ENERGY AND ENVIRONMENTAL IMPACTS OF DEEP LEARNING

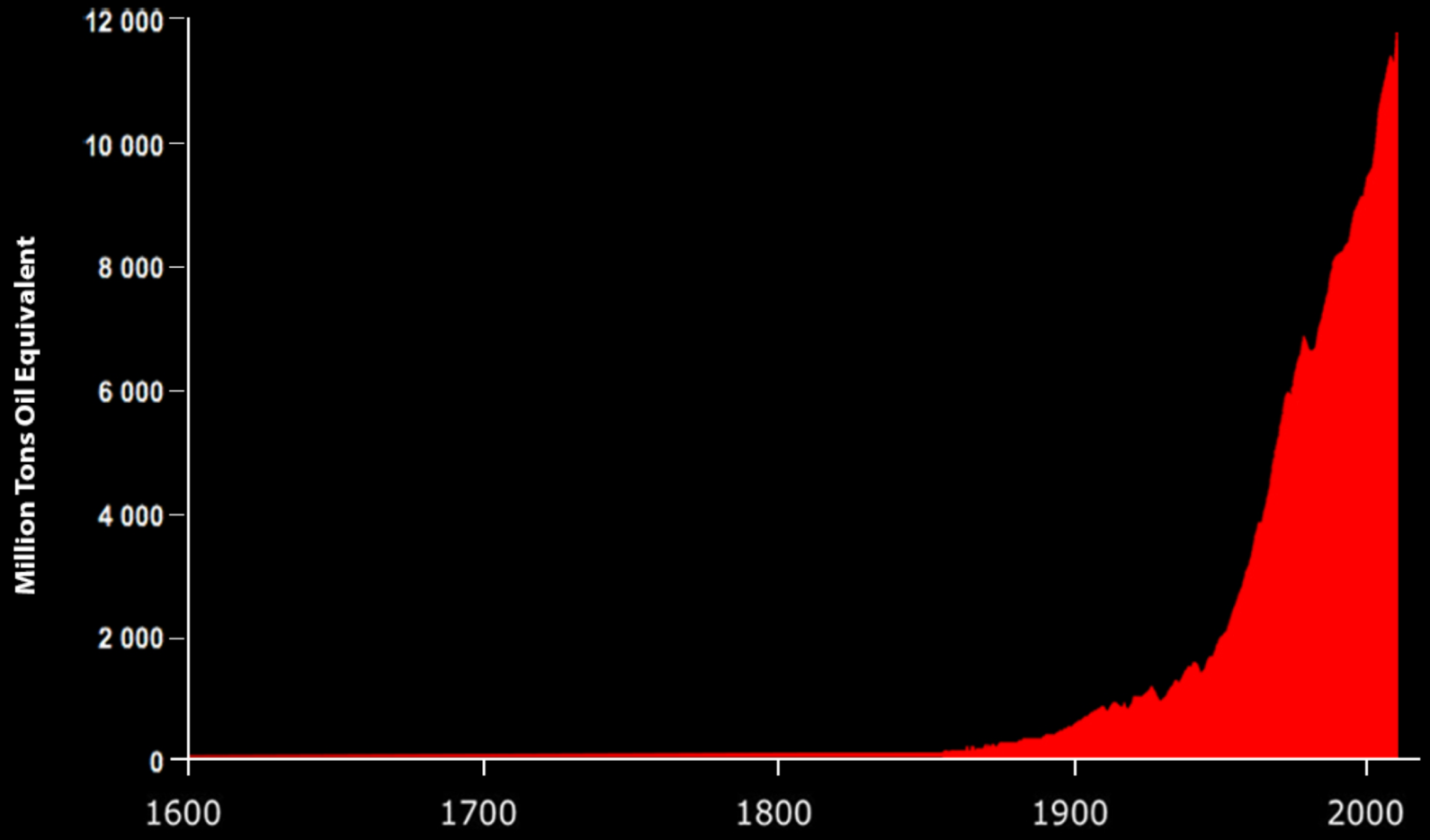


Constance Douwes

Maîtresse de conference en informatique

Centrale Méditerranée / Laboratoire d'Informatique et Systèmes

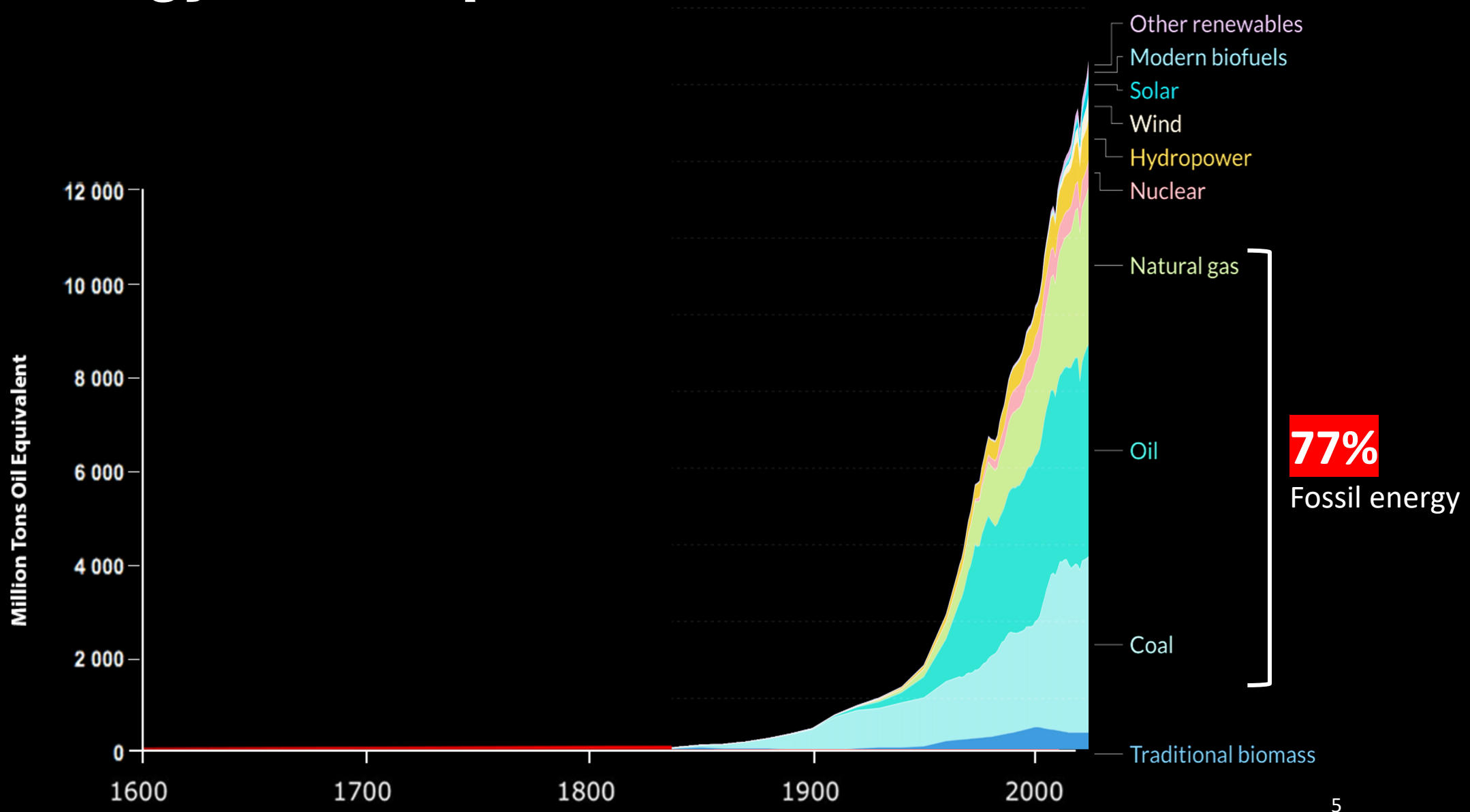
Global energy consumption



Source : Jancovici 2007

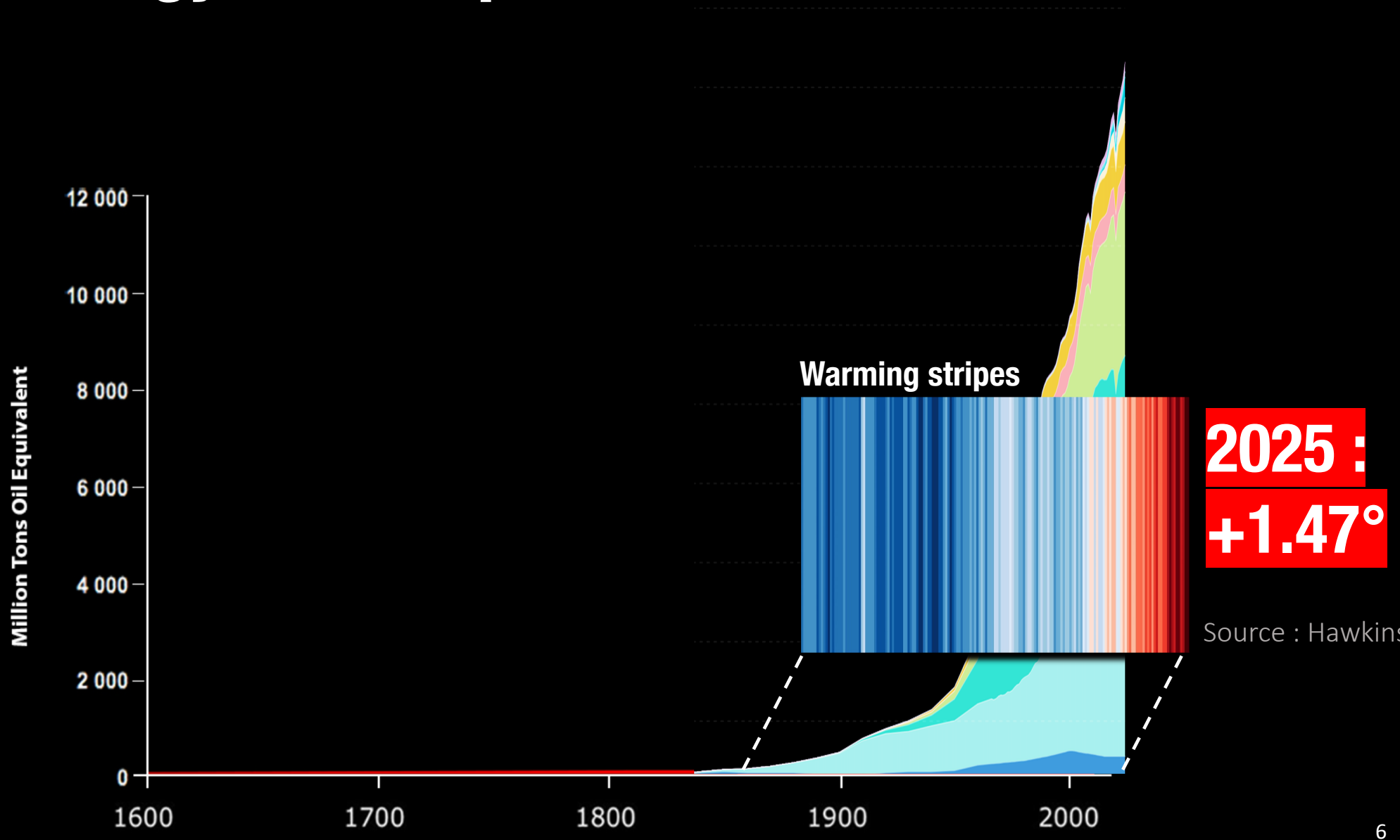
© Constance Douwes

Global energy consumption



© Constance Douwes

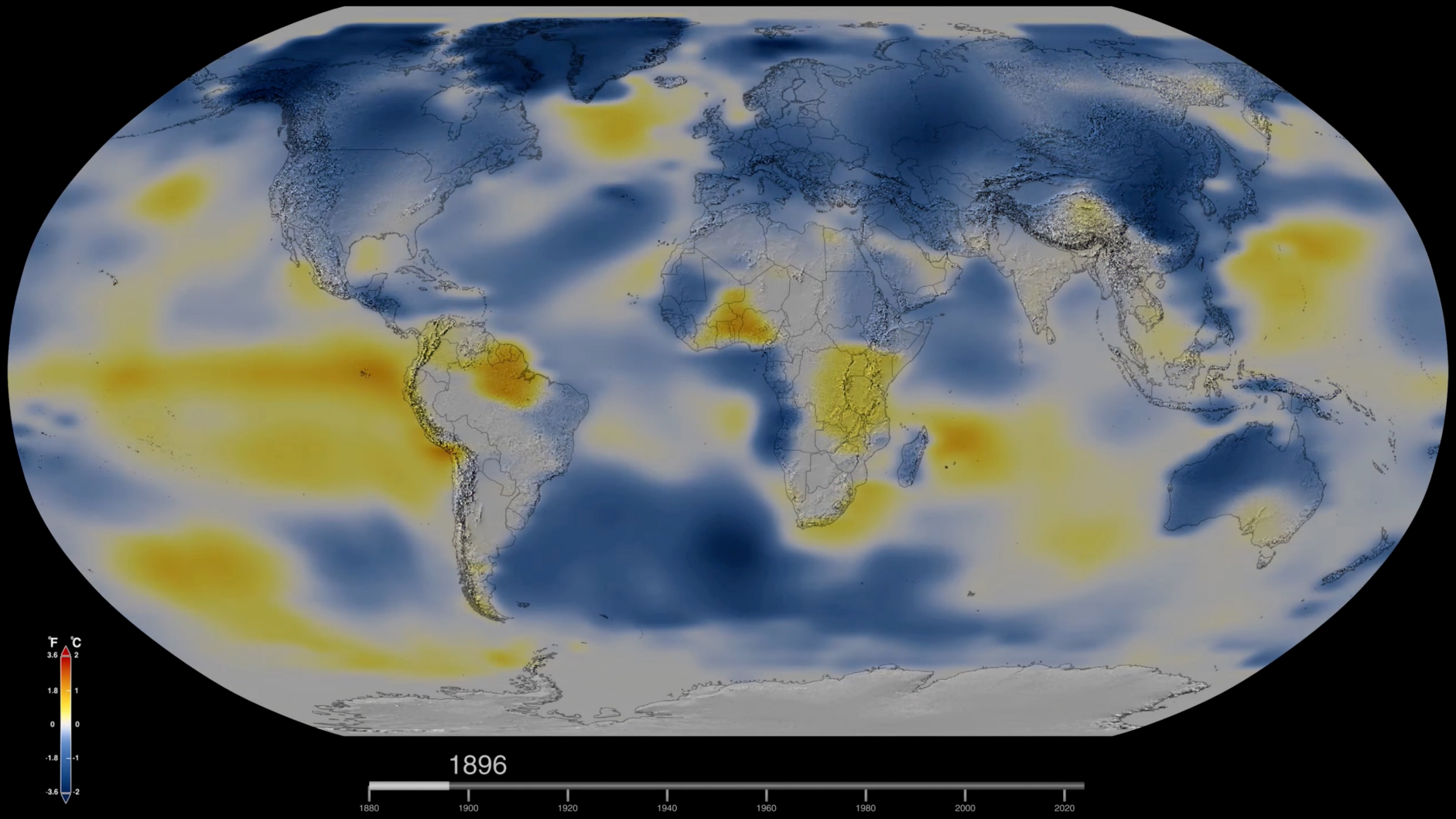
Global energy consumption



Source : Hawkins 2023

© Constance Douwes

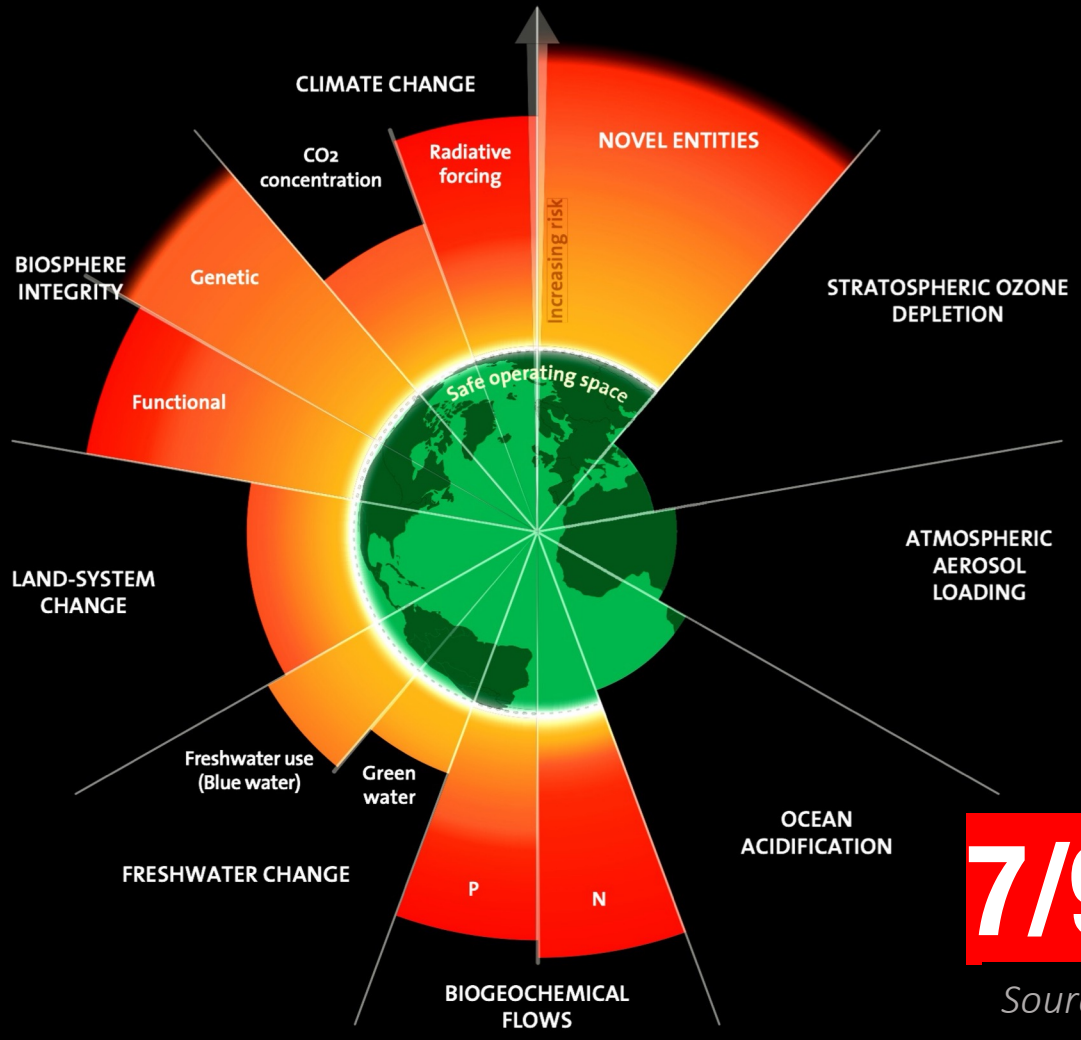
