

11. Green AI

Sustainable Software Engineering
CS4295



Luís Cruz
L.Cruz@tudelft.nl

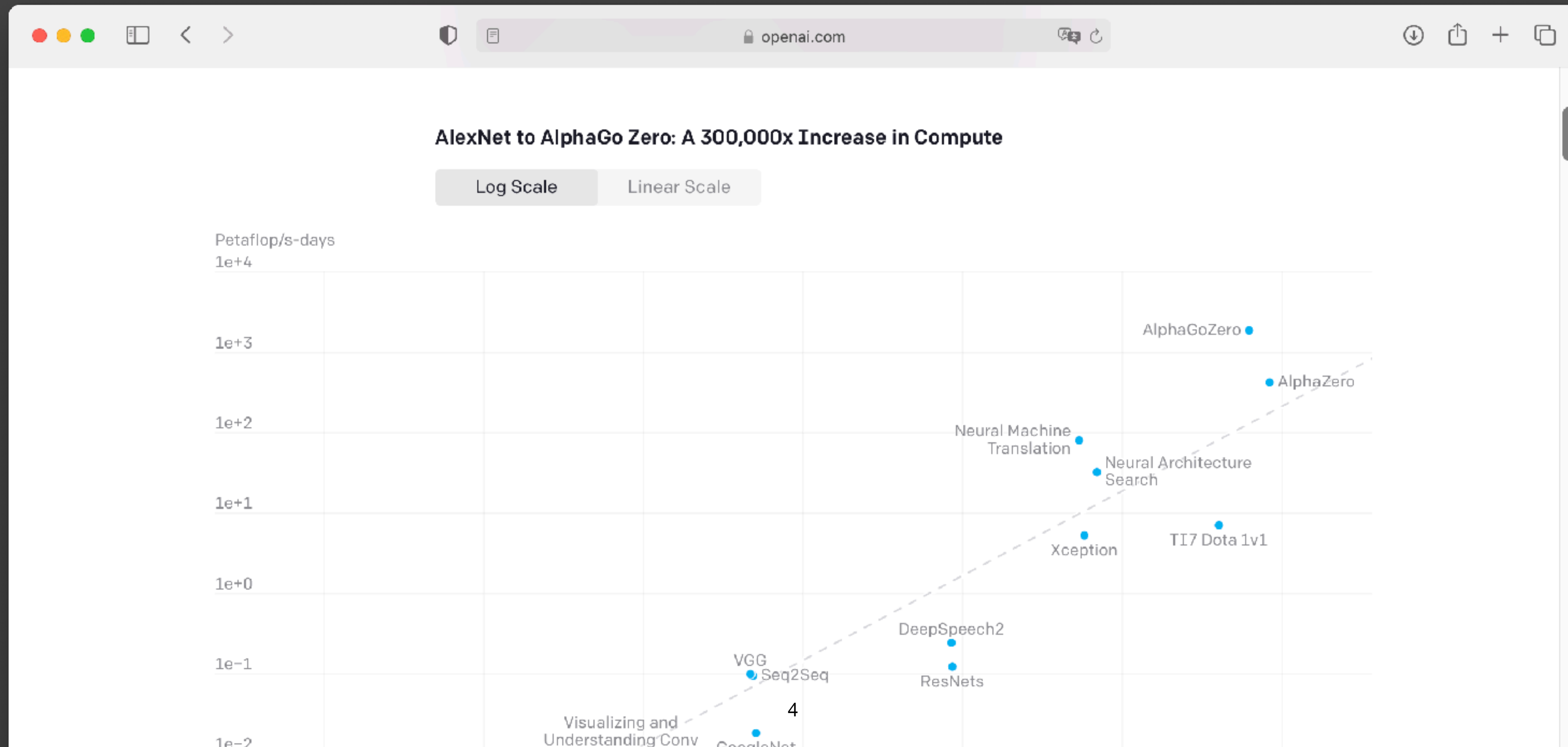
- Overview of Green AI
 - Large language models
- Green data-centric AI
- Model simplification
- Hyper parameter tuning
- Batching for Green AI
- Green AI at Meta

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

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: 0.9999999999

Green AI



- Energy
- Time
- Reproducibility
- Reusage

Research on Green AI

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

On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜

Emily M. Bender*
ebender@uw.edu
University of Washington
Seattle, WA, USA

Timnit Gebru*
timnit@blackinai.org
Black in AI
Palo Alto, CA, USA

Angelina McMillan-Major
aymm@uw.edu
University of Washington
Seattle, WA, USA

Shmargaret Shmitchell
shmargaret.shmitchell@gmail.com
The Author

ABSTRACT

The past 3 years of work in NLP have been characterized by the development and deployment of ever larger language models, especially for English. BERT, its variants, GPT-2/3, and others, most recently Switch-C, have pushed the boundaries of the possible both through architectural innovations and through sheer size. Using these pretrained models and the methodology of fine-tuning them for specific tasks, researchers have extended the state of the art on a wide array of tasks as measured by leaderboards on specific benchmarks for English. In this paper, we take a step back and ask: How big is too big? What are the possible risks associated with this technology and what paths are available for mitigating those risks? We provide recommendations including weighing the environmental and financial costs first, investing resources into curating and carefully documenting datasets rather than ingesting everything on the web, carrying out pre-development exercises evaluating how the planned approach fits into research and development goals and supports stakeholder values, and encouraging research directions beyond ever larger language models.

CCS CONCEPTS

- Computing methodologies → Natural language processing

ACM Reference Format:
Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜. In *Conference on Fairness, Accountability, and Transparency*.

Energy and Policy Considerations for Deep Learning in NLP

Emma Strubell Ananya Ganesh Andrew McCallum
College of Information and Computer Sciences
University of Massachusetts Amherst
{strubell, aganesh, mccallum}@cs.umass.edu

Abstract

Recent progress in hardware and methodology for training neural networks has ushered in a new generation of large networks trained on abundant data. These models have obtained notable gains in accuracy across many NLP tasks. However, these accuracy improvements depend on the availability of exceptionally large computational resources that necessitate similarly substantial energy consumption. As a result these models are costly to train and develop, both financially, due to the cost of hardware and electricity or cloud compute time, and environmentally, due to the carbon footprint required to fuel modern tensor processing hardware. In this paper we bring this issue to the attention of NLP researchers by quantifying the approximate financial and environmental costs of training a variety of recently successful neural network models for NLP. Based on these findings, we propose actionable recommendations to reduce costs and improve equity in NLP research and practice.

Consumption	CO ₂ e (lbs)
Air travel, 1 passenger, NY↔SF	1984
Human life, avg, 1 year	11,023
American life, avg, 1 year	36,156
Car, avg incl. fuel, 1 lifetime	126,000

Training one model (GPU)	CO ₂ e (lbs)
NLP pipeline (parsing, SRL) w/ tuning & experimentation	39 78,468
Transformer (big) w/ neural architecture search	192 626,155

Table 1: Estimated CO₂ emissions from training common NLP models, compared to familiar consumption.¹

NLP models could be trained and developed on a commodity laptop or server, many now require multiple instances of specialized hardware such as GPUs or TPUs, therefore limiting access to these highly accurate models on the basis of finances. Even when these expensive computational resources are available, model training also incurs a substantial cost to the environment due to the energy required to power this hardware for weeks or months at a time. Though some of this energy may come from renewable or carbon credit-offset resources, the high energy demands of these models are still a concern since (1) energy is not currently derived from carbon-neutral sources in many locations, and (2) when renewable energy is available, it is still limited to the equipment we have to produce and store it, and energy spent training a neural network might better be allocated to heating a family's home. It is estimated that we must cut carbon emissions by half over the next decade to deter escalating rates of natural disaster, and based

Introduction

Advances in techniques and hardware for training deep neural networks have recently enabled impressive accuracy improvements across any fundamental NLP tasks (Bahdanau et al., 2015; Luong et al., 2015; Dzout and Manning, 2017; Vaswani et al., 2017), with the most computationally-hungry models obtaining the highest scores (Peters et al., 2018; Devlin et al., 2019; Radford et al., 2019; So et al., 2019). As a result, training a state-of-the-art model now requires substantial computational resources which demand considerable energy, along with the associated financial and environmental costs. Re-

Green AI

DOI:10.1145/3381831

Creating efficiency in AI research will decrease its carbon footprint and increase its inclusivity as deep learning study should not require the deepest pockets.

BY ROY SCHWARTZ, JESSE DODGE, EN ETZIONI

of artificial intelligence (AI) has progress on a broad range of object recognition, game playing, and machine translation.⁴³ Much of this progress has been achieved by increasingly large and intensive deep learning models.⁴ From Amodei et al.,² plots training time for state-of-the-art deep learning models like AlexNet in 2012²⁴ to AlphaZero in 2017, we observe an overall increase of 300,000x, or roughly doubling every few months. An even more recent example is observed in NLP word-embedding models like ELMo²⁴ followed by BERT,¹⁰ Megatron-LM,⁴² T5,³⁶ and GPT-3.⁴ GPT-3 has estimated the carbon footprint of training this model as 550,000 lbs of CO₂.⁴ NLP models and argued this trend is not only unfriendly and prohibitively expensive but also a barrier to participation in NLP research. This work as Red AI.

This article, but our focus is on AI research that relies on deep learning.

DECEMBER 2020 | VOL. 03 | NO. 12

ILLUSTRATION BY LUISA MARRAS

This trend is driven by the strong focus of the AI community on obtaining “state-of-the-art” results,⁹ as exemplified by the popularity of leaderboards,^{33,34} which typically report accuracy (or other similar measures) but omit any mention of cost or efficiency (see, for example, leaderboards.allenai.org).⁵ Despite the clear benefits of improving model accuracy, the focus on this single metric ignores the economic, environmental, and social cost of reaching the reported results.

We advocate increasing research activity in Green AI—AI research that is more environmentally friendly and inclusive. We emphasize that Red AI research has been yielding valuable scientific contributions to the field, but it has been overly dominant. We want to shift the balance toward the Green AI option—to ensure any inspired undergraduate with a laptop has the opportunity to write high-quality papers that could be accepted at premier research conferences. Specifically, we propose making efficiency a more common evaluation criterion for AI papers alongside accuracy and related measures.

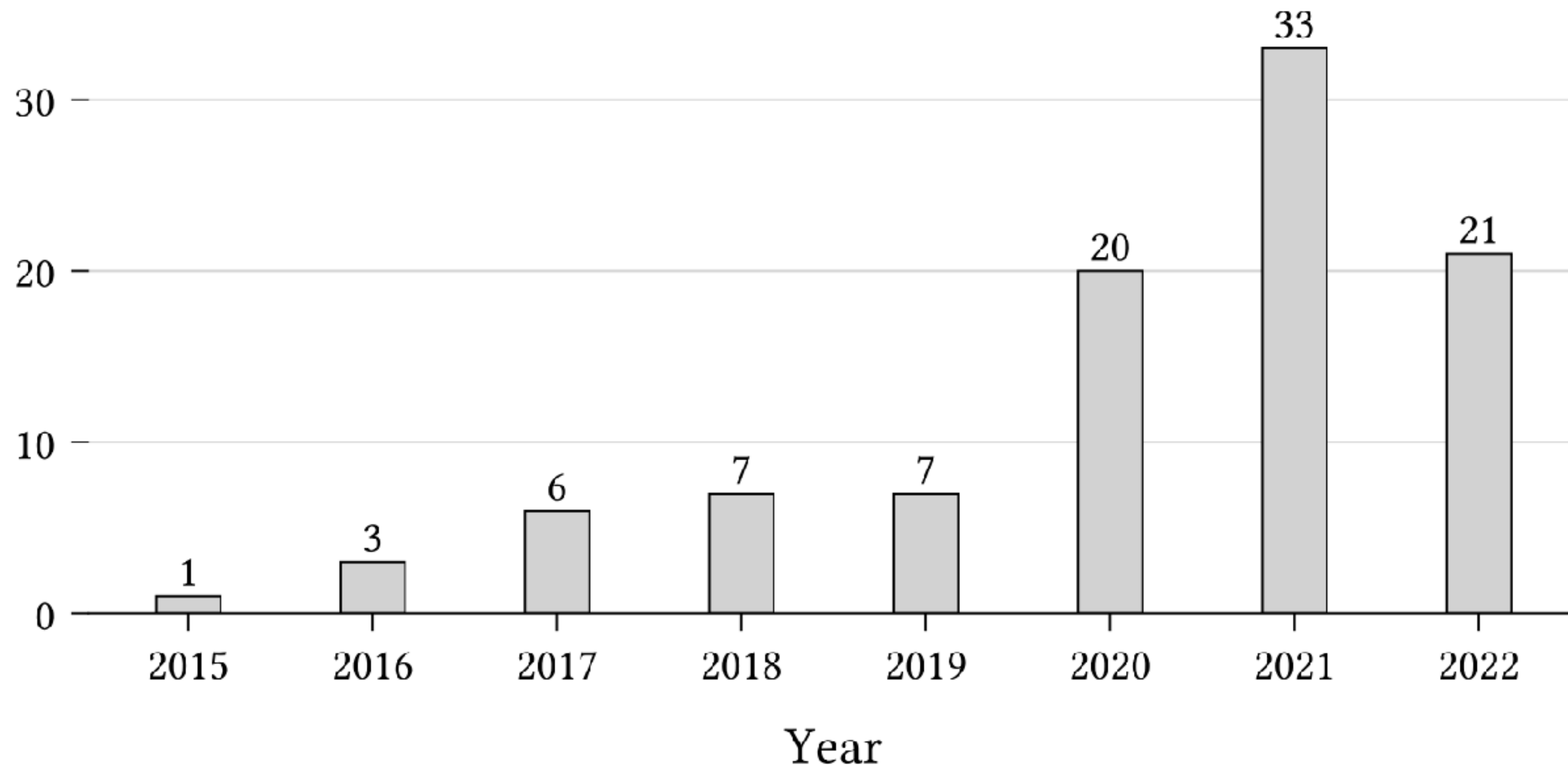
Meaning, in practice, that a system’s accuracy on some benchmark is greater than any previously reported system’s accuracy.

Some leaderboards do focus on efficiency (https://dawn.cs.stanford.edu/benchmark/).

key insights

- The computational costs of state-of-the-art AI research has increased 300,000x in recent years. This trend, denoted Red AI, stems from the AI community’s focus on accuracy while paying attention to efficiency.
- Red AI leads to a surprisingly large carbon footprint, and makes it difficult for academics, students, and researchers to engage in deep learning research.
- An alternative is Green AI, which treats efficiency as a primary evaluation criterion alongside accuracy. To measure efficiency, we suggest reporting the number of floating-point operations required to generate a result.
- Green AI research will decrease AI’s environmental footprint and increase its inclusivity.

