11. Green Al Sustainable Software Engineering **CS4295**



Luís Cruz L.Cruz@tudelft.nl



SustainableSE 2022



Overview of Green Al Green data-centric Al Green Al at Facebook Tim Yarally on Green Al

Tomorrow's class

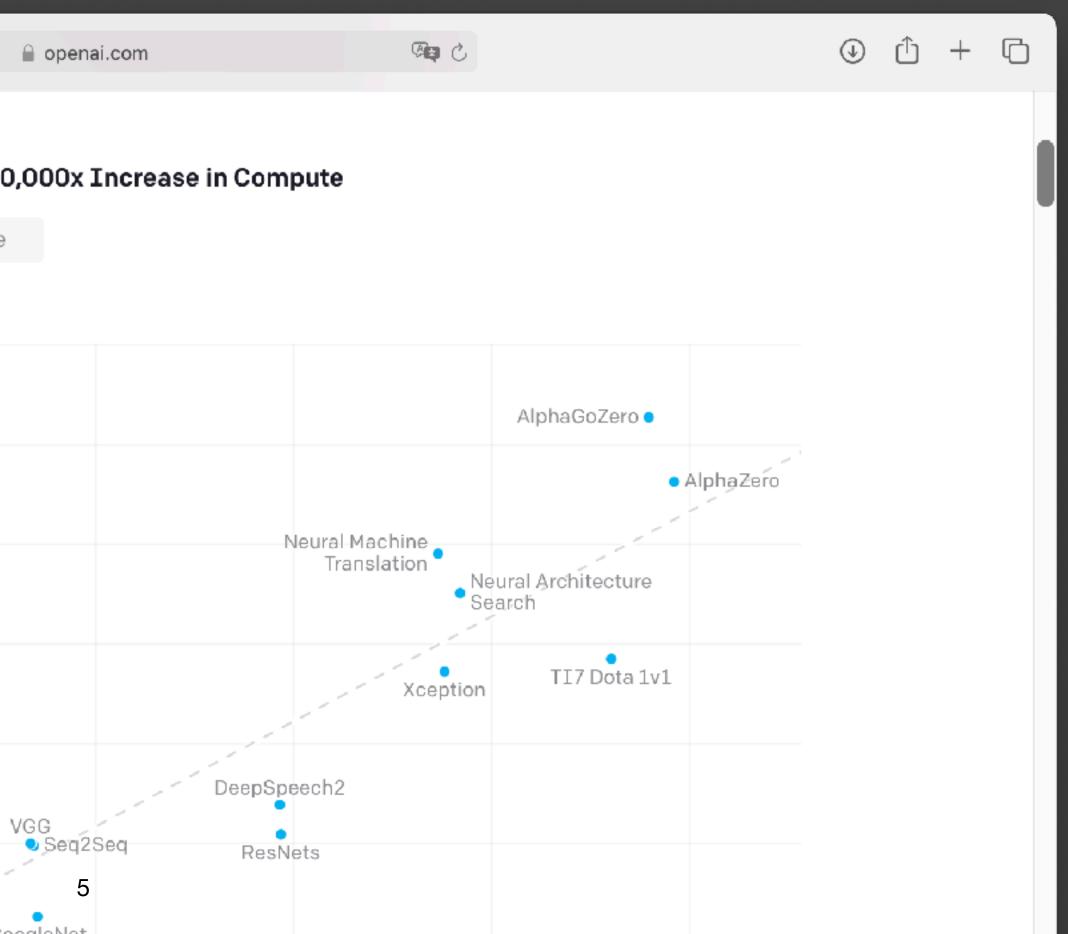
- automating tasks that typically require human intelligence.
- In the past years AI has been widely applied across different domains. Health care.
- To deploy AI systems, we test them against benchmarks (or validation sets).
 - The goal is to outperform the previous existing models.
 - E.g., in Machine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

• Artificial Intelligence (AI) is the branch of computer science that deals with

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

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

•••	=	<	>		0			
					AlexNet 1	o Alpha	Go Zero: /	A 300,
					Log S	cale	Linear	Scale
			Petaflop/s-day 1e+4	/S				
			1e+3					
			1e+2					
			1e+1					
			1e+0					
			1e-1					V
			1e-2			Visua Undersi	alizing and tanding Cor	



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

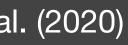
The Equation of Red A



Cost of a single (E)xample Number of (H) yperparameters Size of (D)ataset

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

By Schwartz et al. (2020)



Ssues of Red A

- 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 A

- Google's BERT-large
 - 350 million features
 - Trained for 2.5 days using 512 TPU chips, costing \$60K+
- Open-GPT3
 - 175 billion features
- AlphaGo \bullet
 - 1920 CPUs, 280 GPUs, costing \$35M



