Code Carbon

Track your code's CO2 emissions

AMINE SABONI

MLOps Engineer, Pruna AI
Maintainer at CodeCarbon

Today’s agenda

1. Environmental impact of AI

2. CodeCarbon

3. Mitigation

1. Environmental impact of AI

~2%

Estimated share of Data Centers in global emissions

IEA : https://8billiontrees.com/carbon-offsets-credits/carbon-ecological-footprint-calculators/carbon-footprint-of-data-centers/

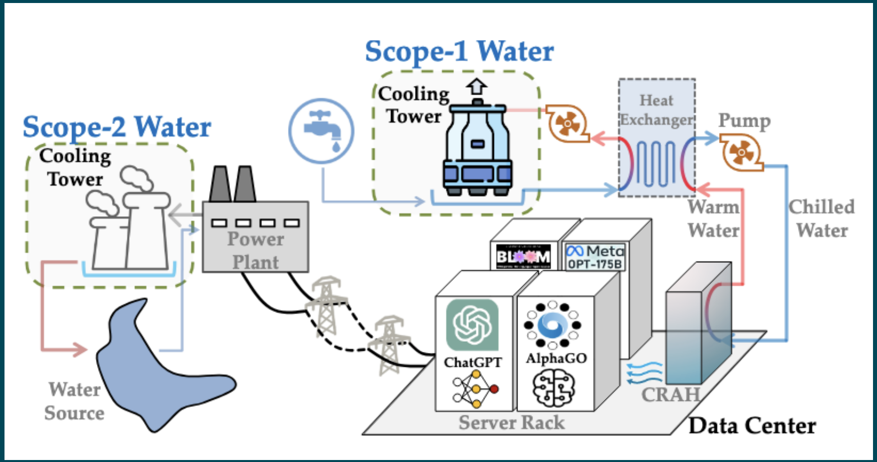
Not only C02

Water usage

Complex water consumption model, with multiple leaking points :

- Scope 1 : datacenter cooling
- Scope 2 : electricity production
- Scope 3 : chips manufacture

→ 1 to 9L of water evaporated for each kWh of electricity.
→ 4.2 to 6 billion m^3 each year (half of UK's consumption)



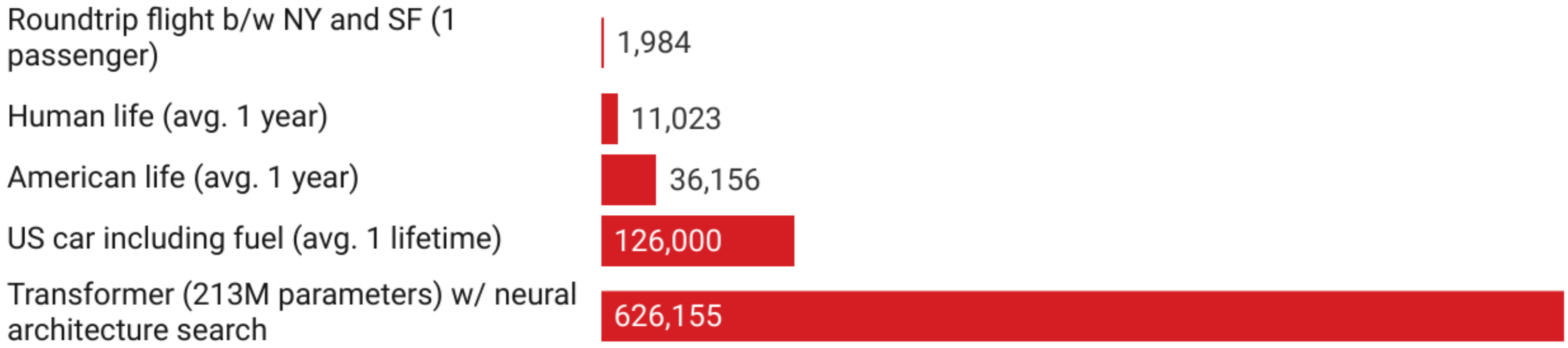
Making AI Less "Thirsty": Uncovering and Addressing the Secret Water Footprint of AI Models, Li & al, March 25

