"FOLLOWING THE SUN" TO MITIGATE THE CARBON FOOTPRINT OF TRAINING AI ALGORITHMS: IS IT WORTH THE PRICE?

Sustainable Software Engineering course – 22nd Feb. 2024

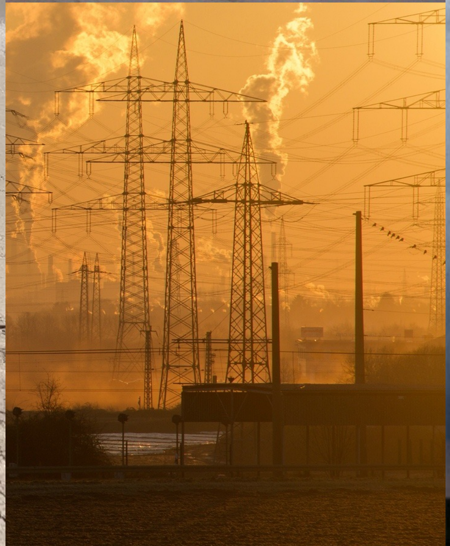
eng. Roberto Vergallo, Ph.D. - University of Salento

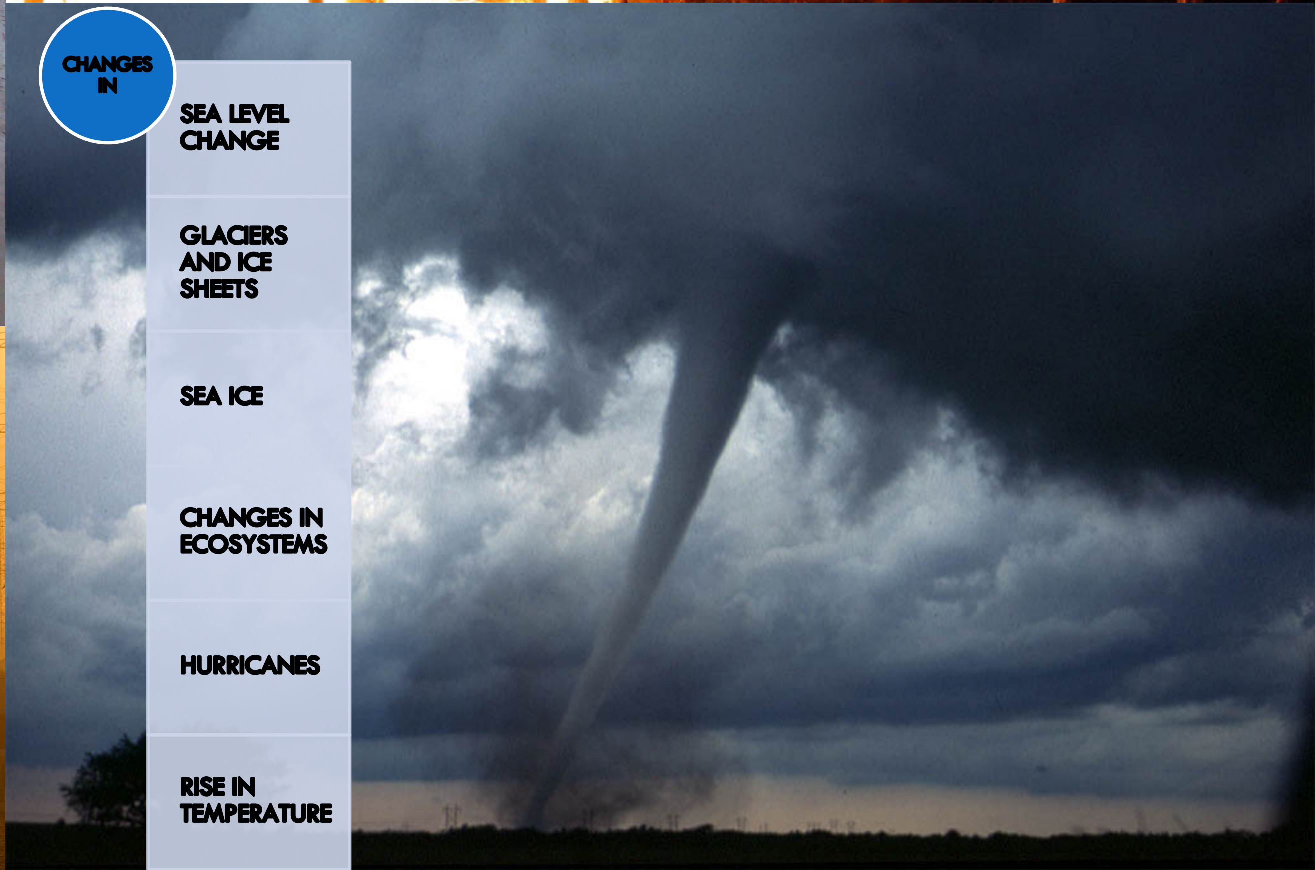
r.vergallo@tudelft.nl

WHOAMI

- Assistant Professor in Computer Science @ University of Salento, Italy
- https://goo.gl/maps/dnANaZjMLRugJ5k2A
- Visiting researcher @ Smart Grid Energy Research Center (UCLA - Los Angeles)
- Main research topics: enhanced CX models in the Fintech field using IoT, AI and DLT, and, obviously, green software
- Co-founder of University spin-off (Vidyasoft s.r.l.)
- Open to collabs (thesis, host students in lab, …)
- https://software.green/