Global warming



© Constance Douwes

https://www.youtube.com/watch?v=8ZdDWQTgP2Y

Planetary bounderies

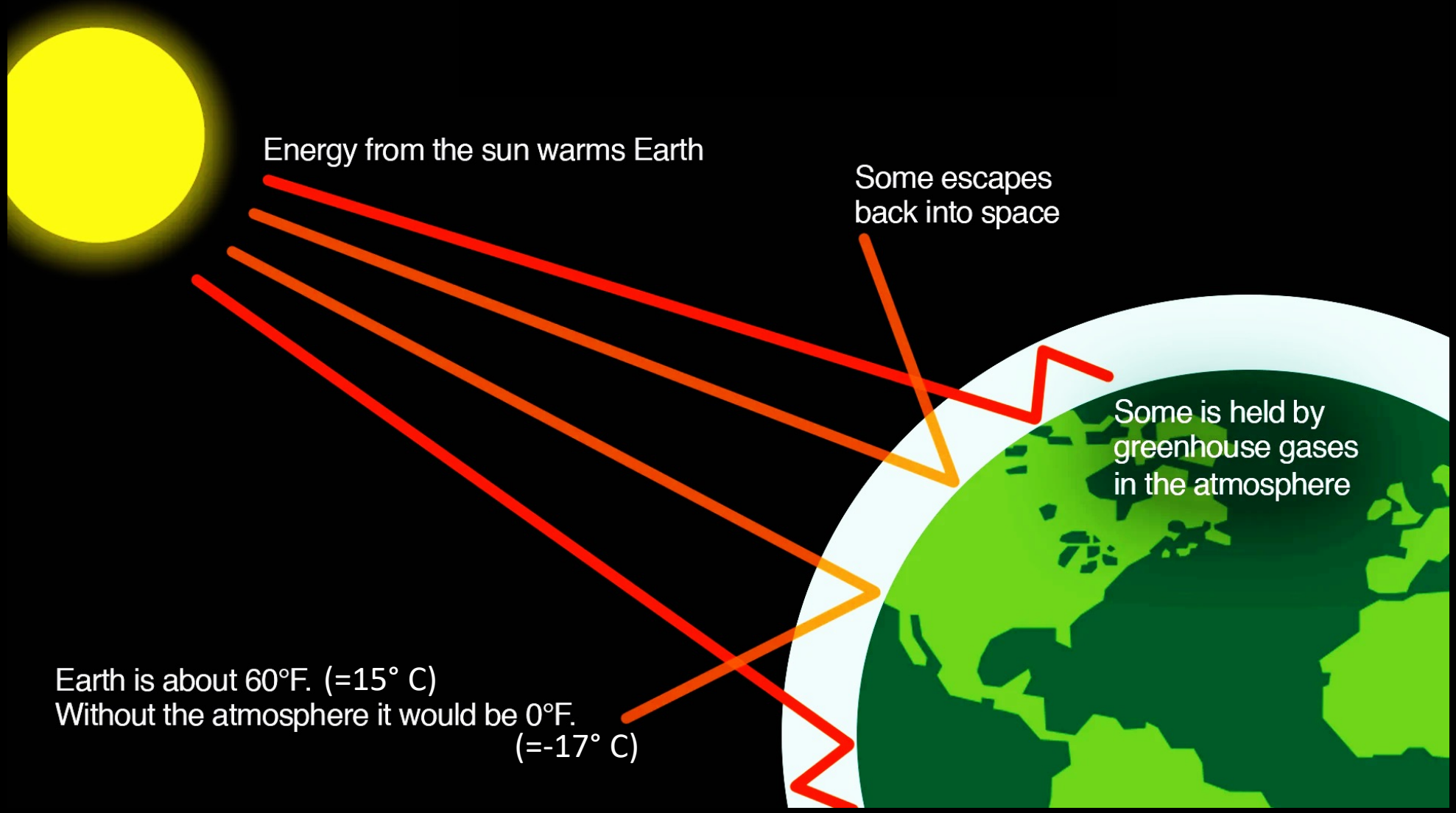


7/9 **Planetary bounderies**

Source : Richardson 2023

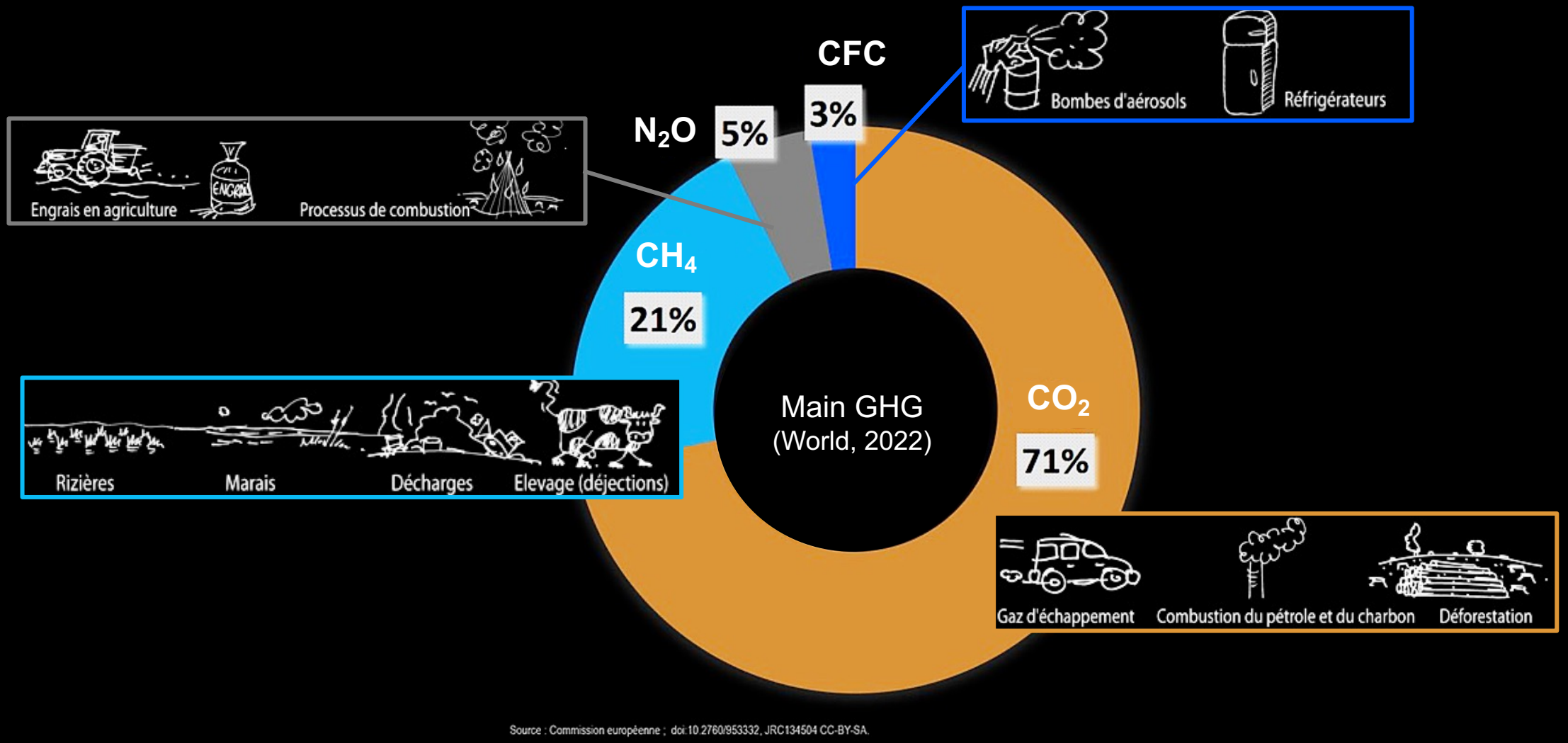
© Constance Douwes

Greenhouse effect



© Constance Douwes

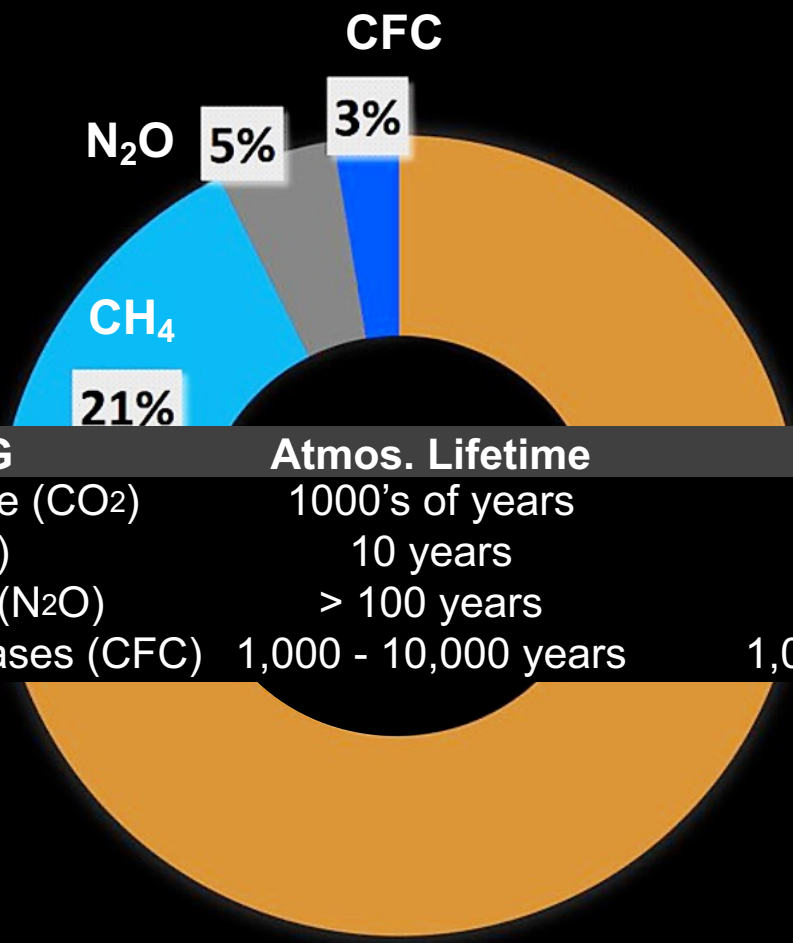
Main greenhouse gas (GHG)



Source : Commission européenne ; doi:10.2760/953332, JRC134504 CC-BY-SA.

© Constance Douwes

Main greenhouse gas (GHG)

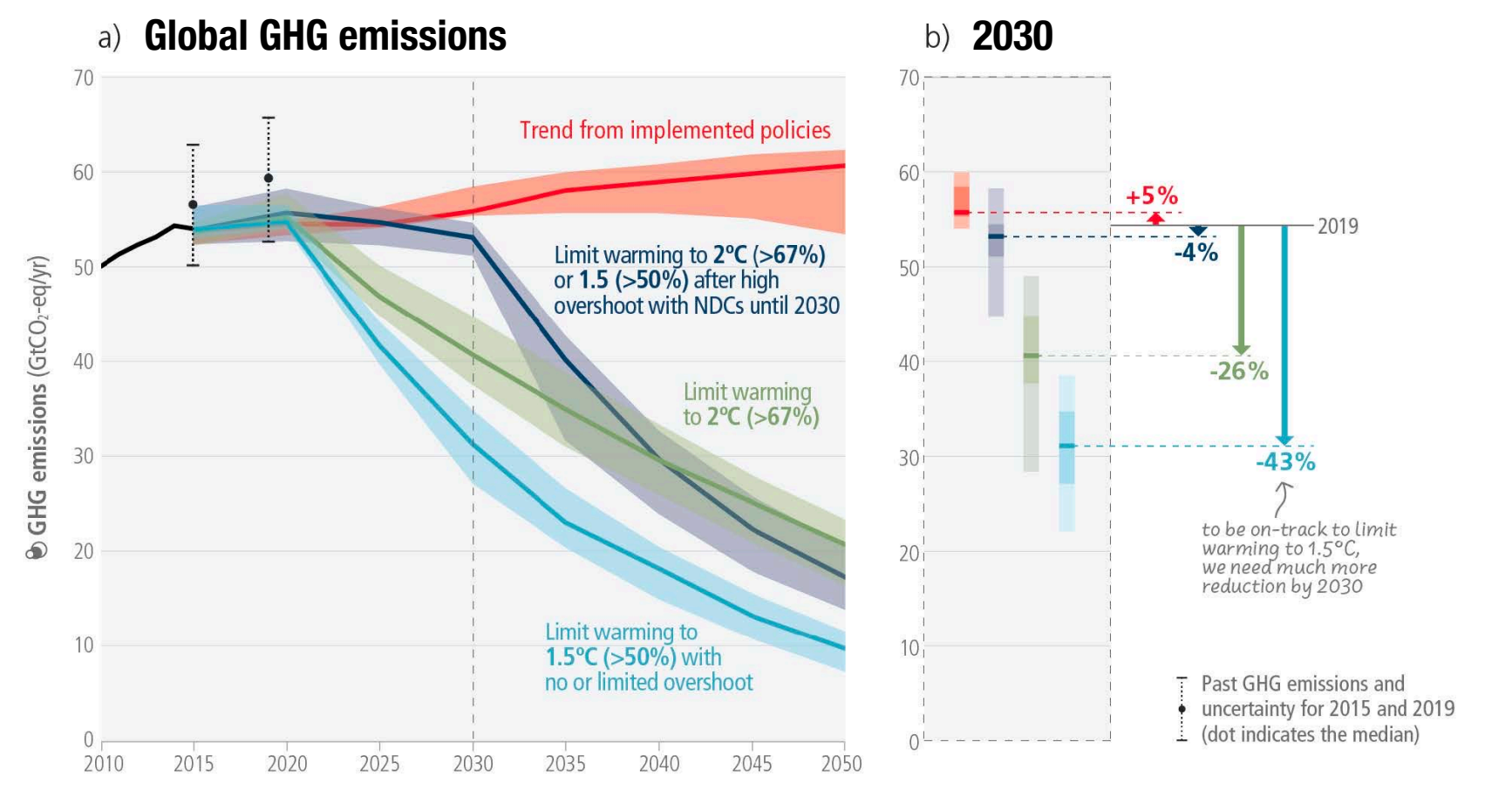


GHG	Atmos. Lifetime	GWP