Publications in Green AI over the years



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

A Systematic Review of Green AI

Roberto Verdecchia
Vrije Universiteit Amsterdam
Amsterdam, The Netherlands
r.verdecchia@vu.nl

Jane Sallou
TU Delft
Delft, The Netherlands
j.sallou@tudelft.nl

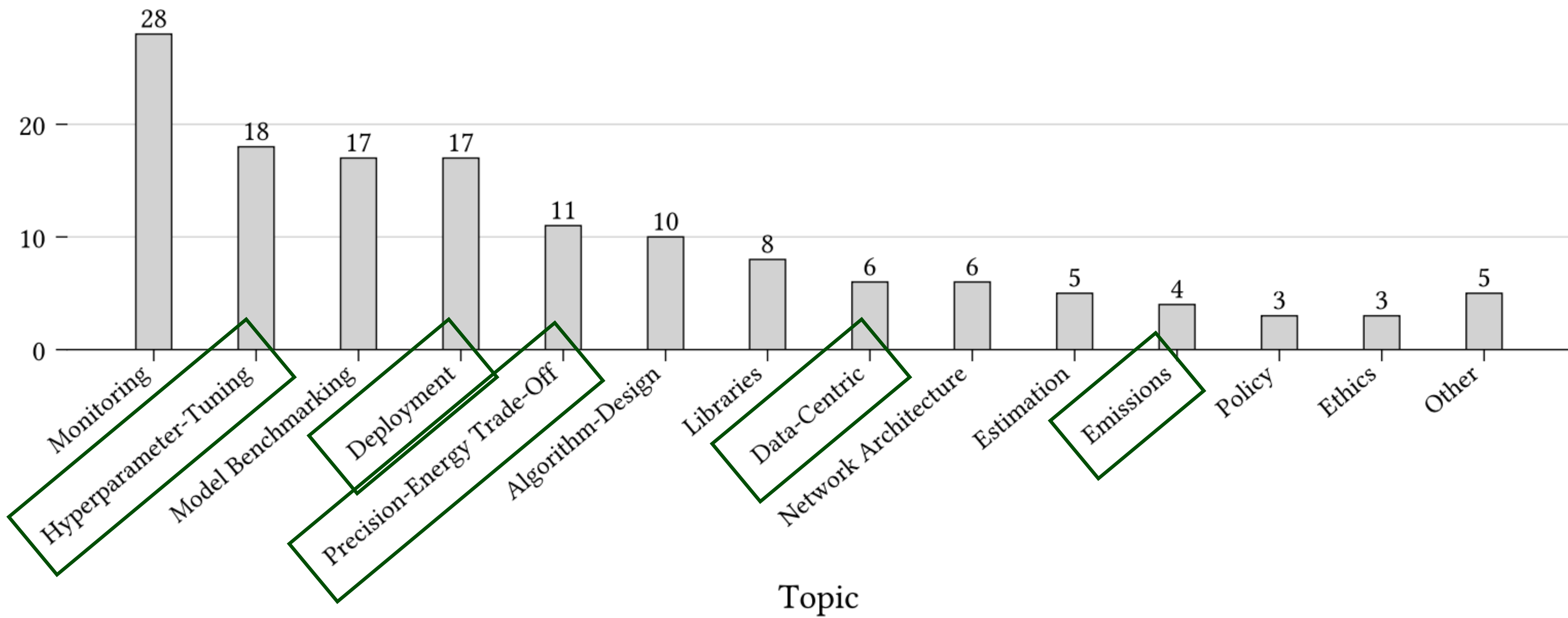
Luis Cruz
TU Delft
Delft, The Netherlands
lcruz@tudelft.nl

Abstract
With the ever-growing adoption of AI-based systems, the carbon footprint of AI is no longer negligible. AI researchers and practitioners are therefore urged to hold themselves accountable for the carbon emission of the AI models they design and use. This led in recent years to the appearance of research tackling AI environmental sustainability, a field referred to as Green AI. Despite the rapid growth of its track in the topic, a comprehensive overview of Green AI research is to date still missing. To address this gap, in this paper, we present a systematic review of the Green AI literature. From the analysis of 94 primary studies, different patterns emerge. The topic experienced a considerable growth from 2020 onward. Most studies consider monitoring AI model footprint, tuning hyperparameters to improve model sustainability, or benchmarking models. A mix of position papers, observational studies, and solution papers are present. Most papers focus on the training phase: use algorithm-specific or study neural networks, and use usage data. Laboratory experiments are the most common research strategy. Reported Green AI energy savings go up to 115%, with savings over 50% being rather common. Industrial parties are involved in Green AI studies, albeit most target academic readers. Green AI tool provisioning is scarce. As a conclusion, the Green AI research field results to have reached a considerable level of maturity. Therefore, from this review emerges that the time is suitable to adapt other Green AI research strategies and put the main emphasis on providing academic results to industrial practice.

1 Introduction
In recent years, the Artificial Intelligence (AI) community has been challenged in being the carbon footprint of AI models to the top of their research agenda. The track paper by Strubell et al. [37] analyzes the carbon impact of training their own state-of-the-art models. Results lead to the conclusion that we need to reduce the carbon footprint of developing and running AI models.
This self-reflection was a response to the AI research community. Many papers followed, calling for a new research direction that would consider this problem. Schwartz et al. coined the term Green AI as "AI research that yields novel results while taking into account the computational cost" [38]. Bender et al. published a position paper highlighting the consequences of continuously increasing the size of AI models [39]. A natural question that is posed is whether we are doing enough as a research community to mitigate the carbon impact of developing and running AI-based software.
AI systems are significantly complex and, as such, Green AI, we need a joint effort that targets all the different stages of an AI system's lifecycle (e.g. data collection, training, maintenance), different artifacts (e.g. data, model, pipeline architecture, hardware), etc. [4].
Given the heterogeneity of the field, it is also difficult to have a broad view of all the Green AI literature that has been published in the past years. To understand the existing research, we conduct a systematic literature review on Green AI. We provide an overview and characterization of the existing research in this field. Moreover, we study how the field has been evolving over the years, pinpoint the main topics, approaches, artifacts, and so on.
This literature review shows that there has been a significant growth in Green AI publications - 798 of the papers have been published since 2020. The most popular topics involve model monitoring, hyperparameter tuning, deployment, and model benchmarking. We also highlight other emerging topics that might lead to interesting solutions - namely, Data-Centric Green AI, Predictive-Energy-Tracking, and so on. The current body of research has already allowed promising results with energy savings from 13% up to 115%.

Graphical Abstract: From a systematic review of the Green AI literature, Green AI results to focus on solutions, and is often not bound to a specific context or algorithm. The Green AI research field results to be mature, i.e. the moment is suitable to port results from academic research to industrial practice.

(Verdecchia, 2022)

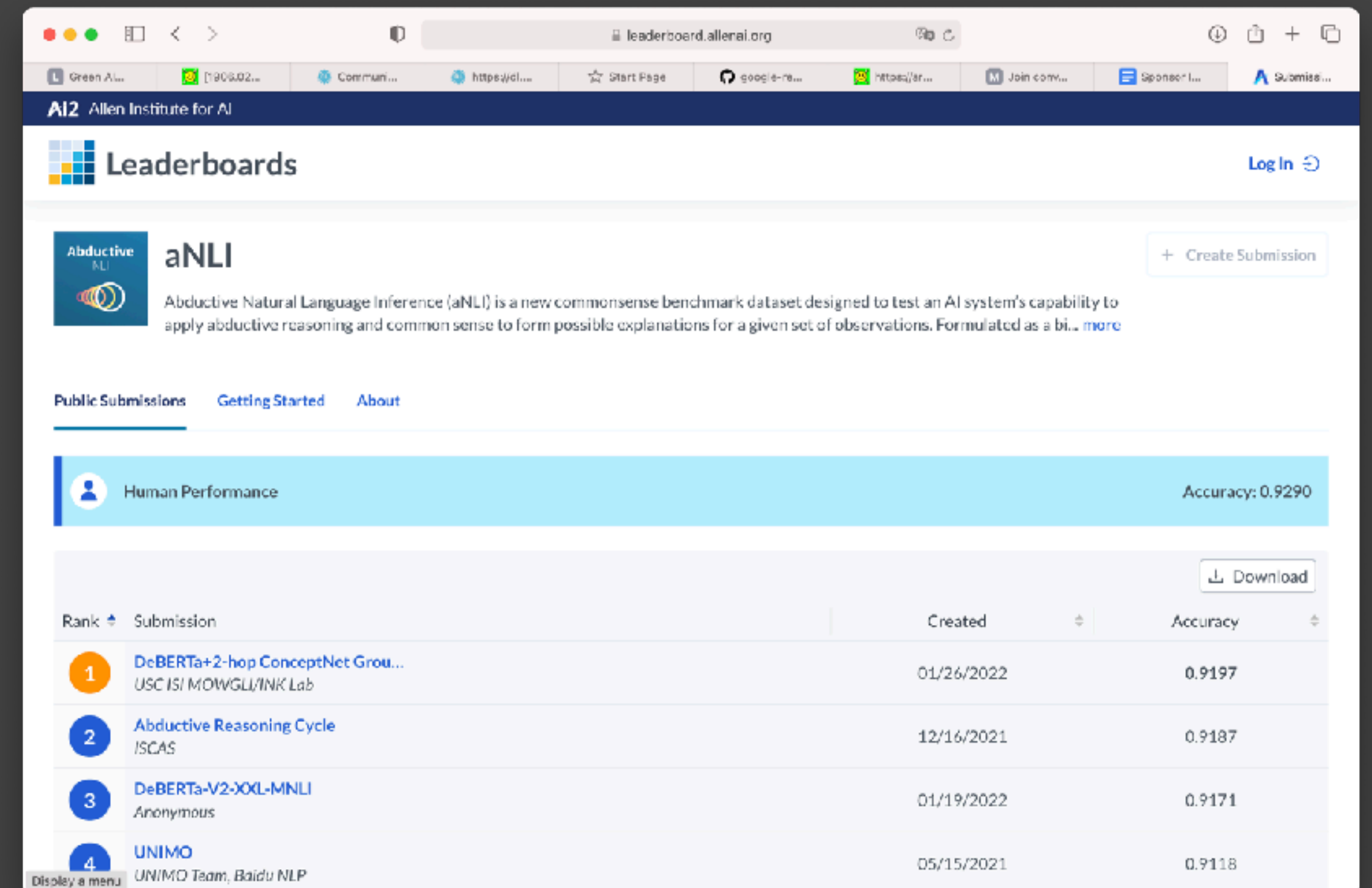


How can we adopt **Green AI**

- **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

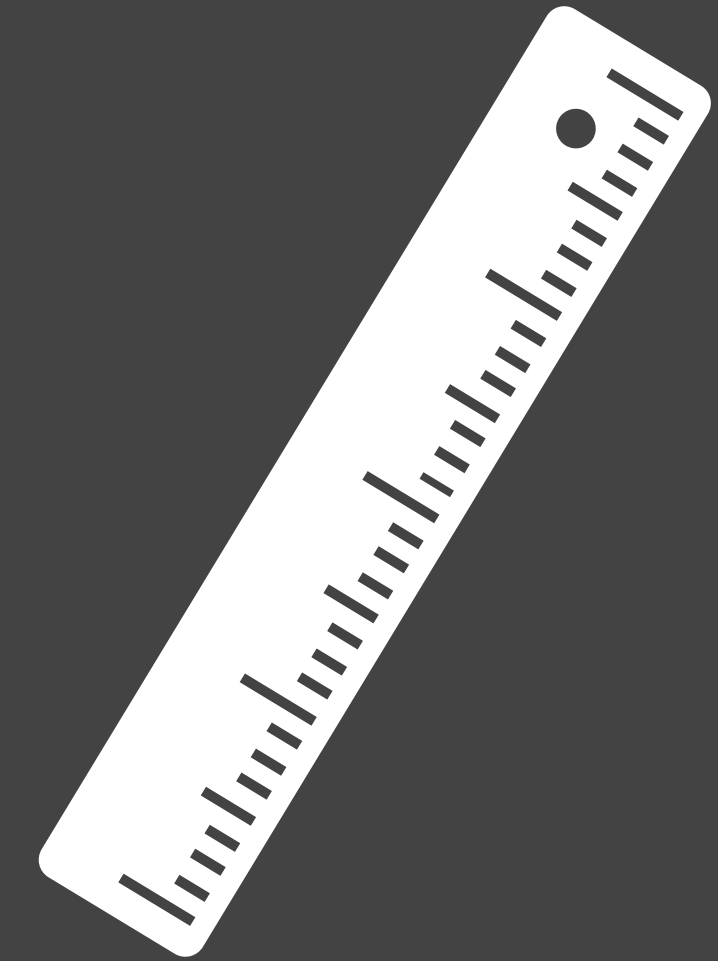


The screenshot shows the AllenAI Leaderboards website for the aNLI benchmark. The page displays the top human performance and a list of public submissions. The table below summarizes the submission data shown in the image.

Rank	Submission	Created	Accuracy
1	DeBERTa+2-hop ConceptNet Group USC ISI MOWGLI/INK Lab	01/26/2022	0.9197
2	Abductive Reasoning Cycle ISCAS	12/16/2021	0.9187
3	DeBERTa-V2-XXL-MNLI Anonymous	01/19/2022	0.9171
4	UNIMO UNIMO Team, Baidu NLP	05/15/2021	0.9118

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

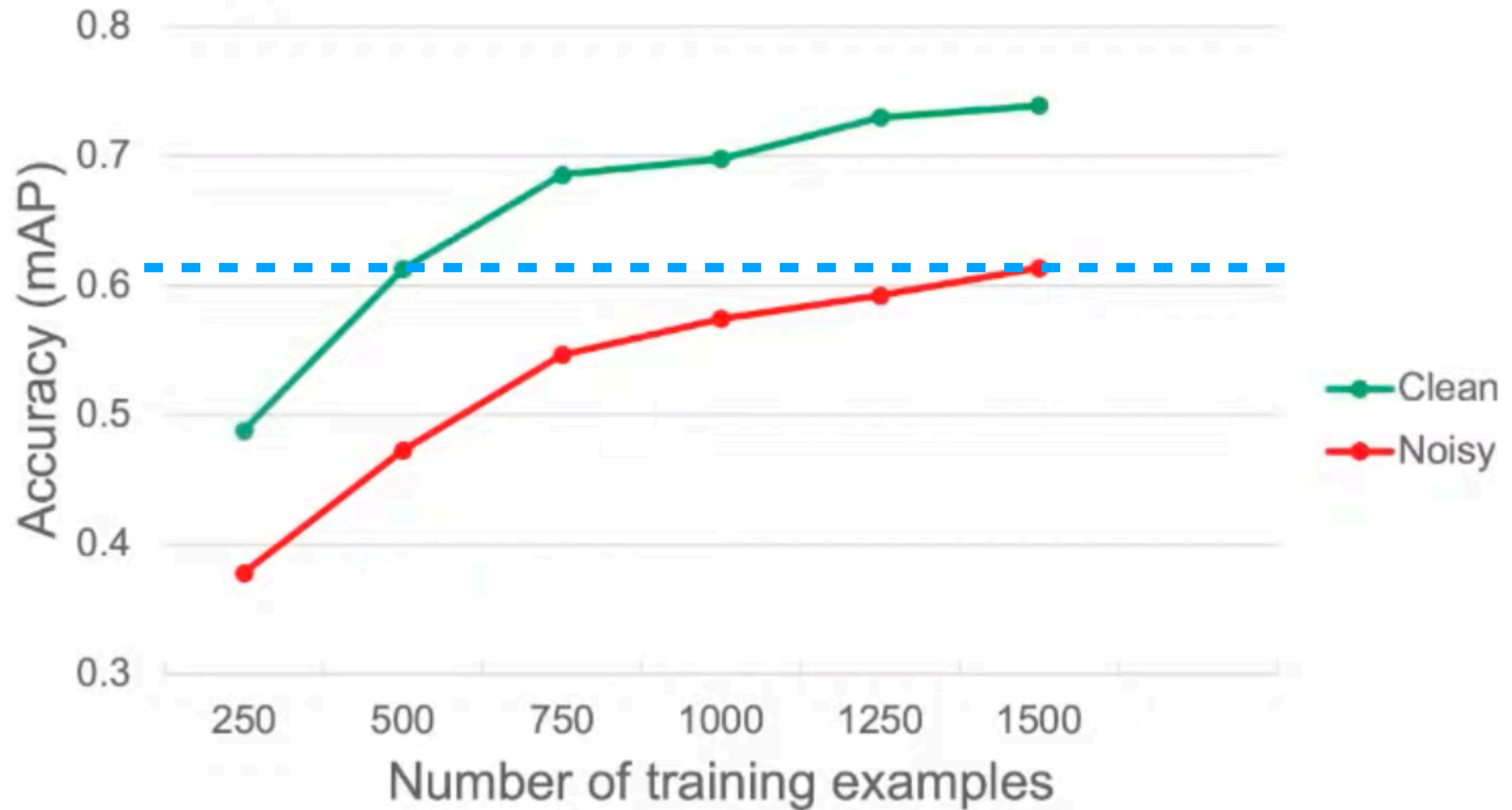


Data-centric AI

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 sub-field 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* [6]. 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>.

