

*The GAISSA project*

# Towards green AI-based software systems: an architecture-centric approach

Silverio Martínez-Fernández

March 16<sup>th</sup>, 2023



# Climate crisis: the carbon footprint of software and AI

- Large AI and software companies shall follow energy efficiency regulations.
  - 2030: EU has agreed ambitious targets for greenhouse gas (GHG) emission reductions, renewable energy and energy efficiency
  - 2050: climate-neutral economy
- IT systems alone already consume up to 10% of global electricity (considering not only computer power, but internet transmission, data centers...)
  - Increase in energy consumption of data centers by 2030
- Need for an ecologic behavior!!
  - The need across the world to mitigate the impacts of climate change

Verdecchia, R., Lago, P., Ebert, C., De Vries, C. (2021). Green IT and Green Software. *IEEE Software* 38(6): 7-15

# Climate crisis: dual role of ML-based systems

- ML-based software systems for energy efficiency
  - They can help reduce the effects of climate crisis
  - Modeling climate change predictions
- Energy efficiency of ML-based systems
  - ML-based systems are themselves a significant emitter of carbon

Dhar, P. (2020). The carbon impact of artificial intelligence. *Nature Machine Intelligence*, 2(8), 423-425

# Software engineering for Energy efficiency (1/2)

## ■ What do companies need?



KPI	Unit of Measure	Example	Rationale	Alignment with Other Reporting Frameworks	EU Policy Reference
Direct GHG emissions from sources owned or controlled by the company (Scope 1)	Metric tons CO <sub>2</sub> e <sup>22</sup>	270 900 tCO <sub>2</sub> e	This KPI ensures companies are accurately measuring their carbon footprints from direct emissions.	TCFD Metrics and Targets, CDP Climate Change Questionnaire, GRI 305, CDSB Framework, SASB, EMAS	EU emissions trading system (ETS) 2030 climate & energy framework

**Further guidance:**

- Companies should disclose 100% of emissions if a company cannot collect reliable data at a figure for 100%. In that case, first disclose the emissions that have been estimated, (2) the remaining emissions and the proportion of emissions for which data have been estimated.
- Companies should, where appropriate, disclose emissions by activity, and by subsidiary. E.g.: Scope 1 emissions from manufacturing

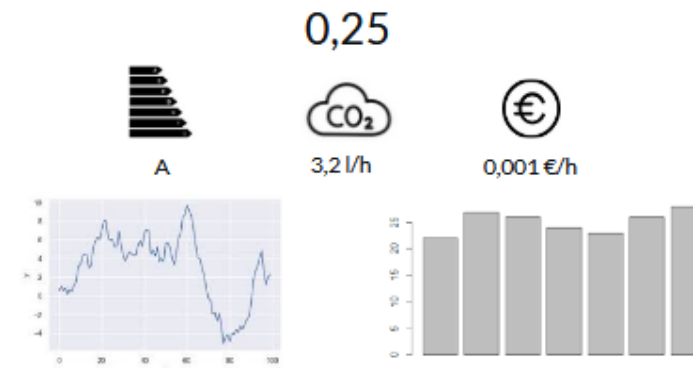
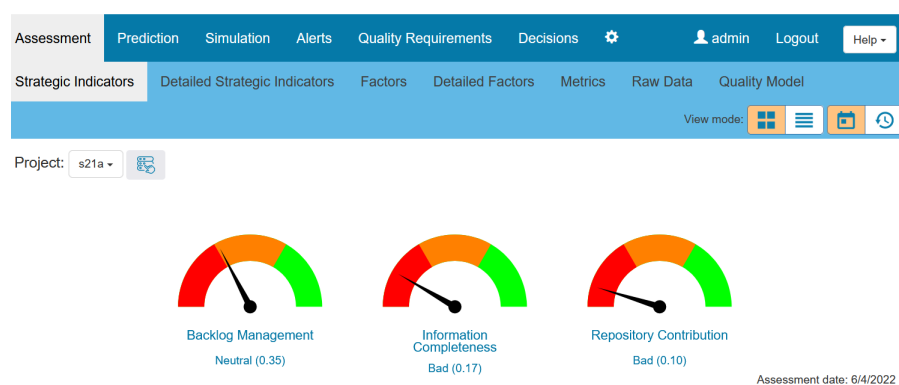
KPI	Unit of Measure	Example	Rationale	Alignment with Other Reporting Frameworks	EU Policy Reference
Total energy consumption and/or production from renewable and non-renewable sources	MWh	292 221 MWh consumed from renewable sources; 1 623 453 MWh consumed from non-renewable sources	Energy consumption and production accounts for an important proportion of GHG emissions.	TCFD Metrics and Targets, CDP Climate Change Questionnaire, GRI 302, CDSB Framework, SASB, EMAS	2030 climate & energy framework; Energy Efficiency Directive

**Further guidance**


- Fuels consumed as feedstock are not combusted for energy purposes and should not be included in calculations for this indicator.
- Include a breakdown of the different sources of renewable energy. Renewable sources of energy are those that can be naturally replenished on a human timescale, such as wind, solar, hydro, geothermal, biomass, etc. This definition excludes all fossil fuels (coal, oil, natural gas) and nuclear fuels. Waste energy should not be included if it is derived from fossil fuels.<sup>23</sup>
- When disclosing non-renewable sources of energy, make a distinction between low carbon sources and other sources of non-renewable energy.