Red Al



Accuracy: 0.999999999

Green Al



- Energy
- Time
- Reproducibility
- Reusage

Research on Green Al

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

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

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:

Eanily 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 Foirness, Accountability, and Trans*- Timnit Gebru* timnit@blackinai.org Black in AI Palo Alto, CA, USA

Shmargaret Shmitchell shmargaret.shmitchell@gmail.com The Aether

alone, we have seen the emergence of BERT and its variants [39, 70, 74, 113, 146], GPT-2 [106], T-NLG [112], GPT-3 [25], and most recently Switch-C [43], with institutions seeningly competing to produce ever larger LMs. While investigating properties of LMs and how they change with size holds scientific interest, and large LMs have shown improvements on various tasks (§2), we ask whether enough thought has been put into the potential risks associated with developing them and strategies to mitigate these risks.

We first consider environmental risks. Echoing a line of recent work outlining the environmental and financial costs of deep learning systems [129], we encourage the research community to prioritize these impacts. One way this can be done is by reporting costs and evaluating works based on the amount of resources they consume [57]. As we outline in §3, increasing the environmental and financial costs of these models doubly punishes marginalized communities that are least likely to benefit from the progress achieved by large LMs and most likely to be harmed by negative environmental consequences of its resource consumption. At the scale we are discussing (outlined in §2), the first consideration should be the environmental cost.

Just as environmental impact scales with model size, so does the difficulty of understanding what is in the training data. In §4, we discuss how large datasets based on texts from the Internet overrepresent hegemonic viewpoints and encode biases potentially damaging to marginalized populations. In collecting ever larger datasets we risk incurring documentation debt. We recommend mitigating these risks by budgeting for curation and documentation

contributed articles

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.

EN ETZIONI

BY ROY SCHWARTZ, JESSE DODGE,



of artificial intelligence (AI) has progress on a broad range of object recognition, game playing, nd machine translation.43 Much of achieved by increasingly large intensive deep learning models.^a from Amodei et al.,² plots training e for state-of-the-art deep learning AlexNet in 2012²⁴ to AlphaZero in vs an overall increase of 300,000x, bling every few months. An even bserved in NLP word-embedding at ELMo34 followed by BERT,8 Megatron-LM,42 T5,36 and GPT-3.4 has estimated the carbon LP models and argued this trend is unfriendly and prohibitively riers to participation in NLP such work as **Red A**I.

this article, but our focus is on AI research that relies on deep

This trend is driven by the strong focus of the AI community on obtaining "state-of-the-art" results,⁶ as exemplified by the popularity of leaderboards,^{53,54} whichtypically report accuracy (or other similar measures) but omit any mertion of cost or efficiency (see, for example, leaderboards.allenai.org).^c Despite the clear benefits of improving model accuracy, the focus on this sirgle 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 irspired undergraduate with a laptop has the opportunity to write highquality 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.

- b 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.stanforc.edu/benchmark/).

key insights

- The computational costs of state-ef-theart 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 seademics, students, and researchers to engage in deep learning research.
- An alternative is Green AI, which treats efficiency as a primary evaluation criterion alonside 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 footprirt and increase its inclusivity.

Energy and Policy Considerations for Deep Learning in NLP

Emma StrubellAnanya GaneshAndrew McCallumCollege of Information and Computer SciencesUniversity 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.

Introduction

dvances in techniques and hardware for traing deep neural networks have recently enoled impressive accuracy improvements across any fundamental NLP tasks (Bahdanau et al., 115; Luong et al., 2015; Dozat and Manng, 2017; Vaswani et al., 2017), with the ost computationally-hungry models obtaining e highest scores (Peters et al., 2018; Devlin et al., 019; Radford et al., 2019; So et al., 2019). As result, training a state-of-the-art model now renires substantial computational resources which emand considerable energy, along with the assciated financial and environmental costs. Re-

Consumption	CO ₂ e (lbs)		
Air travel, 1 passenger, $NY \leftrightarrow SF$	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)

NLP pipeline (parsing, SRL)	39
w/tuning & experimentation	78,468
Transformer (big)	192
w/ neural architecture search	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-neural 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

CEMBER 2020 | VOL. 63 | NO. 12



How can we adopt Green A

- 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**.
- AllenAl leaderboard https://leaderboard.allenai.org
 - No carbon metrics, yet

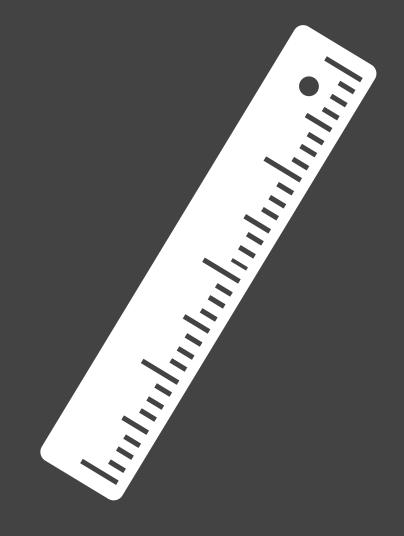
- Report comparable proxies for energy lacksquareconsumption.
 - Learning algorithms behave in a nondeterministic
 - **!** Different data-points lead to different energy consumption