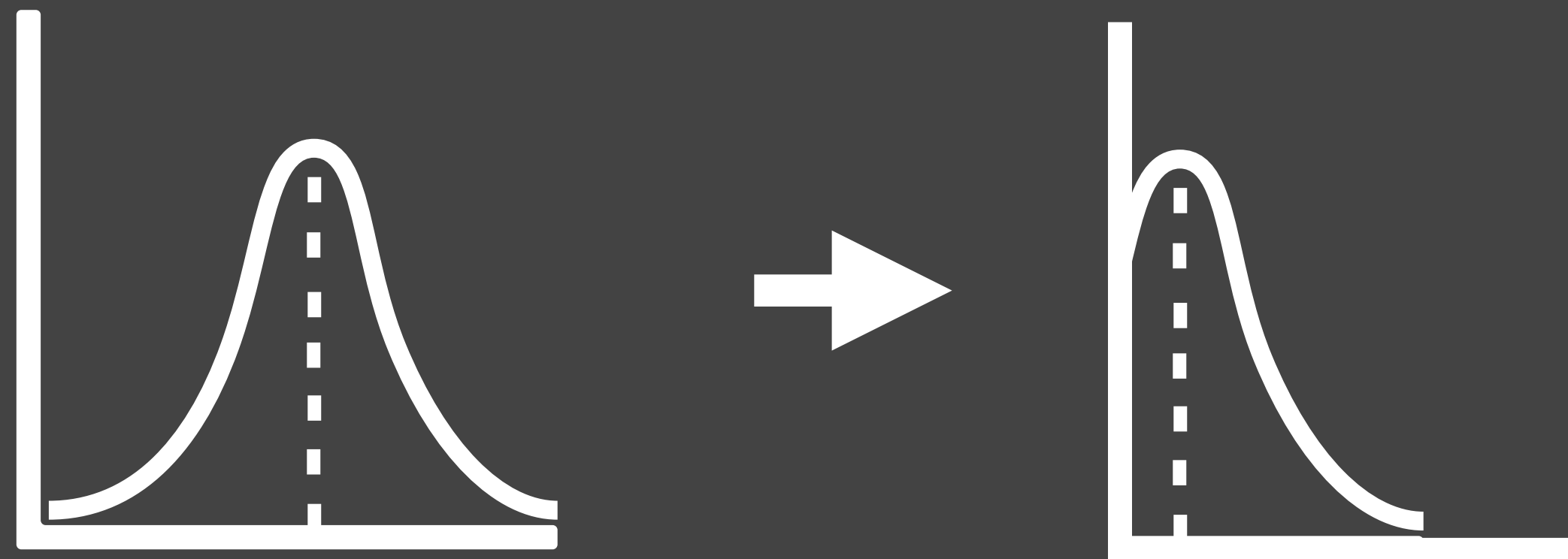
The remainder of this paper is structured as follows. Section II presents the related work on the energy consumption

¹Understanding Data-Centric AI: <https://landing.ai/data-centric-ai/>, Accessed 24th January 2022.

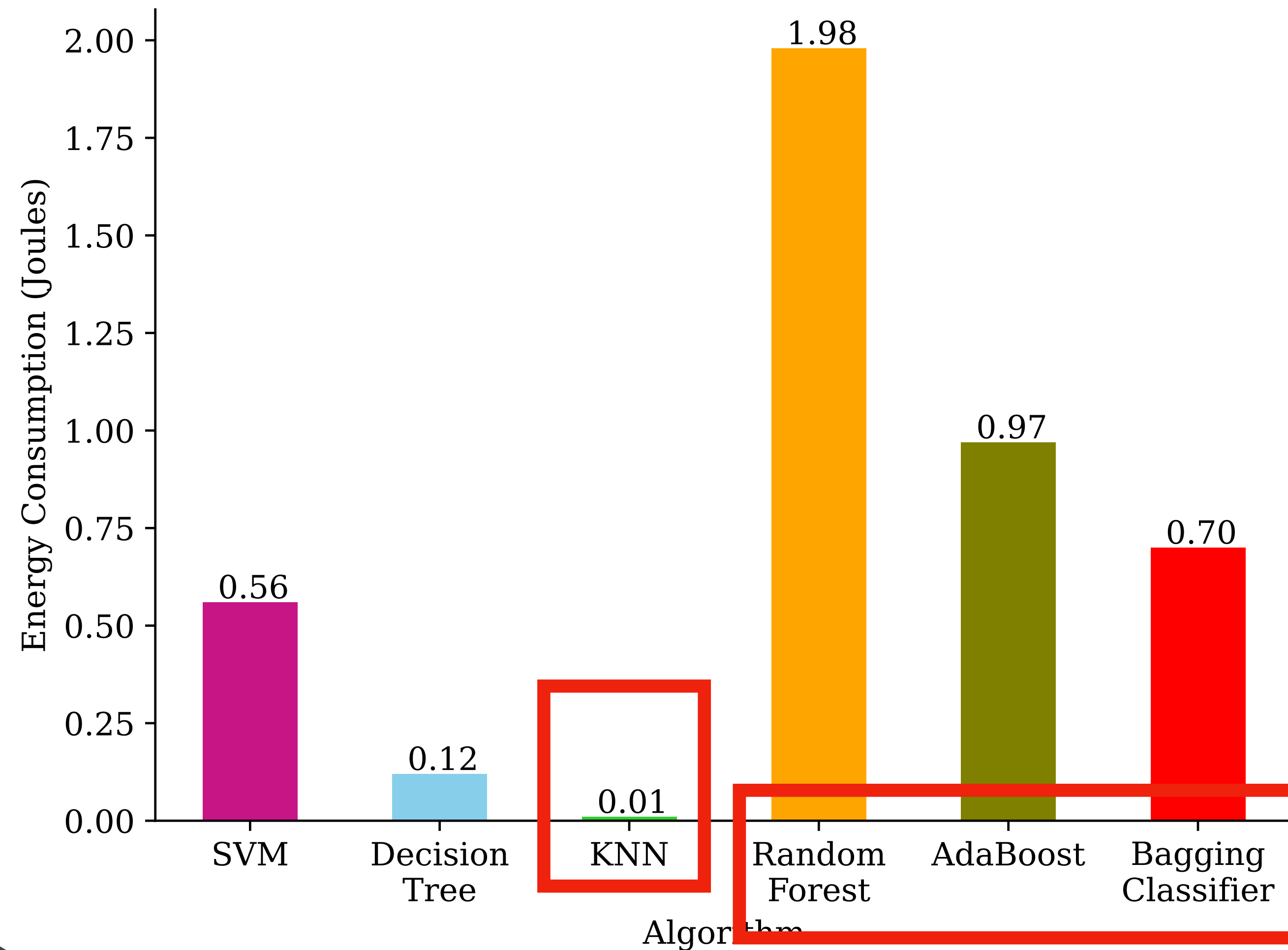
Method

- 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. (?)