CLIMATE CHANGE

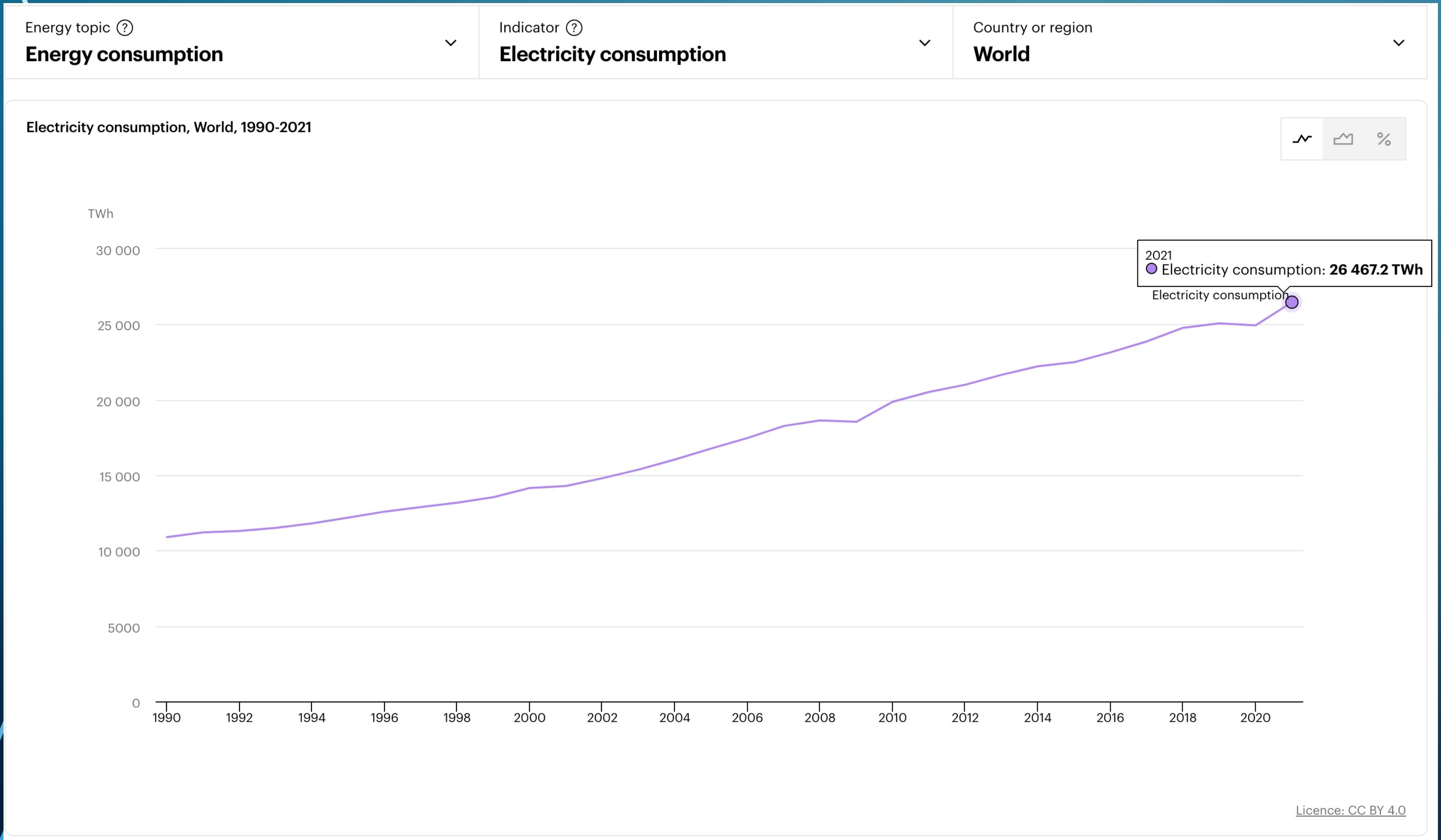




CHANGES IN

- SEA LEVEL CHANGE
- GLACIERS AND ICE SHEETS
- SEA ICE
- CHANGES IN ECOSYSTEMS
- HURRICANES
- RISE IN TEMPERATURE

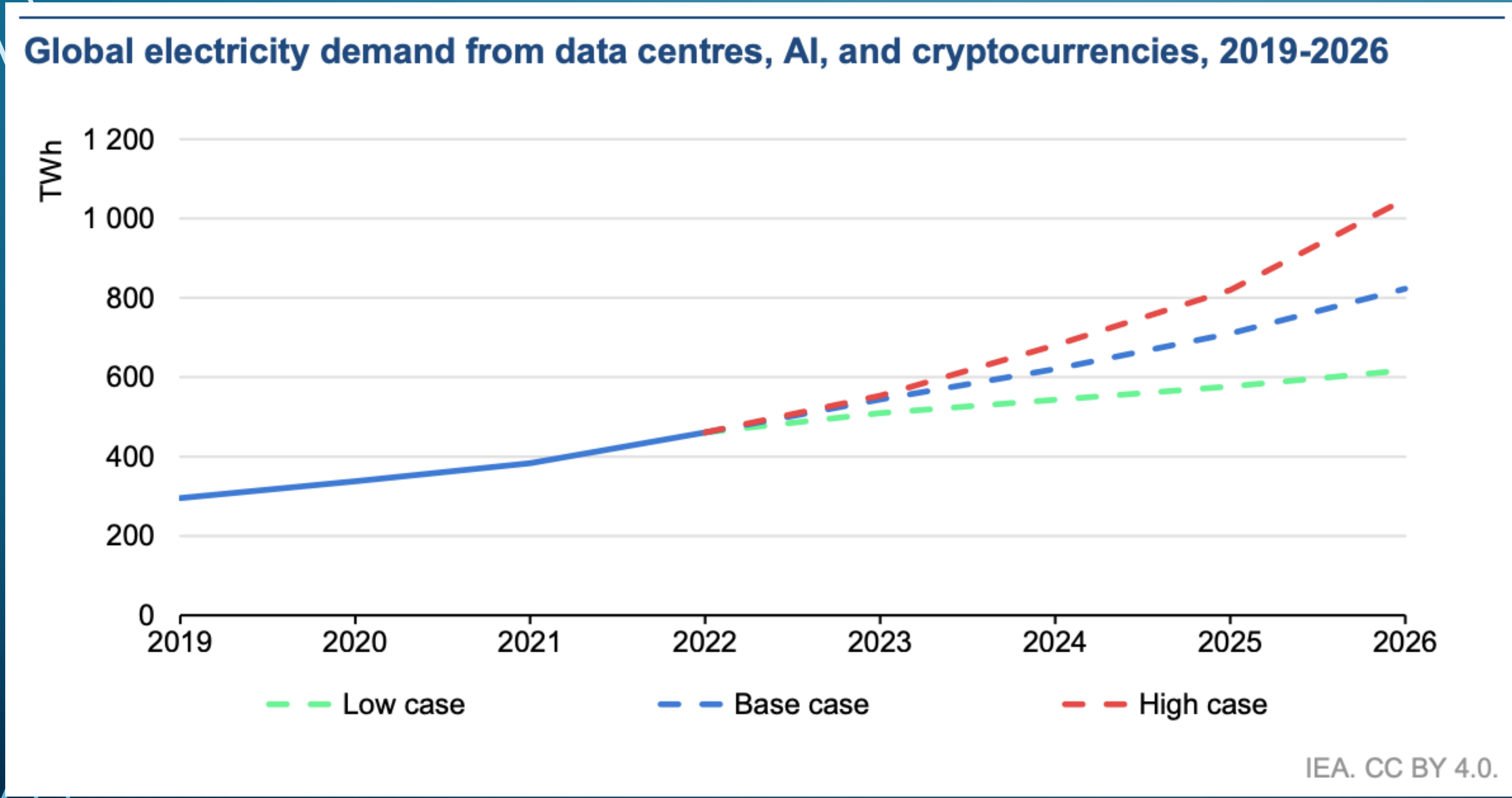
GLOBAL ENERGY CONSUMPTION



The world's electricity consumption has continuously grown over the past half a century, reaching approximately 26,500 terawatt-hours in 2021

Source: https://www.iea.org/data-and-statistics/data-tools/energy-statistics-data-browser?country=WORLD&fuel=Energy%20consumption&indicator=TotElecCons

GLOBAL ENERGY CONSUMPTION



Electricity consumption from data centres, artificial intelligence (AI) and the cryptocurrency sector could double by 2026.

After globally consuming an estimated 460 terawatt-hours (TWh) in 2022, data centres' total electricity consumption could reach more than 1 000 TWh in 2026.

Source: https://www.iea.org/reports/electricity-2024/executive-summary

ARTIFICIAL INTELLIGENCE

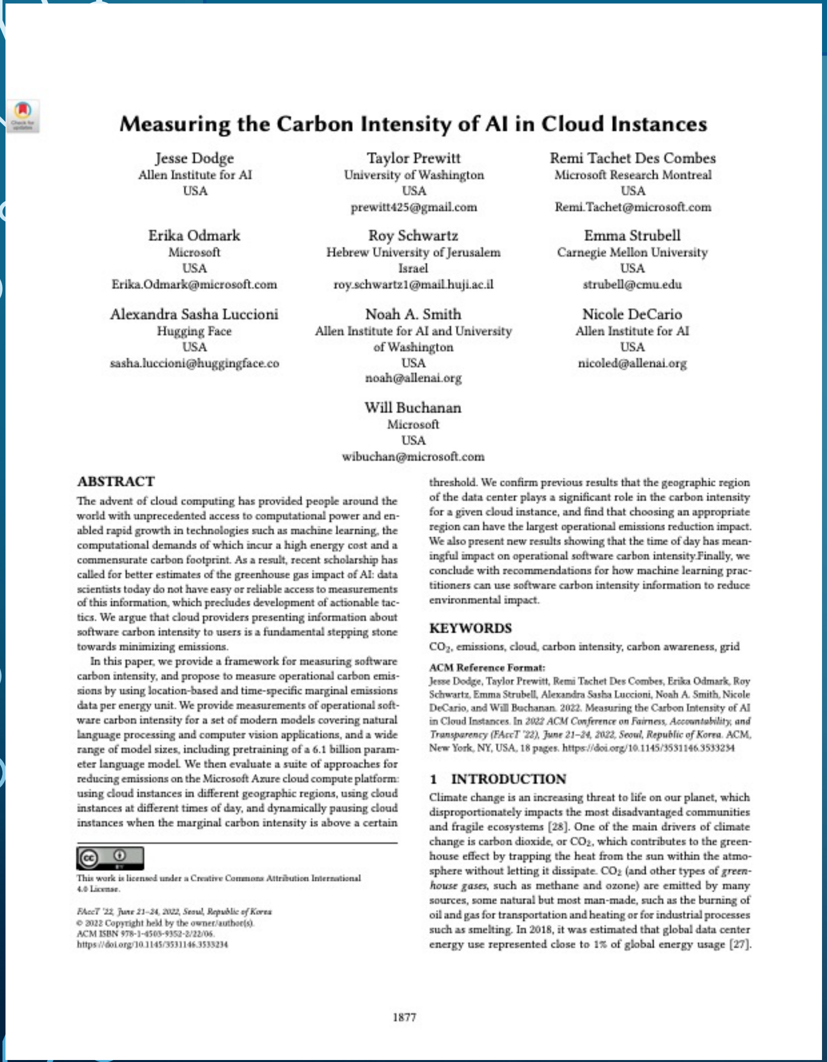
RED AI

Accuracy: 0.999999999

Green AI

- Energy
- Time
- Reproducibility
- Reusage

THE CARBON FOOTPRINT OF TRAINING AI



Measuring the Carbon Intensity of AI in Cloud Instances

Jesse Dodge
Allen Institute for AI
USA

Taylor Prewitt
University of Washington
USA
prewitt425@gmail.com

Remi Tachet Des Combes
Microsoft Research Montreal
USA
Remi.Tachet@microsoft.com

Erika Odmark
Microsoft
USA
Erika.Odmark@microsoft.com

Roy Schwartz
Hebrew University of Jerusalem
Israel
roy.schwartz1@mail.huji.ac.il

Emma Strubell
Carnegie Mellon University
USA
strubell@cmu.edu

Alexandra Sasha Luccioni
Hugging Face
USA
sasha.luccioni@huggingface.co

Noah A. Smith
Allen Institute for AI and University of Washington
USA
noah@allenai.org

Nicole DeCario
Allen Institute for AI
USA
nicoled@allenai.org

Will Buchanan
Microsoft
USA
wibuchan@microsoft.com

ABSTRACT

The advent of cloud computing has provided people around the world with unprecedented access to computational power and enabled rapid growth in technologies such as machine learning, the computational demands of which incur a high energy cost and a commensurate carbon footprint. As a result, recent scholarship has called for better estimates of the greenhouse gas impact of AI: data scientists today do not have easy or reliable access to measurements of this information, which precludes development of actionable tactics. We argue that cloud providers presenting information about software carbon intensity to users is a fundamental stepping stone towards minimizing emissions.

In this paper, we provide a framework for measuring software carbon intensity, and propose to measure operational carbon emissions by using location-based and time-specific marginal emissions data per energy unit. We provide measurements of operational software carbon intensity for a set of modern models covering natural language processing and computer vision applications, and a wide range of model sizes, including pretraining of a 6.1 billion parameter language model. We then evaluate a suite of approaches for reducing emissions on the Microsoft Azure cloud compute platform: using cloud instances in different geographic regions, using cloud instances at different times of day, and dynamically pausing cloud instances when the marginal carbon intensity is above a certain threshold. We confirm previous results that the geographic region of the data center plays a significant role in the carbon intensity for a given cloud instance, and find that choosing an appropriate region can have the largest operational emissions reduction impact. We also present new results showing that the time of day has meaningful impact on operational software carbon intensity.Finally, we conclude with recommendations for how machine learning practitioners can use software carbon intensity information to reduce environmental impact.

