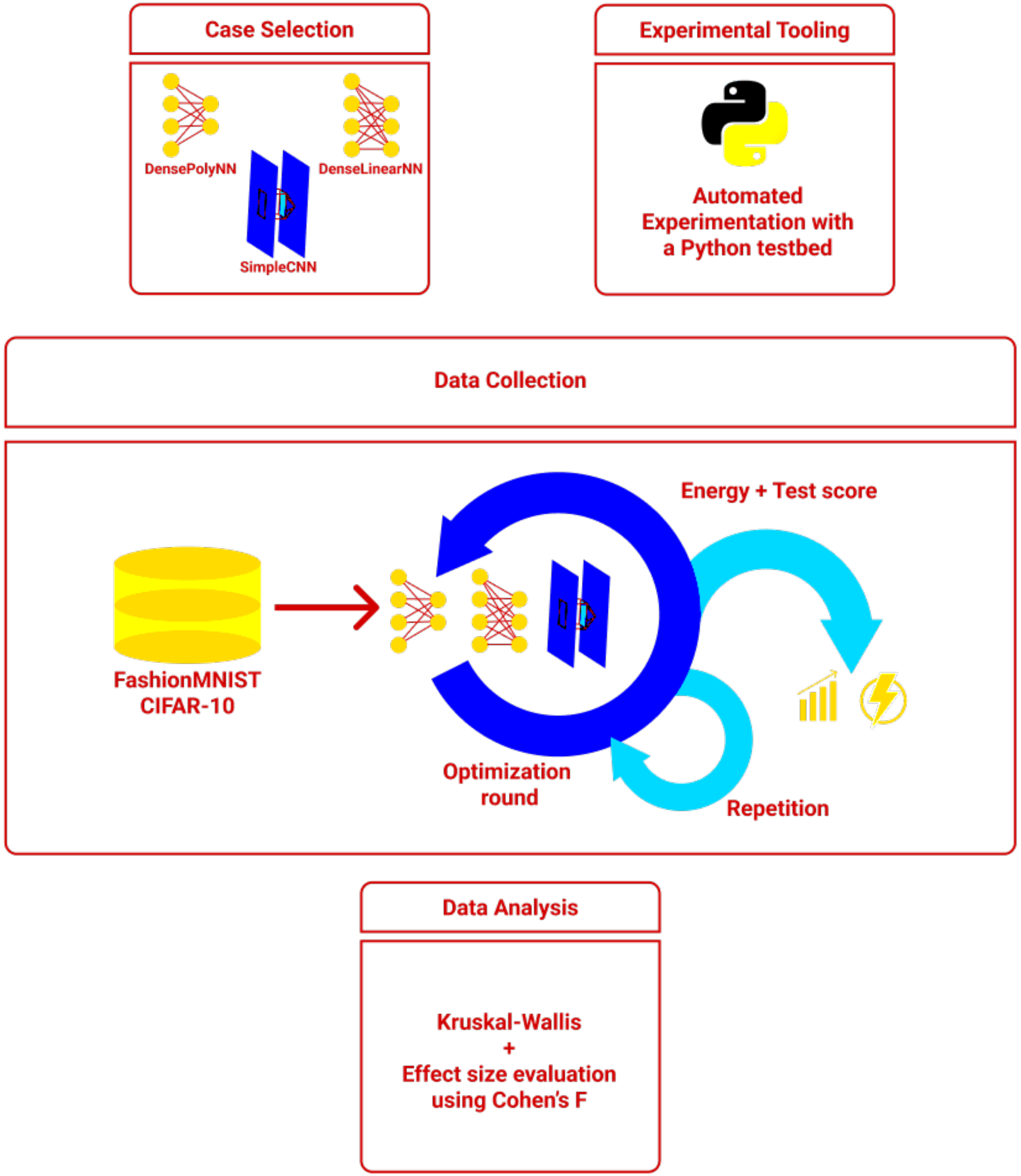
Model simplification out of the box

- Can we apply model simplification to a given model with minimal domain expertise?
- What would be the free gains from applying it to existing open access models? E.g.: http://modelhub.ai/

Hyper parameter tuning

- When training an ML model, there are several **parameters** that need to be **tuned**.
  - E.g., in SVM we have the *Regularization parameter* C, the kernel function, the degree of the kernel function, and depending on the case, many other.
  - The common approach revolves around **grid search**. The user provides a sequence of possible values for each parameter and the pipeline runs **all possible combinations**.
    - **Our question:** Can we save energy with alternative approaches?
    - We studied **Random Search** and **Bayesian Optimisation**.