Artificial intelligence / Machine learning

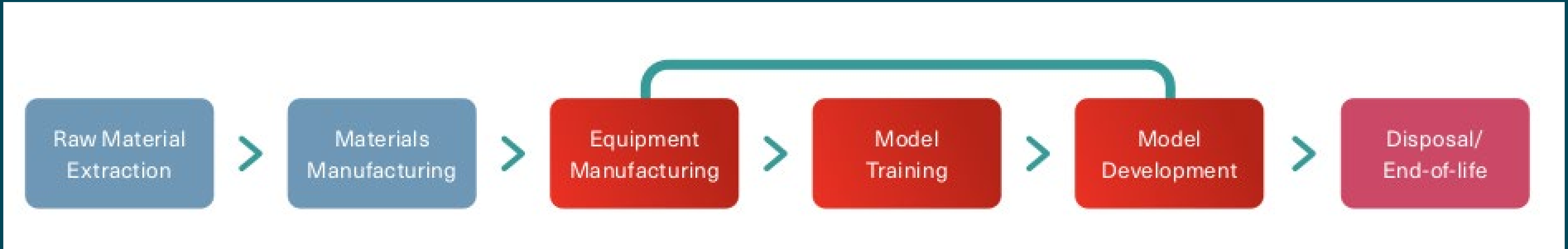
Training a single AI model can emit as much carbon as five cars in their lifetimes