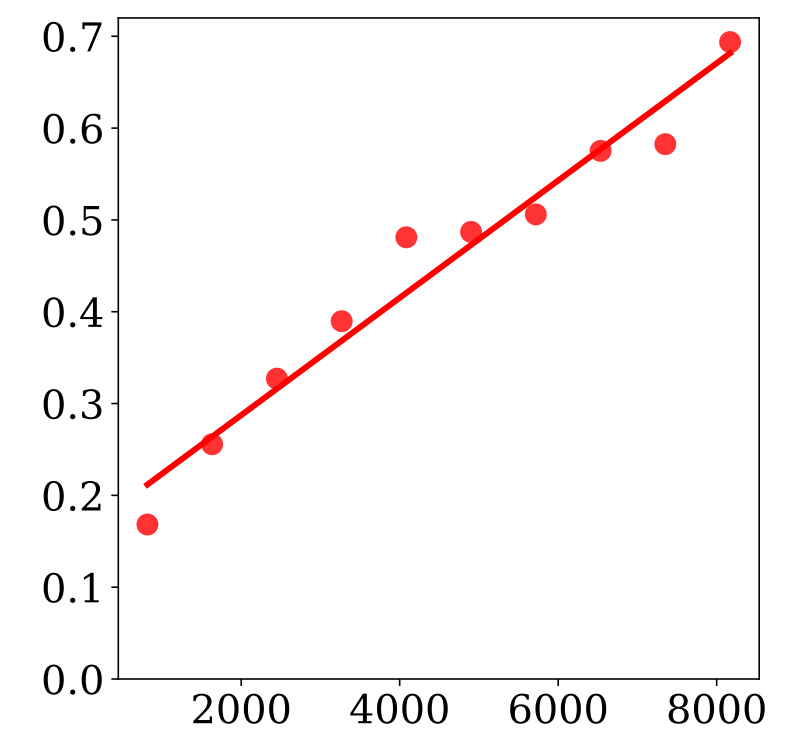
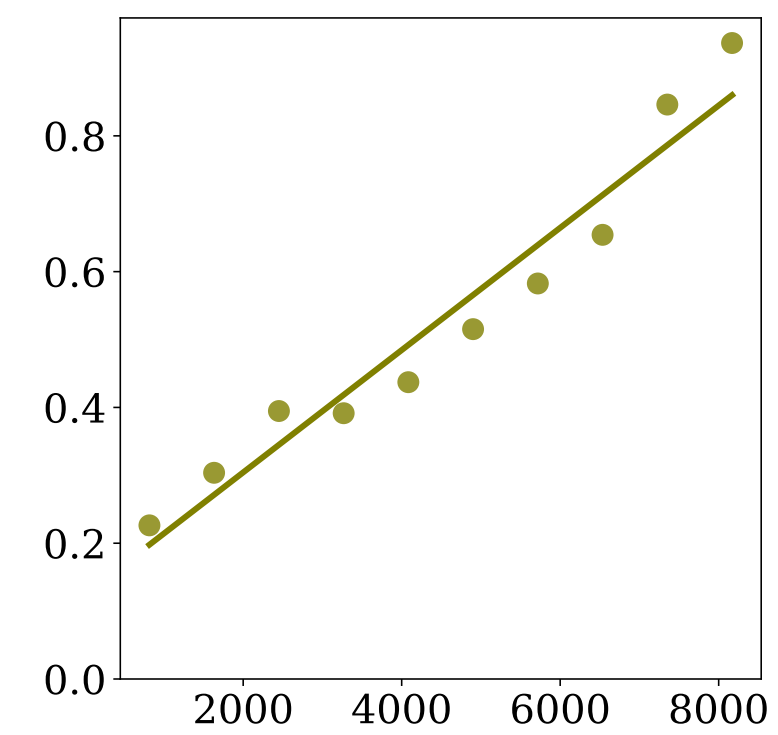
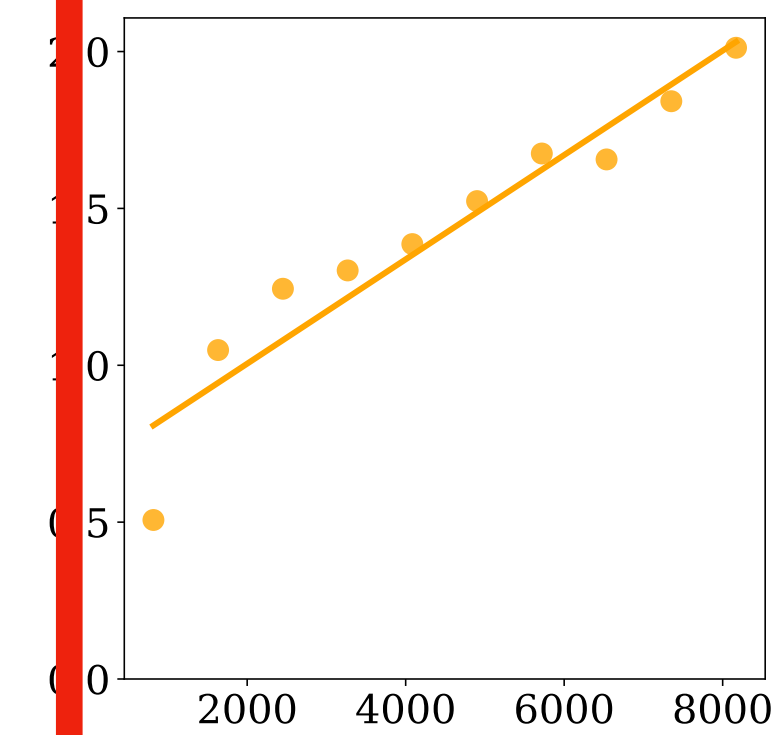
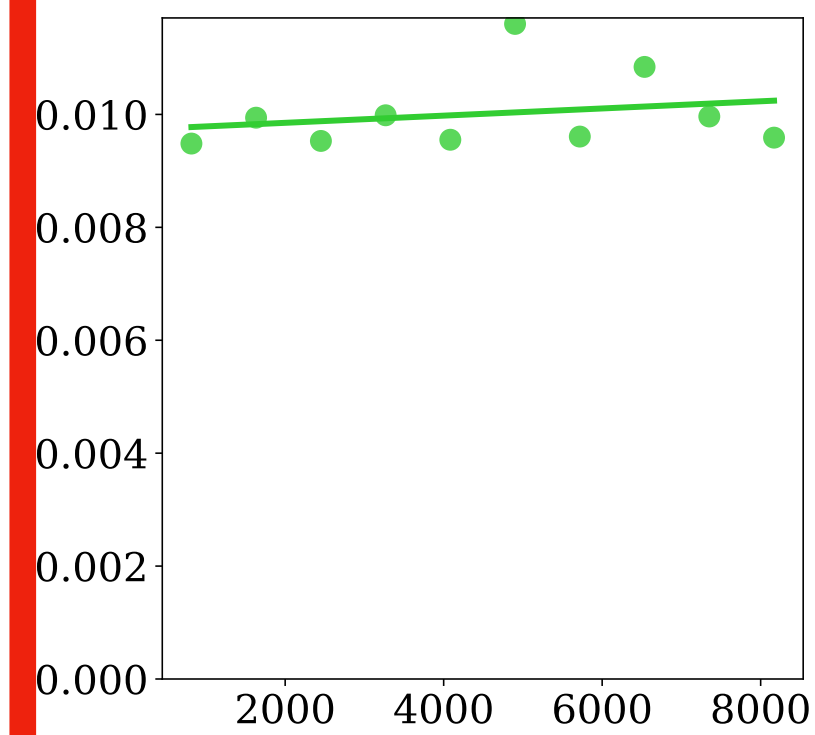
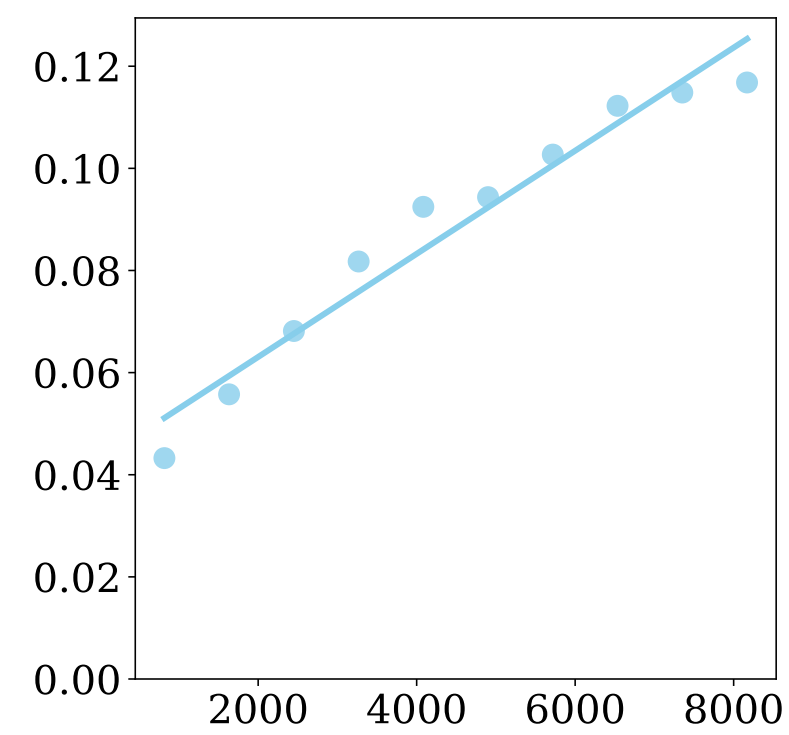
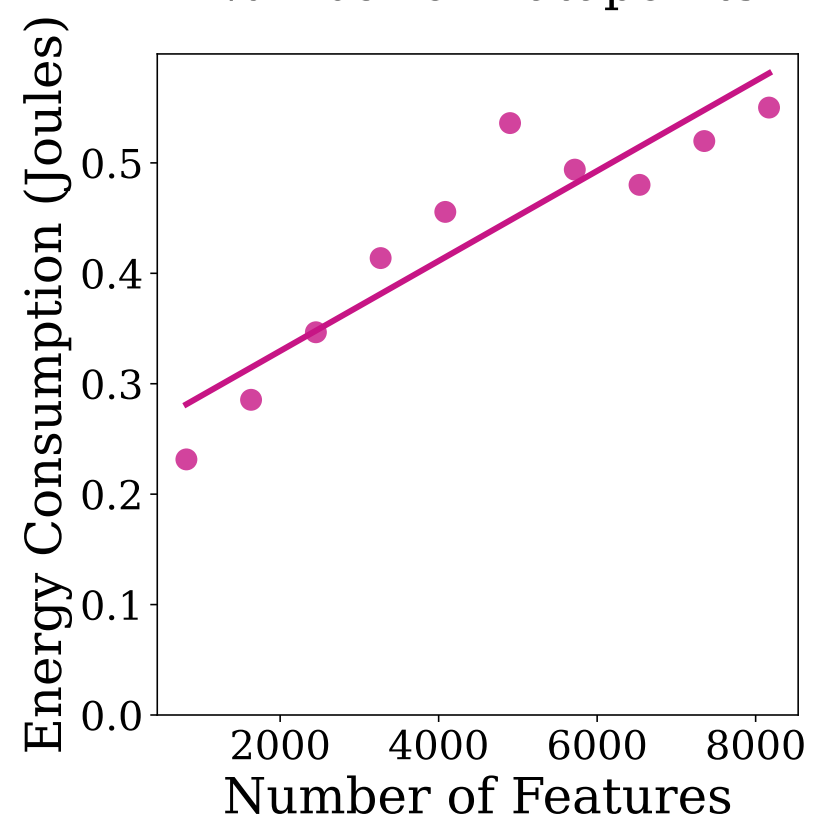
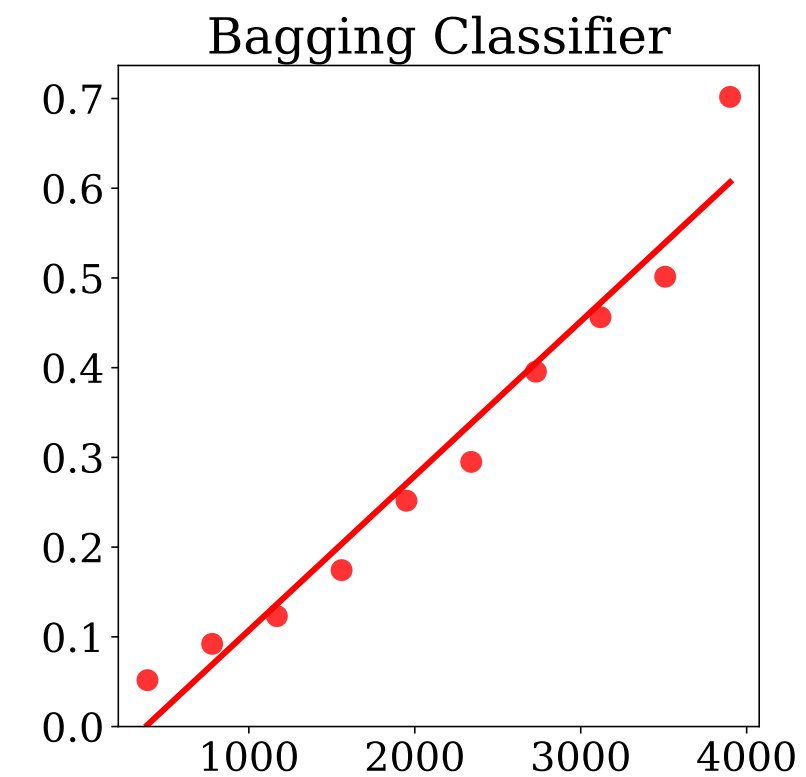
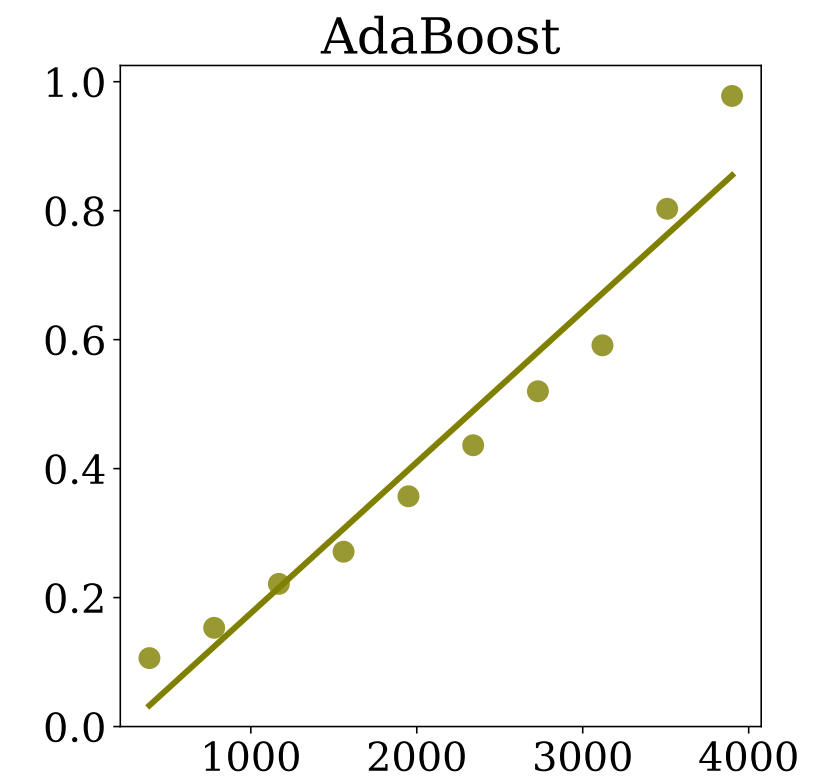
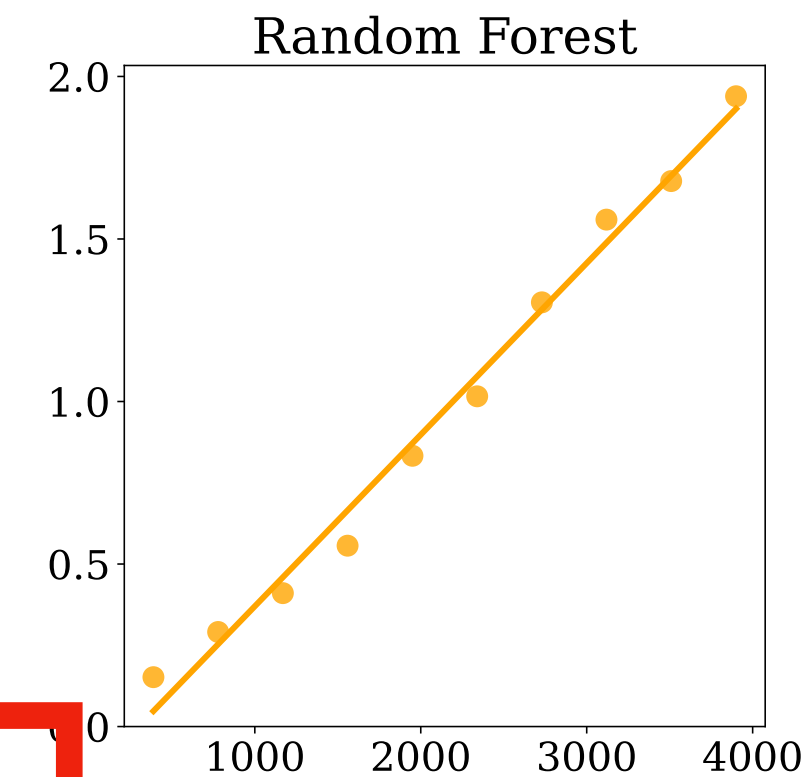
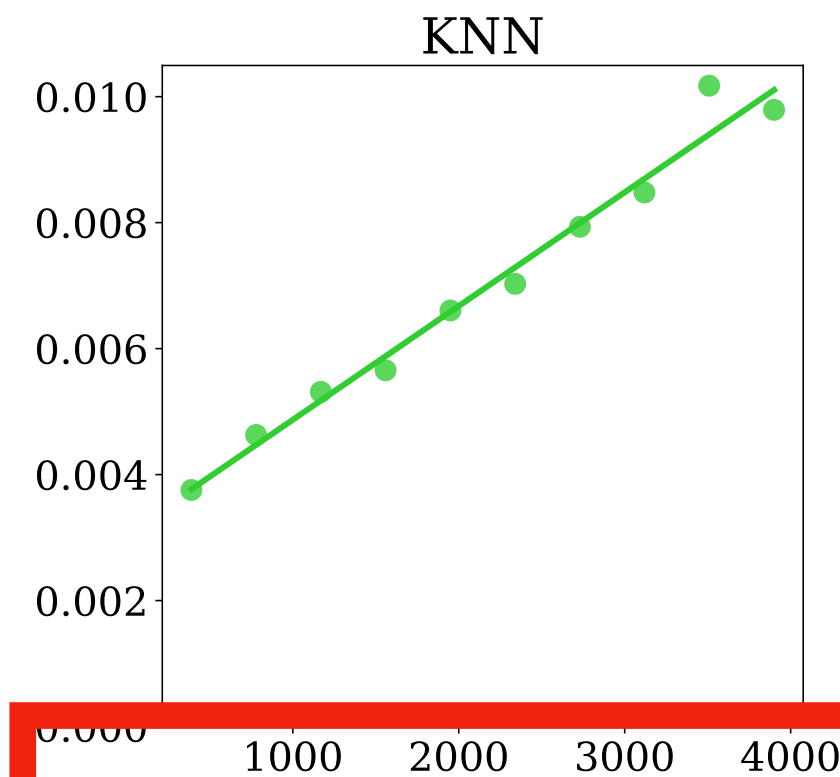
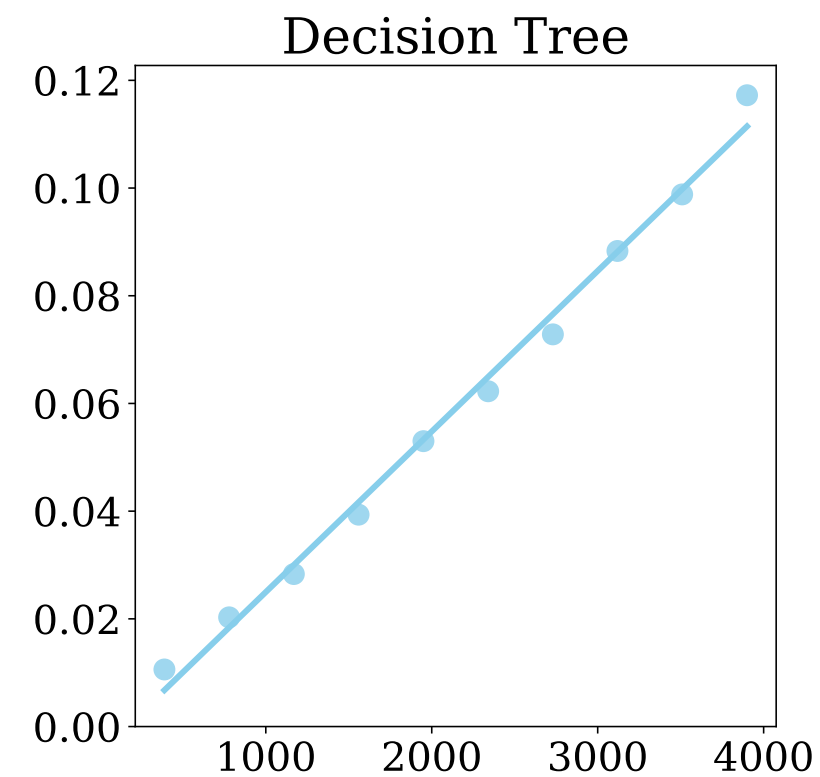
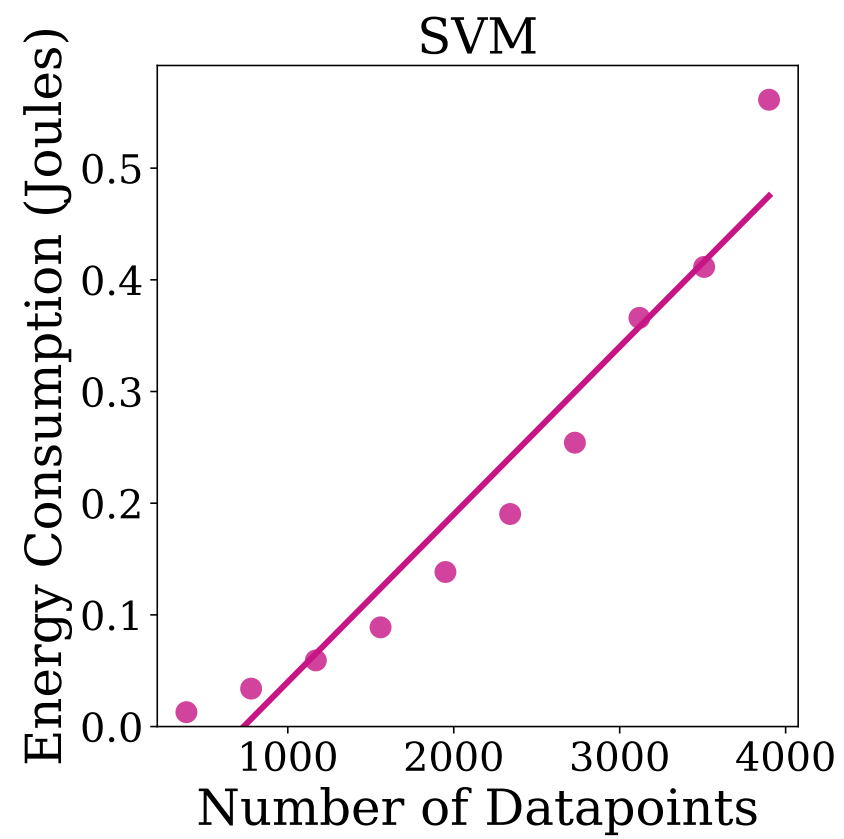
- Repeat 30 times
- Fix random seeds
- ...
- Data was **not Normal** \Rightarrow ^(?) **tailed Normal distribution.**



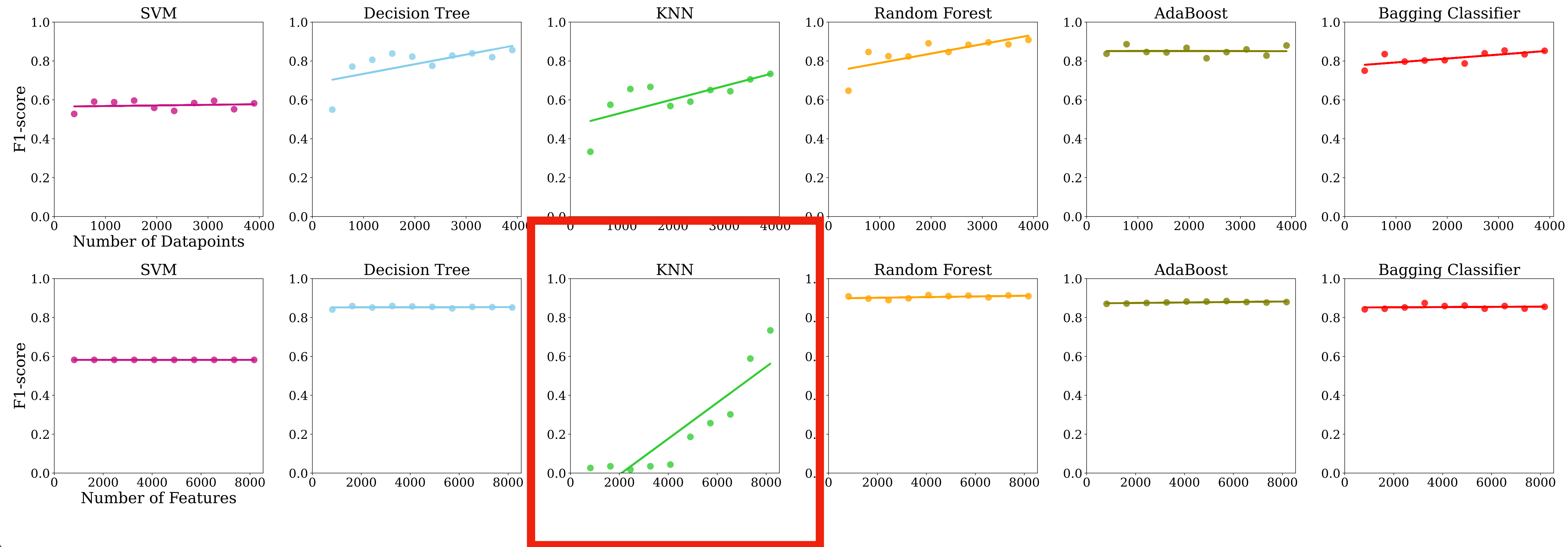
Results: energy consumption of algorithms