Carbon Dioxide (CO_2)	1000's of years	1
Methane (CH_4)	10 years	25
Nitrous Oxide (N_2O)	> 100 years	298
Fluorinated Gases (CFC)	1,000 - 10,000 years	1,000 -10,000

Source : Commission européenne ; doi:10.2760/953332, JRC134504 CC-BY-SA.

© Constance Douwes

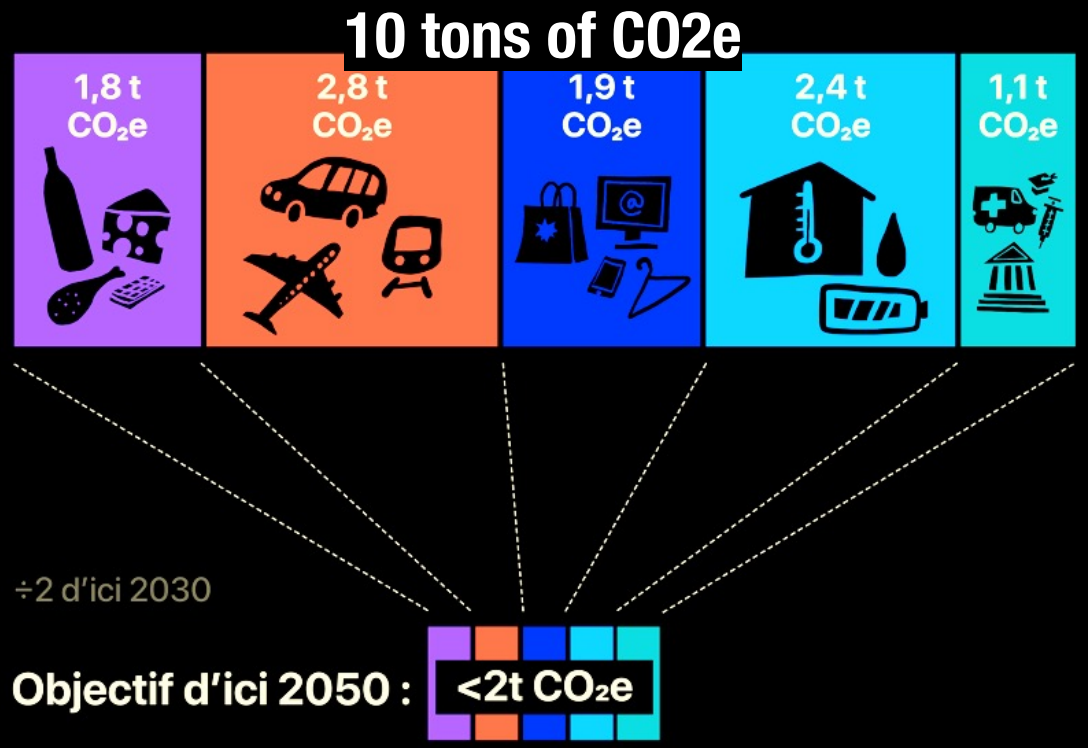
Reducing carbon emissions



Source : Sixth Assessment Report 2023

© Constance Douwes

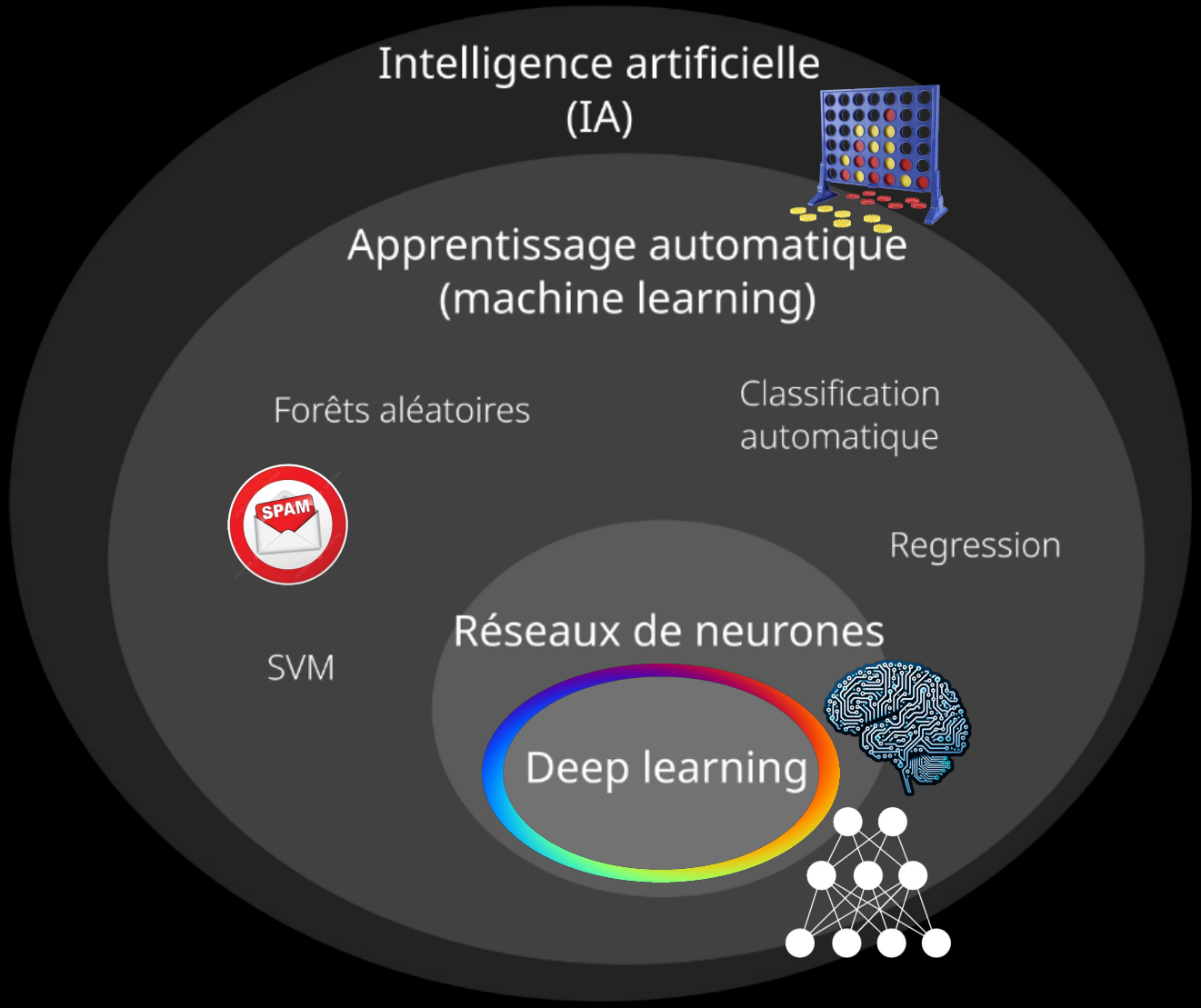
Emissions per year per capita



Sources : Kit Inventons nos vies bas carbone (Fév. 2021), Rapport sur l'état de l'environnement en France (Déc. 2020)