Grid Search
Random Search

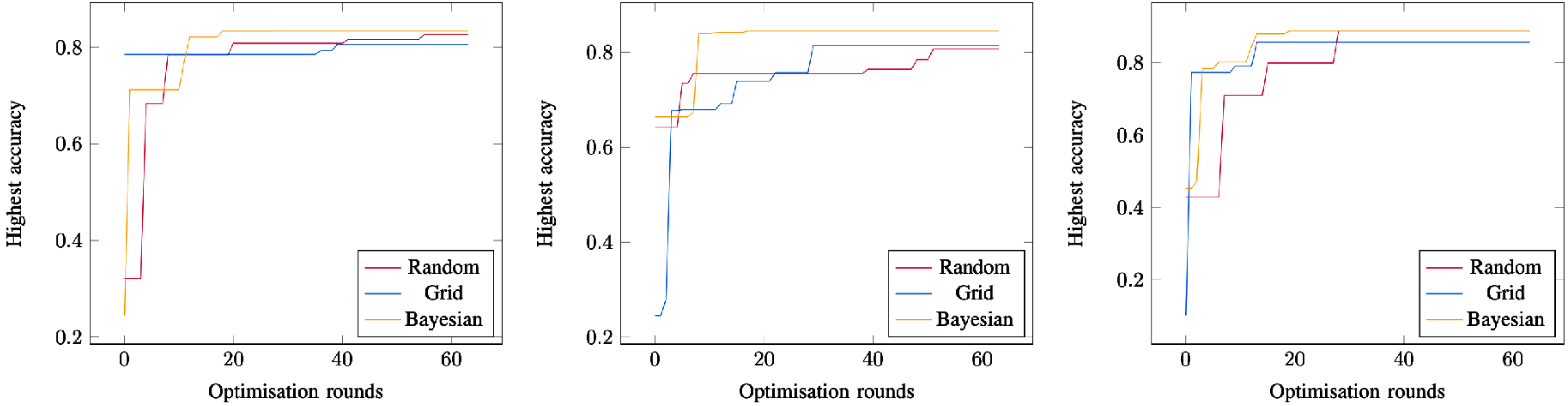


Case Selection
DensePolyNN
DenseLinearNN
SimpleCNN
Experimental Tooling
Automated Experimentation with a Python testbed
Data Collection
Energy + Test score
FashionMNIST
CIFAR-10
Optimization round
Repetition
Data Analysis
Kruskal-Wallis
+
Effect size evaluation using Cohen's F

Results

Conclusions?

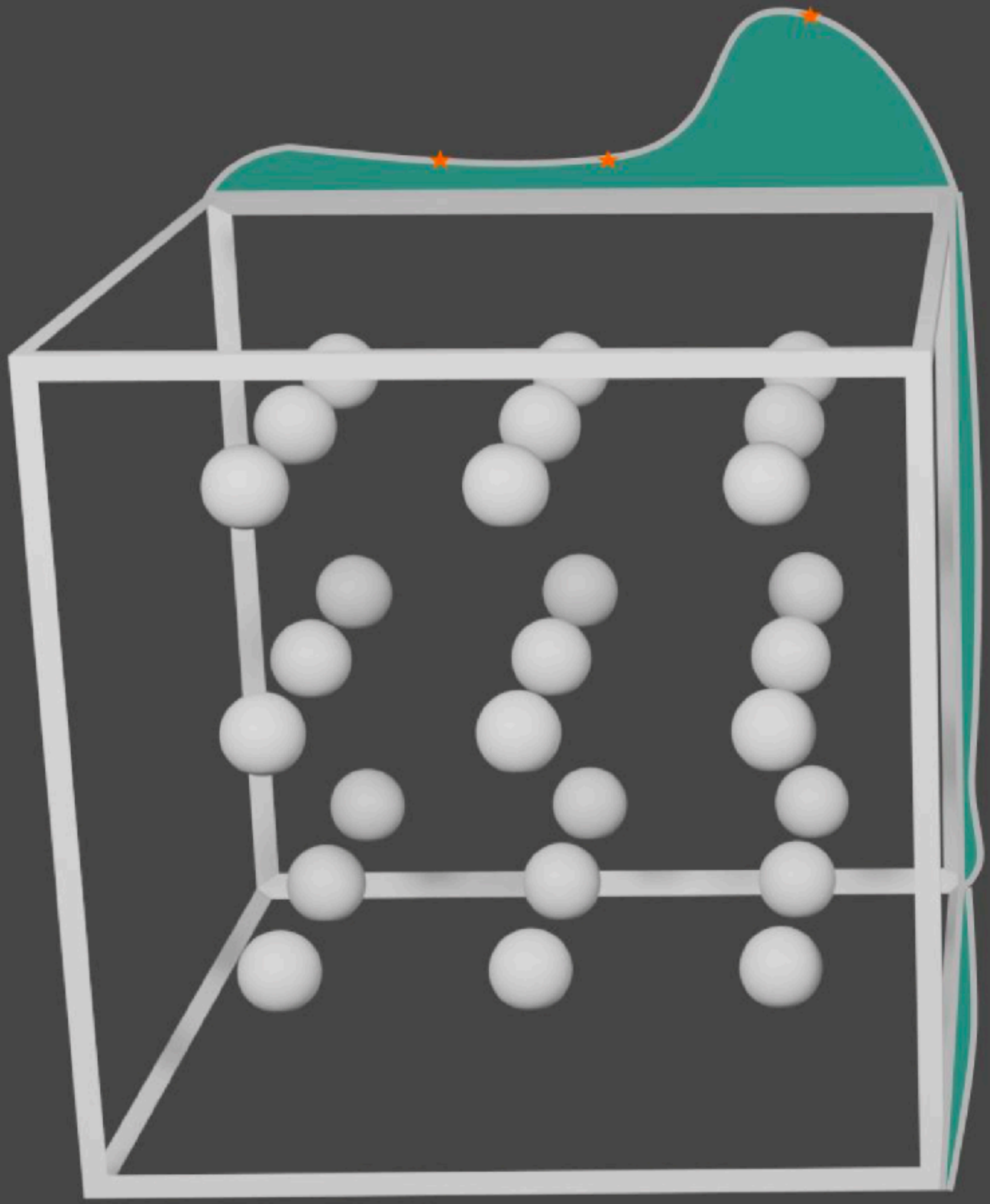
- **Bayesian** converges faster.
- No clear winner between Grid and Random