# Software engineering for Energy efficiency (2/2)

- Leveraging strategic indicators and quality management tools from software engineering to other domains who need to follow EU guidelines by 2030
- Applied research with exploitation of Q-Rapids (H2020) and measurement programs
- Need to move to other domains and study standards → companies that have to report climate-related information)
- Energy profiling tools, e.g., code carbon



# Climate crisis: dual role of ML-based systems

- ML-based software systems for energy efficiency
    - They can help reduce the effects of climate crisis
    - Modeling climate change predictions
  
  - Energy efficiency of ML-based systems
    - ML-based systems are themselves a significant emitter of carbon
-  Our focus

Dhar, P. (2020). The carbon impact of artificial intelligence. *Nature Machine Intelligence*, 2(8), 423-425

# Energy consumption in ML-based systems:

## Motivation (1/4)



- Highly accurate ML-based systems have led to a dramatic growth in AI data volume, models' size, and infrastructure capacity → exponential scaling of ML with significant energy and environmental footprint implications
  - Schwartz et al. report that the cost of training DL models as required by state-of-the-art techniques has increased by a factor of 300.000x in only 6 years, doubling every 3-4 months
  - Wu et al. show that the carbon footprint of training one large ML model for autonomous vehicles is equivalent to 242,231 miles (389,833 km) driven by an average passenger vehicle

# Energy consumption in ML-based systems: Motivation (2/4)

## Common carbon footprint benchmarks

in lbs of CO2 equivalent

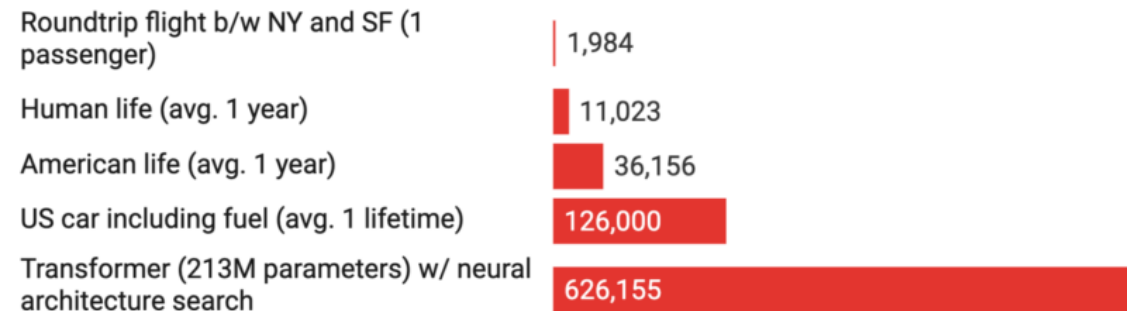
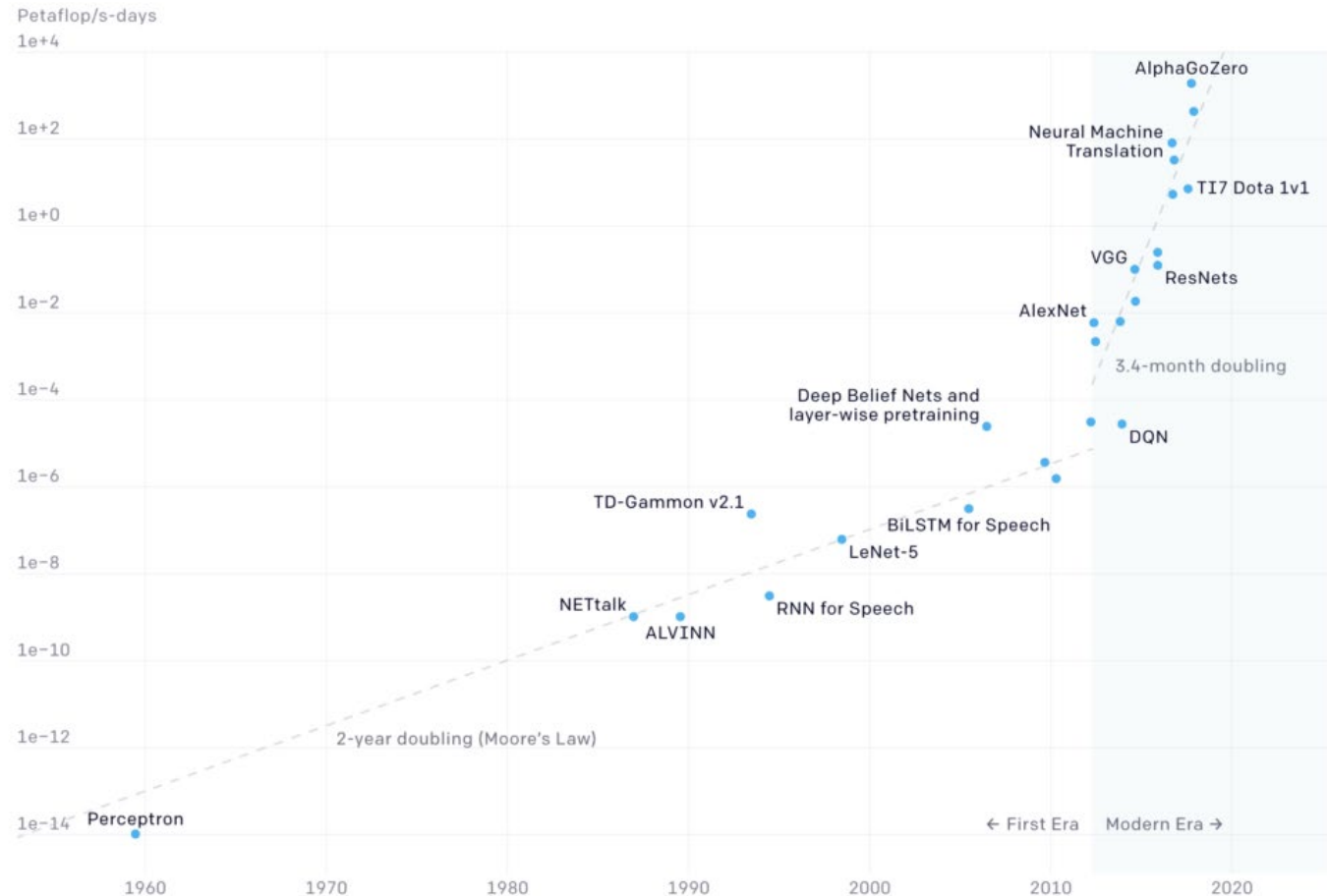