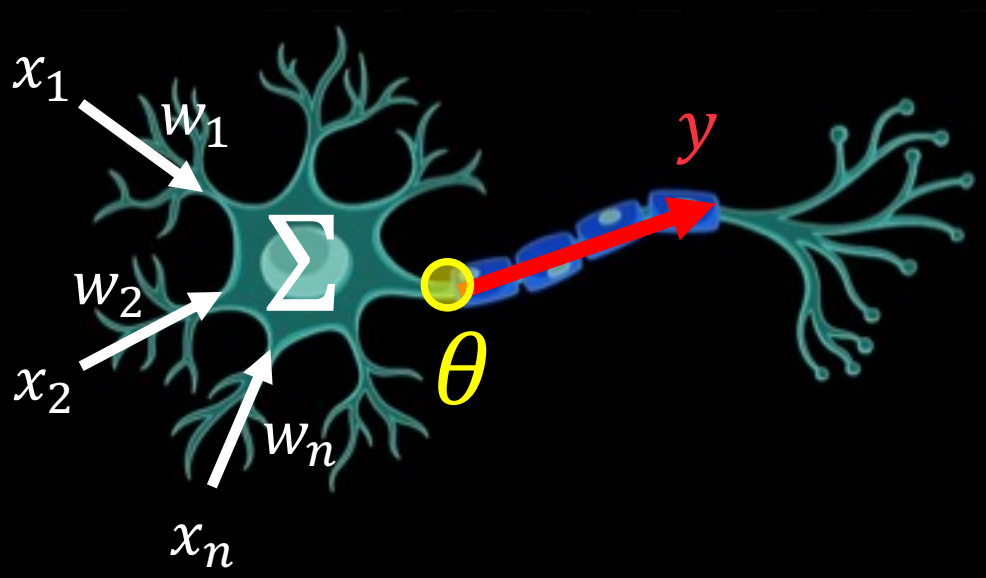
© Constance Douwes

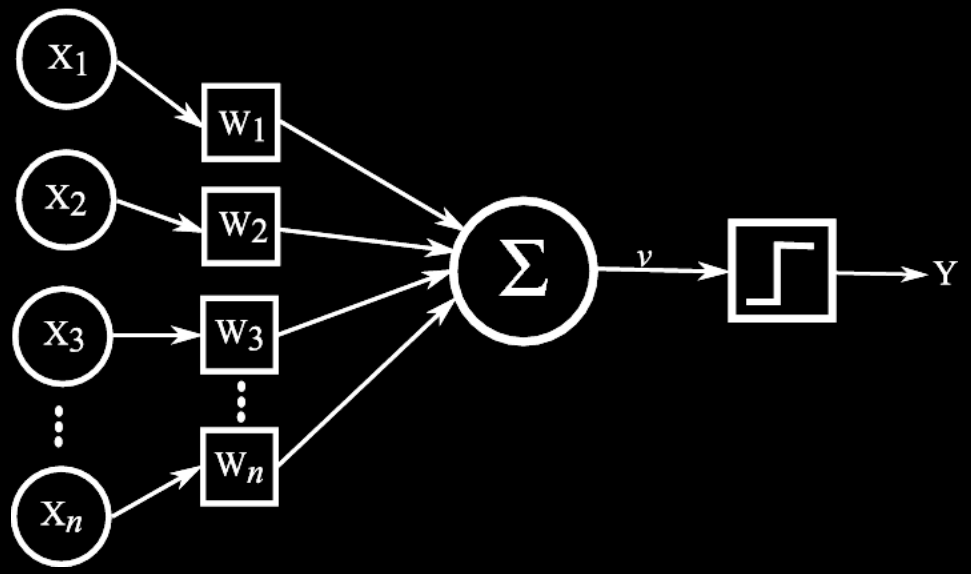
AI family



© Constance Douwes

Artificial neuron



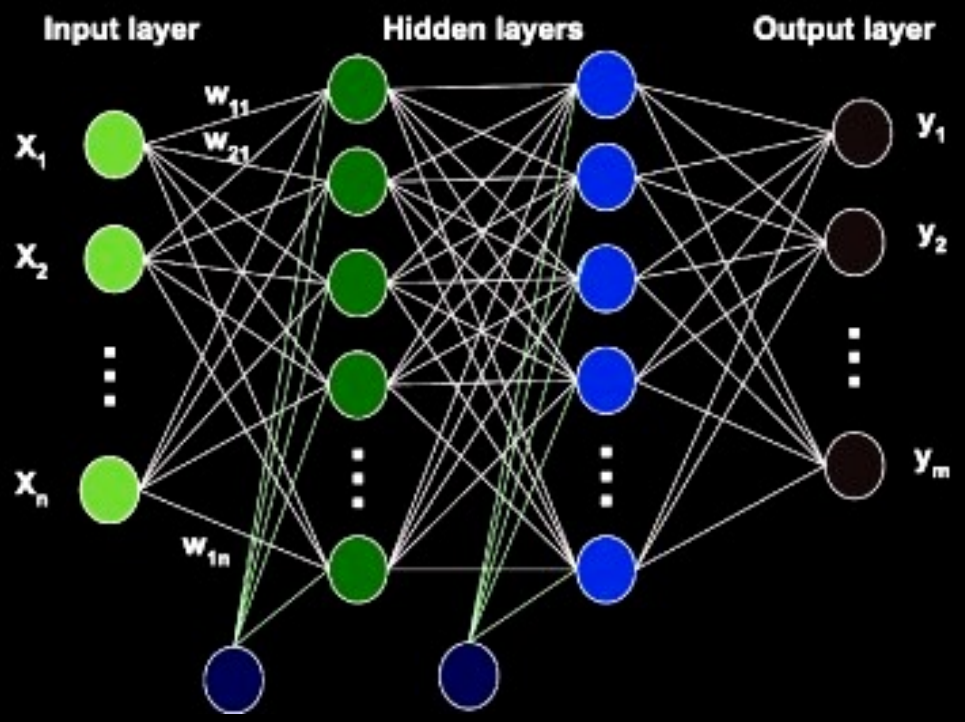


$$y = \begin{cases} 1 & \text{if} \quad \sum_{i=1}^{n} w_i x_i \geq \theta \\ 0 & \text{else.} \end{cases}$$

Source : Rosenblatt 1957

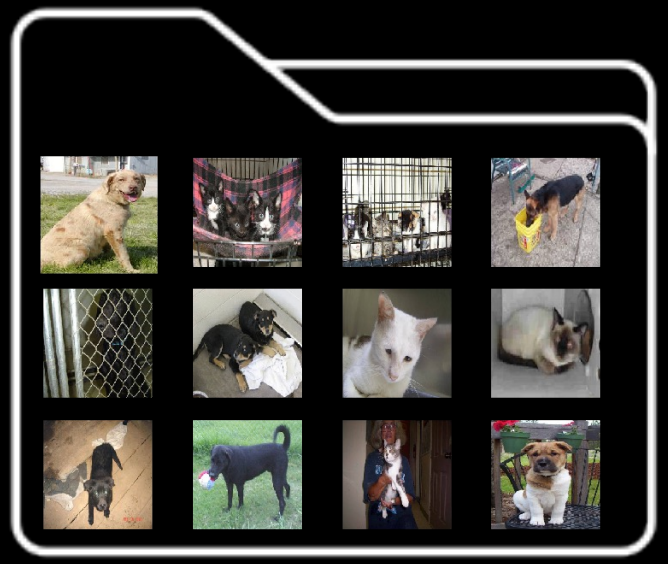
© Constance Douwes

Multi-layer perceptron

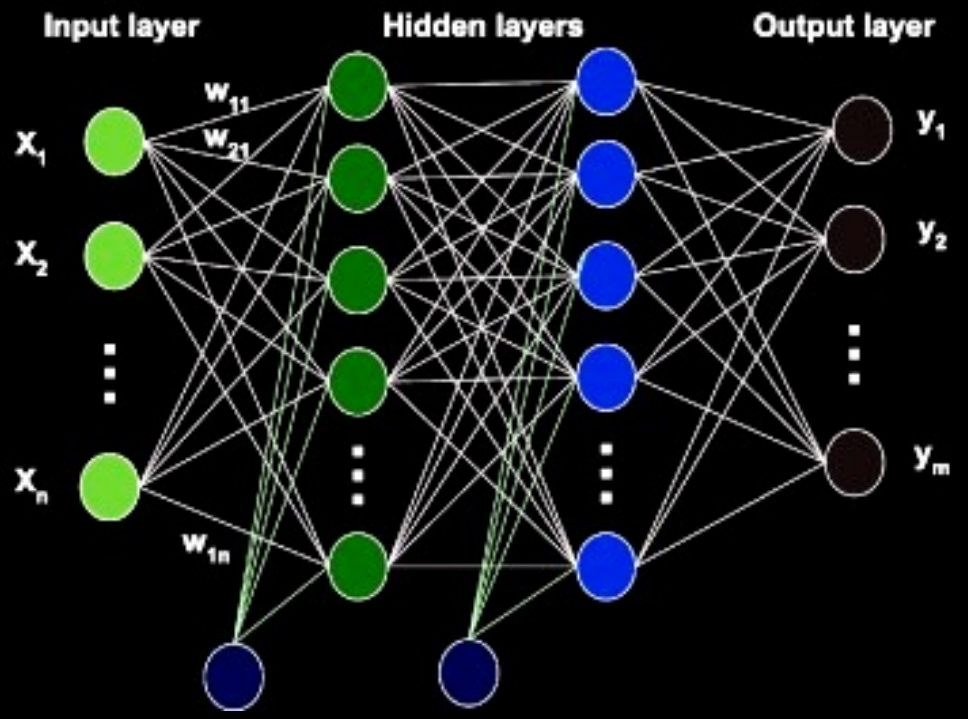


© Constance Douwes

Training



Dataset
Image (pixels) + « True » label



→ Cat or dog ?

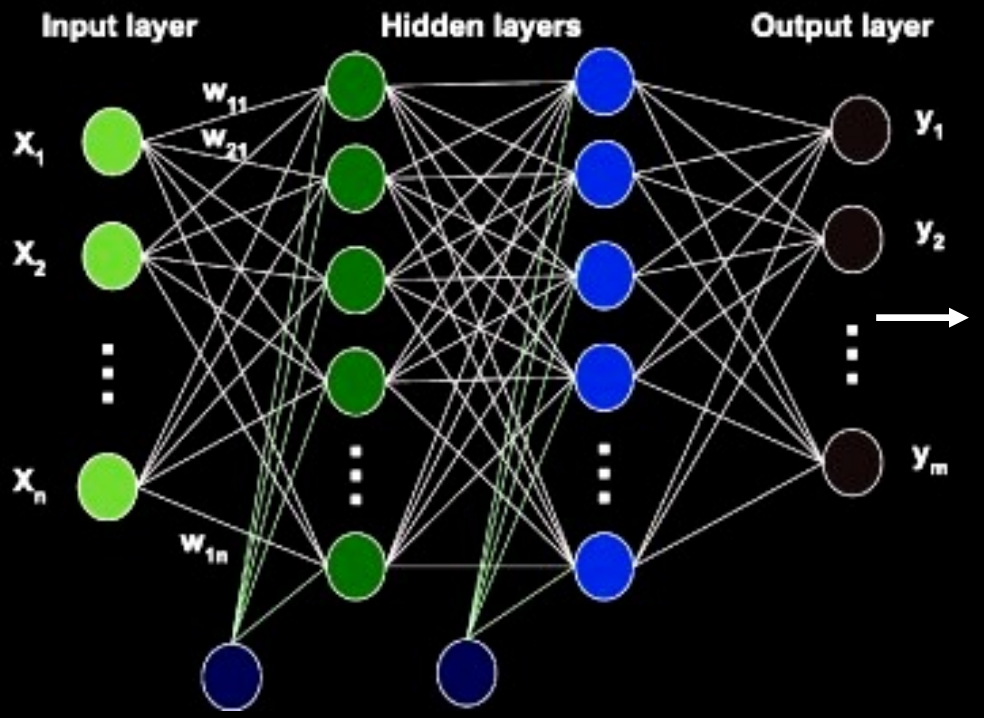
1. Initialization : random weights w_{ij}
2. Adjust weights according to the error
3. Repeat until good performances

© Constance Douwes

Training



Image of the train set





Prédiction

$\hat{Y}$

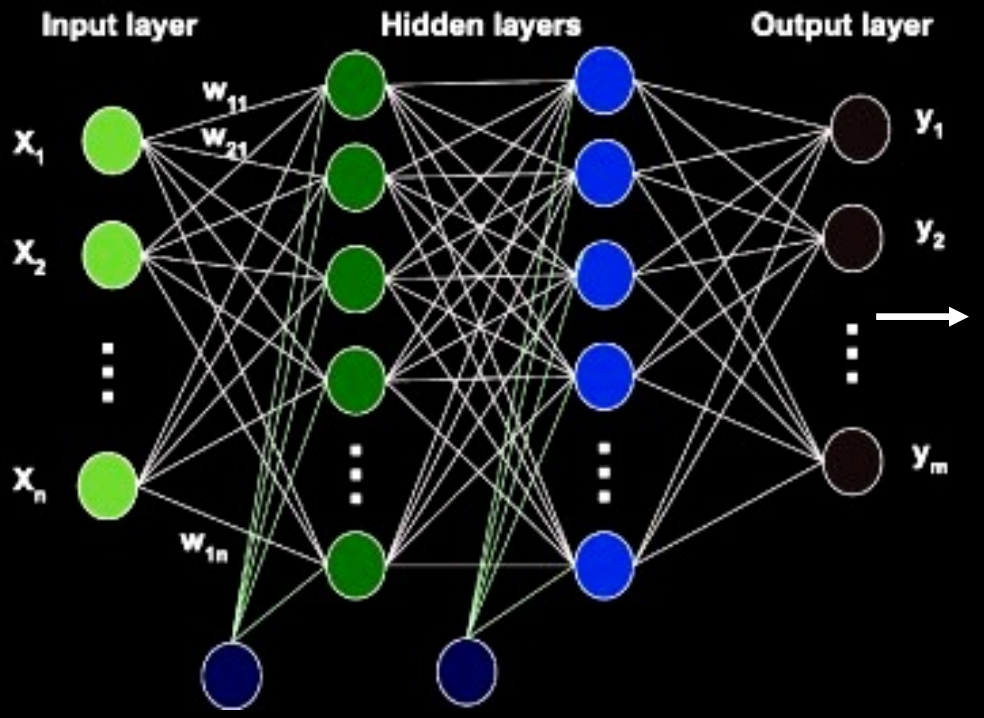
1. Initialization : random weights w_{ij}
2. Adjust weights according to the error
3. Repeat until good performances