•••		O (≣ leaderboa	ard.allenai.org	ිම එ		Œ) ů +
Green AL.	<mark>2</mark> [1906.02	Communi	🥥 https://dl	🖓 Start Page	O google-re	🔀 https://ar	M Join conv	🚍 Sponsor I	🔥 Sub
Ai2 Allen	Institute for Al								
Le	eaderboards	6							Log In
Abductive NU	Abductive Natura				ichmark dataset des ons for a given set o	_		y to	te Submiss
Public Sub	missions Getting Sta	arted About						Accu	racy: 0.92
Rank 🕈	Submission					Crea	ted \$		Downloa cy
1	DeBERTa+2-hop Con USC ISI MOWGLI/INK I					01/26	/2022	0.919	97
2	Abductive Reasoning ISCAS	Cycle				12/16	/2021	0.918	37
3	DeBERTa-V2-XXL-MM Anonymous	NLI				01/19	/2022	0.917	71
A Display a menu	UNIMO UNIMO Team, Baidu Ni	LP				05/15	/2021	0.911	18



Reporting energy/carbon footprint

- Reporting measured energy consumption
 - + Accurate
 - + Easy to map to carbon emissions
 - - Hard to measure
- 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 energy consumption in memory
 - does not reflect carbon emissions



Data-centric Al

Data-centric A

- systems.

 - datasets.

 More about data-centric AI by Andrew Ng: https://www.youtube.com/watch?v=06-AZXmwHjo

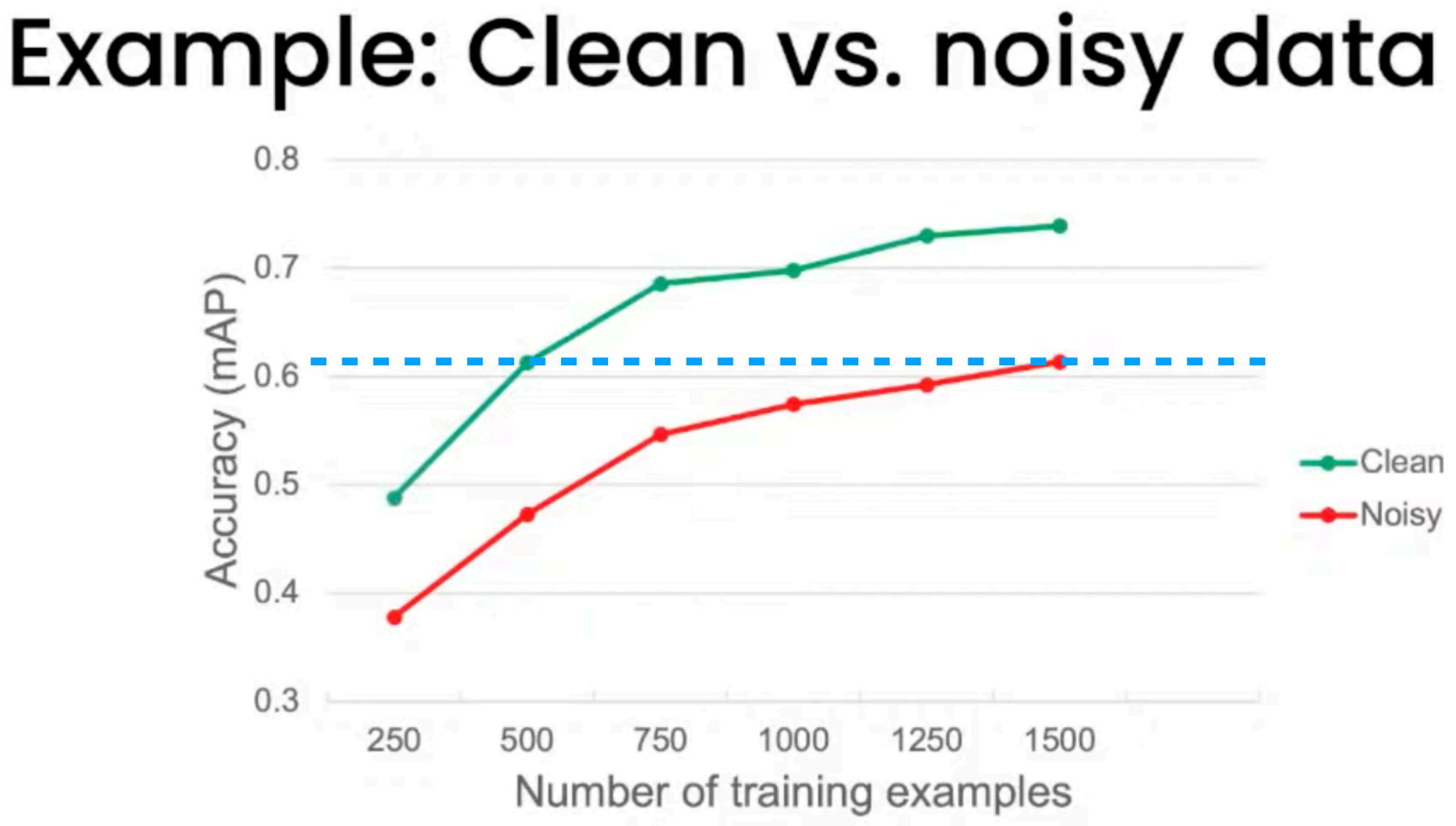
• Emerging discipline that deals with systematically engineering data to build Al