Common carbon footprint benchmarks

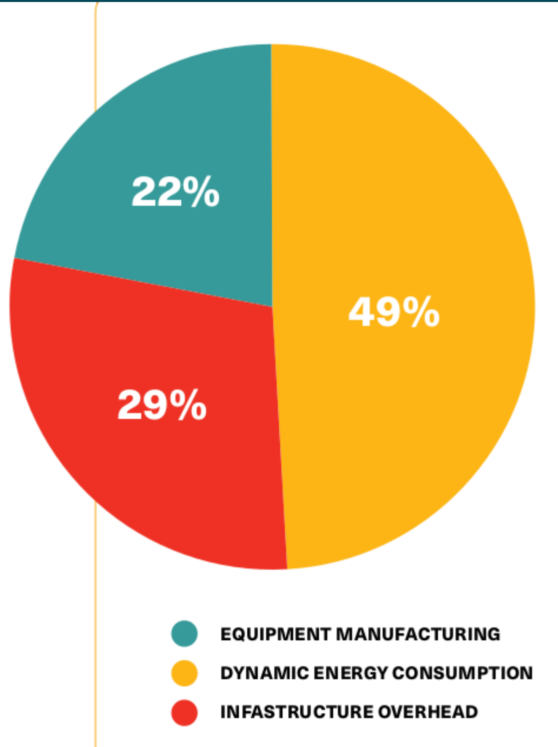
in lbs of CO2 equivalent

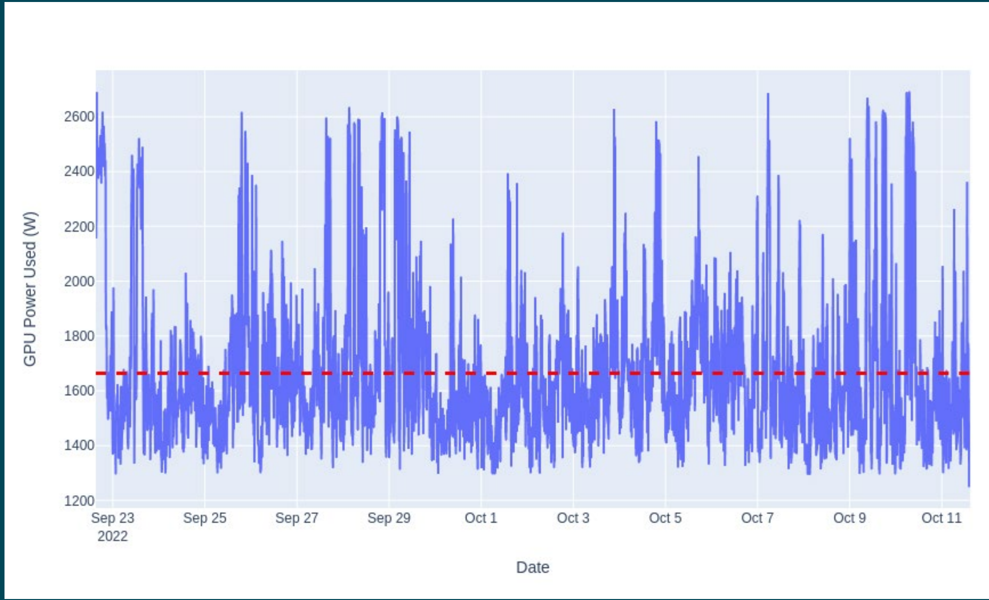


Above and beyond training

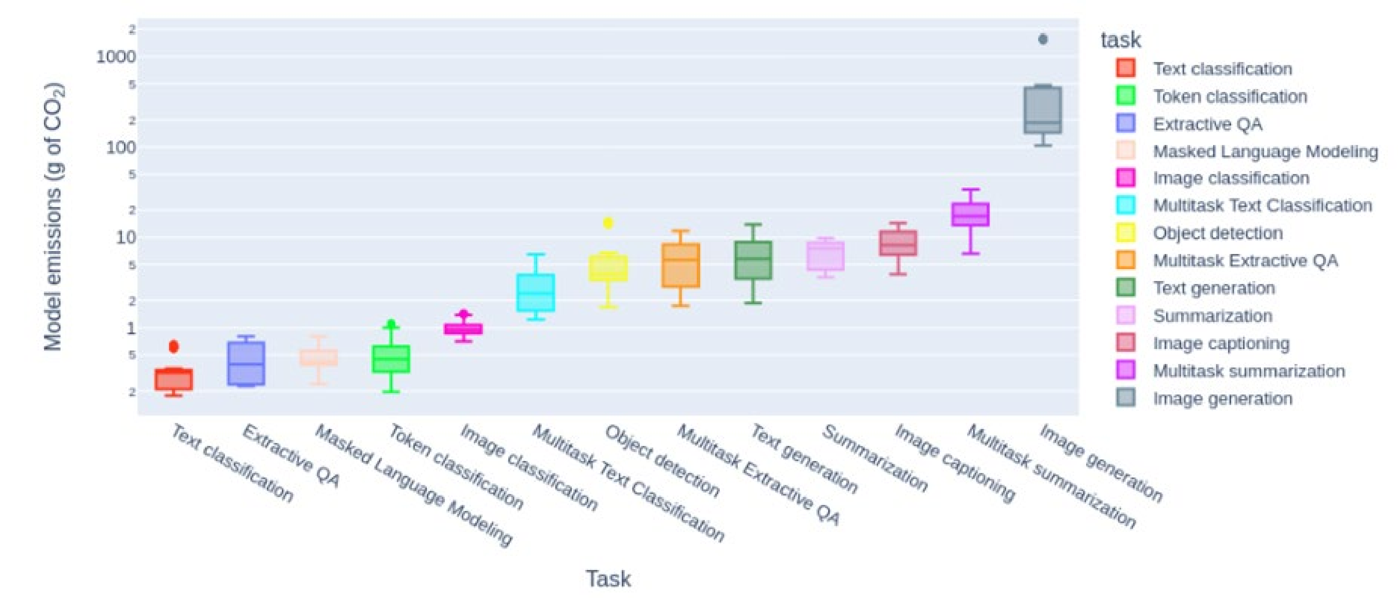


Above and beyond training



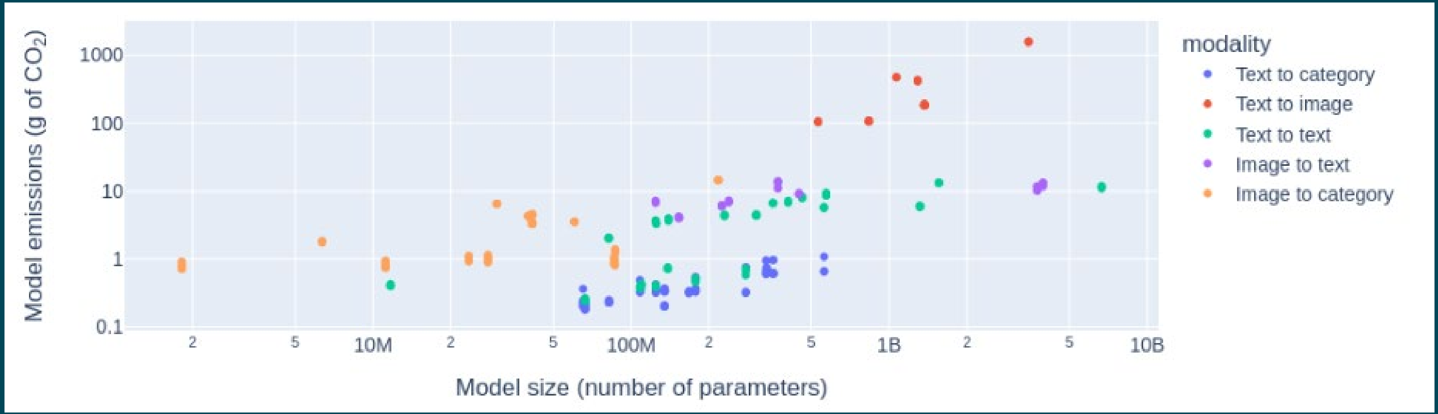


All tasks are not equivalent



[Power Hungry Processing: Watts Driving the Cost of AI Deployment? Sasha Luccioni & al]

Key findings



Modality matters

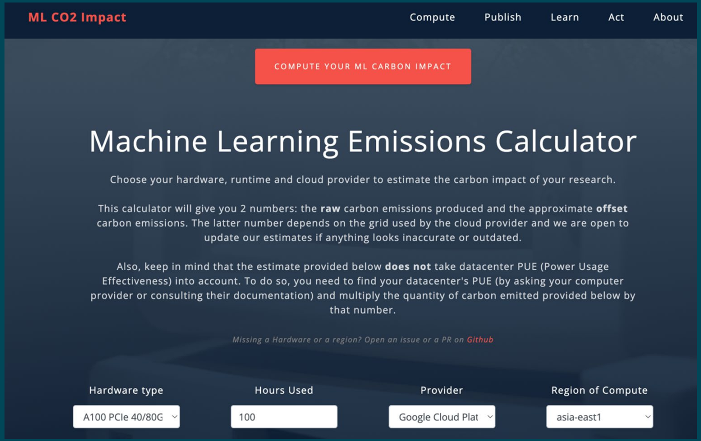
	BLOOMz-7B	BLOOMz-3B	BLOOMz-1B	BLOOMz-560M
Training energy (kWh)	51,686	25,634	17,052	10,505
Finetuning energy (kWh)	7,571	3,242	1,081	543