Results: energy vs data shape



Results: performance vs data shape

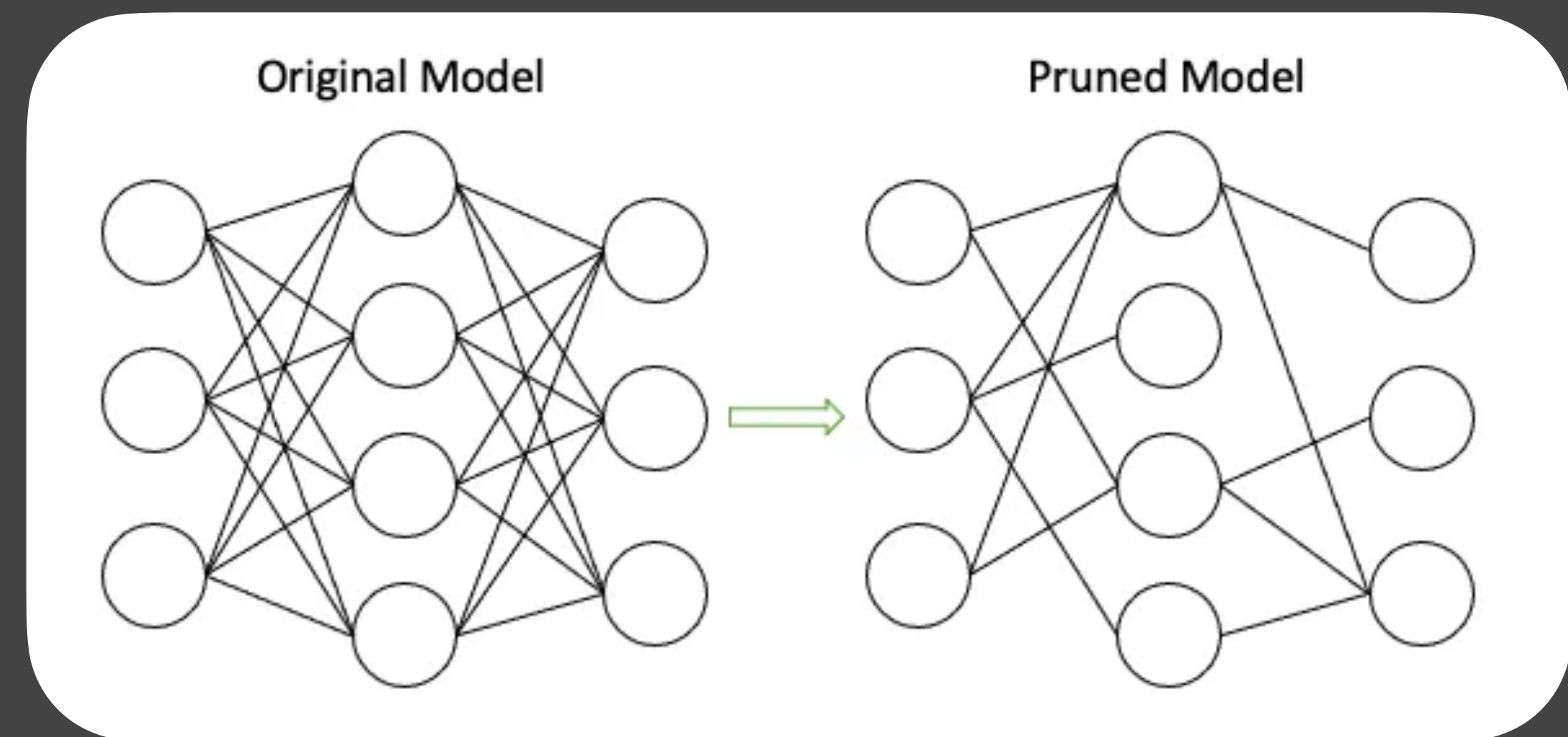
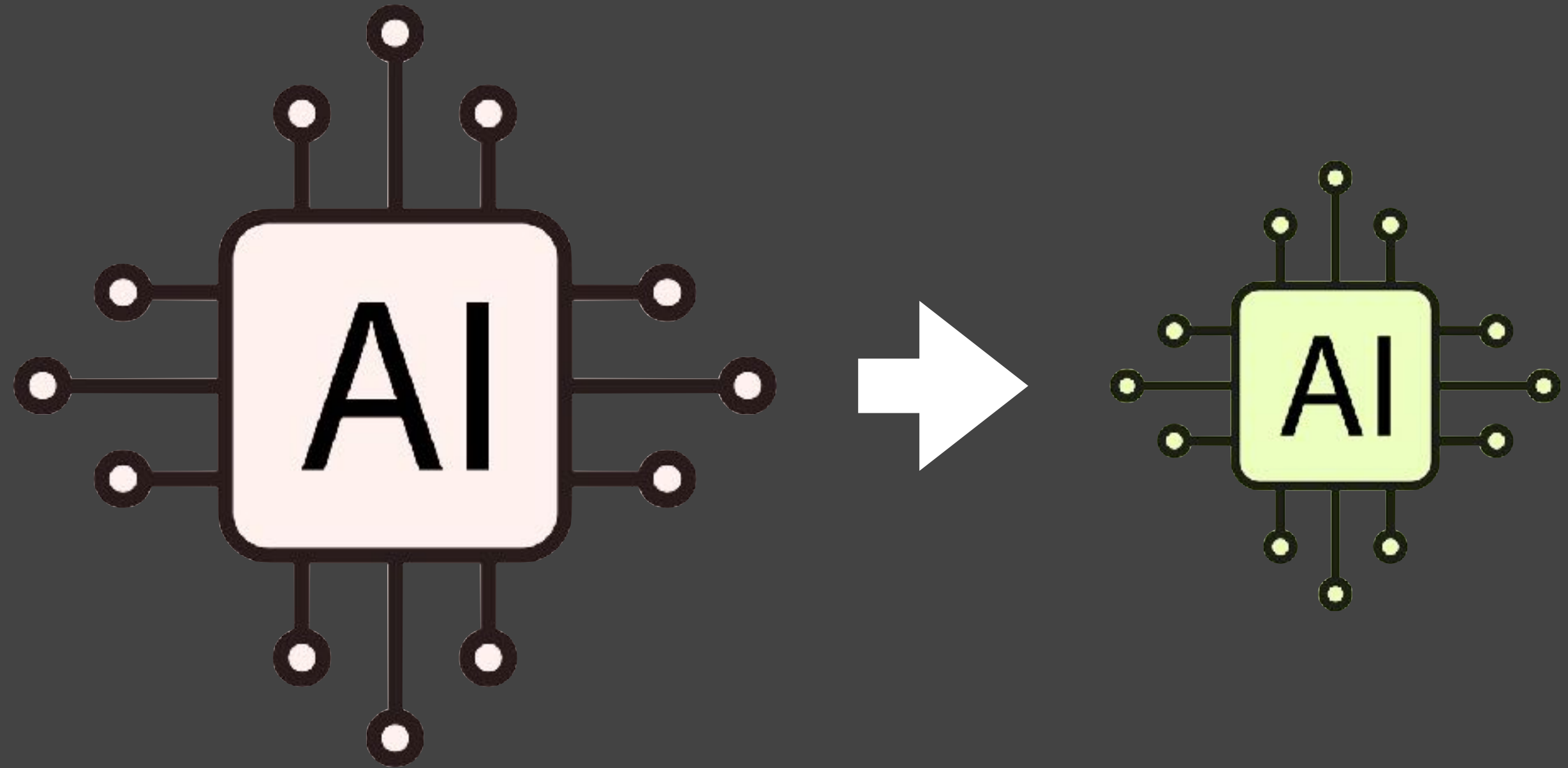