• Shift from improving the training strategy to improving the data.

• It is better to have small but reliable datasets than large but noisy

• => Improve data collection, data labelling, and data preprocessing.







Green Data-centric A

- 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

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 23 to improve AI energy efficiency. Our results call for a research agenda that focuses on data-centric techniques, to further enable 25 and democratize Green-AL

Index Terms-Energy Efficiency, Artificial Intelligence, Green 27 AI, Data-centric, Empirical Experiment

I. INTRODUCTION

We live in the era of artificial intelligence (AI): new intelli-30 gent technologies are emerging every day to change people's 21 lives. Many organizations identified the massive potential of 12 using intelligent solutions to create business value. Hence, in 20 the past years, the modus operandi is collecting as much data ы as possible so that no opportunity is missed. Data science 36 teams are constantly looking for problems where AI can 38 be applied to existing data to train models that can provide 37 more personalized and optimized solutions to their operations 38 customers and operations [1].

39 Nevertheless, the energy consumption of developing AI ap-40 plications is starting to be a concern. Previous studies observed 41 that AI-related tasks are particularly energy-greedy [2], [3]. In 42 fact, since 2012, the amount of computing used for AI training 43 has been doubling every 3.4 months [4]. Hence, a new sub-44 field is emerging to make the development and application of 45 AI technologies environmentally sustainable: Green AI [5].

46 On a related note, AI practitioners have realised that the e current trend of collecting massive amounts of data is not

Abstract-With the growing availability of large-scale datasets, necessarily yielding better models. Being able to collect highquality data is more important than collecting big data - a 49 trend coined as Data-centric Al¹ Instead of creating learning 50 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 mpact 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.

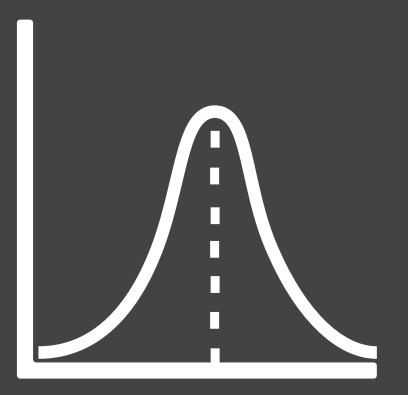
Method

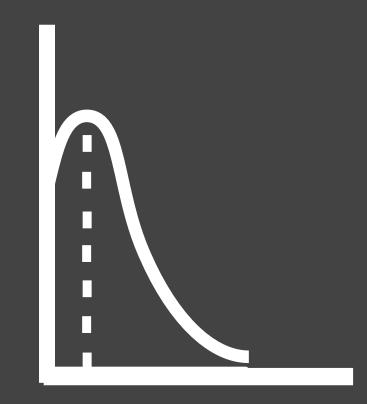
- AdaBoost, Bagging Classifier.
- Reduce the number of rows. 10%, 20%, ..., 100%
 - Stratified random sampling (?)
- Reduce the number of features. 10%, 20%, ..., 100%
 - Feature importance metric based on the Chi-Square Test (Chi2)
- Estimate energy consumption using a RAPL-based tool. (?)

• Single object of study: natural language model to detect spam messages.