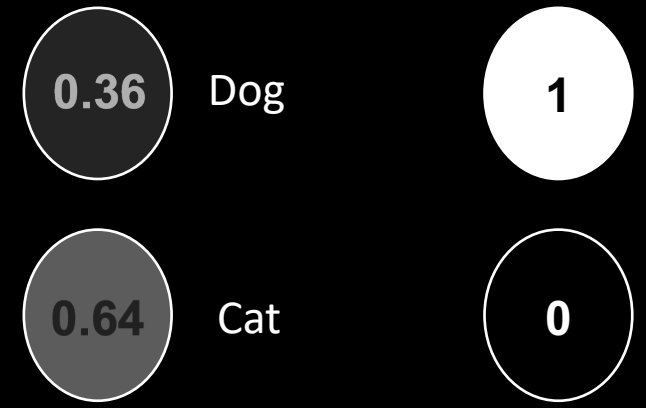
© Constance Douwes

Training



Image of the train set





1. Initialization : random weights w_{ij}
2. Adjust weights according to the error
3. Repeat until good performances

Compute the error

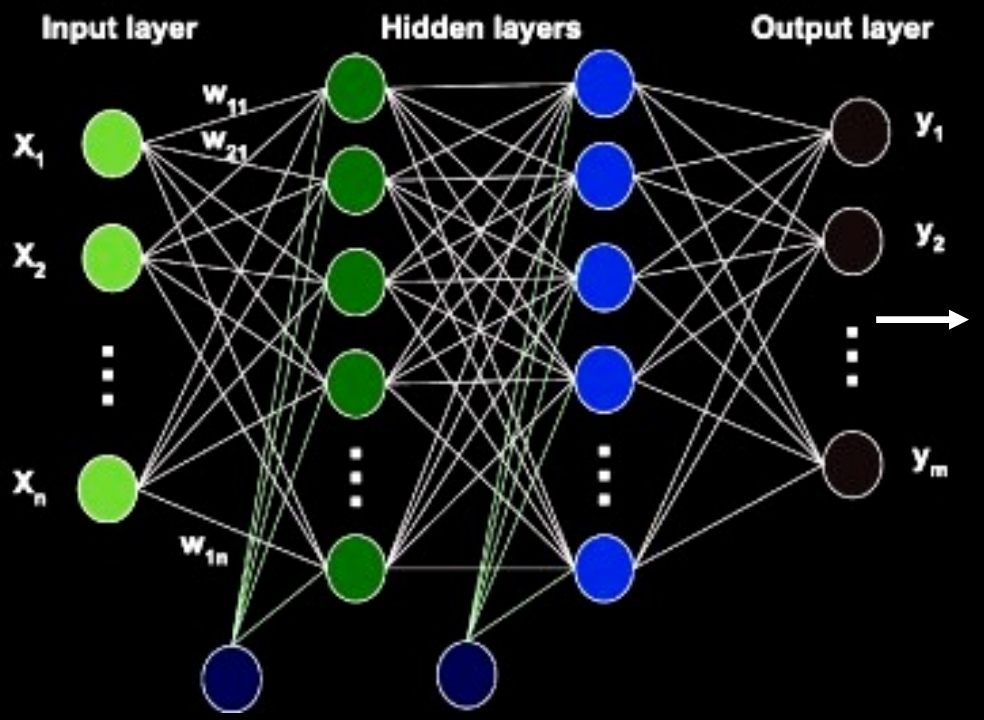
$\mathcal{L}(\hat{Y}, Y)$

© Constance Douwes

Training



Image of the train set





Prédiction

$\hat{Y}$

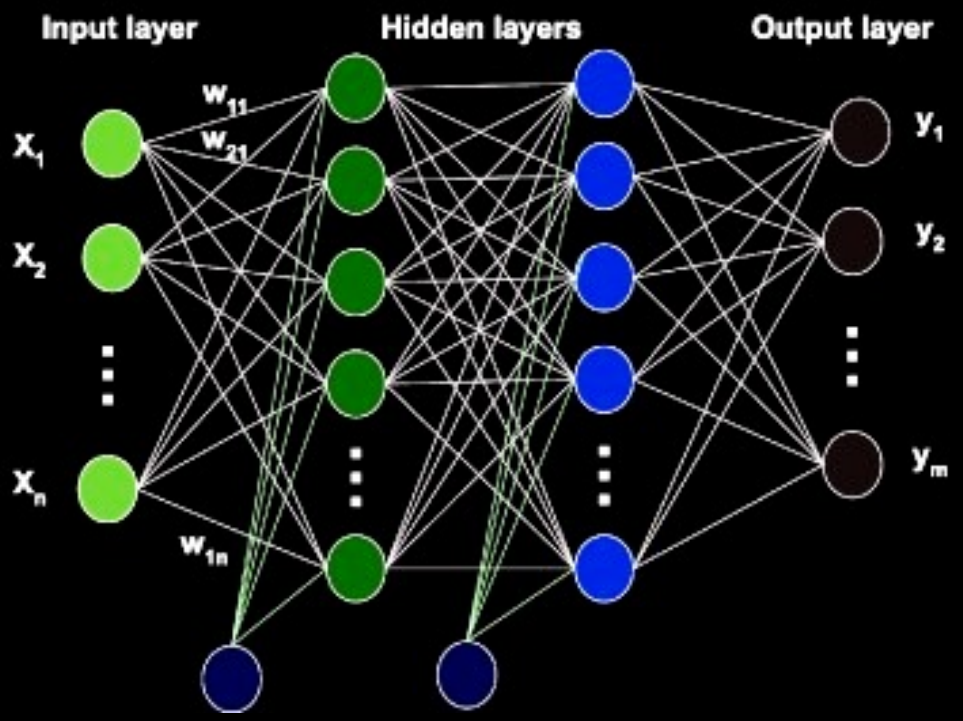
1. Initialization : random weights w_{ij}
2. Adjust weights according to the error
3. Repeat until good performances

© Constance Douwes

Inference



New entries
(never seen by the network)



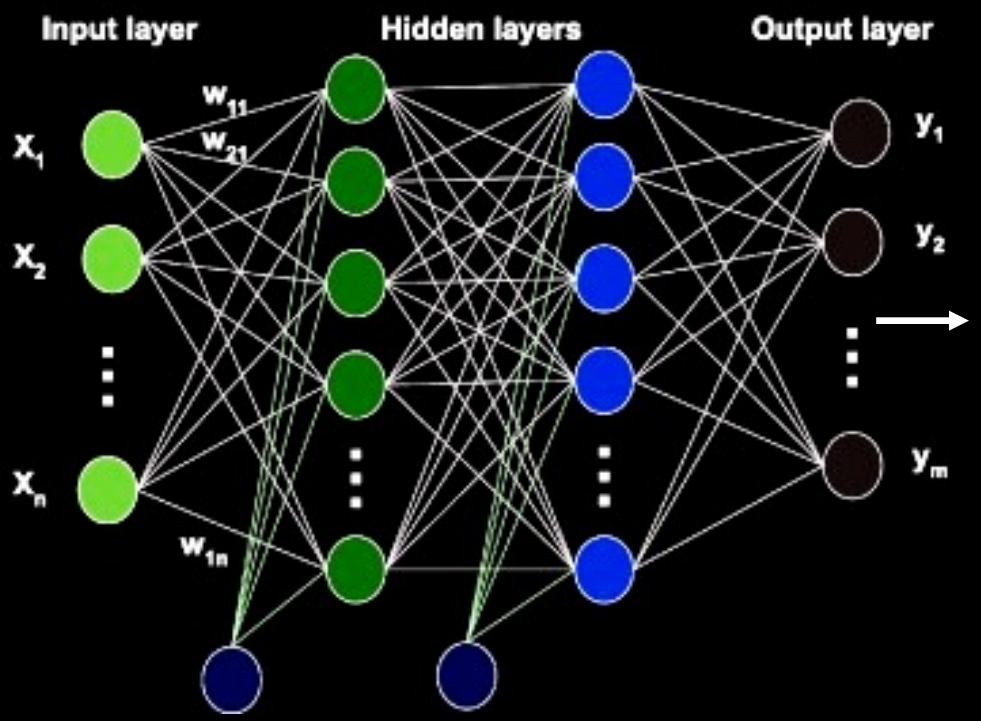
Fixed weights w_{ij}

© Constance Douwes

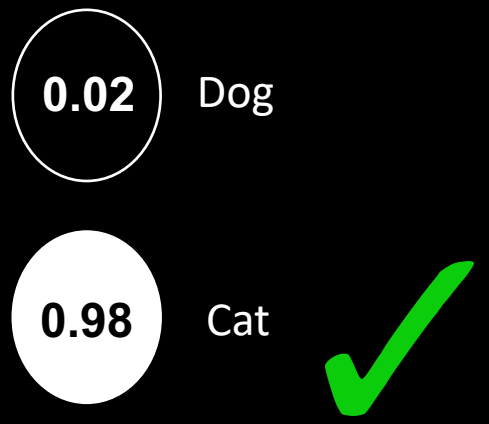
Inference



New entries
(never seen by the network)



Fixed weights w_{ij}



© Constance Douwes

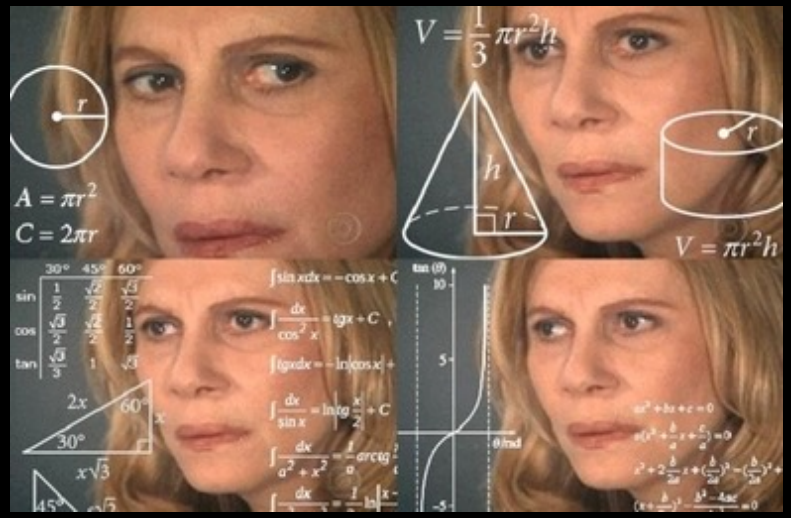
A LOT OF COMPUTATIONS !!!

A LOT OF COMPUTATIONS !!!

A LOT OF COMPUTATIONS !!!

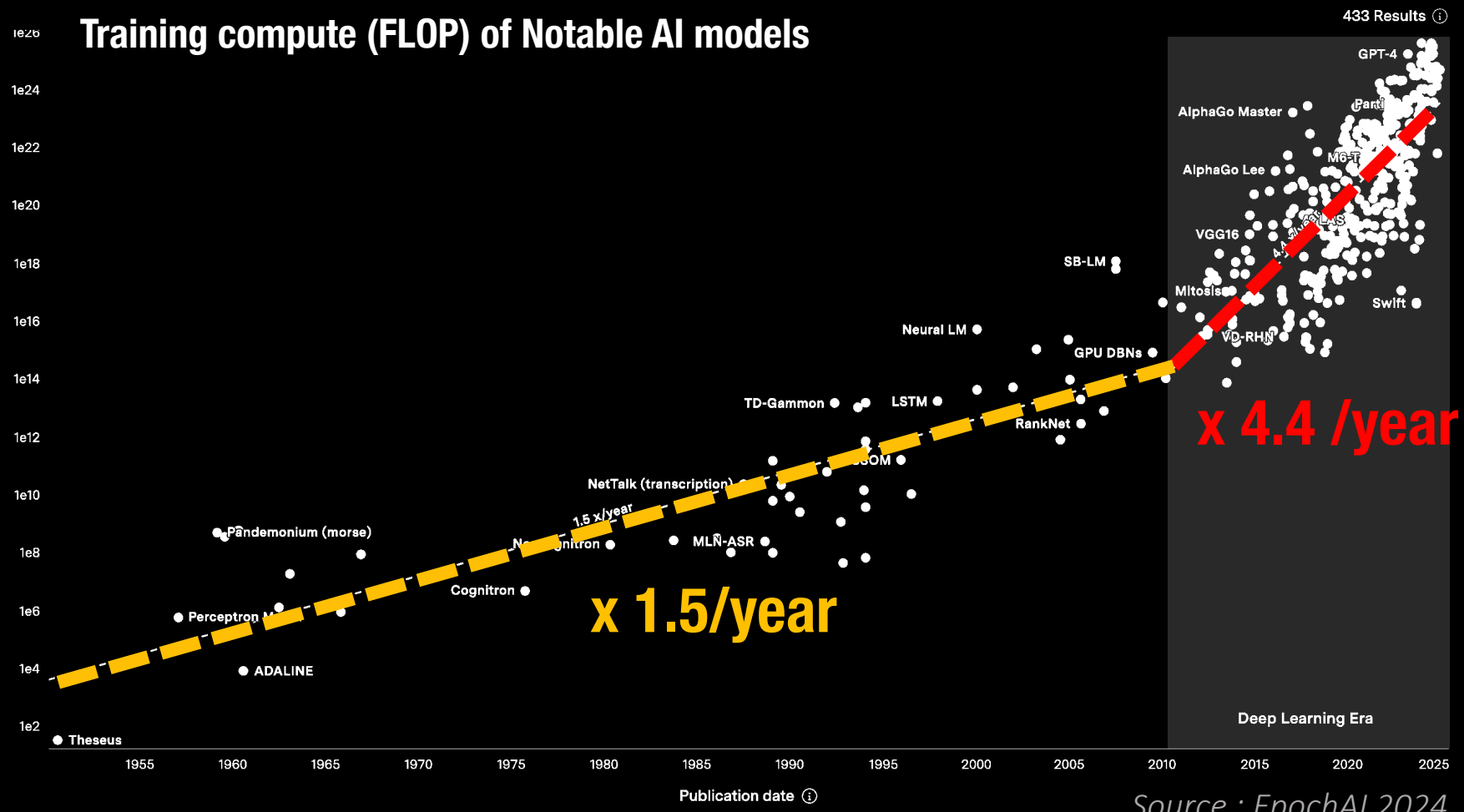
A LOT OF COMPUTATIONS !!!