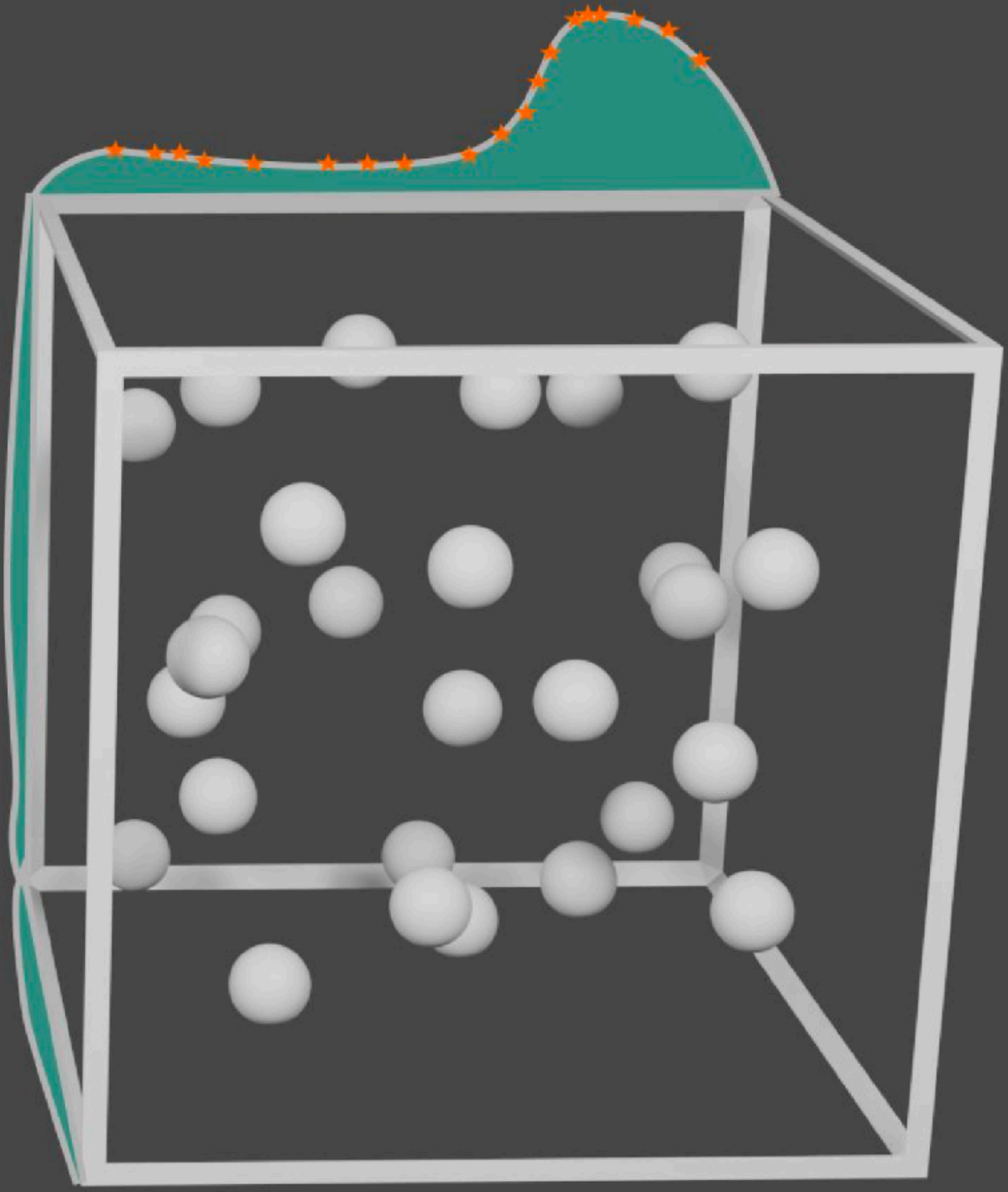
(a) DensePolyNN (b) DenseLinearNN (c) SimpleCNN

Which one to choose?