KEYWORDS

CO_2, emissions, cloud, carbon intensity, carbon awareness, grid

ACM Reference Format:
Jesse Dodge, Taylor Prewitt, Remi Tachet Des Combes, Erika Odmark, Roy Schwartz, Emma Strubell, Alexandra Sasha Luccioni, Noah A. Smith, Nicole DeCario, and Will Buchanan. 2022. Measuring the Carbon Intensity of AI in Cloud Instances. In *2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22), June 21–24, 2022, Seoul, Republic of Korea.* ACM, New York, NY, USA, 18 pages. https://doi.org/10.1145/3531146.3533234

1 INTRODUCTION

Climate change is an increasing threat to life on our planet, which disproportionately impacts the most disadvantaged communities and fragile ecosystems [28]. One of the main drivers of climate change is carbon dioxide, or CO_2, which contributes to the greenhouse effect by trapping the heat from the sun within the atmosphere without letting it dissipate. CO_2 (and other types of *greenhouse gases*, such as methane and ozone) are emitted by many sources, some natural but most man-made, such as the burning of oil and gas for transportation and heating or for industrial processes such as smelting. In 2018, it was estimated that global data center energy use represented close to 1% of global energy usage [27].

FAccT '22, June 21–24, 2022, Seoul, Republic of Korea

ACM ISBN 978-1-4503-9352-2/22/06.
https://doi.org/10.1145/3531146.3533234

How an AI training can be perfomed with a low-environmental impact?

Is focusing on training meaningful?

[1] Dodge, J., Prewitt, T., Tachet des Combes, R., Odmark, E., Schwartz, R., Strubell, E., ... & Buchanan, W. (2022, June). Measuring the carbon intensity of AI in cloud instances. In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (pp. 1877-1894).

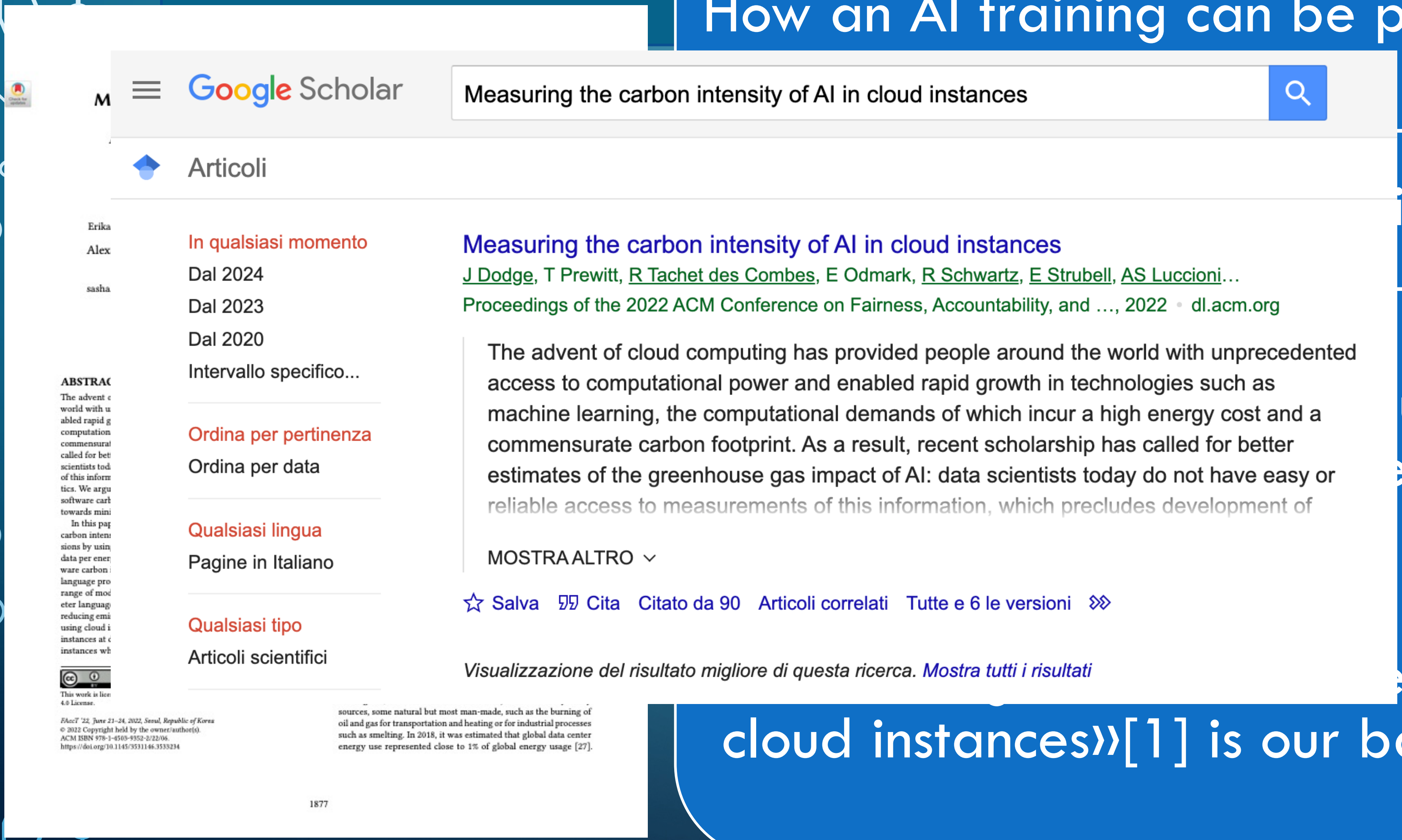
THE CARBON FOOTPRINT OF TRAINING AI

How an AI training can be perfomed with a

ingful

ut low-carbon
echnology is

ensity of AI in
cloud instances»[1] is our baseline.



Google Scholar

Measuring the carbon intensity of AI in cloud instances

Articoli

In qualsiasi momento
Dal 2024
Dal 2023
Dal 2020
Intervallo specifico...

Ordina per pertinenza
Ordina per data

Qualsiasi lingua
Pagine in Italiano

Qualsiasi tipo
Articoli scientifici

Measuring the carbon intensity of AI in cloud instances

J Dodge, T Prewitt, R Tachet des Combes, E Odmark, R Schwartz, E Strubell, AS Luccioni…

Proceedings of the 2022 ACM Conference on Fairness, Accountability, and …, 2022 • dl.acm.org

The advent of cloud computing has provided people around the world with unprecedented access to computational power and enabled rapid growth in technologies such as machine learning, the computational demands of which incur a high energy cost and a commensurate carbon footprint. As a result, recent scholarship has called for better estimates of the greenhouse gas impact of AI: data scientists today do not have easy or reliable access to measurements of this information, which precludes development of

MOSTRA ALTRO

Salva Cita Citato da 90 Articoli correlati Tutte e 6 le versioni

Visualizzazione del risultato migliore di questa ricerca. Mostra tutti i risultati

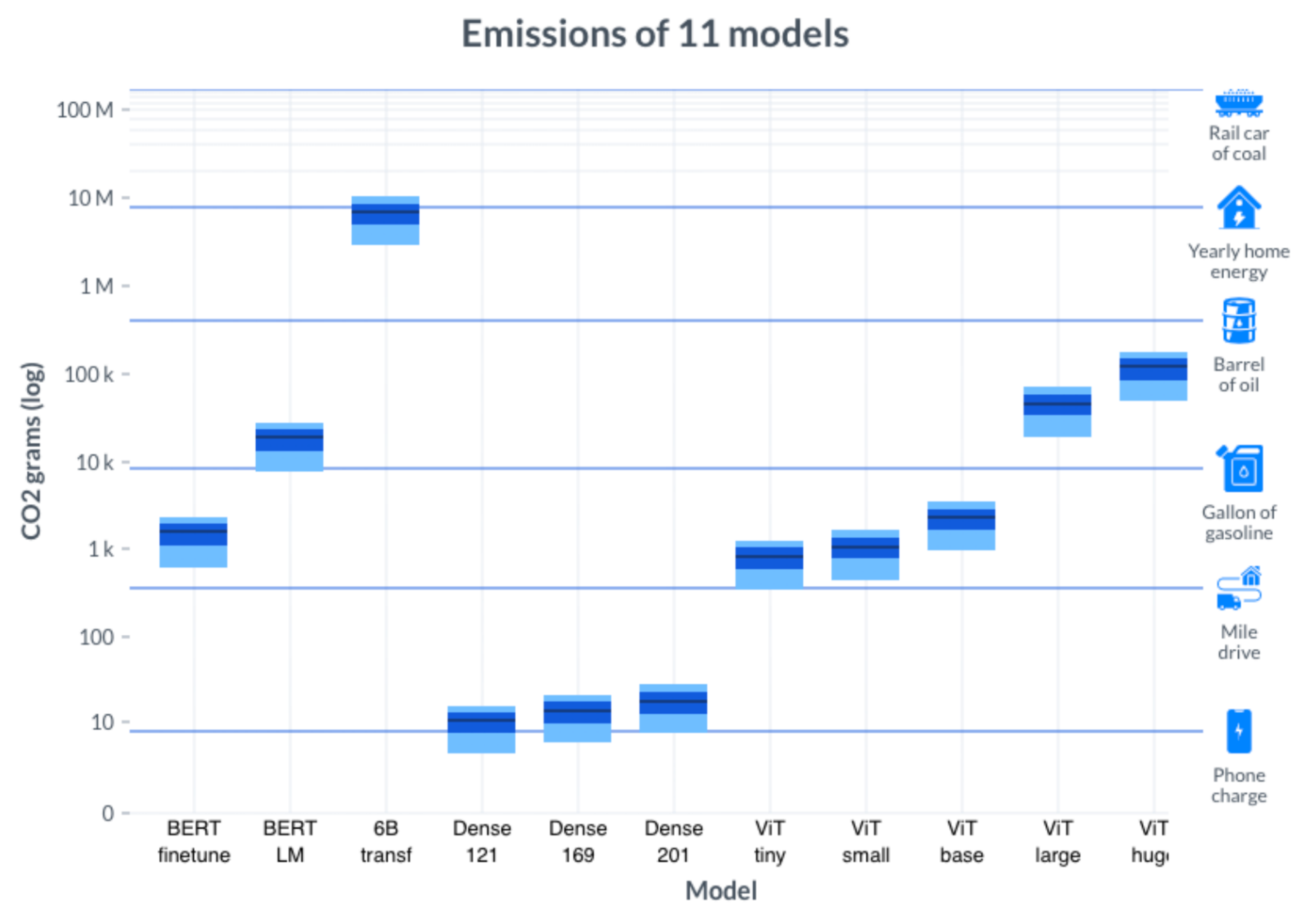
sources, some natural but most man-made, such as the burning of oil and gas for transportation and heating or for industrial processes such as smelting. In 2018, it was estimated that global data center energy use represented close to 1% of global energy usage [27].