A LOT OF COMPUTATIONS !!!



© Constance Douwes

Training compute increase

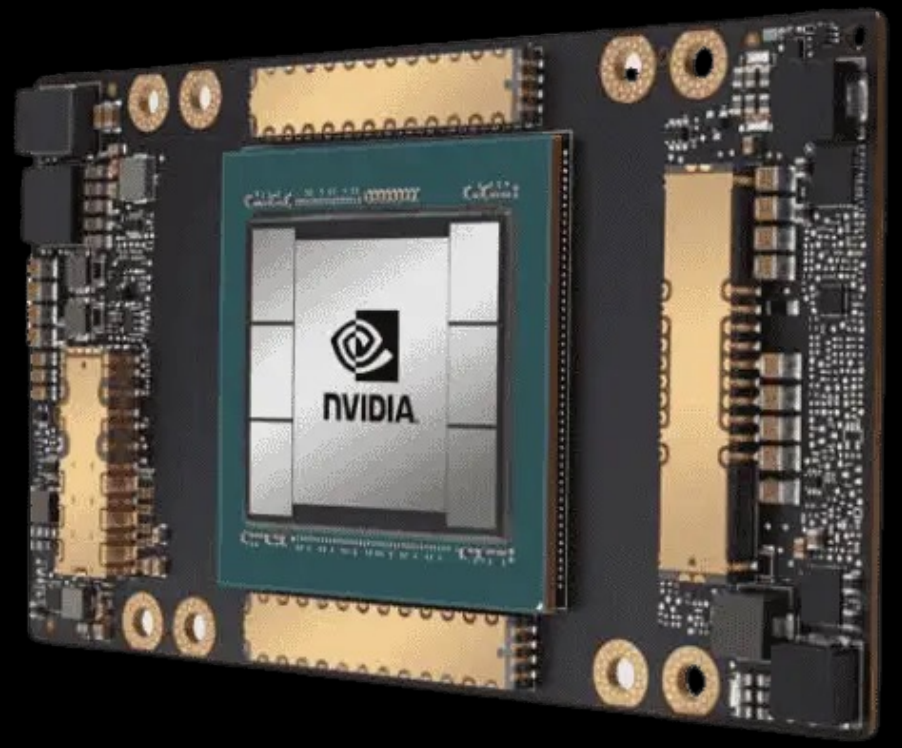


Source : EpochAI 2024

© Constance Douwes

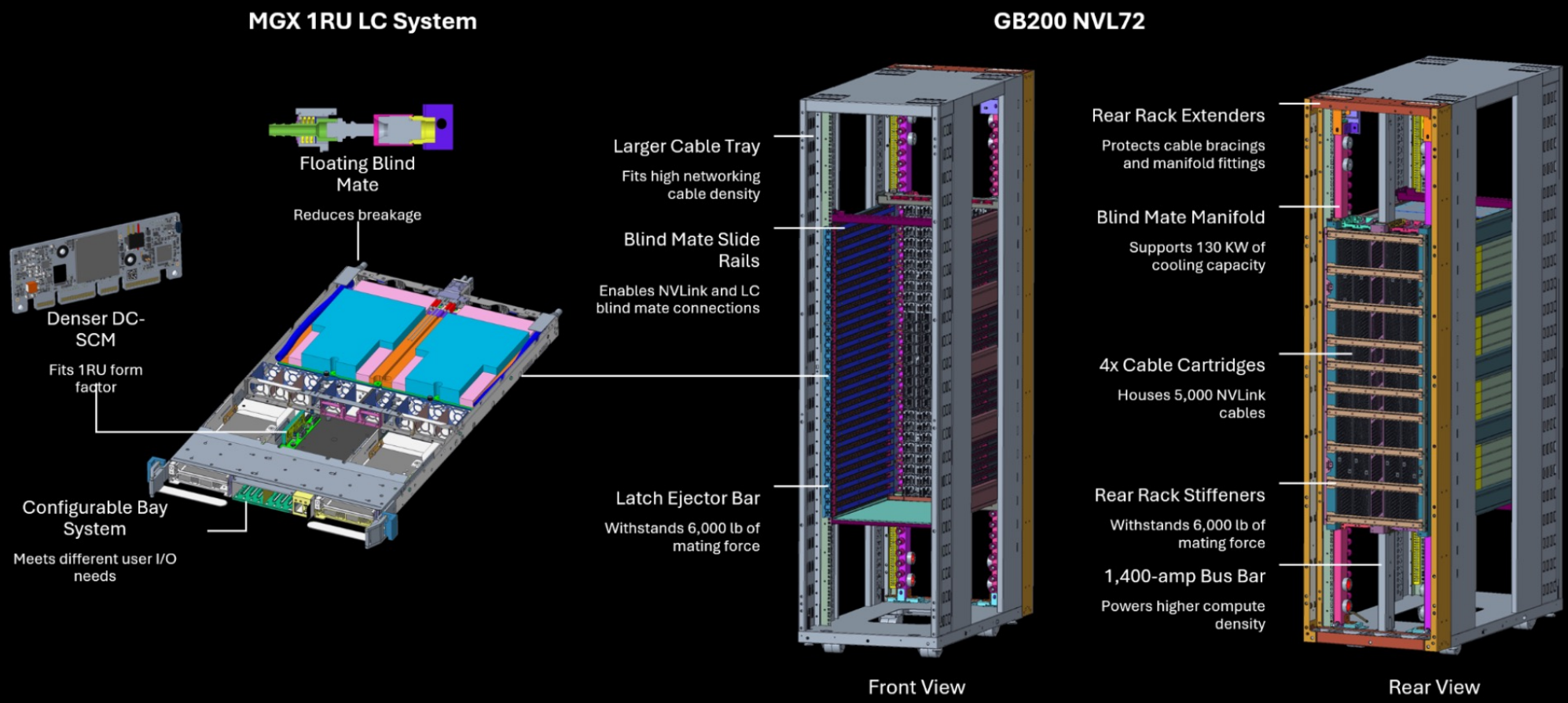
Graphical Processing Units (GPUs)





© Constance Douwes

Rack-scale



Rack example: NVIDIA GB200 NVL72 Rack-Scale Configuration

72 GPUs = 132kWh

© Constance Douwes

Data centers





2 % global CO_2 emissions

→ energy supply **x4** between 2023 and 2035

© Constance Douwes

Source : The Shift Project, 2025

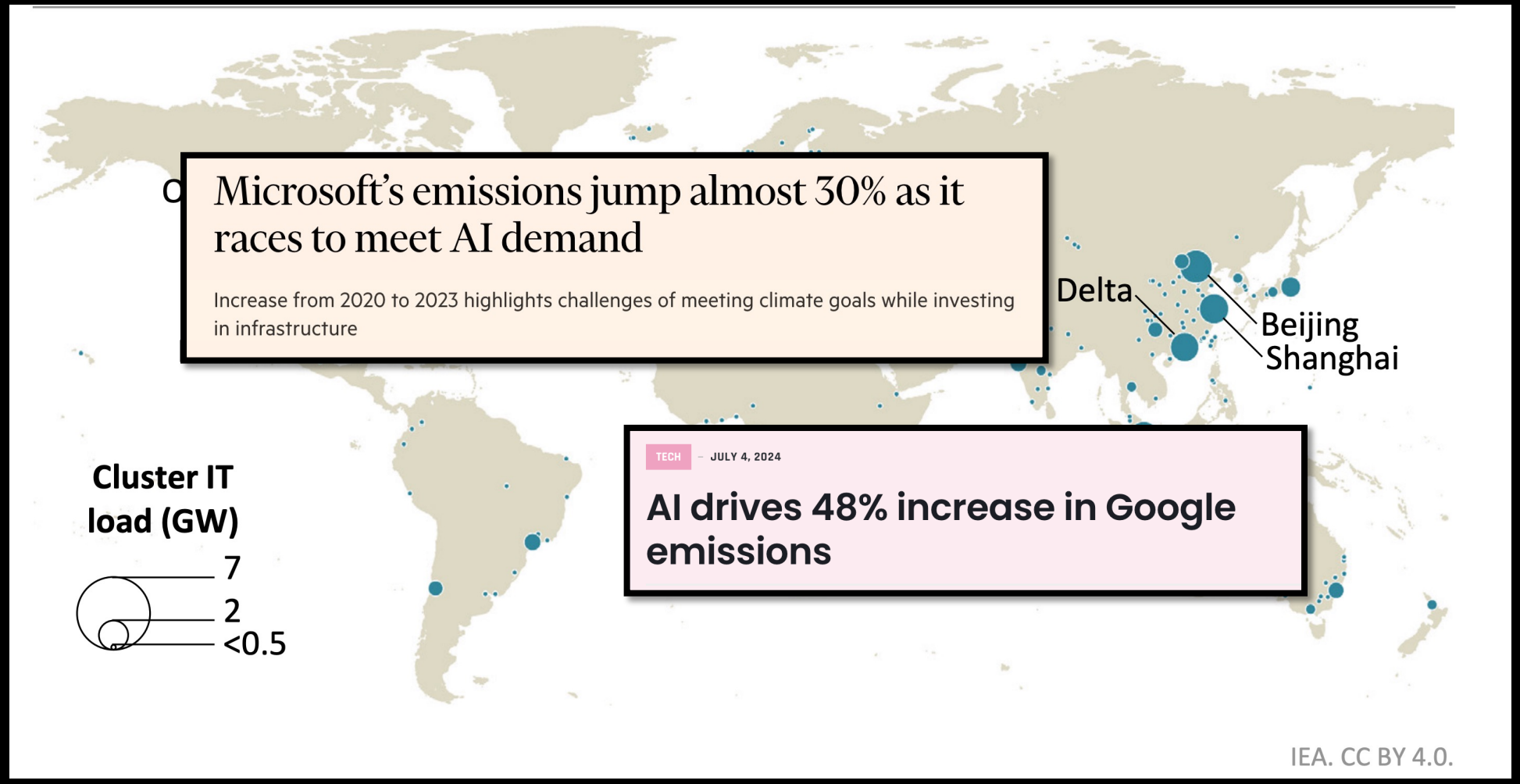
Data centers (now)



Source: IEA, Energy and AI, April 2025

© Constance Douwes

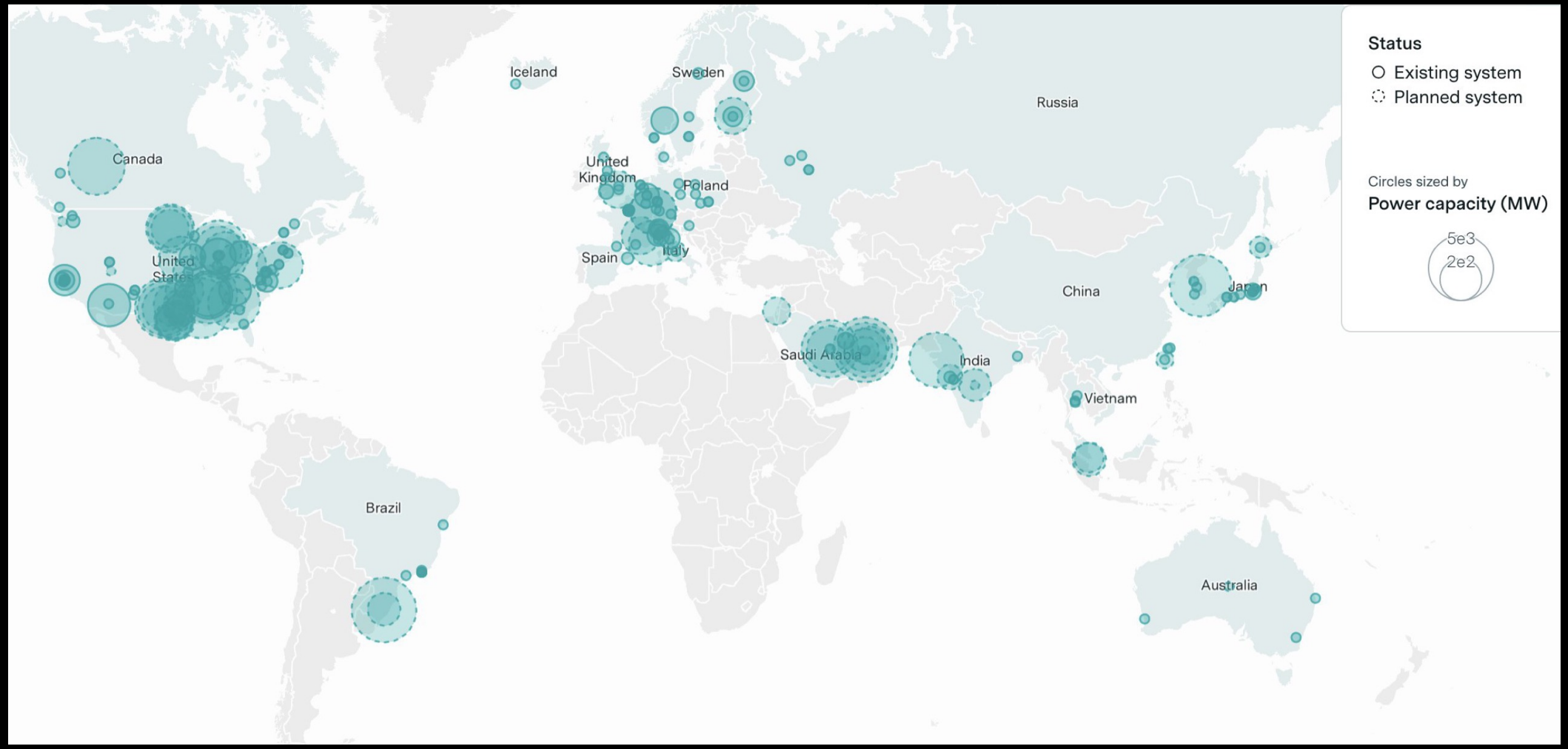
Data centers (now)