Grid Search

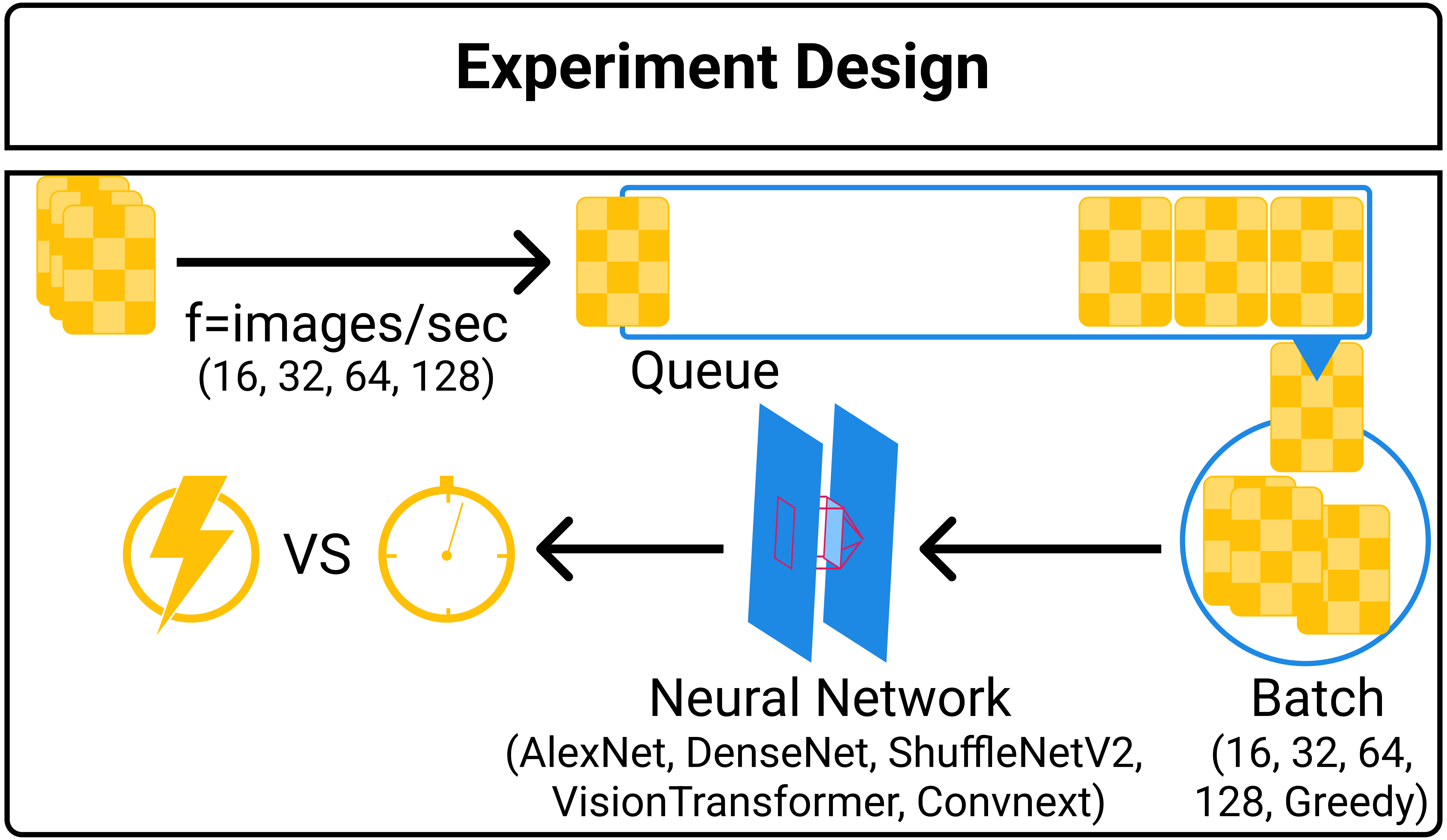


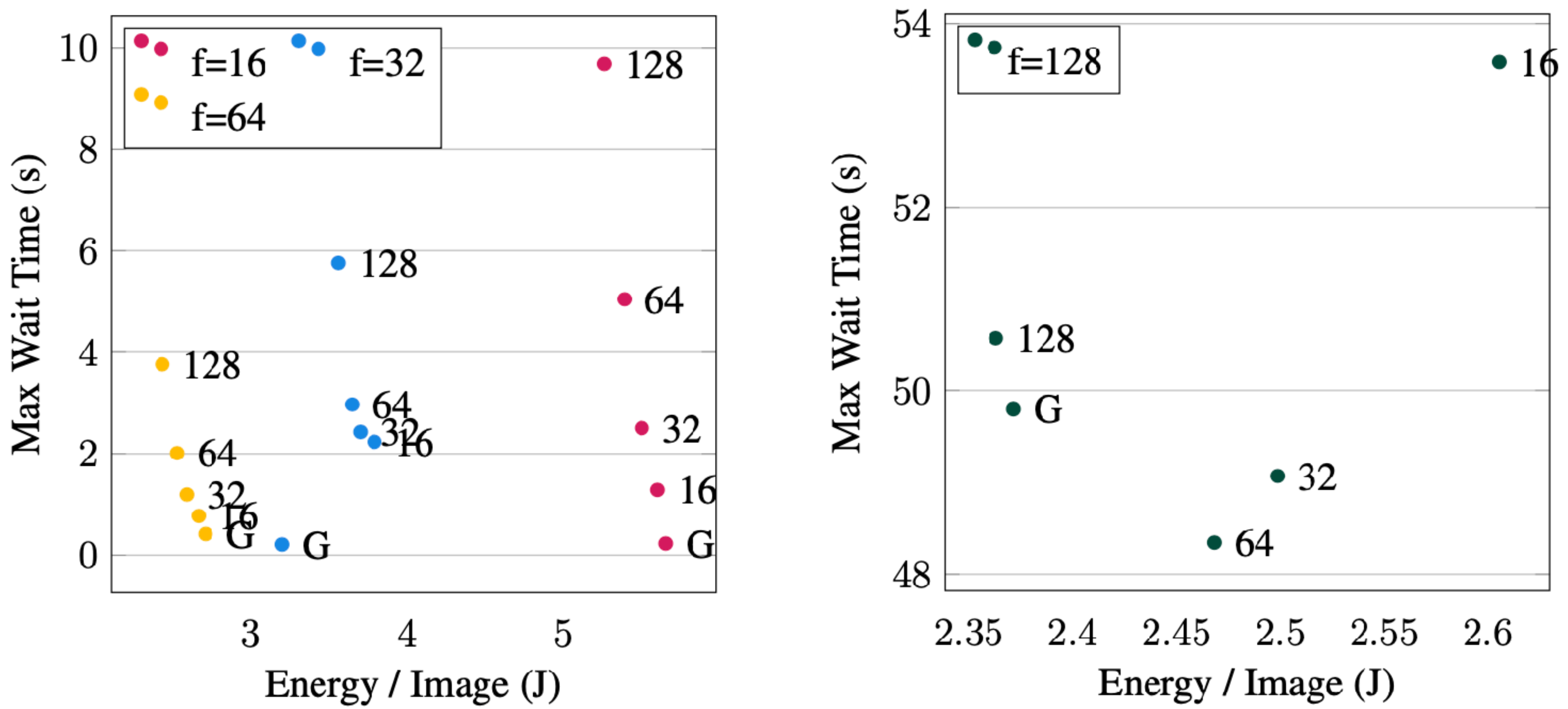
Random Search



Processing inference requests

- Imagine an API for an AI model that allows users to make inference requests. E.g., a license plate reader.
  - What is the most energy-efficient way to bundle these requests?
    - No bundling? (Greedy)
    - Batches of 16? 32?…


Experiment Design
f=images/sec
(16, 32, 64, 128)
Queue
VS
Neural Network
(AlexNet, DenseNet, ShuffleNetV2, VisionTransformer, Convnext)
Batch
(16, 32, 64, 128, Greedy)