FAccT '22, June 21–24, 2022, Seoul, Republic of Korea
© 2022 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-9352-2/22/06.
https://doi.org/10.1145/3531146.3533234

1877

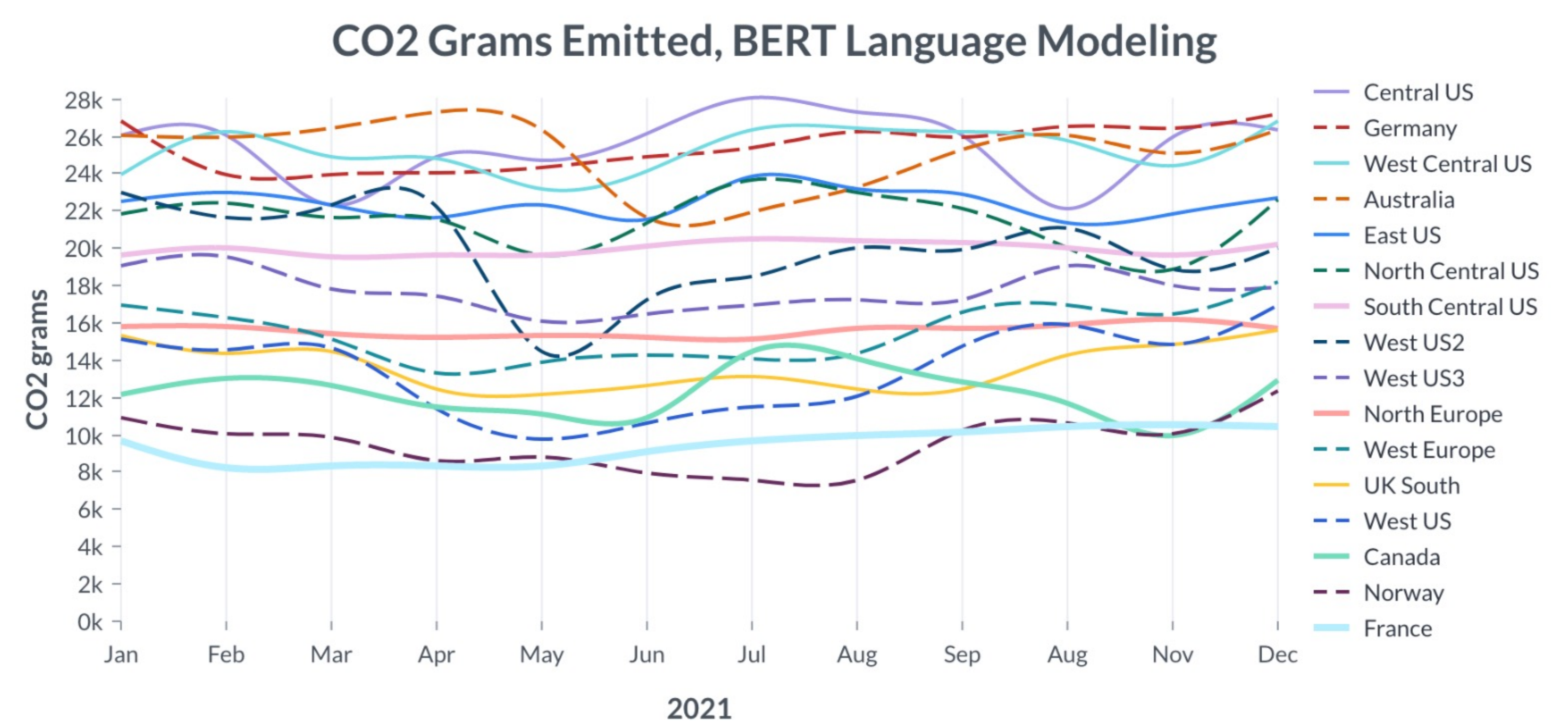
[1] Dodge, J., Prewitt, T., Tachet des Combes, R., Odmark, E., Schwartz, R., Strubell, E., ... & Buchanan, W. (2022, June). Measuring the carbon intensity of AI in cloud instances. In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (pp. 1877-1894).

BENCHMARKING ALGORITHMS



Model	BERT finetune	BERT pretrain	6B Transf.	Dense 121	Dense 169	Dense 201	ViT Tiny	ViT Small	ViT Base	ViT Large	ViT Huge
GPU	4·V100	8·V100	256·A100	1·P40	1·P40	1·P40	1· V100	1·V100	1·V100	4·V100	4·V100
Hours	6	36	192	0.3	0.3	0.4	19	19	21	90	216
kWh	3.1	37.3	13,812.4	0.02	0.03	0.04	1.7	2.2	4.7	93.3	237.6

BENCHMARKING ALGORITHMS



MITIGATION STRATEGY 1

- FLEXIBLE-START

Launch the training at the starting time that would result in the lowest emissions.

00:00	00:05	00:10	00:15	00:20	00:25	00:30	00:35	00:40	00:45	00:50	00:55	01:00	01:05	01:10	01:15	01:20	01:25	01:30	01:35	01:40	01:45	01:50	01:55

MITIGATION STRATEGY 2

- PAUSE AND RESUME

Run the training only during the 5-minute slots with the lowest marginal emissions.

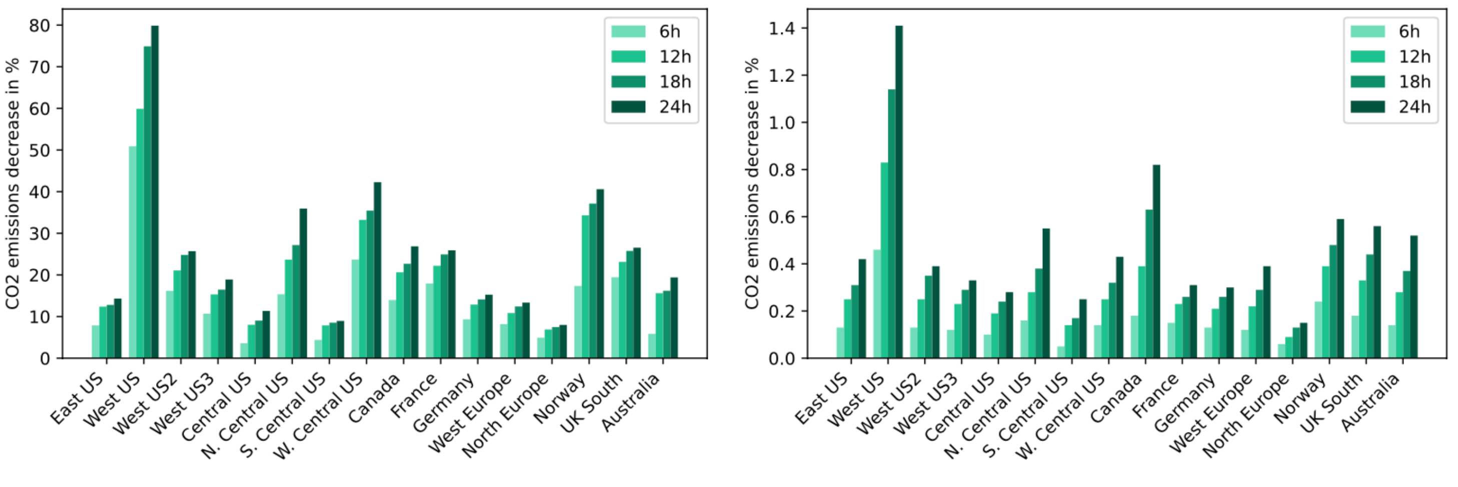
TUNING STRATEGY PARAMETERS

- Find the 5 minute intervals with the lowest marginal emissions during the (N + job_duration) hour window, and select enough intervals to add up to the job duration.
- Then simulate running the job only during those intervals and compute the corresponding emissions
- They explored two sets of values for N:
  - Absolute: N ∈ {6, 12, 18, 24} (hours)
  - Relative: N ∈ {25%, 50%, 75%, 100%} x job_duration