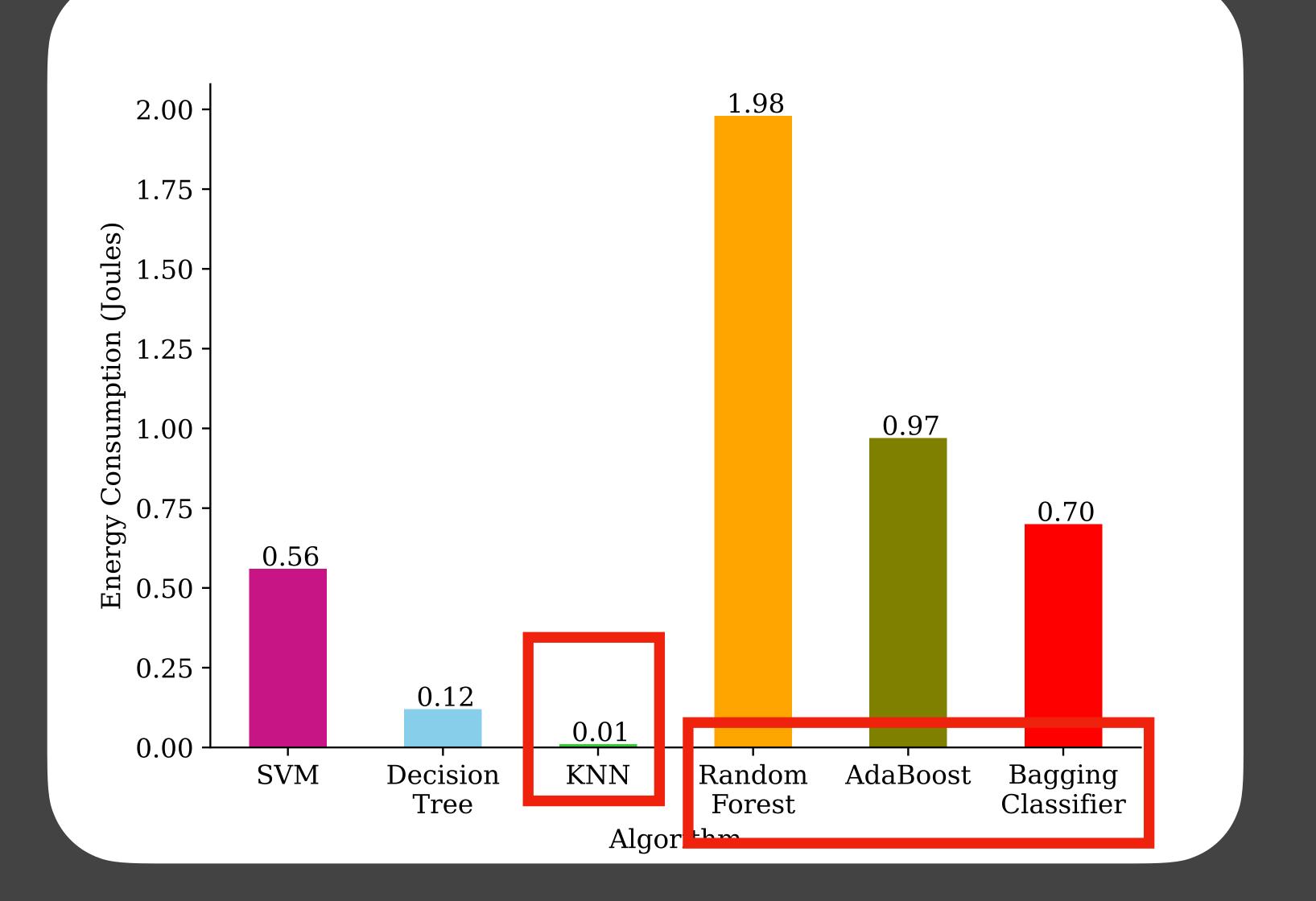
6 machine learning algorithms: SVM, Decision Tree, KNN, Random Forrest,