Source: IEA, Energy and AI, April 2025

© Constance Douwes

Data centers (to come)



Source: Epoch AI, January 2026

© Constance Douwes

In Marseille



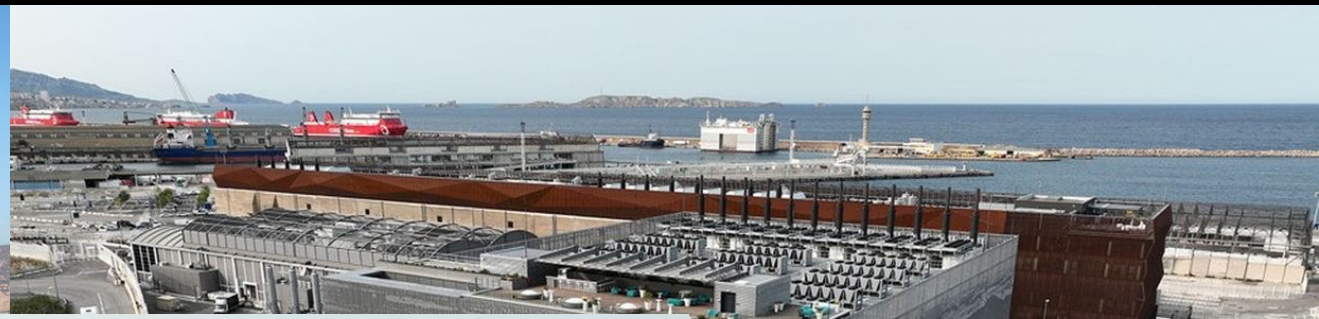


Digital Realty sur l'A55 → direction les Pennes Mirabeau

© Constance Douwes

In Marseille





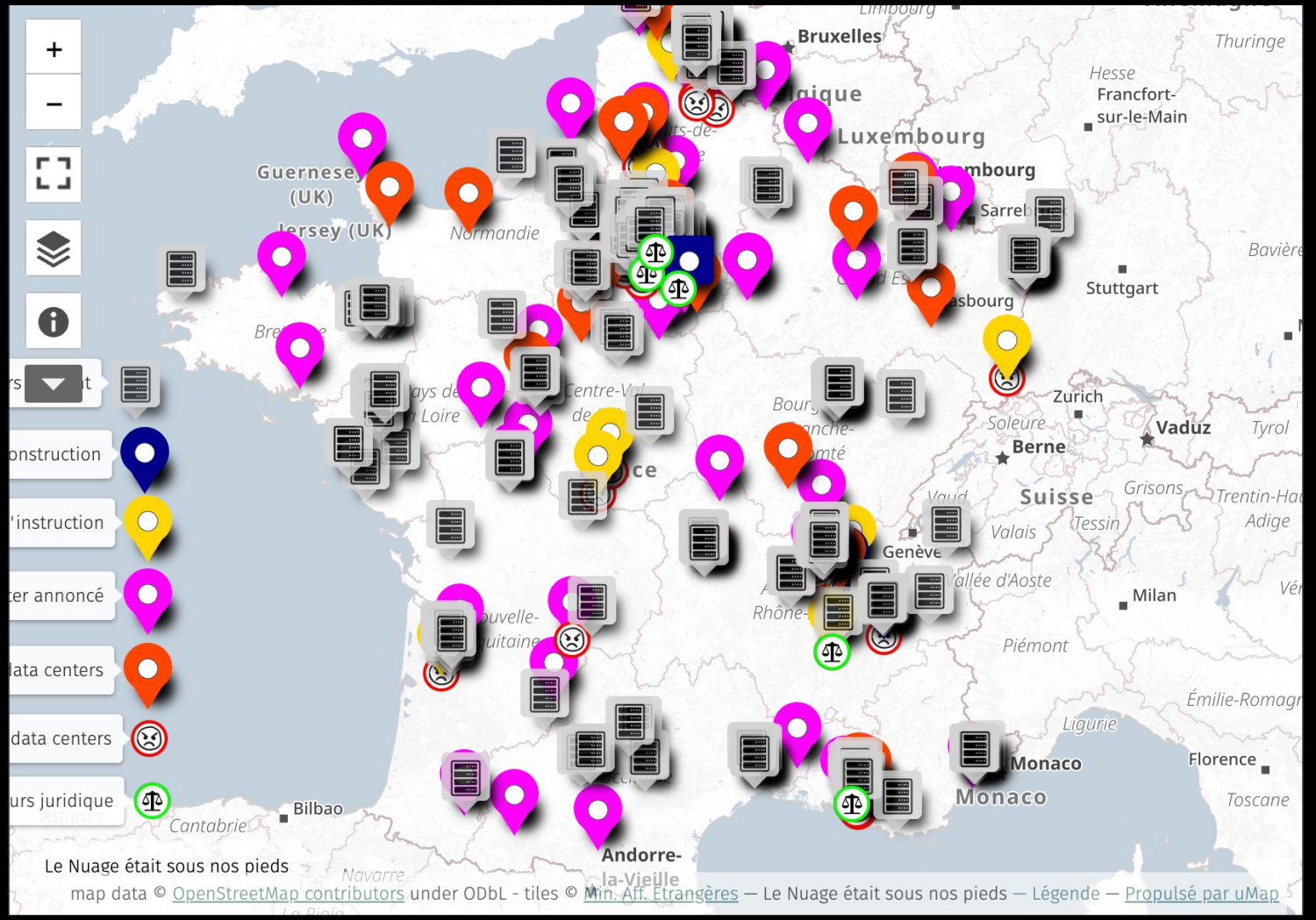




Digital Realty sur l'A55

© Constance Douwes

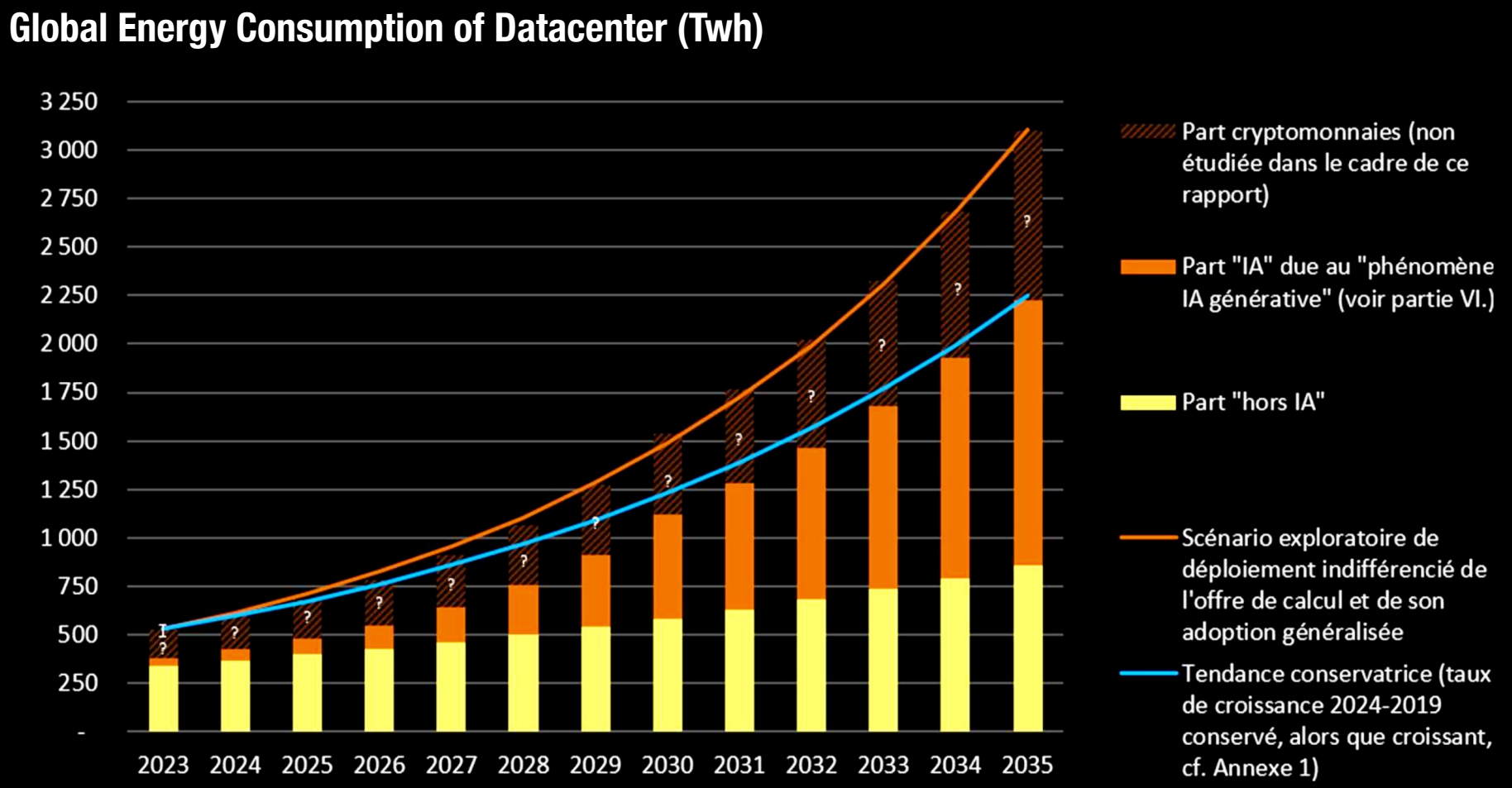
In France



https://lenuageetaitsousnospieds.org/articles/2025-12-11-carte-des-data-centers-des-projets-et-des-contestations-en-france.html

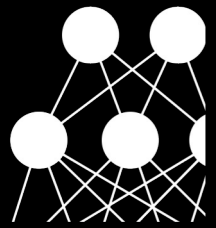
© Constance Douwes

Datacenters prospective



Source : The Shift Project, 2025

© Constance Douwes







How?