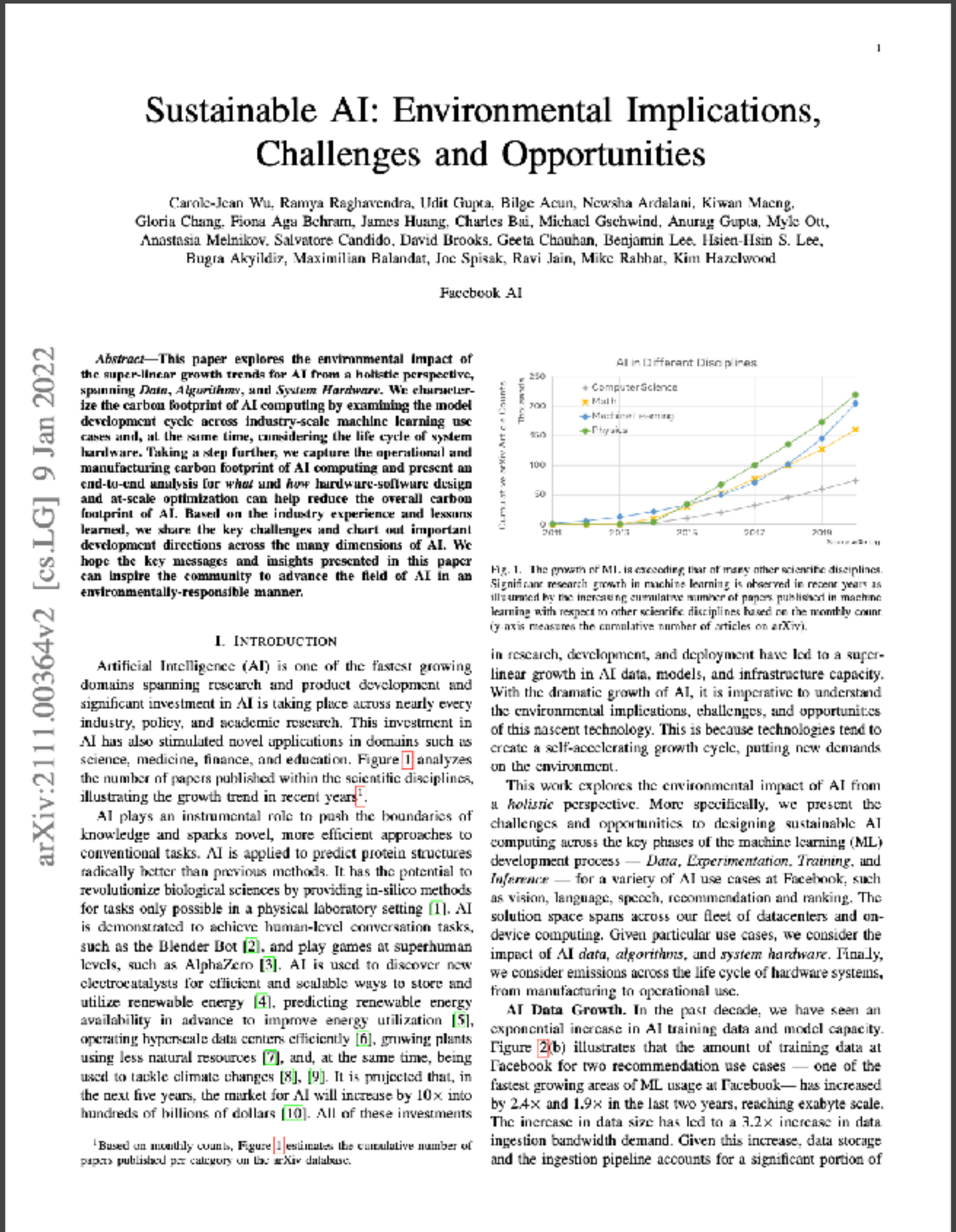


F/color is frequency of upcoming inference requests

Labeled points represent max size of the pool/batch/bundle; G means greedy (process everything that is in the bundle without waiting to reach the max)

Green AI at ~~Facebook~~Meta

Sustainable AI: Environmental Implications, Challenges and Opportunities (2022)



1

Sustainable AI: Environmental Implications, Challenges and Opportunities

Carole-Jean Wu, Ramya Raghavendra, Udit Gupta, Bilge Acun, Newsha Ardalani, Kiwan Maeng, Gloria Chang, Fiona Aga Behram, James Huang, Charles Bai, Michael Gschwind, Anurag Gupta, Myle Ott, Anastasia Melnikov, Salvatore Candido, David Brooks, Geeta Chauhan, Benjamin Lee, Hsien-Hsin S. Lee, Bugra Akyildiz, Maximilian Balandat, Joe Spisak, Ravi Jain, Mike Rabbat, Kim Hazelwood

Facebook AI

arXiv:2111.00364v2 [cs.LG] 9 Jan 2022

***Abstract*—This paper explores the environmental impact of the super-linear growth trends for AI from a holistic perspective, spanning *Data*, *Algorithms*, and *System Hardware*. We characterize the carbon footprint of AI computing by examining the model development cycle across industry-scale machine learning use cases and, at the same time, considering the life cycle of system hardware. Taking a step further, we capture the operational and manufacturing carbon footprint of AI computing and present an end-to-end analysis for *what* and *how* hardware-software design and at-scale optimization can help reduce the overall carbon footprint of AI. Based on the industry experience and lessons learned, we share the key challenges and chart out important development directions across the many dimensions of AI. We hope the key messages and insights presented in this paper can inspire the community to advance the field of AI in an environmentally-responsible manner.**

I. Introduction

Artificial Intelligence (AI) is one of the fastest growing domains spanning research and product development and significant investment in AI is taking place across nearly every industry, policy, and academic research. This investment in AI has also stimulated novel applications in domains such as science, medicine, finance, and education. Figure 1 analyzes the number of papers published within the scientific disciplines, illustrating the growth trend in recent years[1].