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

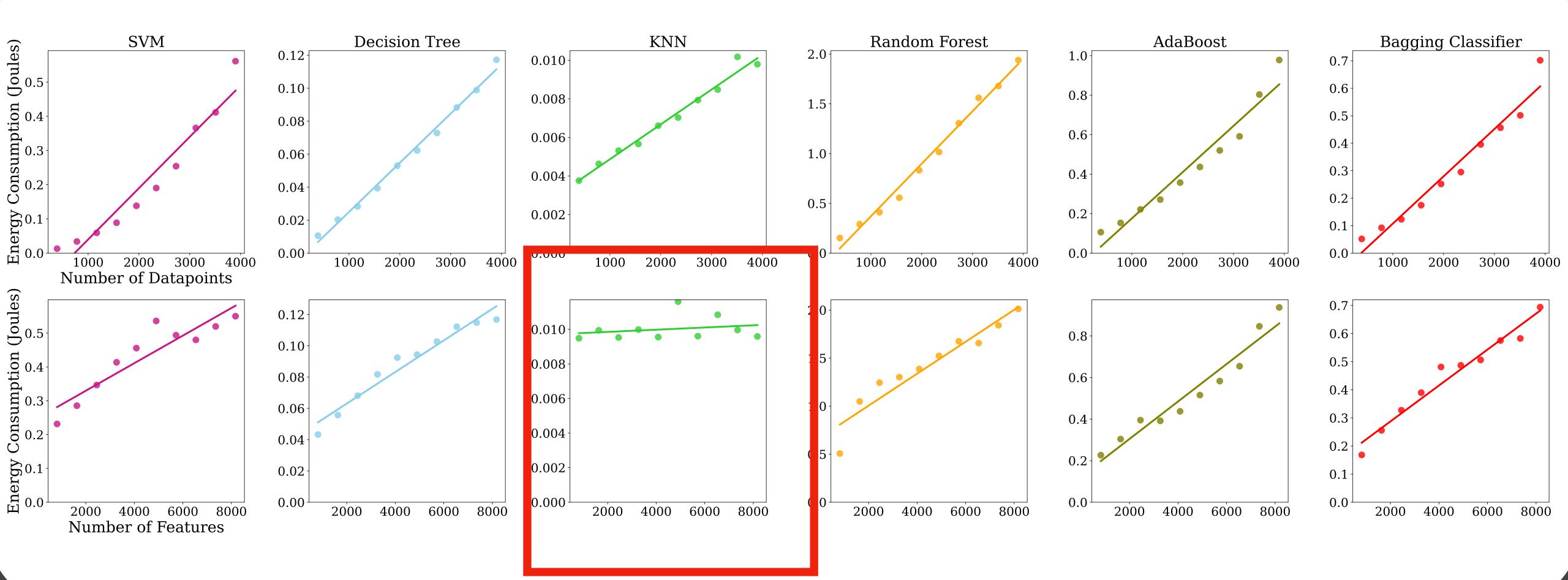




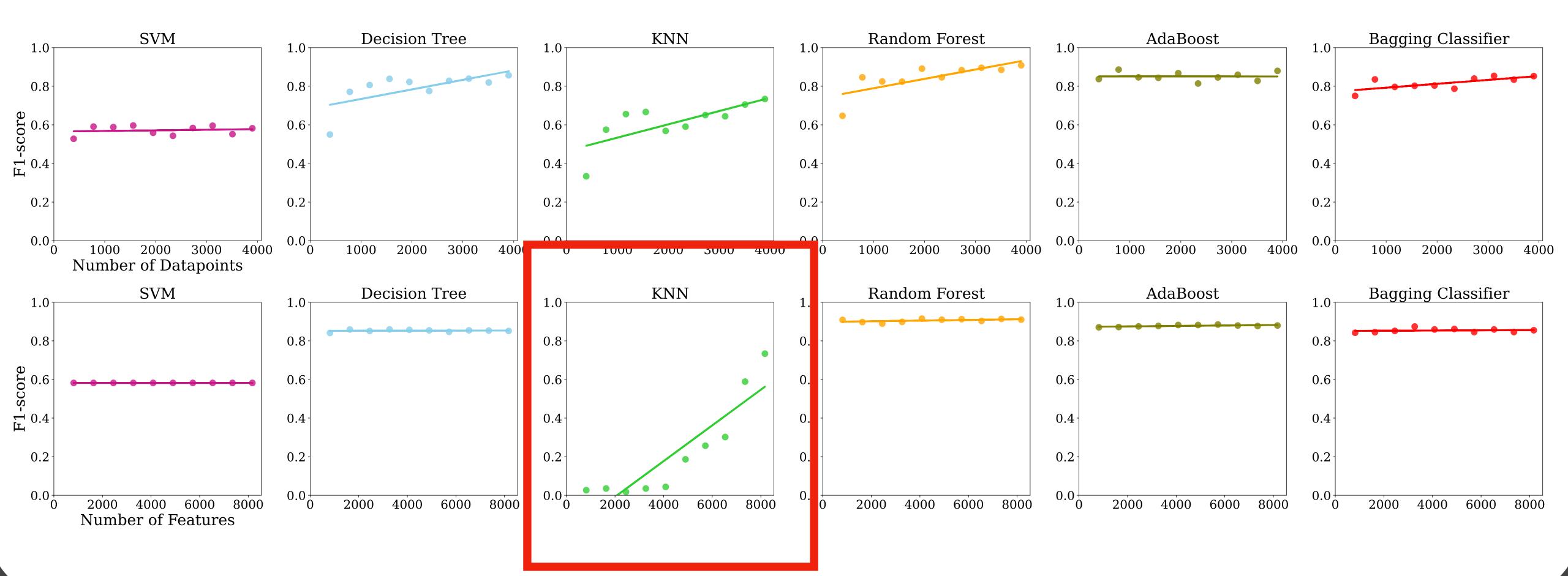
Results: energy consumption of algorithms



Results: energy vs data shape



Results: performance vs data shape



Discussion

- Other data properties should be investigated. \bullet
 - E.g., data types
- \bullet hindering model performance.
- There is a big opportunity in:
 - Model destillation. Distill the knowledge of a complex model into a simpler one.
 - properties)
 - yields similar models)

Reporting energy data is essential. It can lead to different model selection without

• Core set extraction. (extracting the smallest subset that keeps the 623 key dataset

Dataset destillation. (Derive/distill a smaller dataset into a synthetic dataset that

Green Al at Facebook

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

202 6 Ċ \sim .00364v/ 2 Ň

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

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.