Chart: MIT Technology Review • Source: Strubell et al. • Created with Datawrapper

## Two Distinct Eras of Compute Usage in Training AI Systems



Source: <https://openai.com/blog/ai-and-compute/>



# Energy consumption in ML-based systems: Motivation (3/4)



- What to do? Strubell et al. proposed three main actionable recommendations to reduce costs and improve equity of ML-based systems based on natural language processing
  - reporting training time and sensitivity to hyperparameters
  - fostering equitable access to computation resources
  - prioritizing computationally efficient hardware and algorithms

# Energy consumption in ML-based systems: Motivation (4/4)

- ML-based systems have become increasingly pervasive in our lives, requiring **extremely high computational resources**
  - **Red** AI (cost of executing the model on a single example, size of the training dataset, number of hyperparameters experiments)
  - Examples:
    - Google's BERT-large (350 million features, trained for 2.5 days using 512 TPU chips, costing \$60K+)
    - Open-GPT3 (175 billion features)
    - AlphaGo (1920 CPUs, 280 GPUs, costing \$35M)
- When modelling and developing green AI-based systems, energy efficiency must be better understood, defined, reported, and managed in order to **deliver AI-based systems with less demanding computational power needs**
  - **Green** AI (yielding novel results without increasing computational cost, and ideally reducing it)
- Considering not only accuracy, but also energy, time, reproducibility, reuse

Schwartz, R., Dodge, J., Smith, N. A., Etzioni, O. (2020). Green AI. *Communications of the ACM* 63(12): 54-63

# The GAISSA project

# The GAISSA project

<b>Project number: TED2021-130923B-I00</b>		<b>Project acronym: GAISSA</b>
<b>Title</b>	<b>Towards green AI-based software systems: an architecture-centric approach</b>	
Start date	1-Dec-2022	
Duration	24 months (until 30-Nov-2024)	
Call	Spanish Program with Next Generation EU funds Transición Ecológica y Digital	
Keywords	Green AI, energy efficient ML systems, sustainable software engineering	
Budget	277.035 Euros	
Project Coordinators	Silverio Martínez-Fernandez, Xavier Franch (UPC)	
Website	<a href="https://gaissa.upc.edu/en">https://gaissa.upc.edu/en</a>	