Discussion

- 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.**

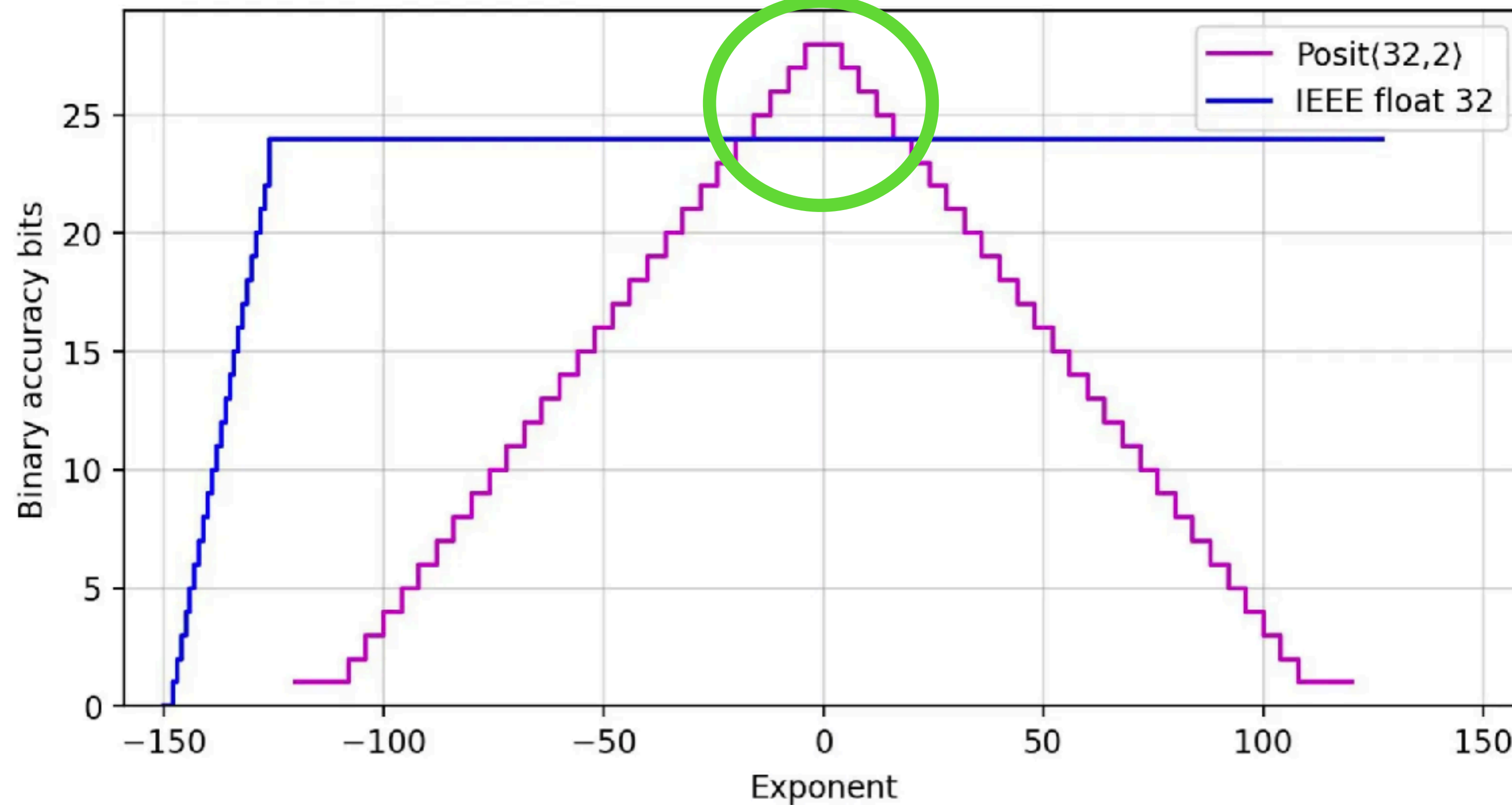
Data/Model Simplification

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



Posit vs Float

Better for DL use cases



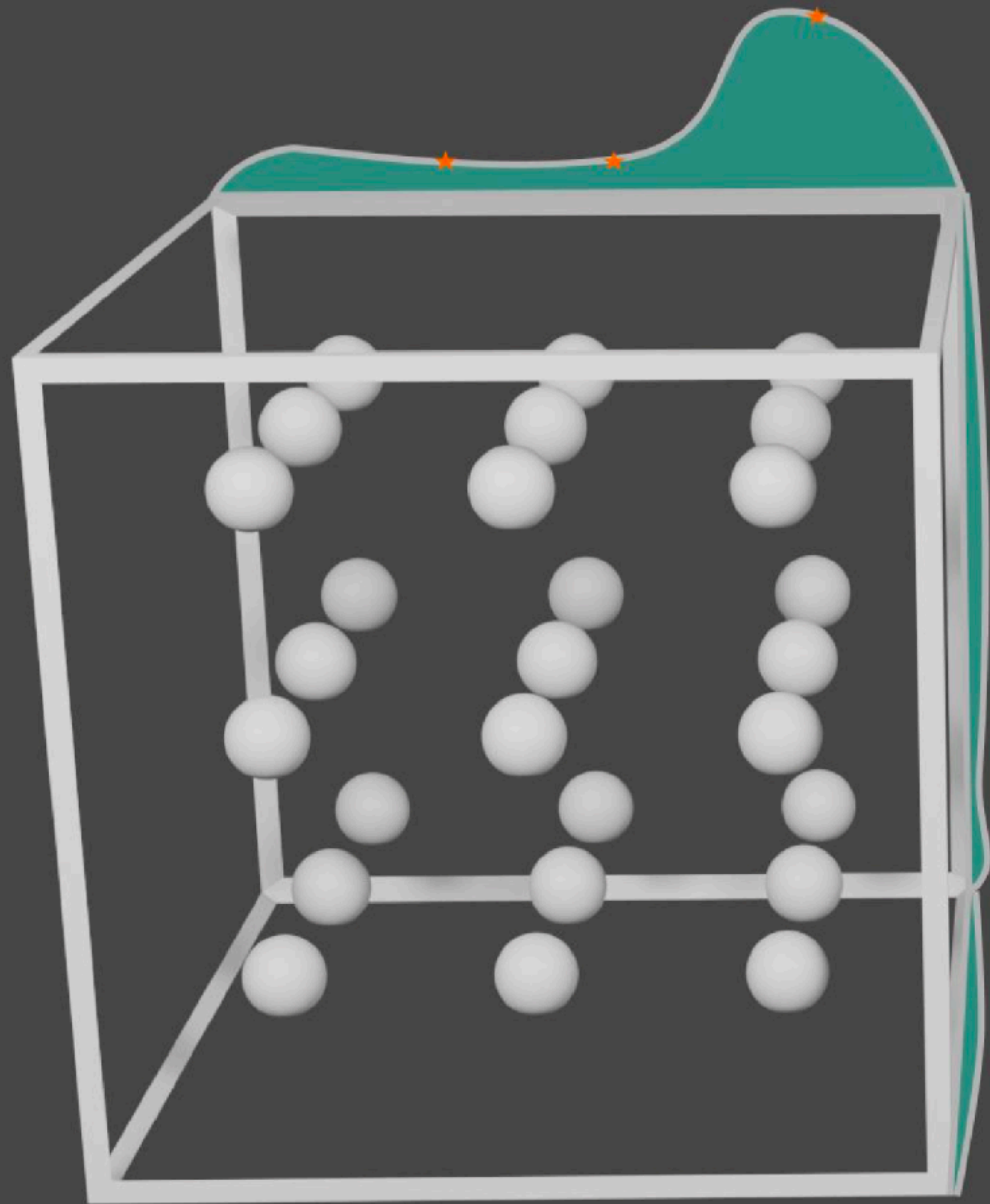
Model simplification out of the box

- 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/>

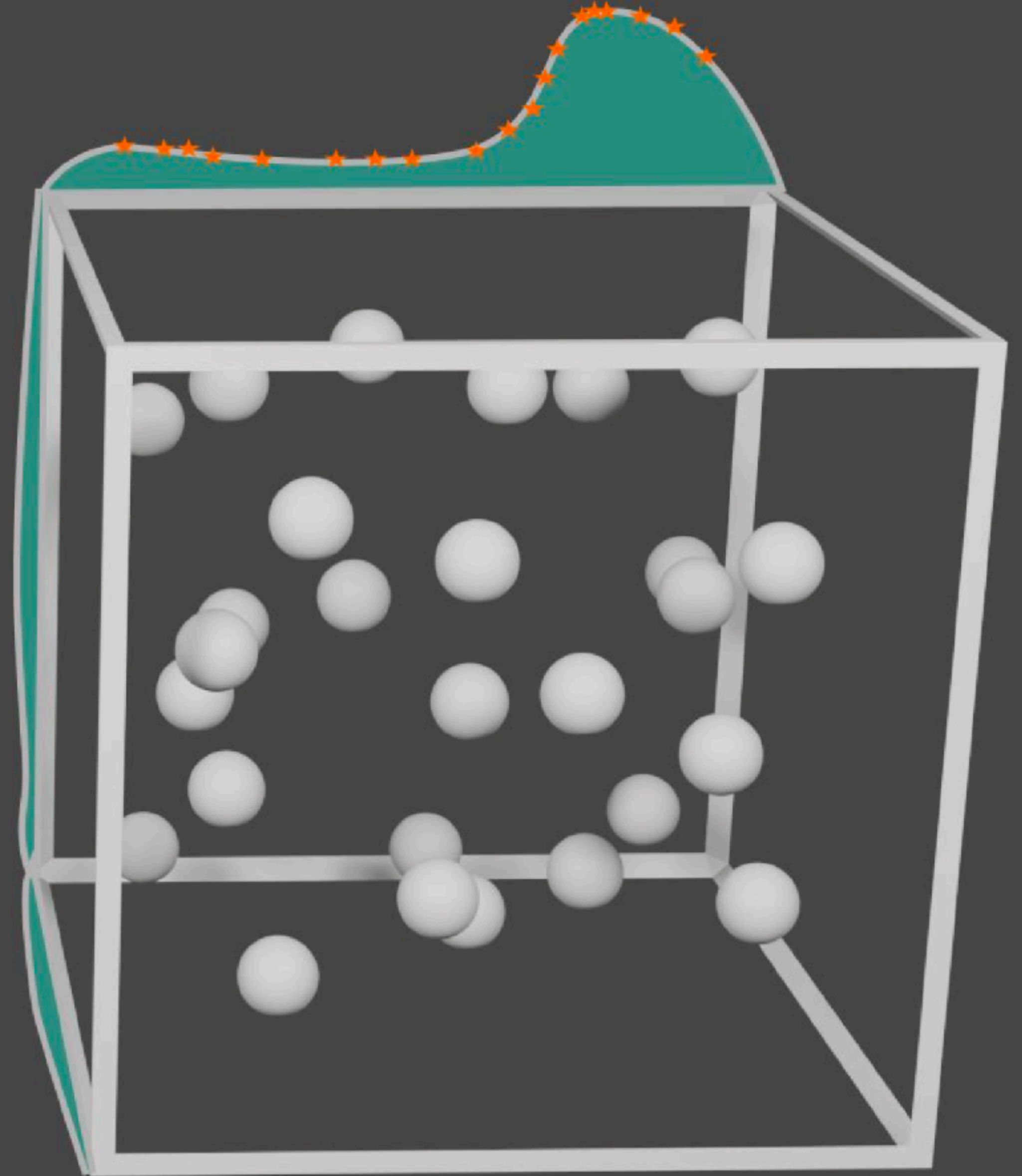
Hyper parameter tuning

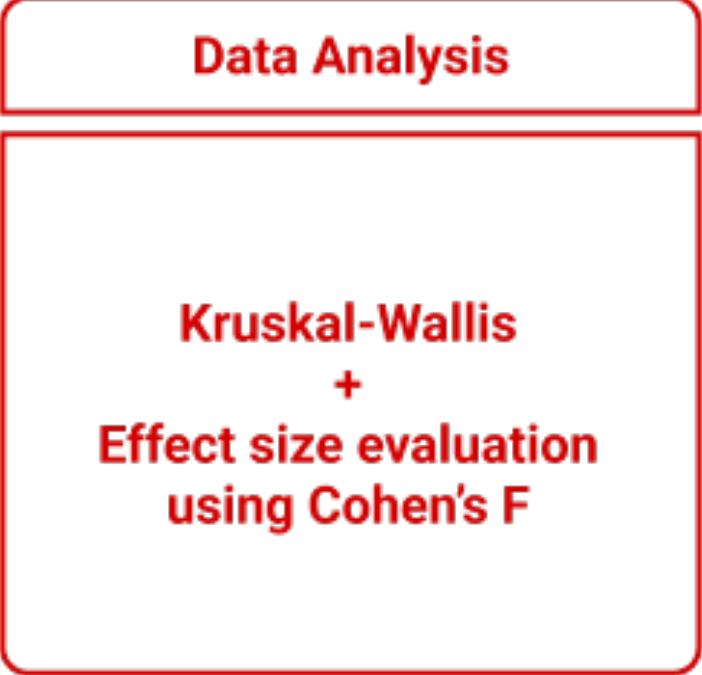
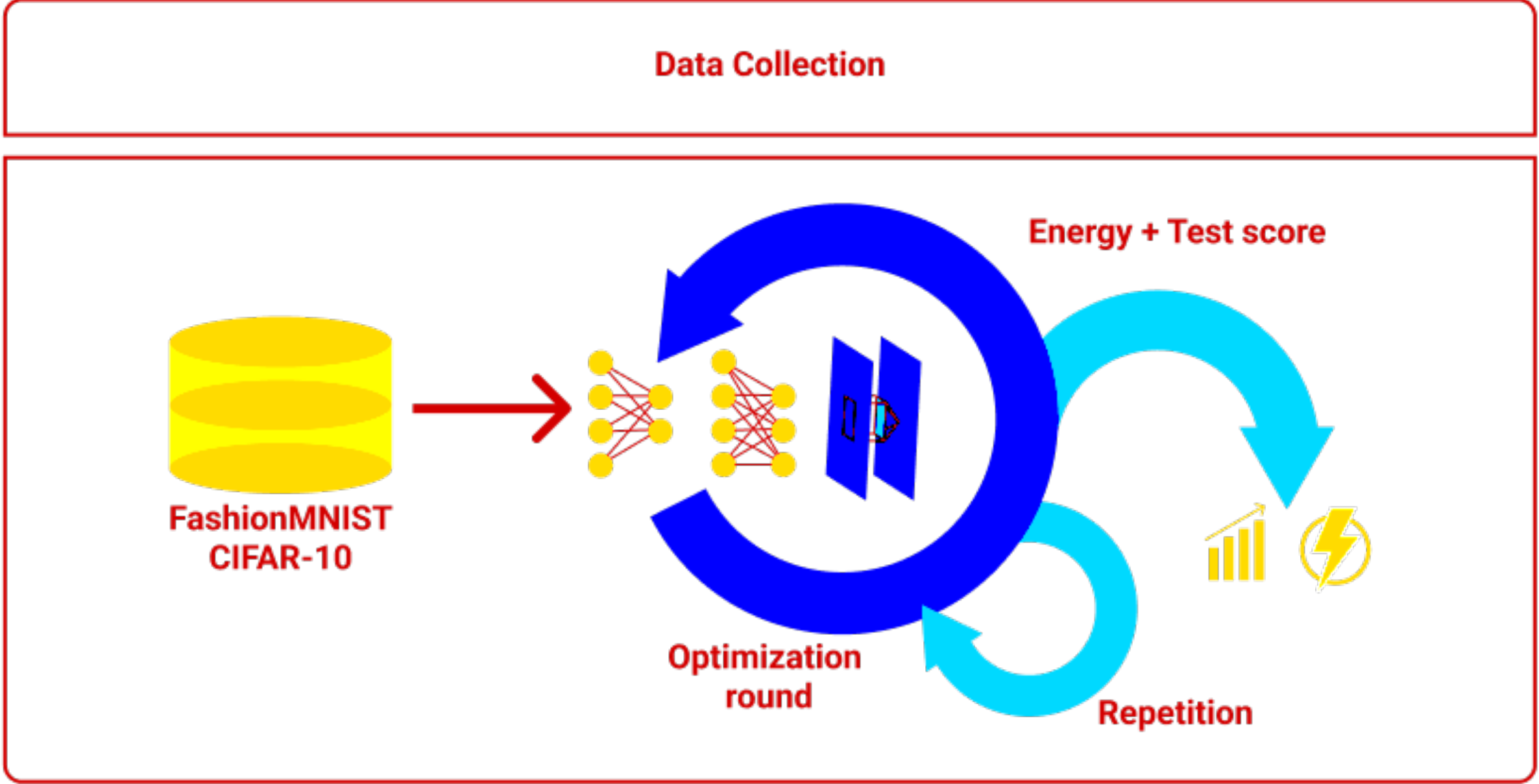
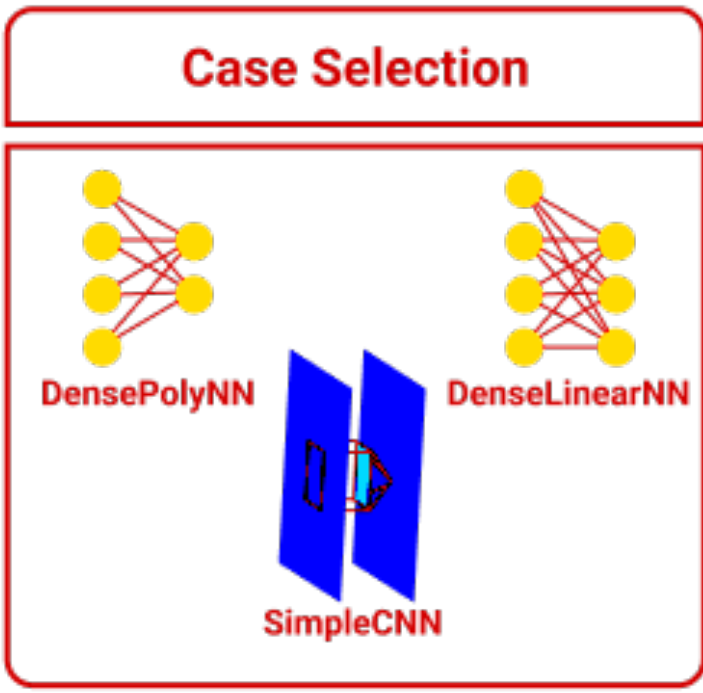
- When training an ML model, there are several **parameters** that need to be **tuned**.
 - E.g., in SVM we have the *Regularization parameter* C , the kernel function, the degree of the kernel function, and depending on the case, many other.
 - The common approach revolves around **grid search**. The 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

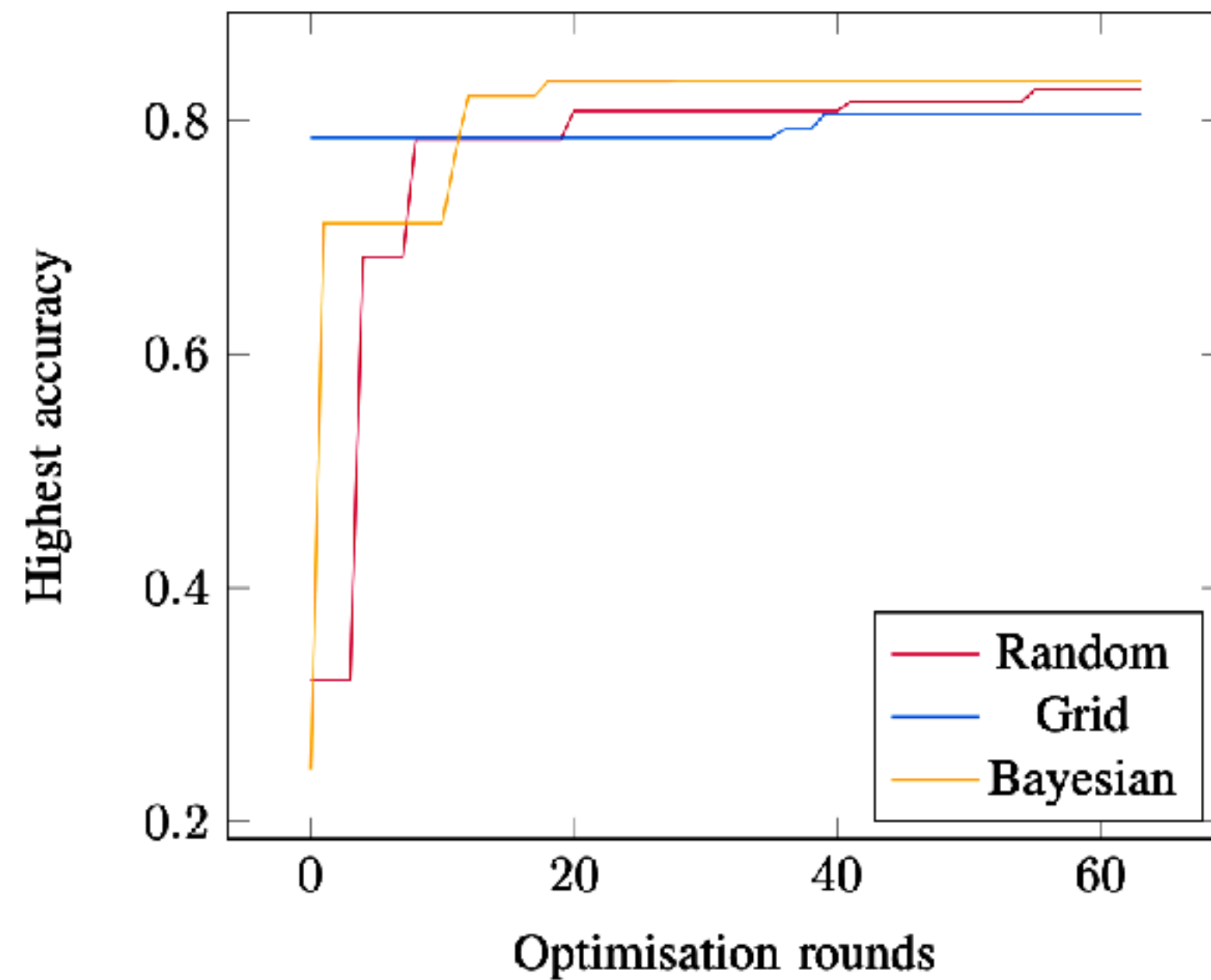




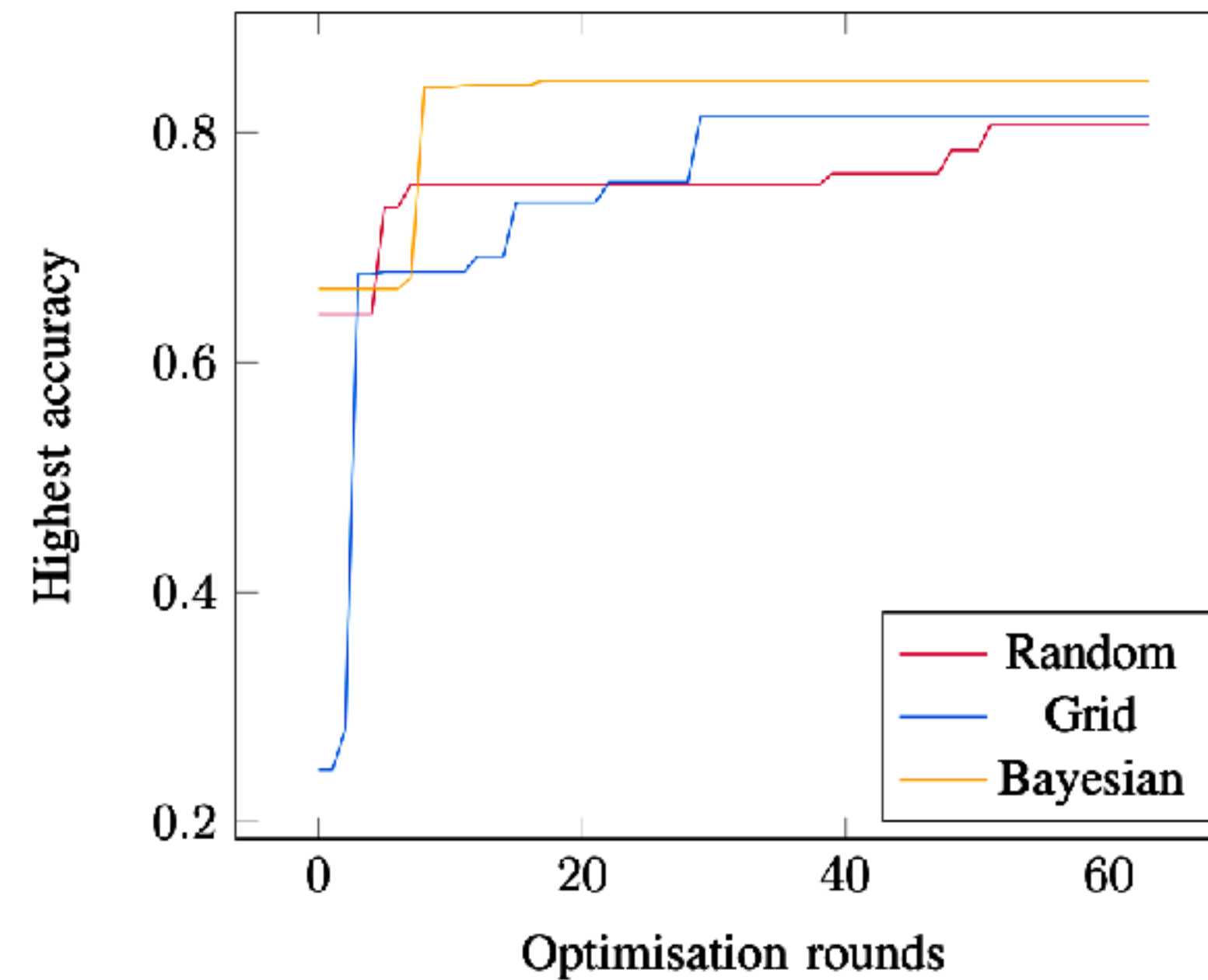
Results

Conclusions?