Al 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 $10 \times$ into hundreds of billions of dollars [10]. All of these investments

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

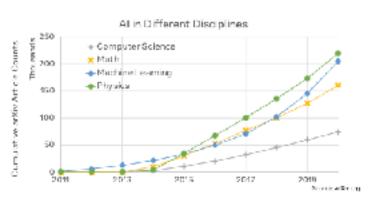


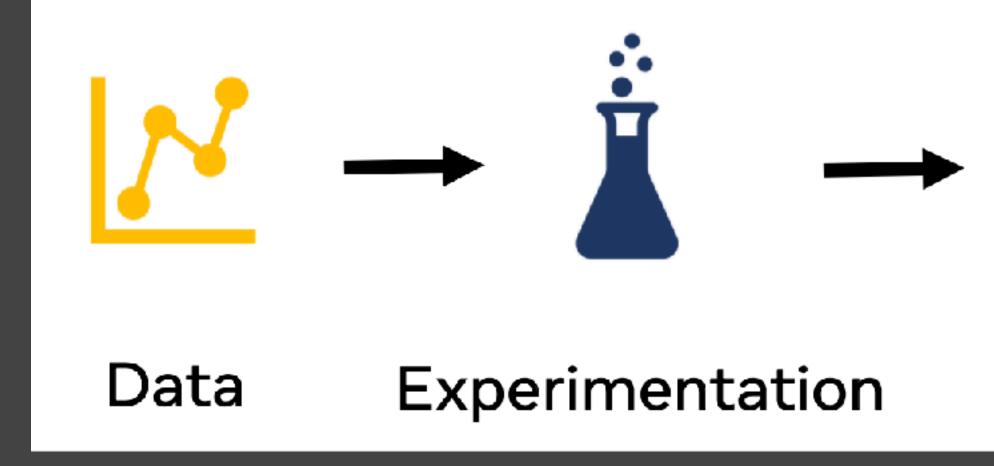
Fig. 1. The growth of ML is exceeding that of many other scientific disciplines. Significant research growth in machine learning is observed in recent years as illustrated by the increasing cumulative number of papers published in machine. learning with respect to other scientific disciplines based on the monthly count. $(\mathbf{v}, \mathbf{ax})$ is measures the cumulative number of articles on \mathbf{arX} iv).

in research, development, and deployment have led to a superlinear 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.

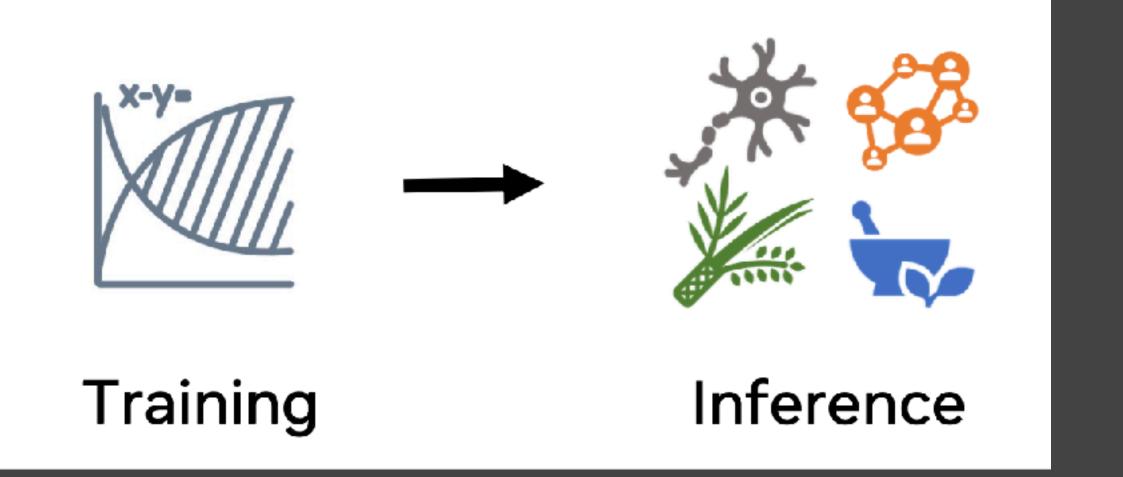
This work explores the environmental impact of AI from a holistic perspective. More specifically, we present the challenges and opportunities to designing sustainable AI computing across the key phases of the machine learning (ML) development process — Data, Experimentation, Training, and Inference — 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 ondevice computing. Given particular use cases, we consider the impact of AI data, algorithms, and system hardware. Finaly, 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 $2.4 \times$ and $1.9 \times$ 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