# Hypothesis & Goal

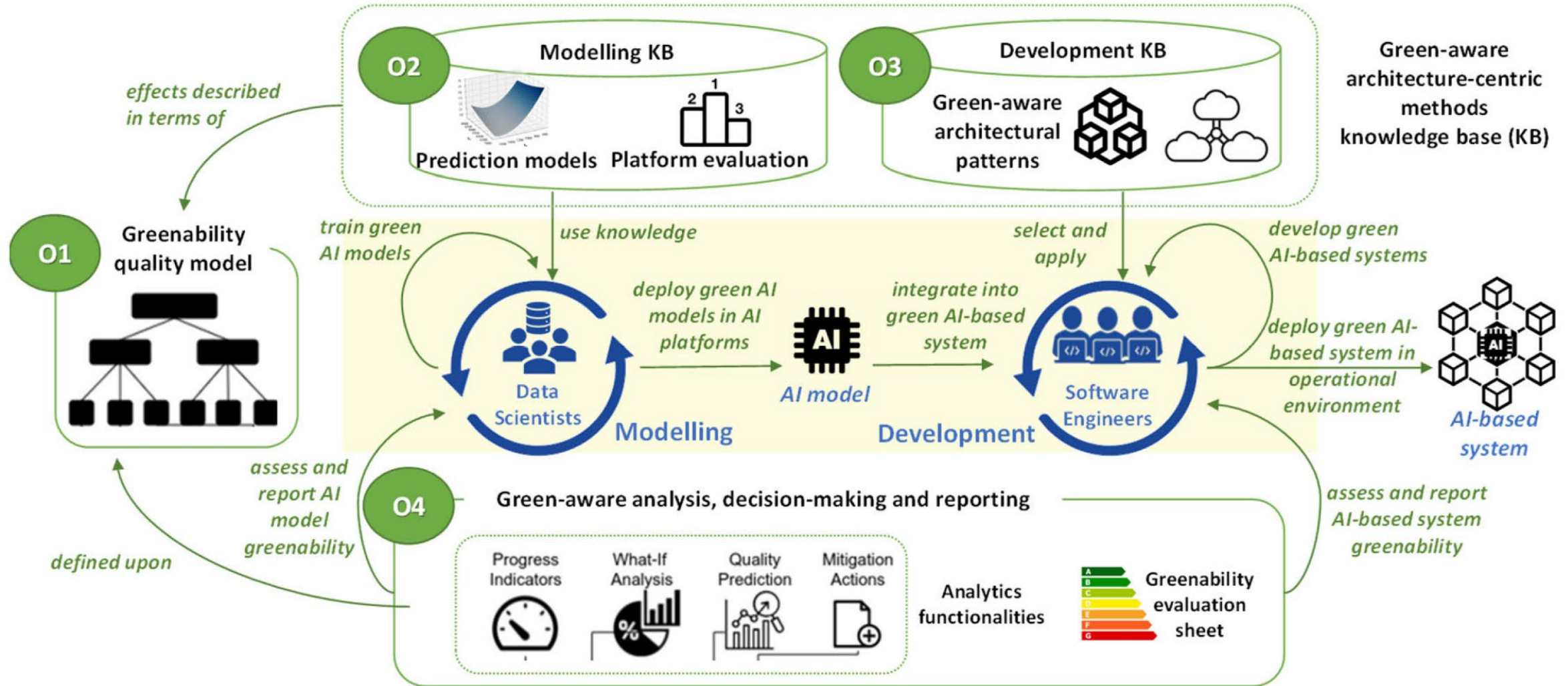
## Hypothesis

- When modelling and developing green AI-based systems,
- the **impact of architectural decisions on energy efficiency** must be **better understood, defined, reported, and managed**
- in order to **deliver AI-based systems with less demanding computational power needs**

## Goal

- Provide data scientists and software engineers tool-supported, architecture-centric methods for the modelling and development of green AI-based systems

# Roadmap



# Objectives

## Main objective

To provide **data scientists** and **software engineers** tool-supported, **architecture-centric methods** for the modeling and development of green **AI-based systems**.

## Specific Objective 1

Define, implement, and evaluate **a quality model for measuring the greenability** of AI-based systems.

## Specific Objective 2

Define, implement and evaluate **architecture-centric methods to guide the training and deployment of green AI models** and measure energy efficiency of AI platforms.

## Specific Objective 3

Define, implement and evaluate **architecture-centric methods to guide the development of green AI-based systems**.

## Specific Objective 4

Define, implement and evaluate **analytic tools to support greenability-driven analysis and decision-making** during the modeling and development of AI-based systems.

# First results: Energy efficiency of training neural networks architectures

Xu, Y., Martínez-Fernández, S., Martínez, M., & Franch, X. (2023). Energy Efficiency of Training Neural Network Architectures: An Empirical Study. Proceedings of the 56th Hawaii International Conference on System Sciences, pp. 781-790.