Inference energy (kWh)	1.0×10^{-4}	7.3×10^{-5}	6.2×10^{-5}	5.4×10^{-5}
Cost parity (# inferences)	592,570,000	395,602,740	292,467,741	204,592,592

Table 5. The BLOOMz models from our study with their training energy cost (from [31]), finetuning energy cost (from [34]), inference cost (from the present study), and cost parity, as the number of inferences required to sum to the training cost.

Inference adds up

2. CodeCarbon

A bit of History

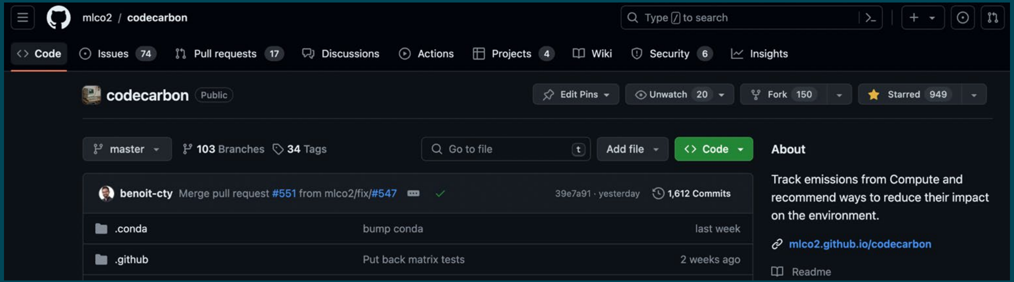


2019 : MLC02, online tool to quantify ML emissions by researchers from MILA Research Institute