AI plays an instrumental role to push the boundaries of knowledge and sparks novel, more efficient approaches to conventional tasks. AI is applied to predict protein structures radically better than previous methods. It has the potential to revolutionize biological sciences by providing in-silico methods for tasks only possible in a physical laboratory setting [1]. AI is demonstrated to achieve human-level conversation tasks, such as the Blender Bot [2], and play games at superhuman levels, such as AlphaZero [3]. AI is used to discover new electrocatalysts for efficient and scalable ways to store and utilize renewable energy [4], predicting renewable energy availability in advance to improve energy utilization [5], operating hyperscale data centers efficiently [6], growing plants using less natural resources [7], and, at the same time, being used to tackle climate changes [8], [9]. It is projected that, in the next five years, the market for AI will increase by 10× into hundreds of billions of dollars [10]. All of these investments in research, development, and deployment have led to a super-linear growth in AI data, models, and infrastructure capacity. With the dramatic growth of AI, it is imperative to understand the environmental implications, challenges, and opportunities of this nascent technology. This is because technologies tend to create a self-accelerating growth cycle, putting new demands on the environment.

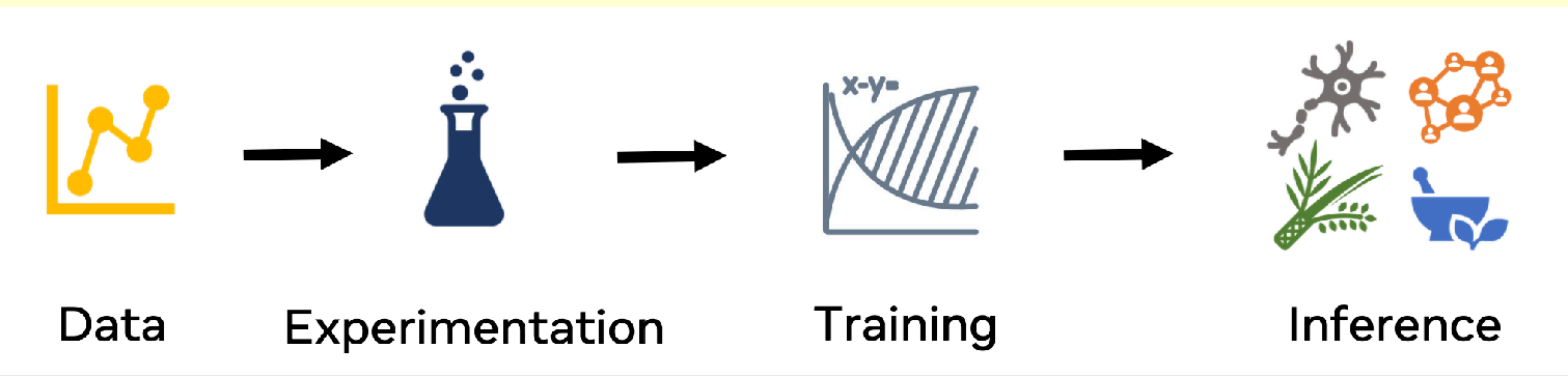
Fig. 1. The growth of ML is exceeding that of many other scientific disciplines. Significant research growth in machine learning is observed in recent years as illustrated by the increasing cumulative number of papers published in machine learning with respect to other scientific disciplines based on the monthly count (y-axis measures the cumulative number of articles on arXiv).