CHOOSE REGIONS WISELY

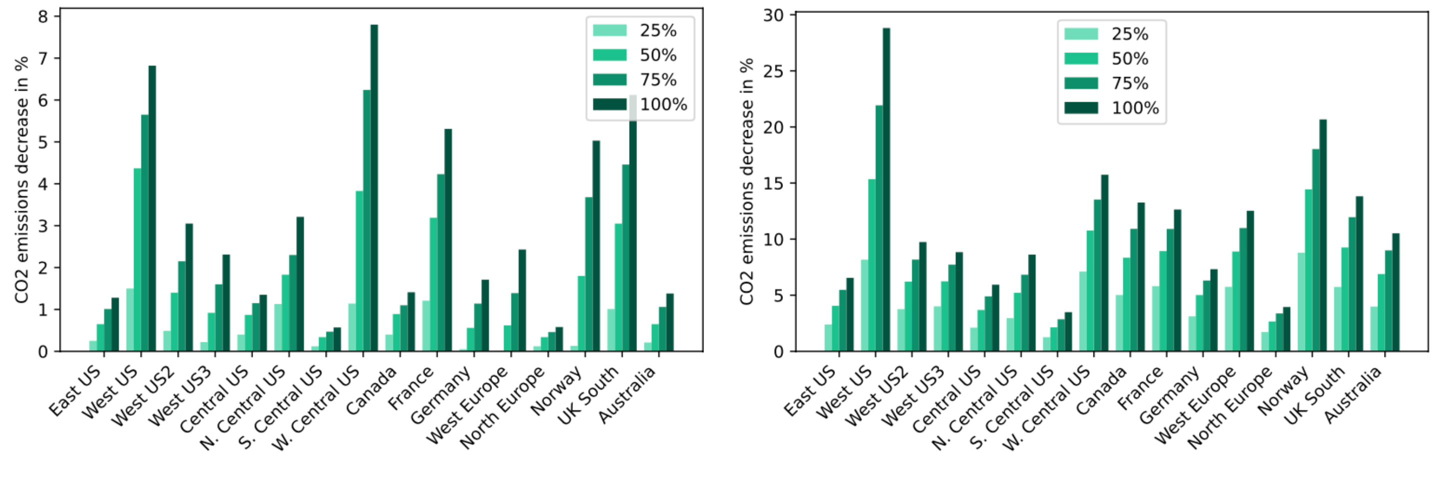
The region that the algorithms are evaluated in has a significant impact for both strategies

For example, the West US region varies frequently throughout a single day, and thus **Pause and Resume** can lead to significant reductions.



(a) ***Flexible Start*** **optimization for Dense 201.**

(b) ***Flexible Start*** **optimization for 6B parameters Transformer.**



(a) ***Pause and Resume*** **optimization for Dense 201.**

(b) ***Pause and Resume*** **optimization for 6B parameters Transformer.**

EVALUATION

- In order to account for daily variations (weather, electricity demand, etc.), they report the average emissions decrease computed over 5 different start times in each month, giving a total of 60 data points.
- FLEXIBLE START
  - Significant emissions reductions for shorter jobs (e.g., the DenseNet experiments
  - Minimal savings for jobs longer than a day
  - Useful for use cases where an AI workload needs to run regularly, but the practitioner has some flexibility on when it runs (e.g. where models are re-trained on a regular schedule to incorporate new data over time)

EVALUATION

- PAUSE AND RESUME
  - short experiments only see emissions reductions smaller than 10%
  - the 6 billion transformer training sees the largest decrease in emissions
  - Useful for use cases where an AI workload can be increased in duration by some proportion of the original run time

EVALUATION

Model	BERT finetune	BERT LM	6B Transf.	Dense 121	Dense 169	Dense 201	ViT Tiny	ViT Small	ViT Base	ViT Large	ViT Huge
FS	14.5%	3.4%	0.5%	26.8%	26.4%	25.9%	5.6%	5.3%	4.2%	1.3%	0.5%
P&R	19.0%	8.5%	2.5%	27.7%	27.3%	27.1%	12.5%	12.3%	11.7%	4.7%	2.4%
Pauses / hr	0.23	0.3	0.15	0.06	0.07	0.08	0.3	0.3	0.3	0.23	0.14

$(14.5+3.4+0.5+26.8+26.4+25.9+5.6+5.3+4.2+1.3+0.5)/11 = 10.4\%$

$(19+8.5+2.5+27.7+27.3+27.1+12.5+12.3+11.7+4.7+2.4)/11 = 14.5\%$

OPEN ISSUES

FLEXIBLE-START

- more efficient for short workloads.
  - 3.4% emission reductions on BERT LM

PAUSE AND RESUME

- more efficient for workloads longer than a day.

  8.5% emission reductions on BERT LM

Both strategies are based on the temporal management of training during

Training completion can be delayed by up to 24 hours or even more: you have to be patient!

DEMAND SHIFTING

On the Effectiveness of the 'Follow-the-Sun' Strategy in Mitigating the Carbon Footprint of AI in Cloud Instances

Roberto Vergallo[a], Alessio Errico[a], Luca Mainetti[a]

[a] *University of Salento, Dept. of Innovation Engineering, via per Monteroni, 165, Lecce, 73100, LE, Italy*

Abstract

"Follow-the-Sun" (FtS) is a theoretical computational model aimed at minimizing the carbon footprint of computer workloads. It involves dynamically moving workloads to regions with cleaner energy sources as demand increases and energy production relies more on fossil fuels. With the significant power consumption of Artificial Intelligence (AI) being a subject of extensive debate, FtS is proposed as a strategy to mitigate the carbon footprint of training AI models. However, the literature lacks scientific evidence to measure the actual advantages of applying FtS to AI workloads. In this paper, we present the results of an experiment conducted to address this research gap.

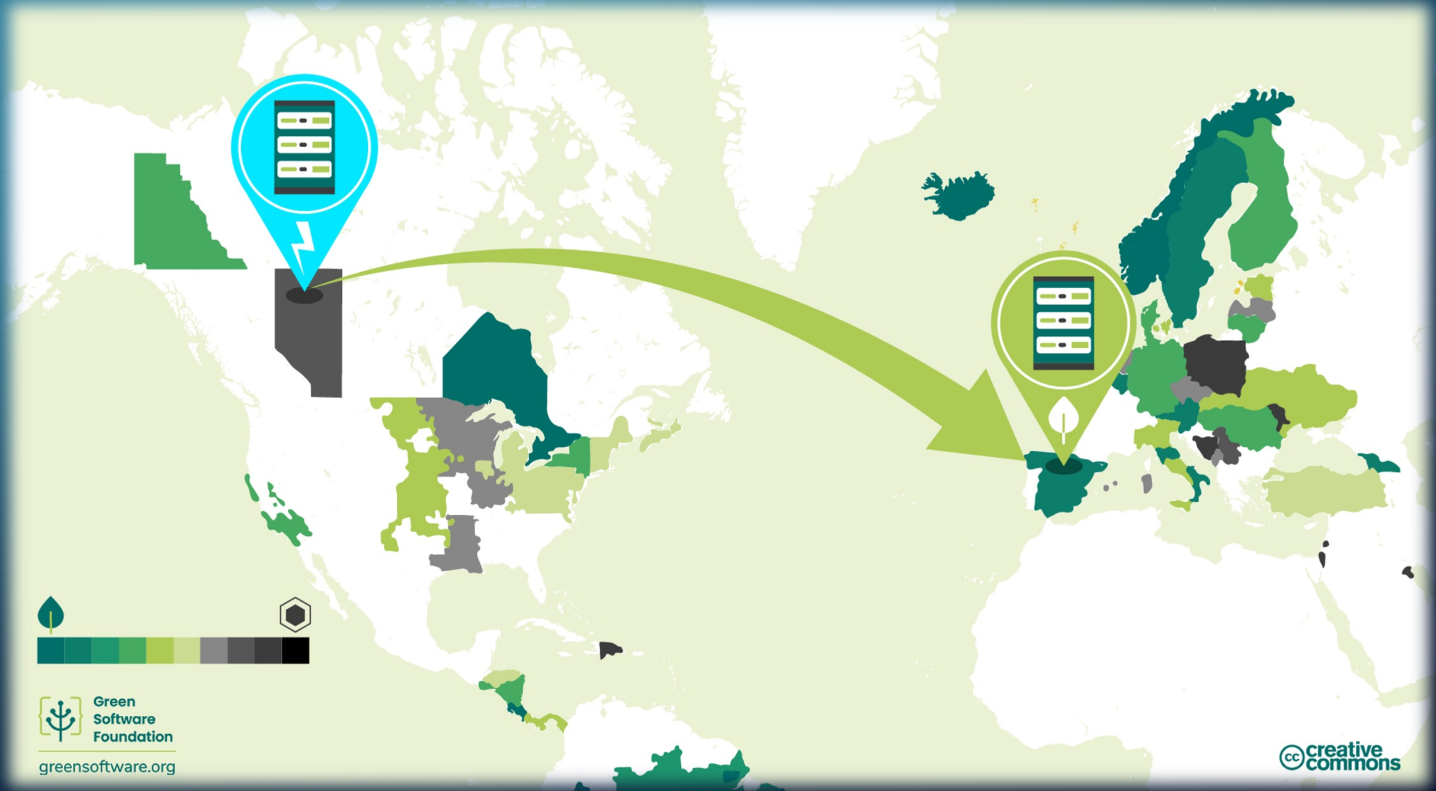
We benchmarked four AI algorithms in the anomaly detection domain and measured the differences in carbon emissions across four scenarios: no strategy, FtS, and two strategies previously introduced in the state-of-the-art, namely Flexible Start and Pause and Resume. To conduct our experiment, we utilized historical carbon intensity data from the year 2021 for seven European cities. Our results demonstrate that the FtS strategy not only achieves average reductions of up to 14.58% in carbon emissions (with peaks of 16.3%) but also helps in preserving the time needed for training.