- collection, experimentation, training, inference.
- At Facebook, recommendation systems split energy consumption evenly between
- Operational/embodied cost split: **30%/70%**



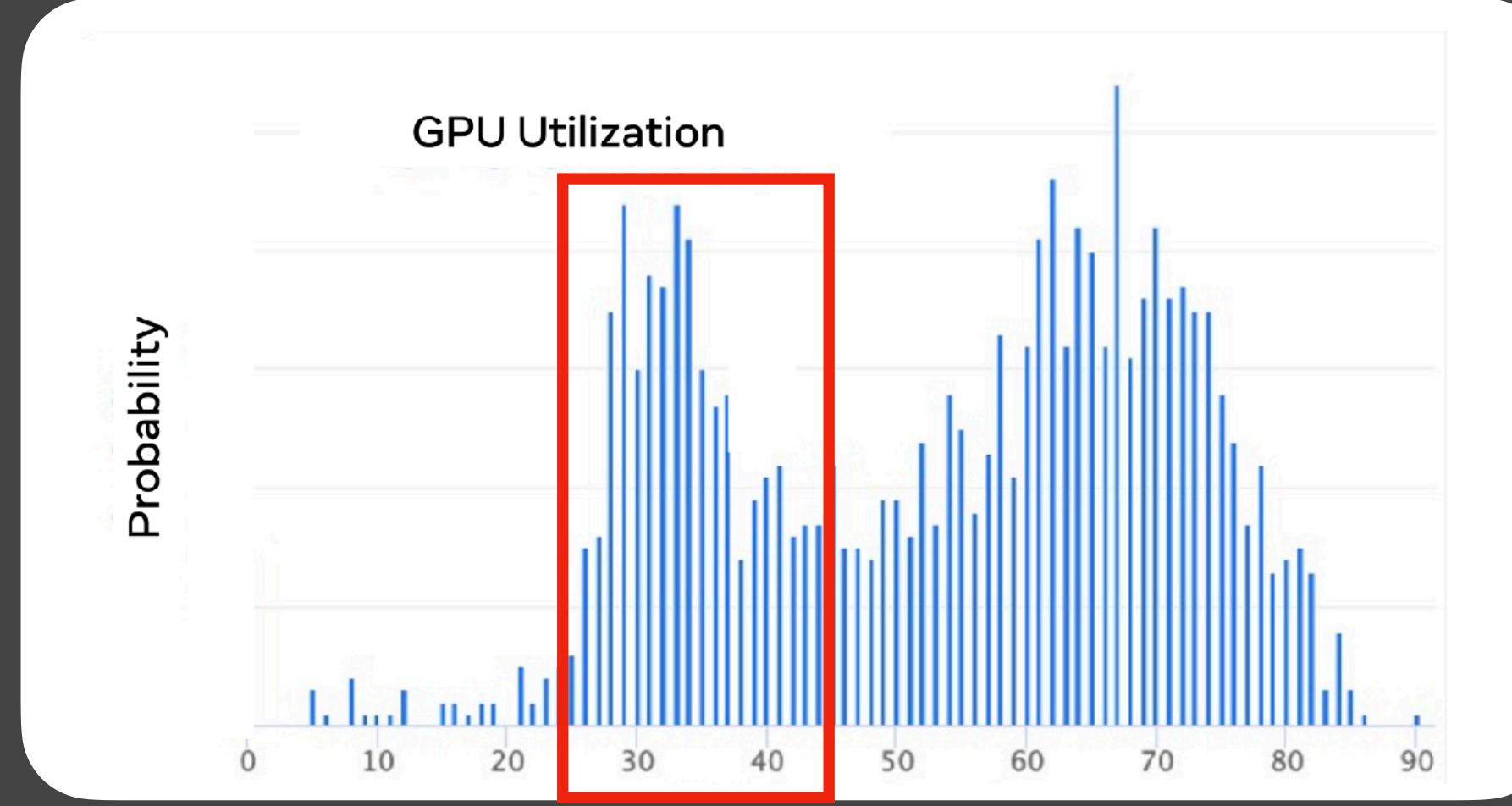
• There are 4 main overarching stages where carbon emissions need to be isolated: data

training and inference; text translation models have a 35%/65% split. (Operational cost)



Open issues according to Meta

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



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.

Federated learning might not be the solution

Most of the carbon footprint stems from communications ightarrow