This work explores the environmental impact of AI from a *holistic* perspective. More specifically, we present the challenges and opportunities to designing sustainable AI computing across the key phases of the machine learning (ML) development process — *Data*, *Experimentation*, *Training*, and *Inference* — for a variety of AI use cases at Facebook, such as vision, language, speech, recommendation and ranking. The solution space spans across our fleet of datacenters and on-device computing. Given particular use cases, we consider the impact of AI *data*, *algorithms*, and *system hardware*. Finally, we consider emissions across the life cycle of hardware systems, from manufacturing to operational use.

AI Data Growth. In the past decade, we have seen an exponential increase in AI training data and model capacity. Figure 2(b) illustrates that the amount of training data at Facebook for two recommendation use cases — one of the fastest growing areas of ML usage at Facebook— has increased by 2.4× and 1.9× in the last two years, reaching exabyte scale. The increase in data size has led to a 3.2× increase in data ingestion bandwidth demand. Given this increase, data storage and the ingestion pipeline accounts for a significant portion of

[1] Based on monthly counts, Figure 1 estimates the cumulative number of papers published per category on the arXiv database.

Carbon footprint mapped to the AI lifecycle

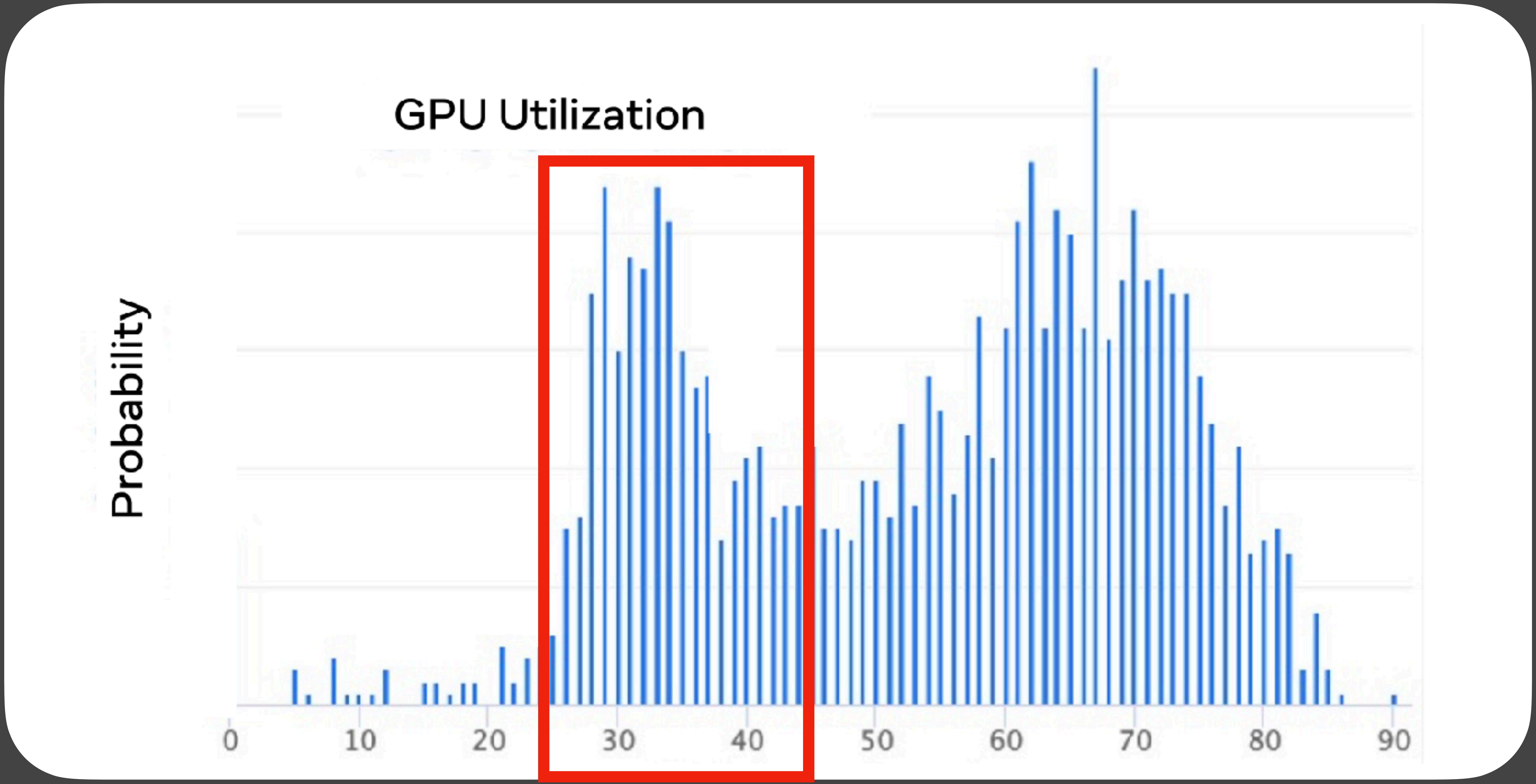


- There are 4 main overarching stages where carbon emissions need to be isolated: **data collection**, **experimentation**, **training**, **inference**.
- At Facebook, recommendation systems split energy consumption **evenly between training and inference**; text translation models have a **35%/65%** split. (Operational cost)
- Operational/embodied cost split: **30%/70%**

Open issues according to Meta

- A vast portion of projects only use **GPUs at 30%**.
  Should be higher to attenuate embodied carbon.

Based on 10K AI projects

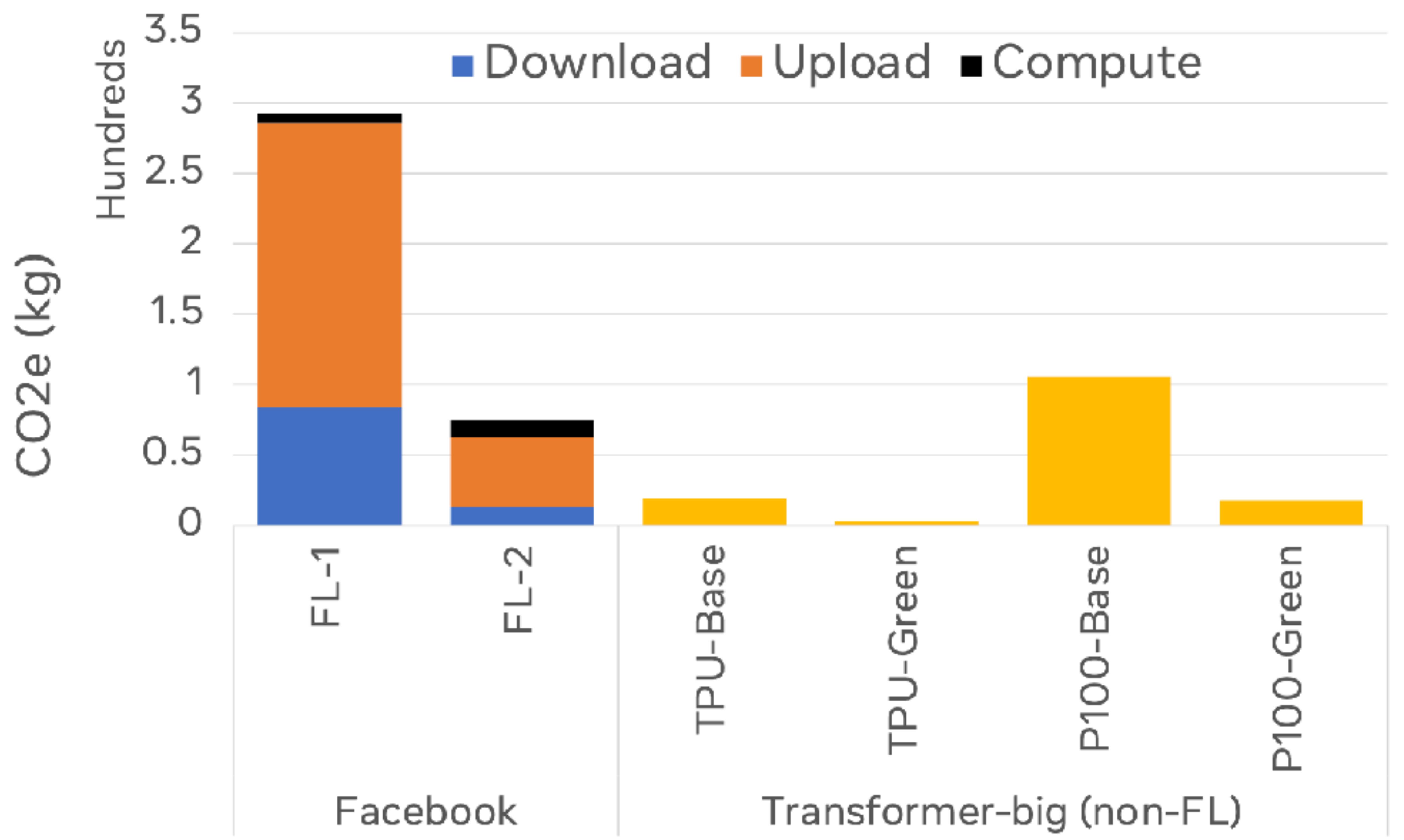


Federated learning

- Federated learning consists of training a ML model across **multiple decentralized edge devices** holding local data samples.
- Federated-learning is a nice solution for **devices with limited energy resources.** E.g., IoT.

Is federated learning a solution for Green AI?

- Most of the carbon footprint stems from **communications**



recap