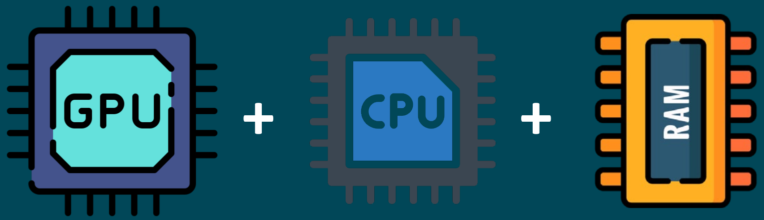
2020 : CodeCarbon launch, open source python package to measure any code (not just ML) emissions

2021 : Development of the CodeCarbon API and Dashboard with Data For Good

2023 : Creation of a non-profit in France to support the project



Hardware energy consumption



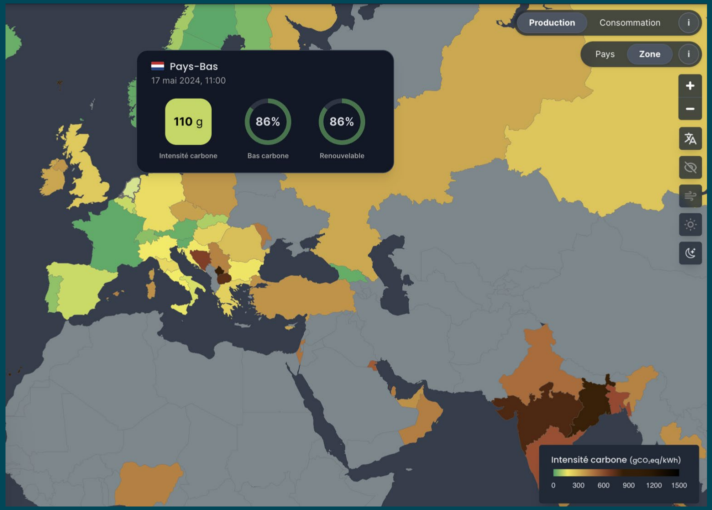
nvidia-smi
RocM (AMD)