<https://hdl.handle.net/10125/102727>



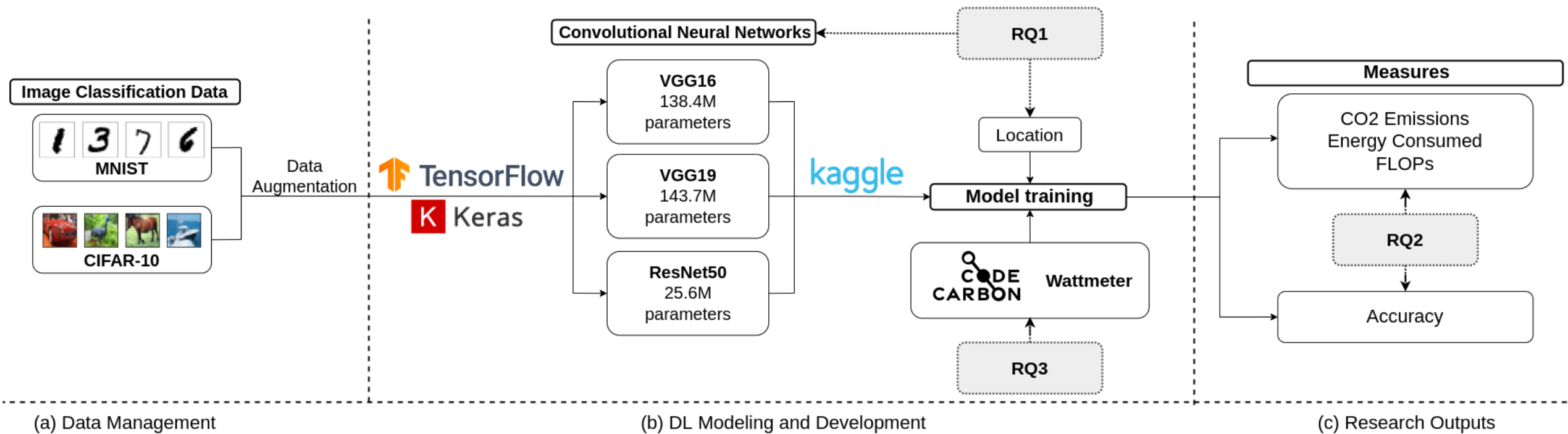
# Research goal

- *Analyze* convolutional neural networks architectures
- *with the purpose of* measuring their energy efficiency
- *with respect to* the model training
- *from the point of view of* the AI practitioner
- *in the context of* creating an image classification model for computer vision.

# Research questions

- RQ1: Does the **CNN architecture** have an impact on **energy consumption**?
- RQ2: What is the **relationship between model accuracy and the energy** needed to train the model?
- RQ3: What are the **differences between software-based and hardware-based methods of measuring** the energy efficiency of a model?

# Research: energy efficiency of ML-based systems



Y. Xu, S. Martínez-Fernández, M. Martínez, and X. Franch,  
 Energy efficiency of training neural network architectures: An empirical study, HICSS 2023

**Table 1. Independent, dependent and other variables of the study.**

<b>Class</b>	<b>Name</b>	<b>Description</b>	<b>Scale</b>	<b>Operationalization</b>
Independent	Architecture Type	The deep CNN architecture	nominal	See section 3.3.1
	Measuring instrument	Energy measuring method (by hardware or software)	nominal	See section 3.3.1
Dependent	Emissions	Carbon dioxide (CO <sub>2</sub> ) emissions, expressed as kilograms of CO <sub>2</sub> -equivalents (CO <sub>2</sub> -eq)	numerical	Profiled
	Energy consumption	Net power supply consumed during the compute time, measured as kWh	numerical	See measuring method
	Floating-point operations	Number of floating point operations per second (FLOP)	numerical	Retrieved from modeling
	Accuracy	Validation accuracy obtained after training	numerical	Retrieved from modeling
Others	Dataset	The input dataset used to train the models	nominal	See section 3.3.3
	Hardware	GPU and CPU type	nominal	Profiled
	Location	Province/State/City where the compute infrastructure is hosted	nominal	Profiled

# Energy in ML-based systems: Key metrics (1/5)

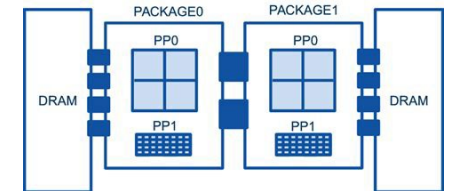
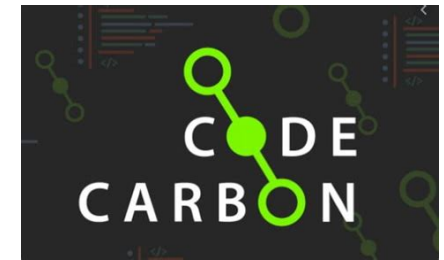
- Real measured energy consumption
  - Net power supply consumed during the compute time
    - measured as kWh
  - Benefits: accurate, easy to map to carbon emissions
  - Drawbacks: needed hardware to measure, hard to measure
- How to measure?
  - E.g., with wattmeters