Keywords: demand shifting, follow the sun, carbon footprint, green AI, training workload
PACS: 07.05.Bx, 07.05.Mh
2000 MSC: 68M20, 68T01

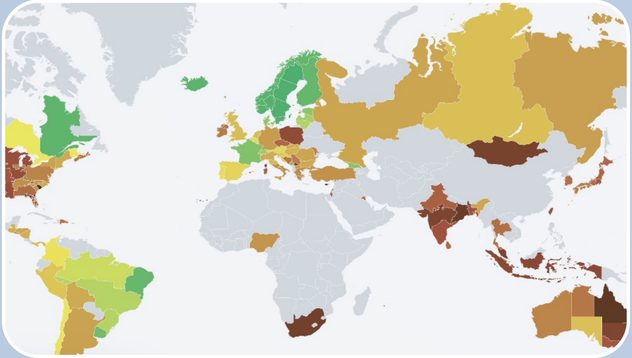
Preprint submitted to Journal of Computational Sciences *September 1, 2023*

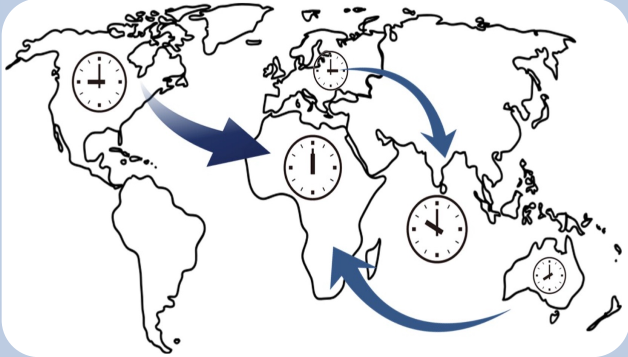
Vergallo, R., Errico, A., & Mainetti, L. On the Effectiveness of the'Follow-the-Sun'Strategy in Mitigating the Carbon Footprint of AI in Cloud Instances. *Available at SSRN 4566638*.

Strategy of moving workloads to regions or times when resources are constrained.



MOTIVATIONS





Emissions depend not only on the time of day but also on the grid region where training is performed

“Follow the Sun”[2] is an approach applied to various problems, but there is no scientific validity or evidence regarding its effectiveness

Current state-of-art strategies don't preserve time

[2] Follow the Sun, GitHub. https://follow-the-sun.github.io

OBJECTIVES

Develop a new green AI training approach by leveraging the benefits of Cloud technology.

Compare the proposed strategy with other strategies related to the same problem

The research is part of AMEDEA project (Assessment and Mitigation of the Environmental impact of DL Algorithms) Cod. IsCa7_AMEDEA 107C

BENCHMARK

FRAUD-DETECTION WORKLOADS

-AUTOENCODER

-HF-SCA [3]

-SVM

-ISOLATION FOREST

SANDBOX

BANCA D'ITALIA

BANK TRANSACTION DATASET PROVIDED BY AN ITALIAN BANK WITHIN THE REGULAMENTARY SANDBOX INITIATIVE

[3] *Distante, C., Fineo, L., Mainetti, L., Manco, L., Taccardi, B., & Vergallo, R. (2022). HF-SCA: Hands-Free Strong Customer Authentication Based on a Memory-Guided Attention Mechanisms. Journal of Risk and Financial Management, 15(8), 342.*

CARBON INTENSITY DATA

WattTime

UNIVERSITÀ DEL SALENTO

Green Software Foundation

HISTORICAL CARBON INTENSITY DATA FOR YEAR 2021 FOR THE FOLLOWING REGIONS:

-MILAN

-PARIS

-FRANKFURT

-ZARAGOZA

-LONDON

-DUBLIN

-STOCKHOLM

EXPERIMENTAL SET-UP

Workload Runner

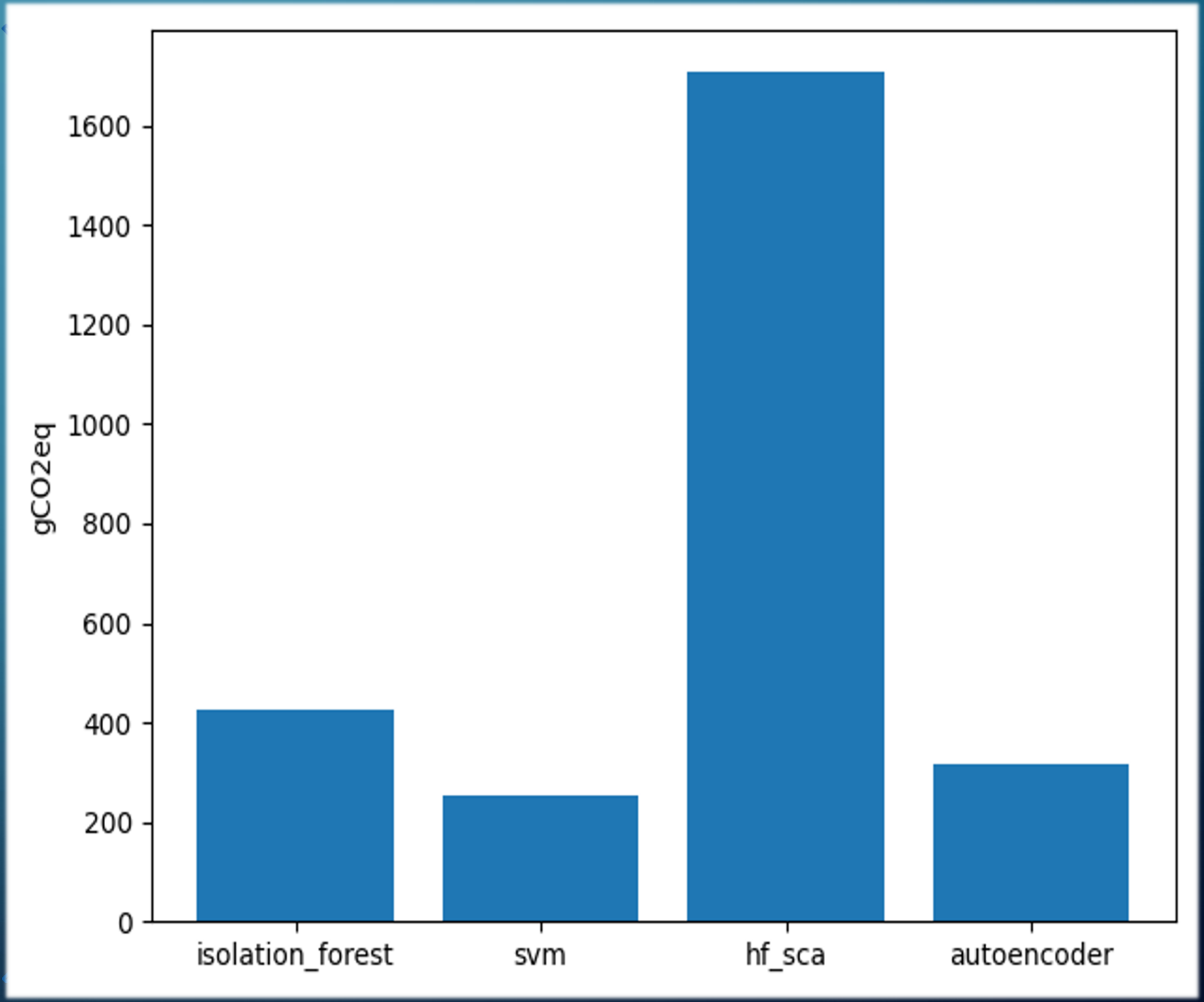
- Perform training and track energy consumptions

Strategy Launcher

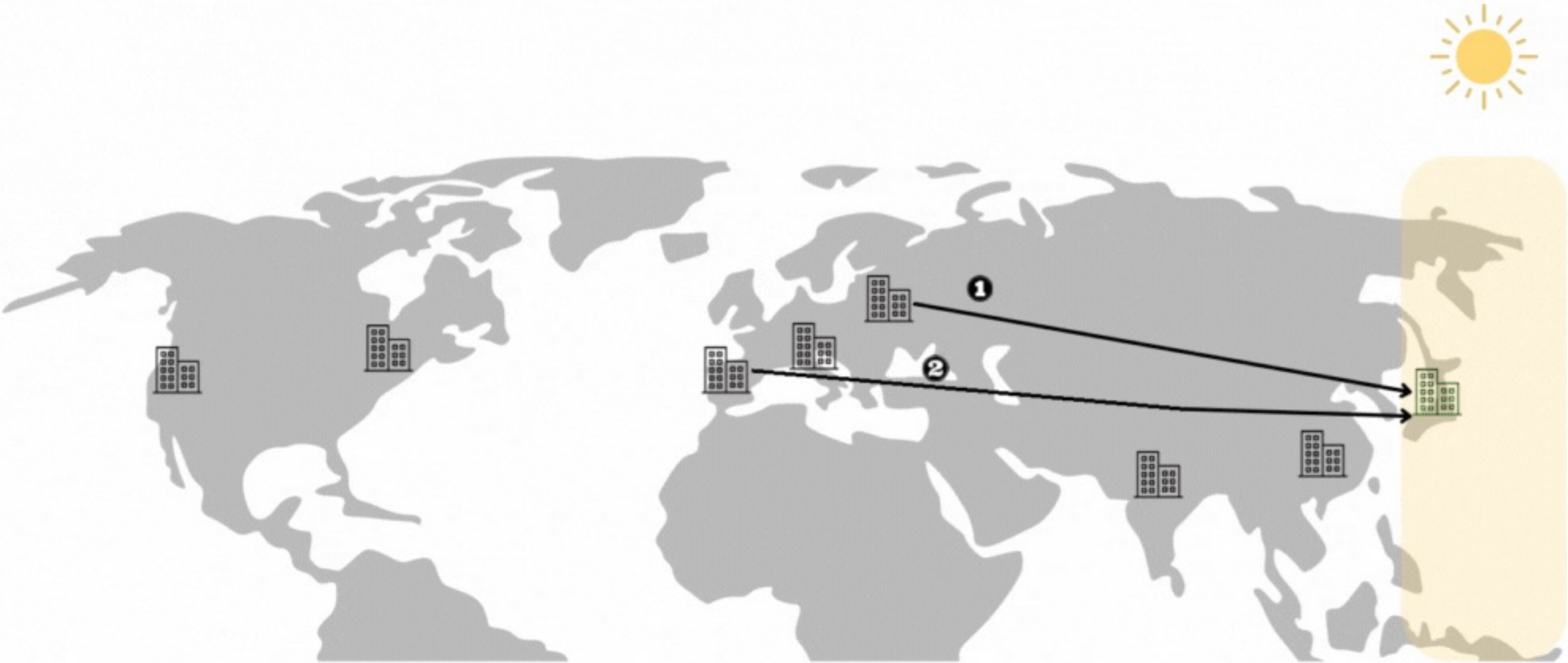
- Compute emissions for different strategies