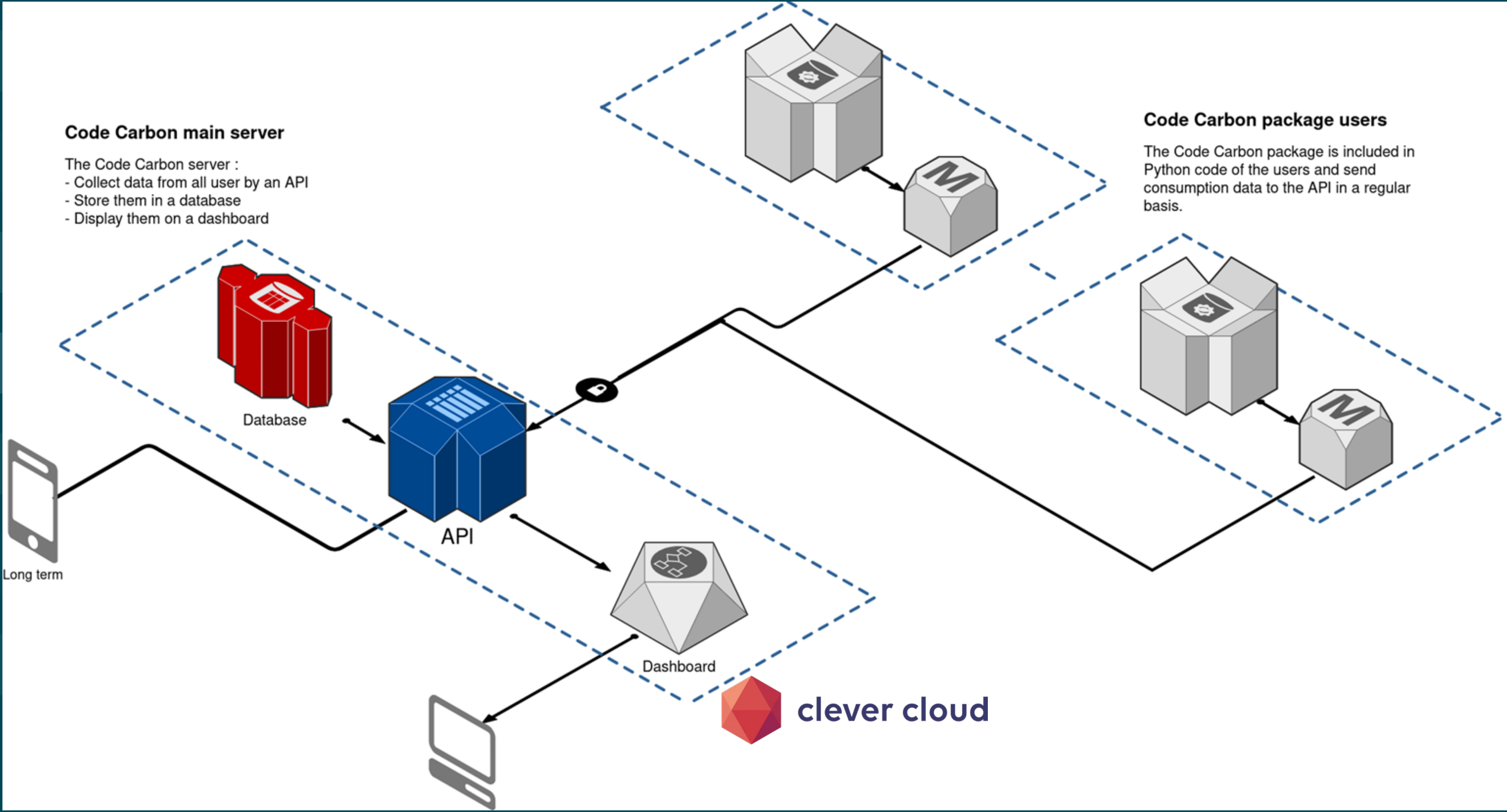
RAPL
psutils

RAPL
psutils

x

Electrical mix of workload data center

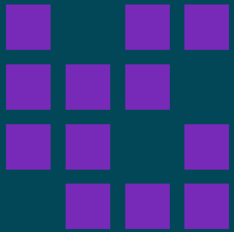



Code Carbon main server
The Code Carbon server :
- Collect data from all user by an API
- Store them in a database
- Display them on a dashboard
Database
API
Long term
Dashboard
clever cloud
Code Carbon package users
The Code Carbon package is included in Python code of the users and send consumption data to the API in a regular basis.

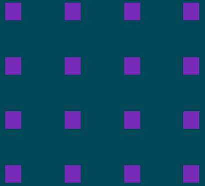
3. Mitigation

Model compression & optimization

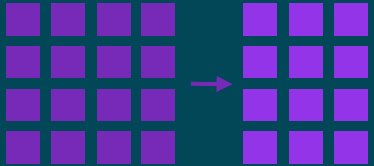
Pruning



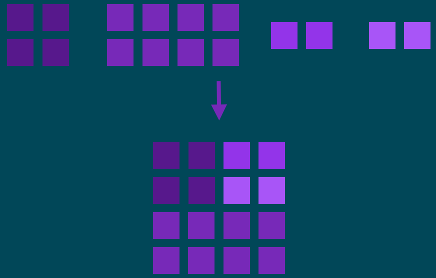
Quantization



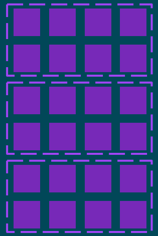
Distillation



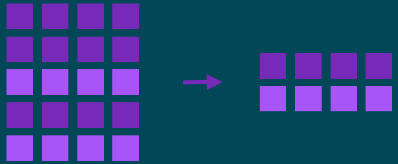
Compilation



Batching



Caching



Reducing Code's emissions

As a developer

Build lighter websites
Pause Copilot when not needed
Keep hardware longer

As a data scientist

Use fine-tuned models
Location and time of trainings
Use optimization tools like Zeus or Pruna
Keep hardware longer

As a user

Use AI systems responsibly
- Smallest model for your need

Keep you hardware longer

In IT Operations

Pick hosting location according to carbon intensity
Keep hardware longer

Thanks !

See you on https://github.com/mlco2/codecarbon !

Data For Good

clever cloud

moz://a

Mozilla Tech Fund 2024