Luís Cruz, Sustainable Software Engineering

# Energy in ML-based systems: Key metrics (2/5)

- Estimated energy consumption
  - Net power supply consumed during the compute time, measured as kWh
  - Benefits: easy to measure, correlates with energy consumption in most cases.
  - Drawbacks: difficult to compare with measurements from other setups
- How to measure it?
  - E.g., CodeCarbon profiler



Luís Cruz. Sustainable Software Engineering

# Energy in ML-based systems: Key metrics (3/5)

- Code Carbon
  - Python package that enables us to track emissions in order to estimate the carbon footprint of an experiment
  - uses RAPL for measuring the energy consumed by the CPU and RAM, and NVIDIA Management Library (NVML) for the energy consumption of the GPU
  - The package logs the data of each experiment into an `emissions.csv` file:
    - Duration of the compute (in seconds)
    - Emissions as CO<sub>2</sub>-equivalents (in kg)
    - Energy consumed (in kWh)



```
from codecarbon import EmissionsTracker

tracker = EmissionsTracker()
tracker.start()
# GPU Intensive code goes here
tracker.stop()
```

# Energy in ML-based systems: Key metrics (4/5)

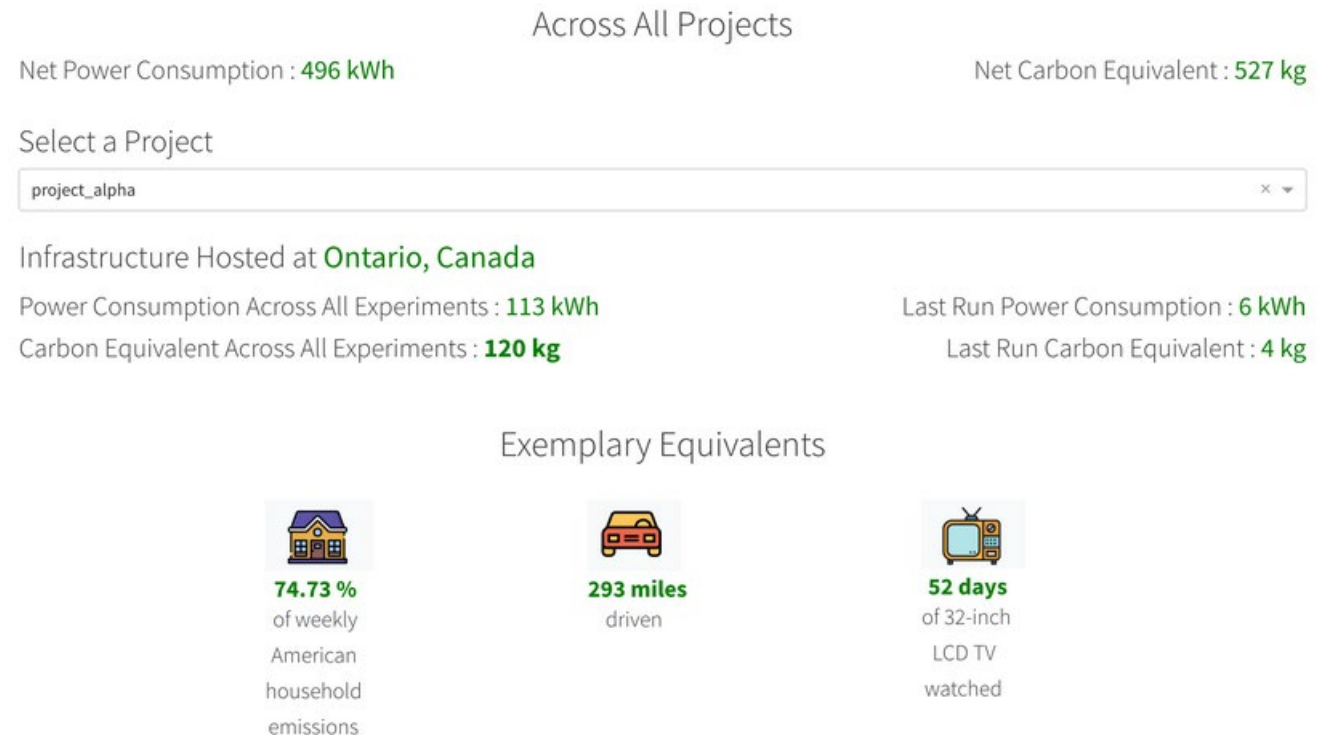
- Floating point operations (FLOPS)
  - Number of floating point operations per second
  - Benefits: comparable across different setups, cheap
  - Drawbacks: does not factor in energy consumption in memory, does not reflect carbon emissions
  - How to measure it?
    - Computing the FLOPs required for the training of the model
    - E.g., use the keras-flops4 package for TensorFlow
      - <https://github.com/tokusumi/keras-flops>