NVIDIA DGX A100

CodeCarbon python library for energy consumptions each 5 minutes



PROPOSED STRATEGY: FOLLOW THE SUN



PROPOSED STRATEGY: FOLLOW THE SUN

STRATEGY VERSIONS

STATIC-START
FOLLOW THE SUN

FLEXIBLE-START
FOLLOW THE SUN

DATA TRANSFER VERSIONS

UPSTREAM

- Transfer dataset to all Cloud Instances before the training start

IN-TRAINING

- Transfer data during the training

ASSUMPTIONS

Energy consumed for data transfer: 0.023 kWh/Gb from 2015 [4]

COMPRESSED DATASET SIZE: 0.320GB

WORKLOADS NOT PARTICULARLY LARGE

NEGLIGLIBLE TRANSFER TIME

NEGLIGIBLE EMISSIONS FOR TRAINING STATE DATA TRANSFER

[4] Malmodin, J., & Lundén, D. (2018). The energy and carbon footprint of the global ICT and E&M sectors 2010-2015. Sustainability, 10(9), 3027.

PROPOSED STRATEGY: FOLLOW THE SUN

GENERAL IDEA

- Checking-time: how often to designate the new region to transfer the training to
- The workload is divided into k slots based on the selected checking time
- Each slot corresponds to a training segment that will be executed on the region with the least environmental impact

PROPOSED STRATEGY: FOLLOW THE SUN

1° version

Static-start Follow The Sun

- The workload will start at the specified starting time, and at each checking time, it will be moved to the greenest region to continue with the training.

REGION 1	00:00	00:05	00:10	00:15	00:20	00:25	00:30	00:35	00:40	00:45	00:50	00:55	01:00	01:05	01:10	01:15	01:20	01:25	01:30	01:35	01:40	01:45	01:50	01:55
REGION 2	00:00	00:05	00:10	00:15	00:20	00:25	00:30	00:35	00:40	00:45	00:50	00:55	01:00	01:05	01:10	01:15	01:20	01:25	01:30	01:35	01:40	01:45	01:50	01:55
REGION 3	00:00	00:05	00:10	00:15	00:20	00:25	00:30	00:35	00:40	00:45	00:50	00:55	01:00	01:05	01:10	01:15	01:20	01:25	01:30	01:35	01:40	01:45	01:50	01:55

PROPOSED STRATEGY: FOLLOW THE SUN

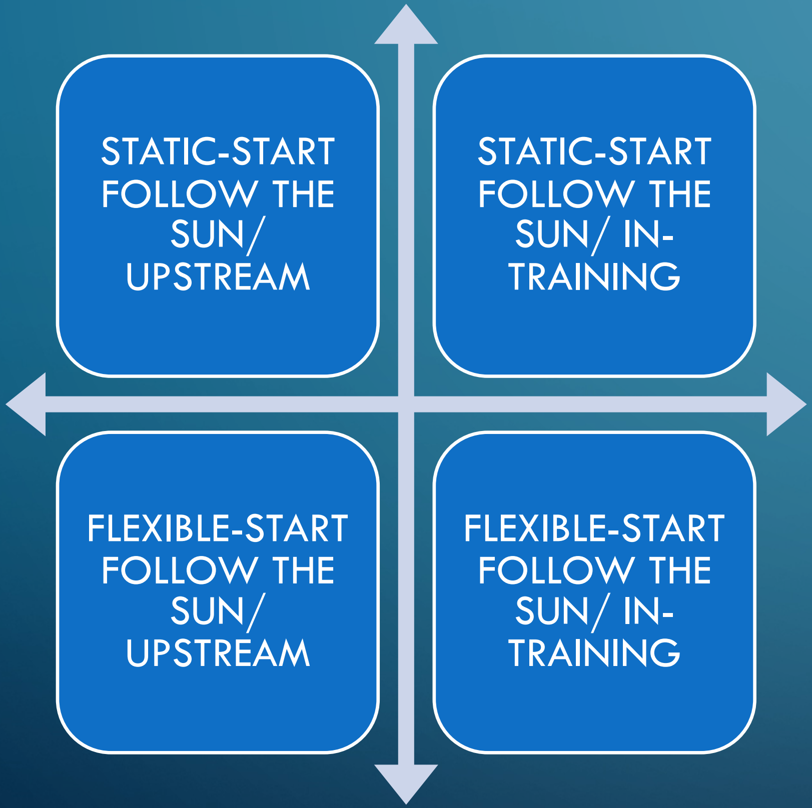
2° version

Flexible-start Follow The Sun

- **Starting at the time that minimizes emissions without interrupting the execution**

REGION 1	00:00	00:05	00:10	00:15	00:20	00:25	00:30	00:35	00:40	00:45	00:50	00:55	01:00	01:05	01:10	01:15	01:20	01:25	01:30	01:35	01:40	01:45	01:50	01:55
REGION 2	00:00	00:05	00:10	00:15	00:20	00:25	00:30	00:35	00:40	00:45	00:50	00:55	01:00	01:05	01:10	01:15	01:20	01:25	01:30	01:35	01:40	01:45	01:50	01:55
REGION 3	00:00	00:05	00:10	00:15	00:20	00:25	00:30	00:35	00:40	00:45	00:50	00:55	01:00	01:05	01:10	01:15	01:20	01:25	01:30	01:35	01:40	01:45	01:50	01:55

PROPOSED STRATEGY: FOLLOW THE SUN



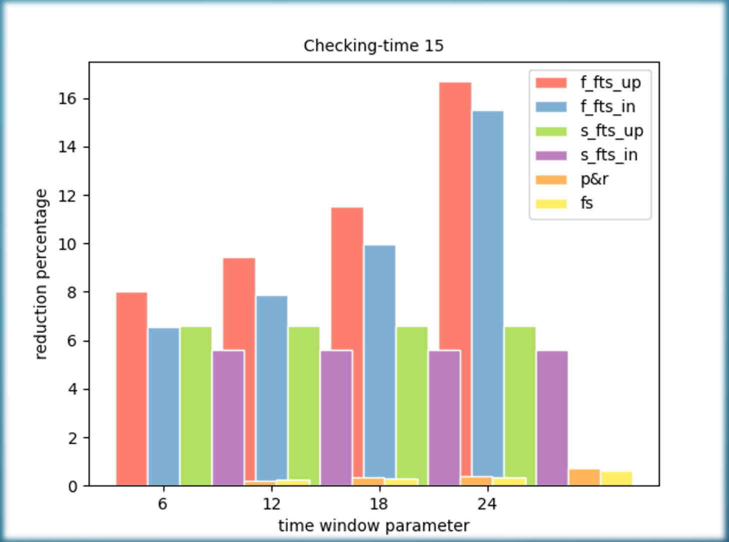
Strategies evaluations under the same conditions

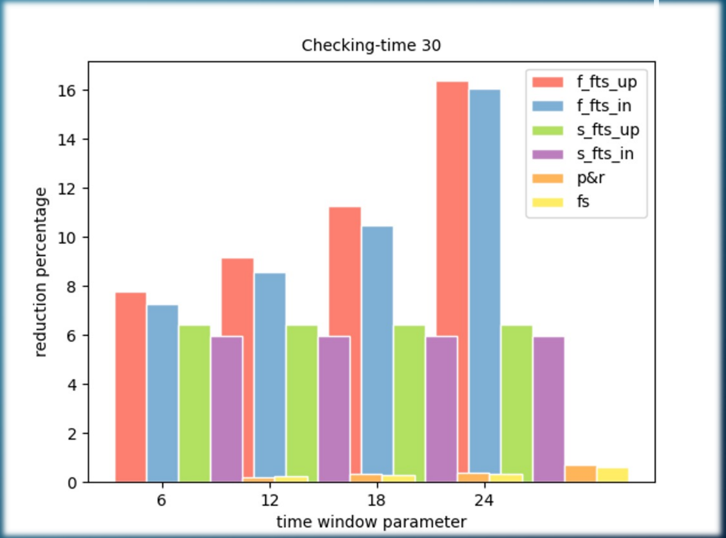
- 6 starting times for each month
- Percentage reduction averaged for the entire year and all regions
- Set for time window
  - (6h, 12h, 18h, 24h)
- A set for checking-time
  - (15m, 30m, 60m, 120m)