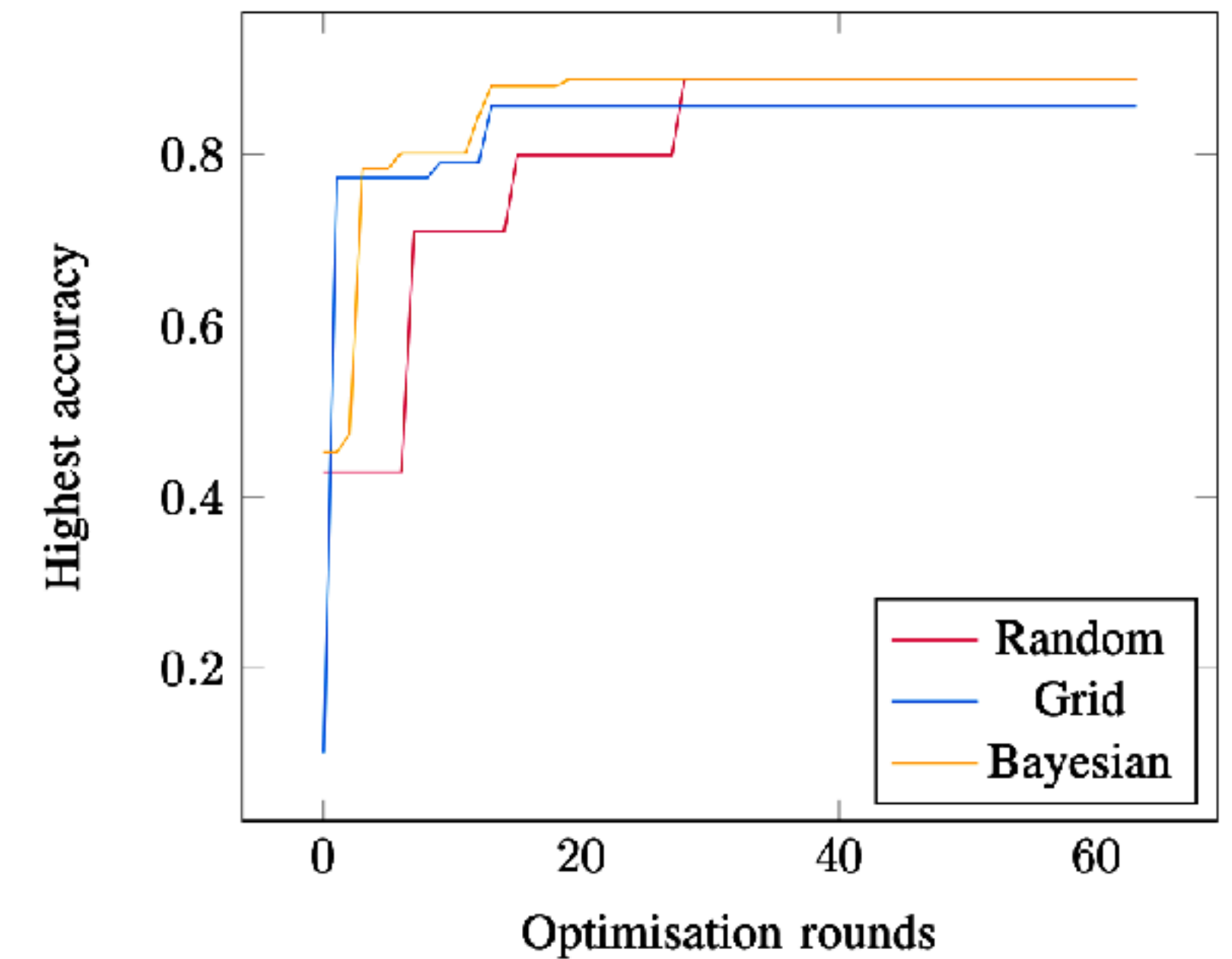
- **Bayesian** converges faster.
- No clear winner between Grid and Random



(a) DensePolyNN



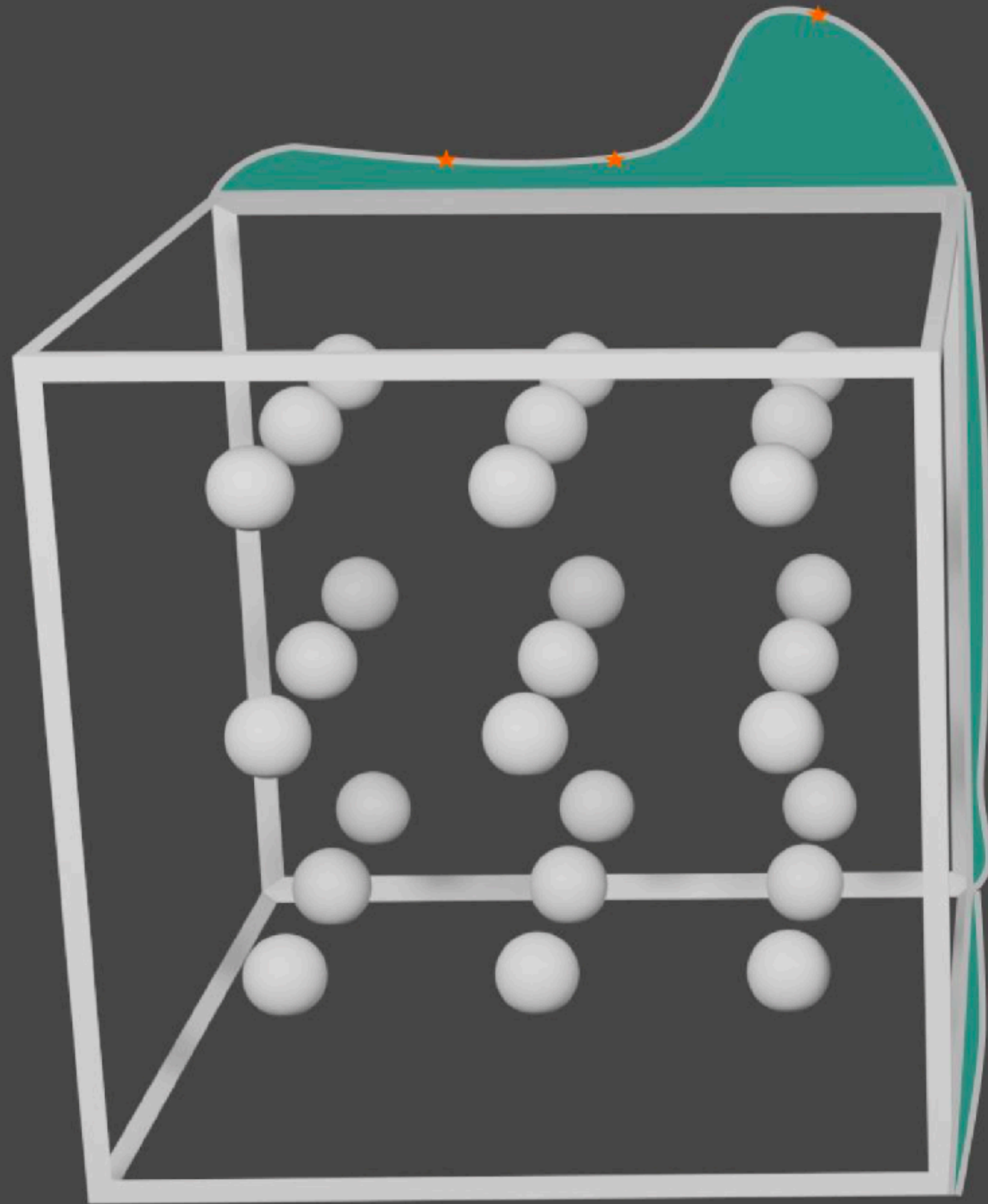
(b) DenseLinearNN



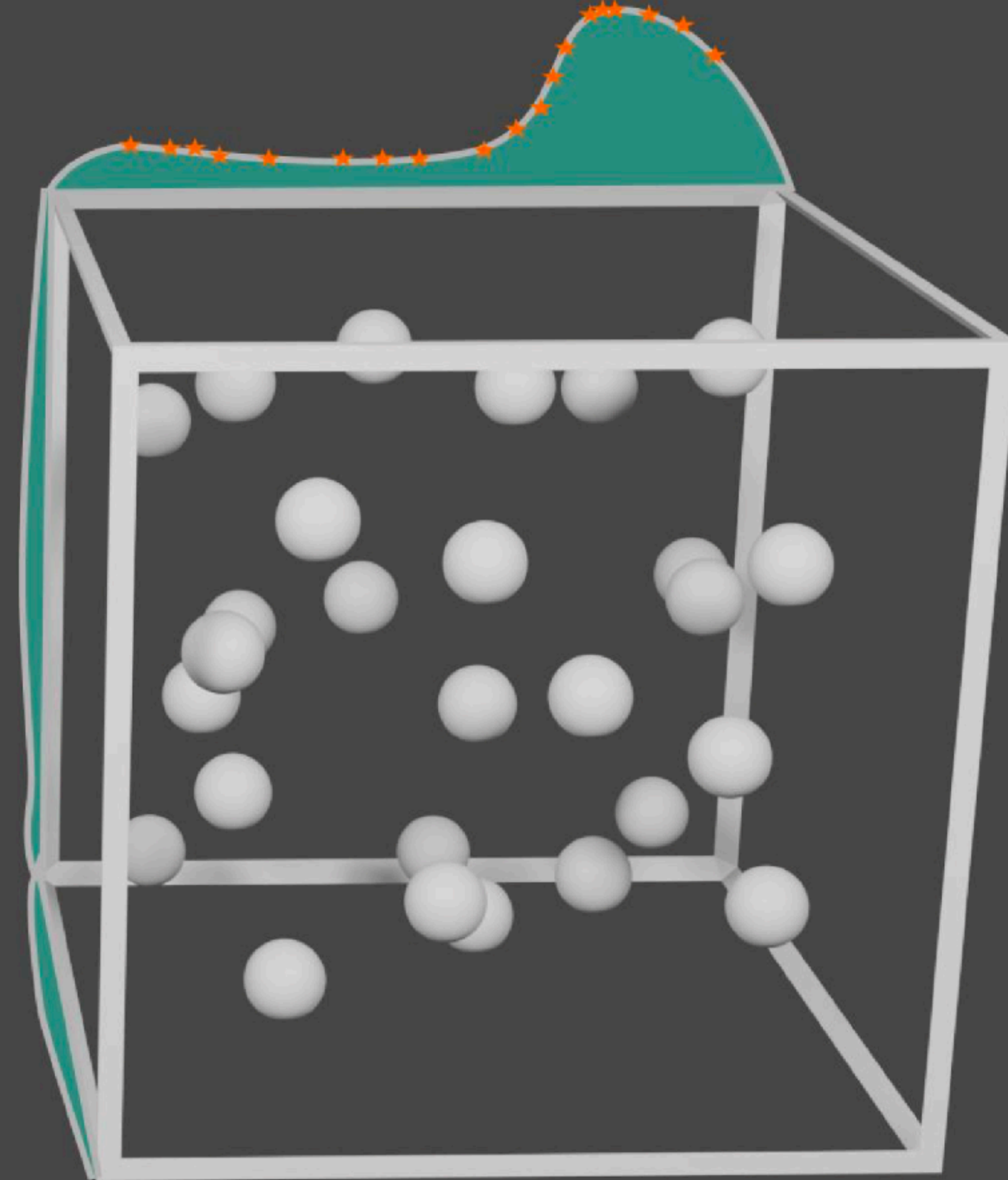
(c) SimpleCNN

Which one to choose?