# Energy in ML-based systems: Key metrics (5/5)

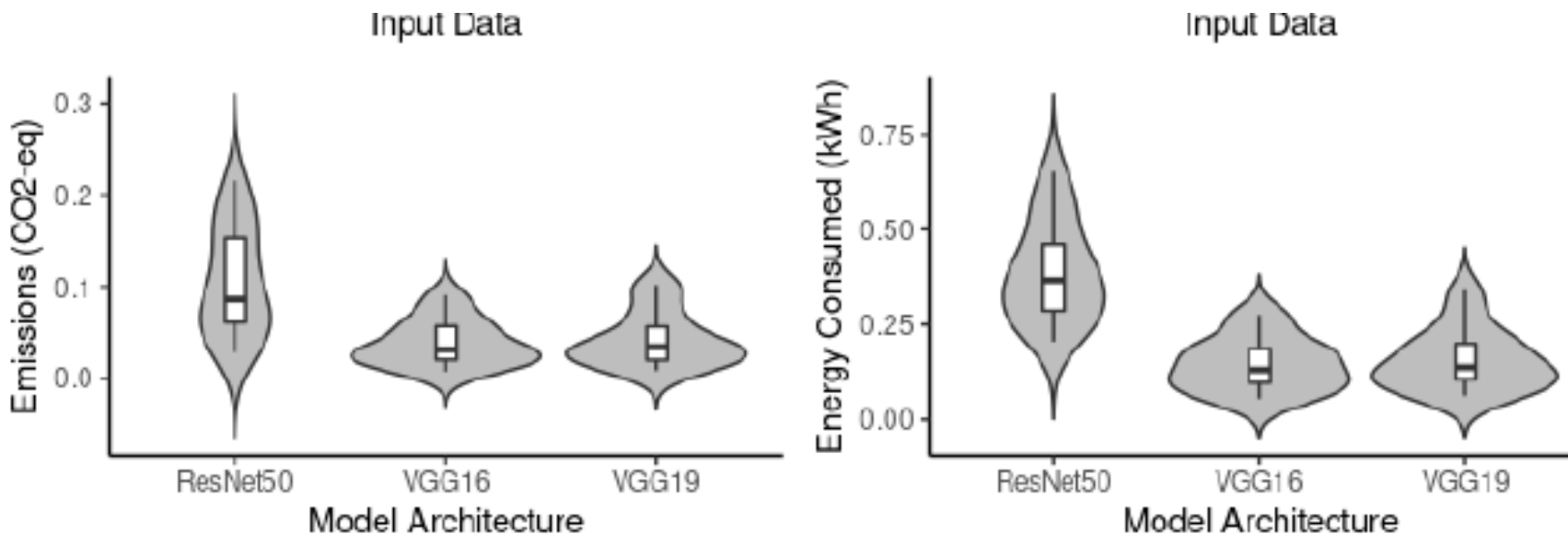
## Emissions

- Carbon dioxide (CO<sub>2</sub>) emissions
  - expressed as kilograms of CO<sub>2</sub>-equivalents (CO<sub>2</sub>-eq)
- CodeCarbon estimates the net carbon intensity
  - using the international energy mixes derived from the United States' Energy Information Administration's Emissions & Generation Resource Integrated Database (eGRID)



# RQ1: Does the CNN architecture have an impact on energy consumption?

- Yes, although it is also intuitive



# RQ2: What is the relationship between model accuracy and the energy needed to train the model?

- **Score = Accuracy/Energy**

By choosing the architecture with highest Score, we obtain either

- (i) an improvement in both accuracy and energy efficiency (e.g., models trained using MNIST dataset), or
- (ii) an improvement in energy efficiency with a detriment (small such in the case on CIFAR10 on South Carolina) on accuracy.

**Table 6. Scores of the different experiment configurations. Accuracy: validation accuracy from last epoch of training. Energy: Kilowatt per hour. Score = Accuracy/Energy.**

Location	Data	Architecture	Accuracy	Energy	Score
Oregon	CIFAR10	VGG16	0.6189	0.0583	10.63
		VGG19	0.6018	0.0493	<b>12.64</b>
		ResNet50	0.3021	0.1057	4.11
	MNIST	VGG16	0.9429	0.0879	<b>11.02</b>
		VGG19	0.9395	0.0932	10.44
		ResNet50	0.8858	0.1893	7.64
S.Carolina	CIFAR10	VGG16	0.6167	0.0667	9.26
		VGG19	0.6157	0.0574	10.88
		ResNet50	0.1	0.1224	1.17
	MNIST	VGG16	0.9459	0.0920	10.42
		VGG19	0.9384	0.1137	8.26
		ResNet50	0.8883	0.2171	6.36
Taipei	CIFAR10	VGG16	0.6191	0.0567	10.99
		VGG19	0.6147	0.0637	9.80
		ResNet50	0.2169	0.1347	2.48

# RQ3: What are the differences between software-based and hardware-based methods of measuring the energy efficiency of a model?