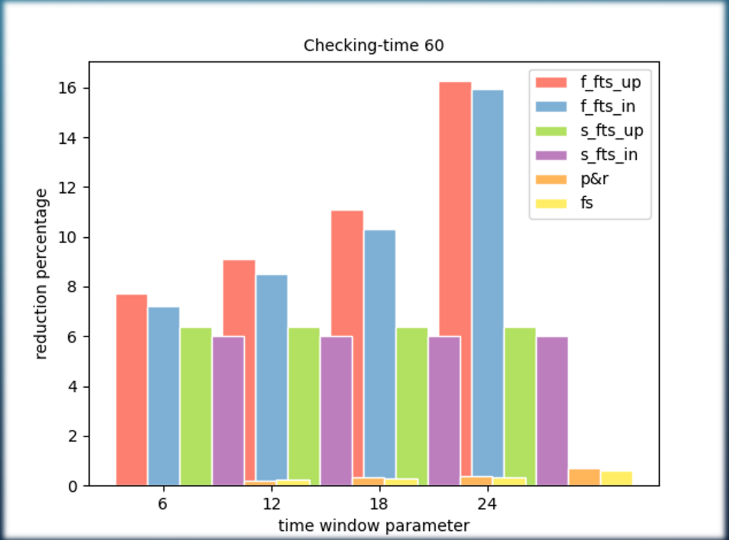
EVALUATION

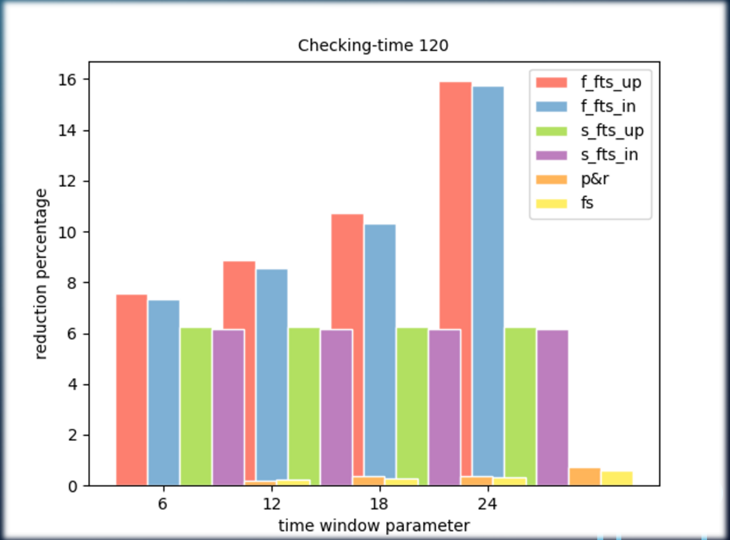
Reductions increase as the time window increases

As the checking time decreases, the reductions increase









EVALUATION

STATIC-START FOLLOW THE SUN

- Average reduction percentages ranging from 5% to 7%

FLEXIBLE-START FOLLOW THE SUN

- Average reduction is between 14-16% for the shorter workloads and almost 10% for the longest one
- Showed peaks of reductions beyond 81%
- Flexible-Start is a lower-bound: we can only do better!

EVALUATION

STATIC-START FOLLOW THE SUN

PRESERVE THE ENTIRE WORKLOAD DURATION

Strategy	HF-SCA	Autoencoder	SVM	Isolation Forest
No-Strategy	16h	3:30h	2:30h	4:15h
Flexible Start	19:15h	12:56h	12:45h	13:25h
Pause and Resume	21:33h	16:57h	15:50h	17:34h
Static-Start Follow the Sun	16h	3:30h	2:30h	4:15h
Flexible-Start Follow The Sun	37:32h	17:31h	17:47h	16:22h

EVALUATION

EMISSION REDUCTION

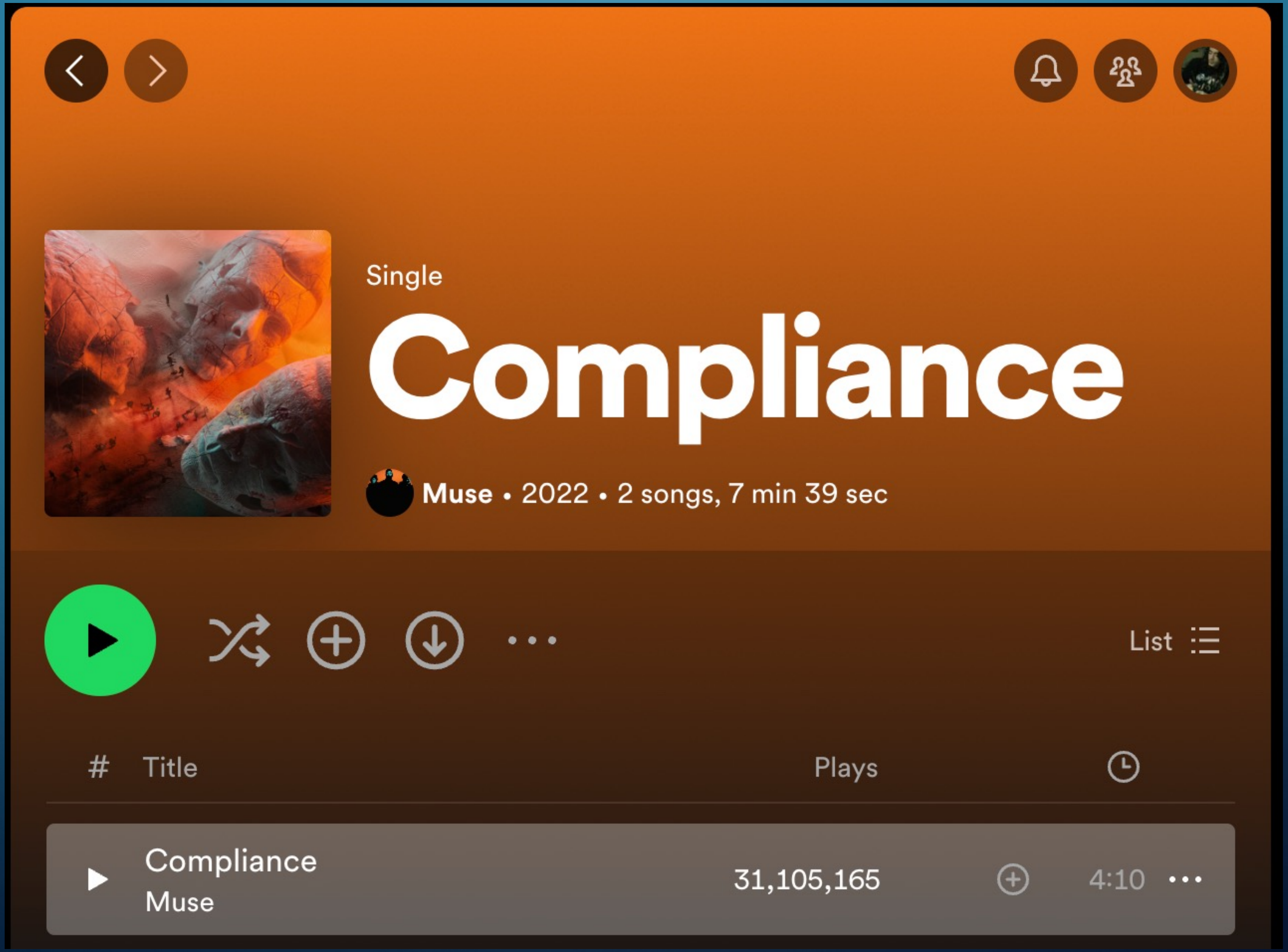
- Flexible-Start Follow the sun has the best average percentage reduction between all strategies

TIME SAVING

- Static-Start version perserve the workloads length. Other strategies does not consider this opportunity at all

Strategy	Avg time dilatation	Avg carbon reduction
Flexible-Start[1]	7:54h	5.72%
Pause and Resume[1]	11:15h	6.51%
Static-Start FtS	No dilatation	5.925%
Flexible-Start FtS	15:36h	13.85%

PROBLEMS?


Single
Compliance
Muse • 2022 • 2 songs, 7 min 39 sec
List
Title Plays
Compliance
Muse
31,105,165
4:10

EVALUATION

ROBUSTNESS WITH RESPECT TO THE SET OF REGIONS

- the Follow the sun strategy gives the same result no matter what the starting region is

GDPR LIMITATIONS

- The proposed strategy could be subjected to these kind of constrain

ARCHITECTURE COMPLEXITY

- The strategy proposed requires a more complex infrastructure

Strategy	GDPR limitations	Complex architecture required	Robustness wrt regions
Flexible-Start[1]	No	No	No
Pause and Resume[1]	No	No	No
Static-Start FtS	Yes	Yes	Yes
Flexible-Start FtS	Yes	Yes	Yes

ACCURACY vs SUSTAINABILITY

HF-SCA	Autoencoder	SVM	Isolation Forest
0.97	0.73	0.51	0.56

Consideration on the experiment

- In this case HF-SCA has much better AUC score and it is preferable despite high emissions
- In cases of more comparable performance, prefer the greener model

CONCLUSION

EMISSIONS REDUCTION

- 13.85% of average reductions between workloads, vs 6.51% for state of art strategies

TIME SAVING

- You can avoid wasting time for reducing emissions

ROBUSTNESS WRT STARTING REGION

- Same results regardless the starting region

THANK YOU!

Roberto Vergallo

r.vergallo@tudelft.nl