Grid Search



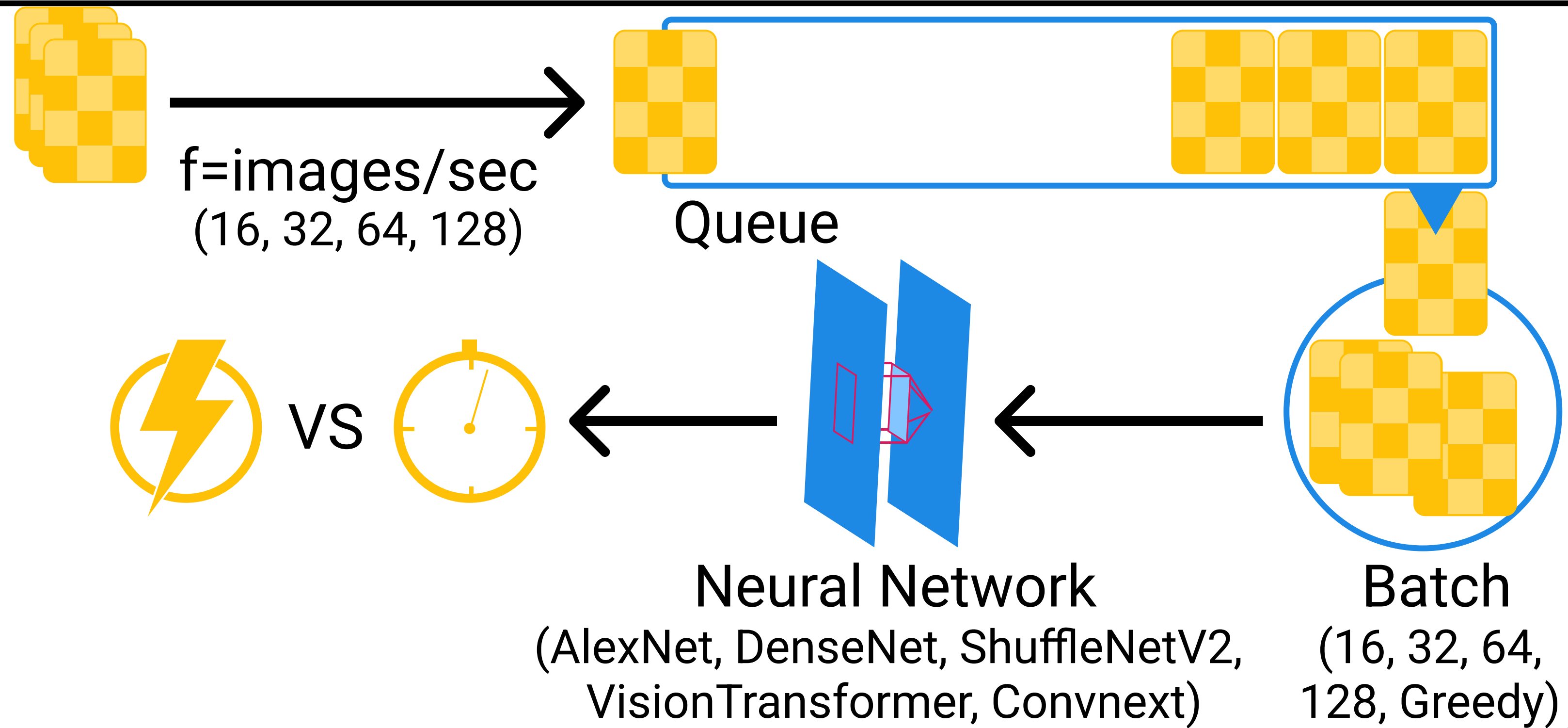
Random Search

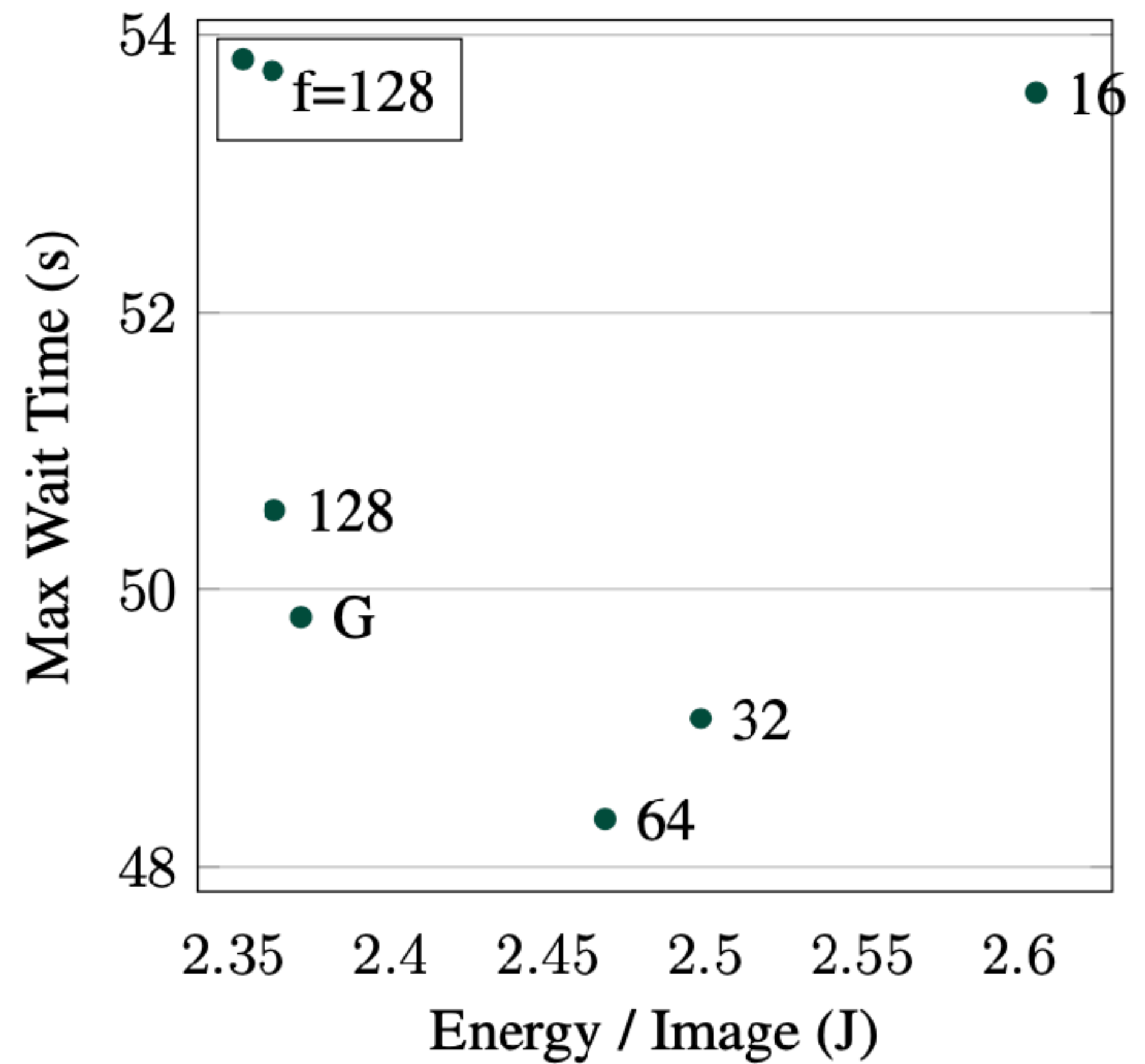
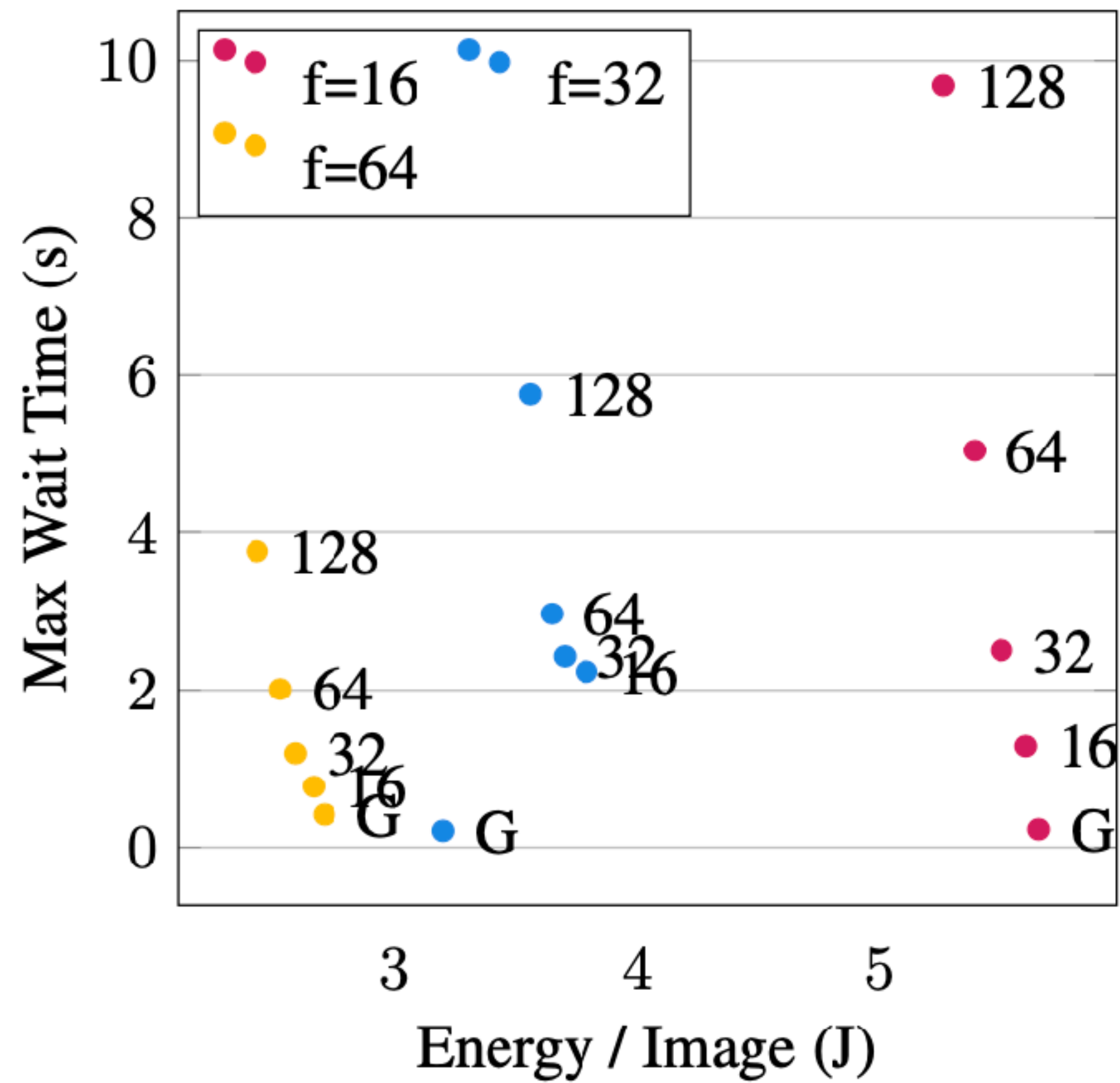


Processing inference requests

- 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/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 FacebookMeta

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

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

Sustainable AI: Environmental Implications, Challenges and Opportunities

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

Facebook AI

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 years¹.

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

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

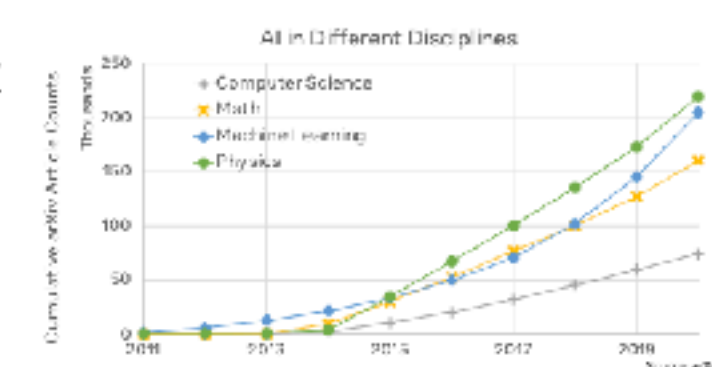


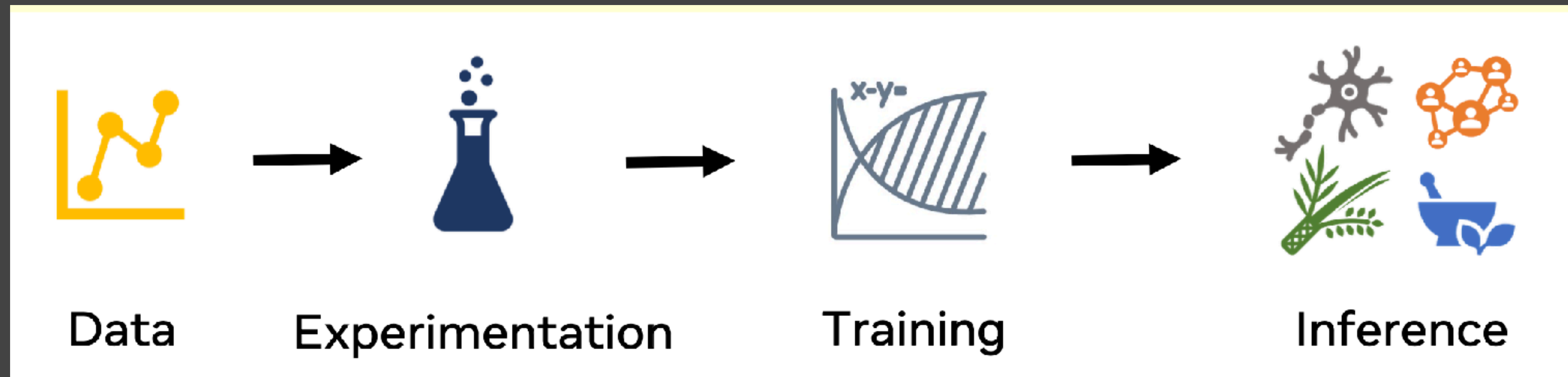
Fig. 1. The growth of AI is exceeding that of many other scientific disciplines. Significant research growth in machine learning is observed in recent years as illustrated by the increasing cumulative number of papers published in machine learning with respect to other scientific disciplines based on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

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.

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

Carbon footprint mapped to the AI lifecycle

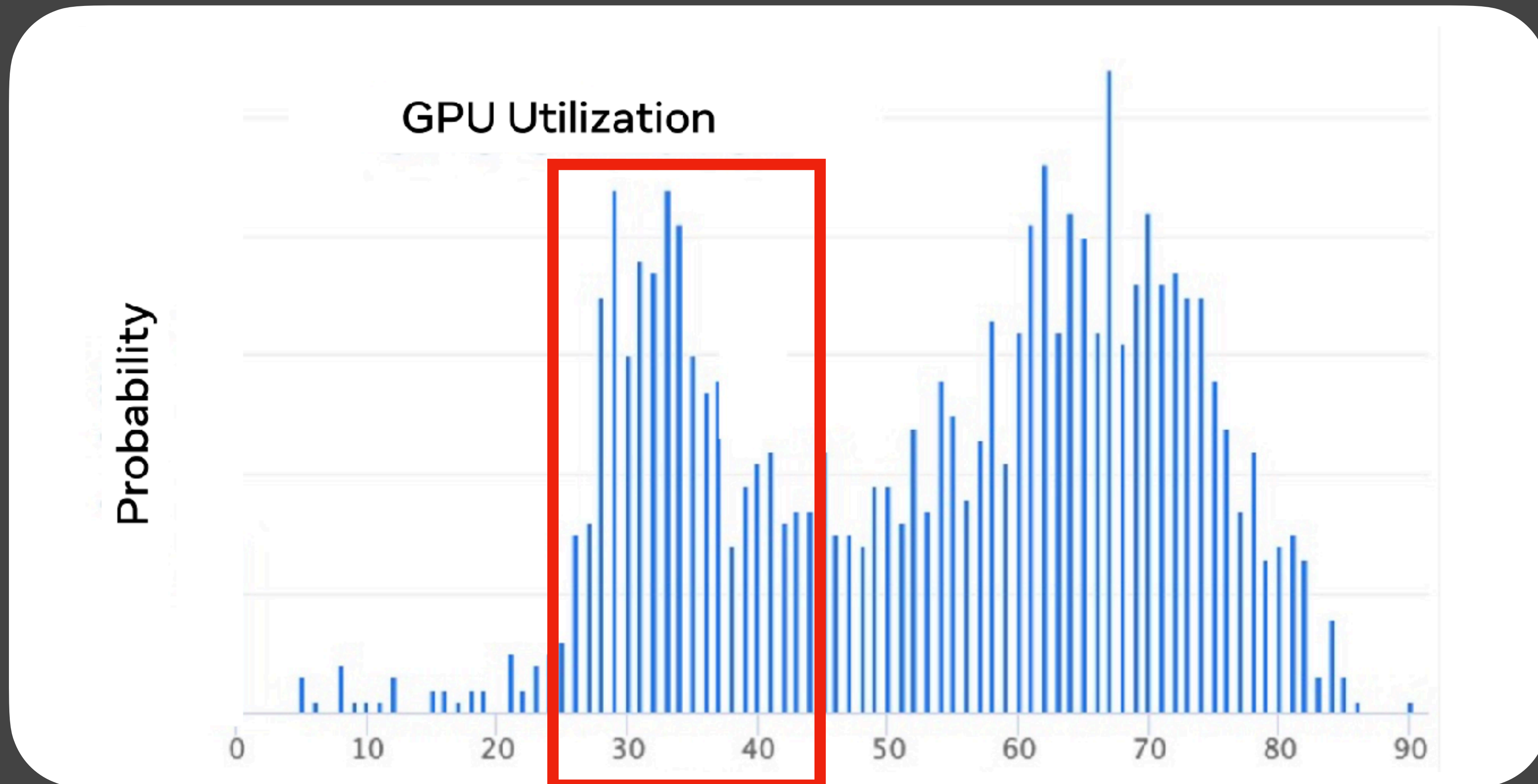


- There are 4 main overarching stages where carbon emissions need to be isolated: **data collection, experimentation, training, inference**.
- At Facebook, recommendation systems split energy consumption **evenly between training and inference**; text translation models have a **35%/65%** split. (Operational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

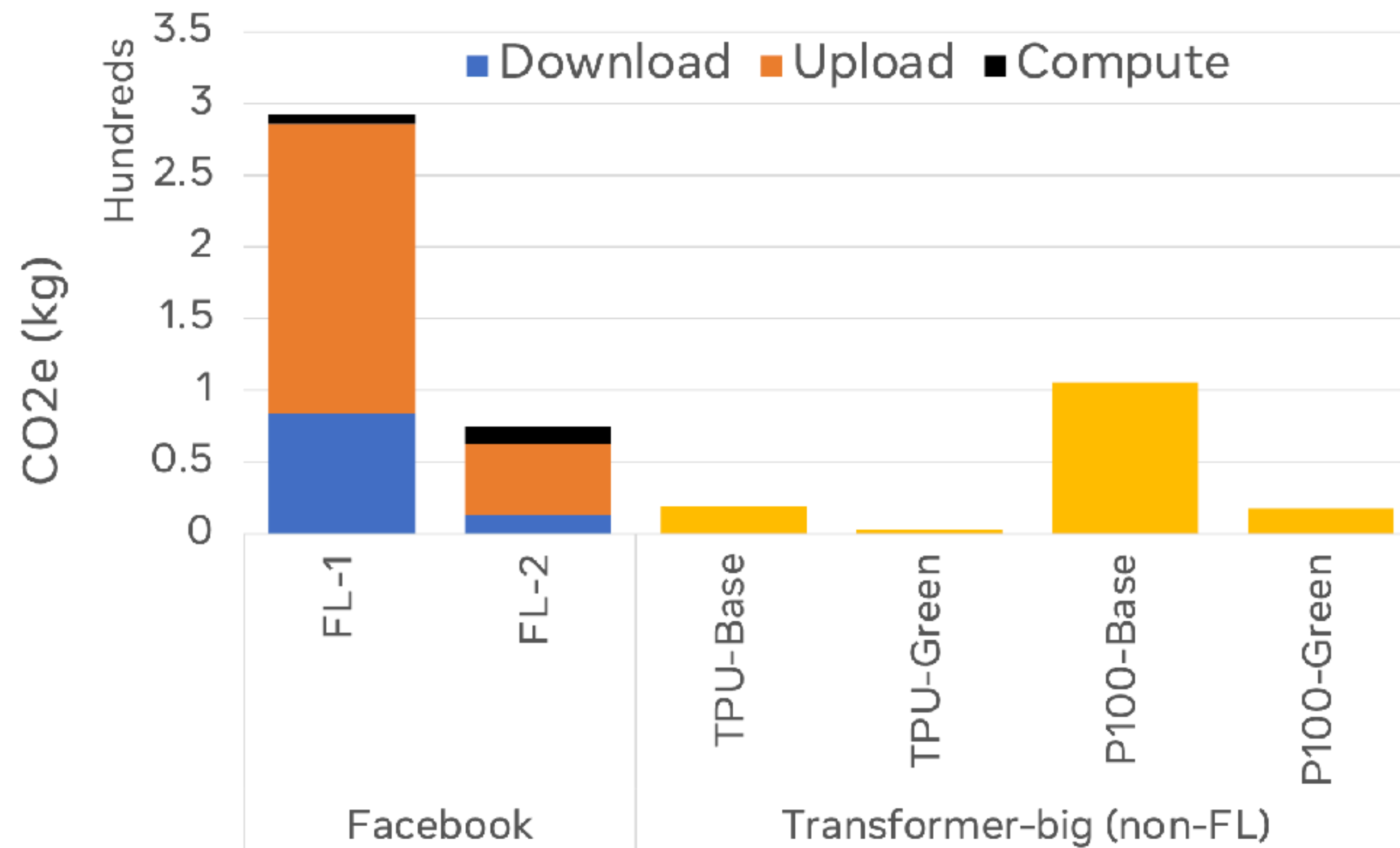


Federated learning

- 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