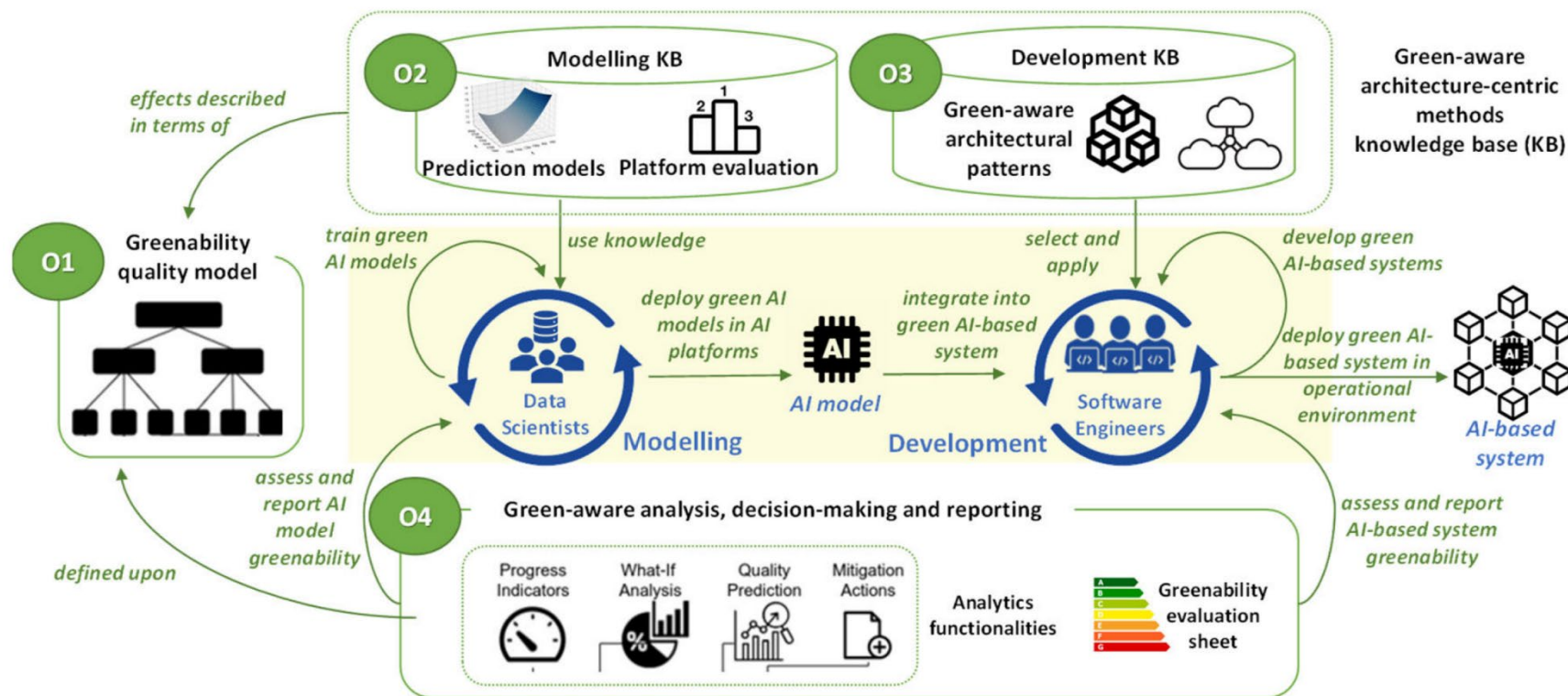
- Energy consumption returned by the wattmeter is larger than the total energy consumed reported by the profiler in the same amount of time, going from 42% to 46%. Two reasons:
  - A profiler is not analog to a power meter
  - Wattmeter also includes the energy consumed by all components from devices connected to the wattmeter

**Table 7. Energy consumption obtained using a wattmeter and a profiler, expressed in kWh. For the profiler, we present the energy consumption and the total reported by the profiler.**

Data	Archit.	Watt. (kWh)	CodeCarbon (kWh)			
			CPU	GPU	RAM	TOTAL
MNIST	VGG16	2.25	0.04	0.77	0.41	1.21
	VGG19	2.54	0.09	0.85	0.47	1.40
	ResNet50	3.03	0.05	1.08	0.59	1.72
CIFAR10	VGG16	1.48	0.06	0.52	0.28	0.86
	VGG19	1.73	0.05	0.61	0.32	0.98
	ResNet50	1.70	0.01	0.64	0.35	0.99

# Conclusions and future work

- We can definitely look for a tradeoff on accuracy and energy consumption! Future work: an architecture-centric approach



# Thank you!

- Contact: [silverio.martinez@upc.edu](mailto:silverio.martinez@upc.edu)
- The GAISSA project (TED2021-130923B-I00) is funded by MCIN/AEI/10.13039/501100011033 and by the European Union "NextGenerationEU"/PRTR.