© Constance Douwes

Life-cycle phases

Data
Collection, Processing

Model
Develoment, Training

Model
Deployment, Inference



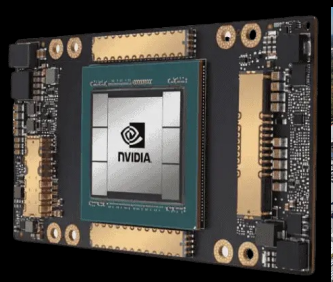
© Constance Douwes

Life-cycle phases

Data
Collection, Processing

Model
Develoment, Training

Model
Deployment, Inference





© Constance Douwes

Life-cycle phases

Data
Collection, Processing

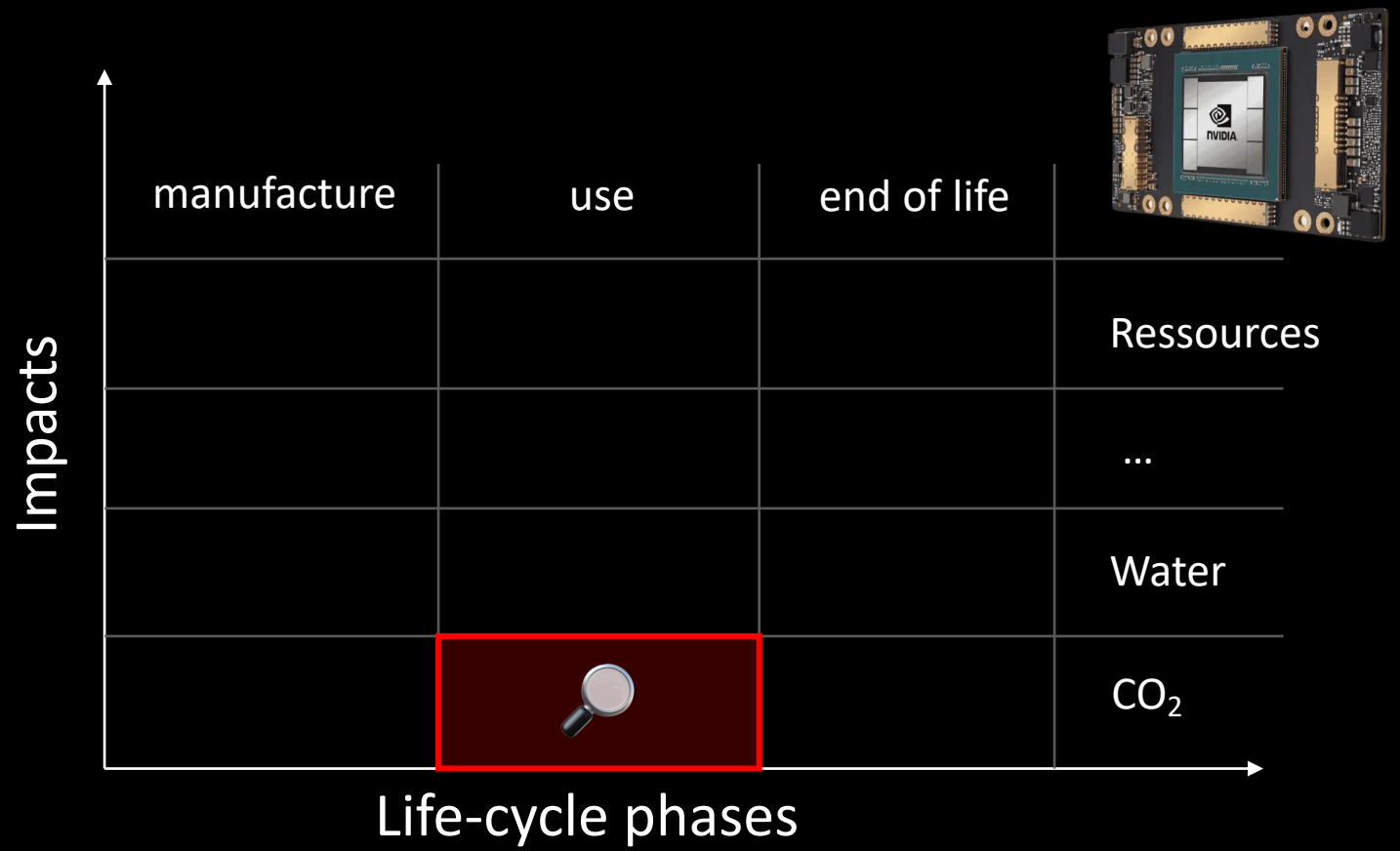
Model
Develoment, Training

Model
Deployment, Inference



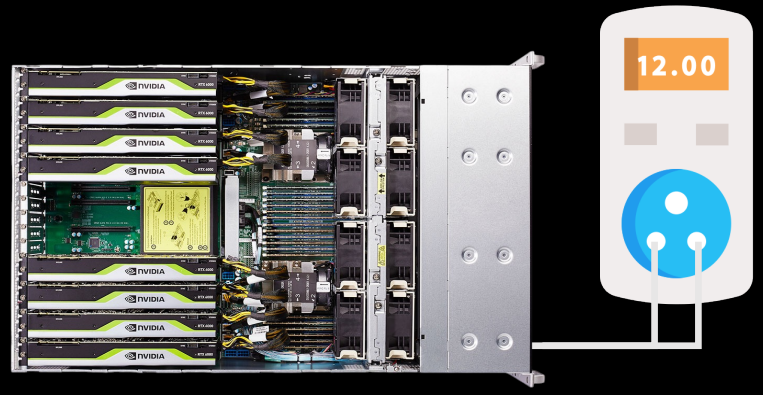
© Constance Douwes

Life-cycle analysis



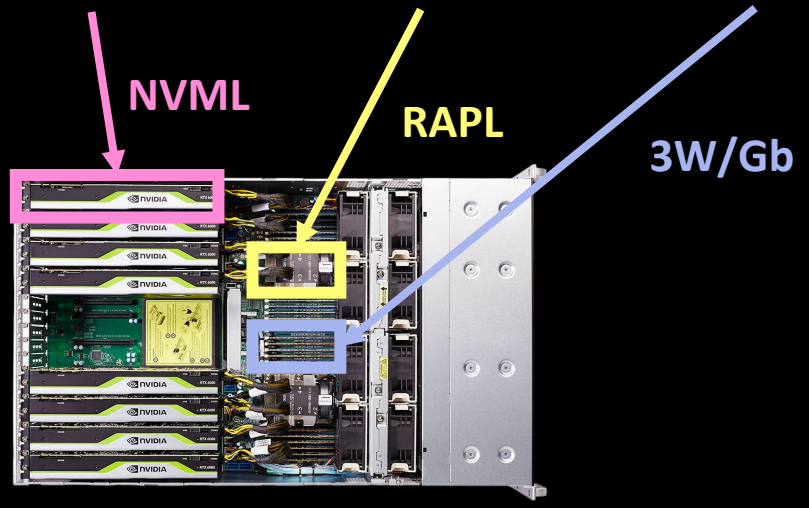
© Constance Douwes

$$\text{⚡Energy} = \sum_{t=0}^{T} \text{Power}(t) \times \Delta t$$



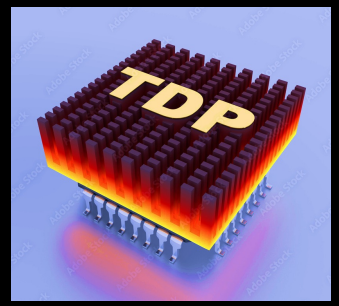
© Constance Douwes

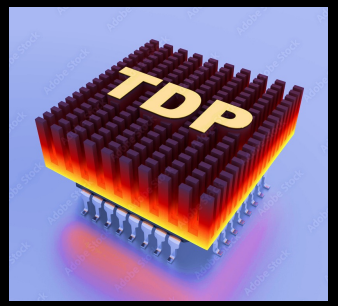
$$\text{⚡Energy} = \sum_{t=0}^{T} (P_{GPU} + P_{CPU} + P_{MEM})(t) \times \Delta t$$



© Constance Douwes

$$\text{⚡Energy} = (P_{GPU} + P_{CPU} + P_{MEM}) \times \text{Time}$$





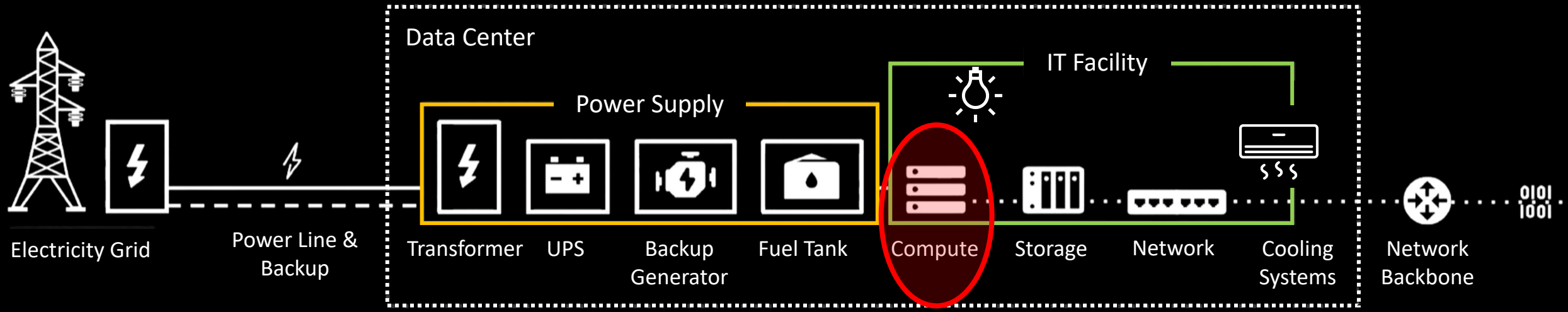
3W / 8Gb

TDP : Thermal Design Power

© Constance Douwes

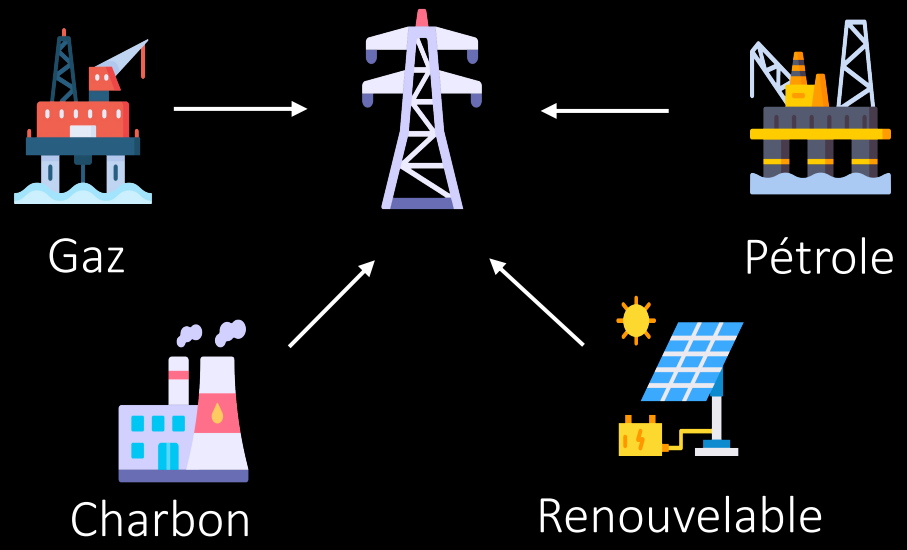
⚡Energy × PUE

Power Usage Effectiveness

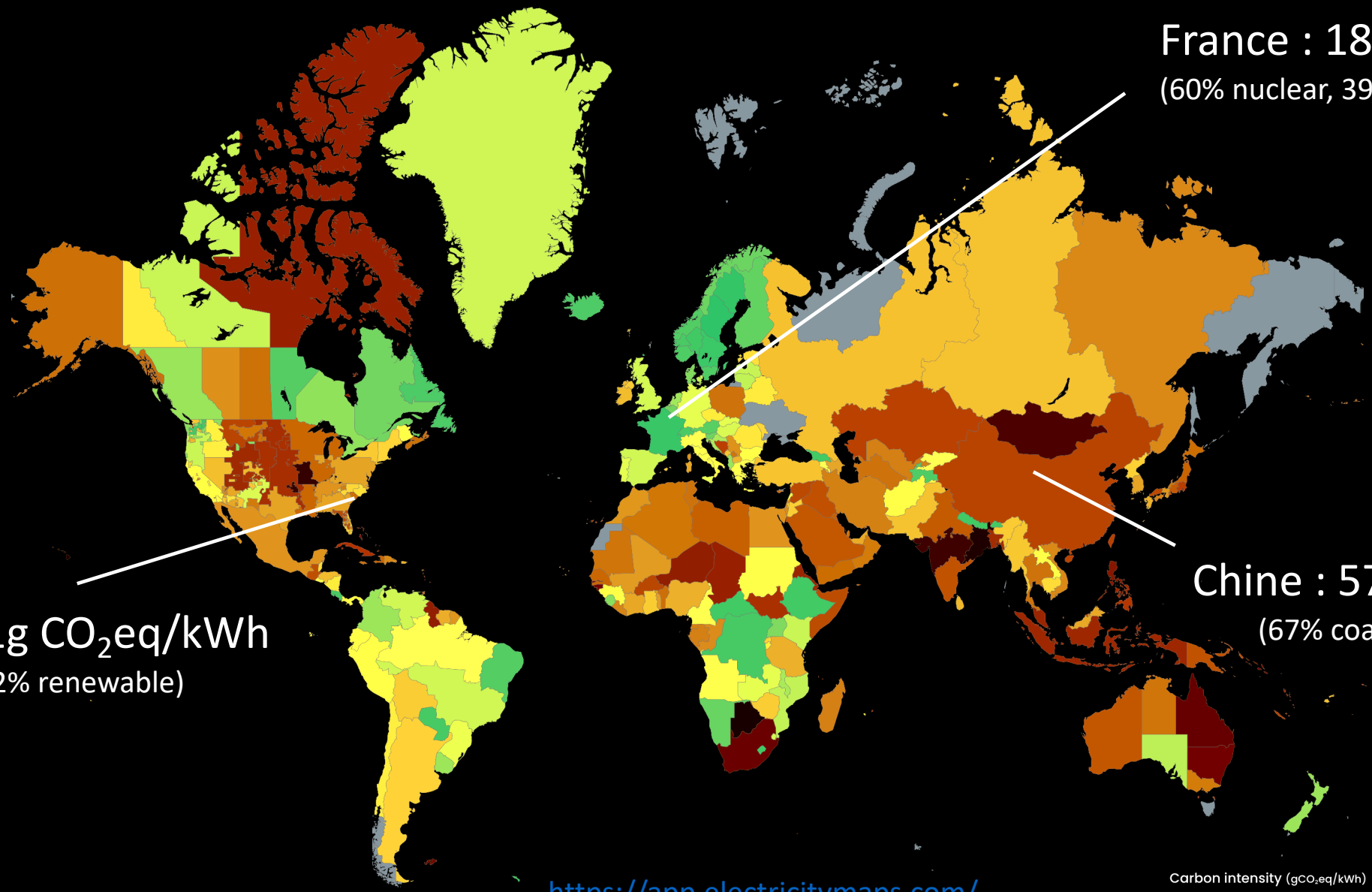


© Constance Douwes

⚡Energy × PUE × Carbon intensity = CO_2eq



© Constance Douwes



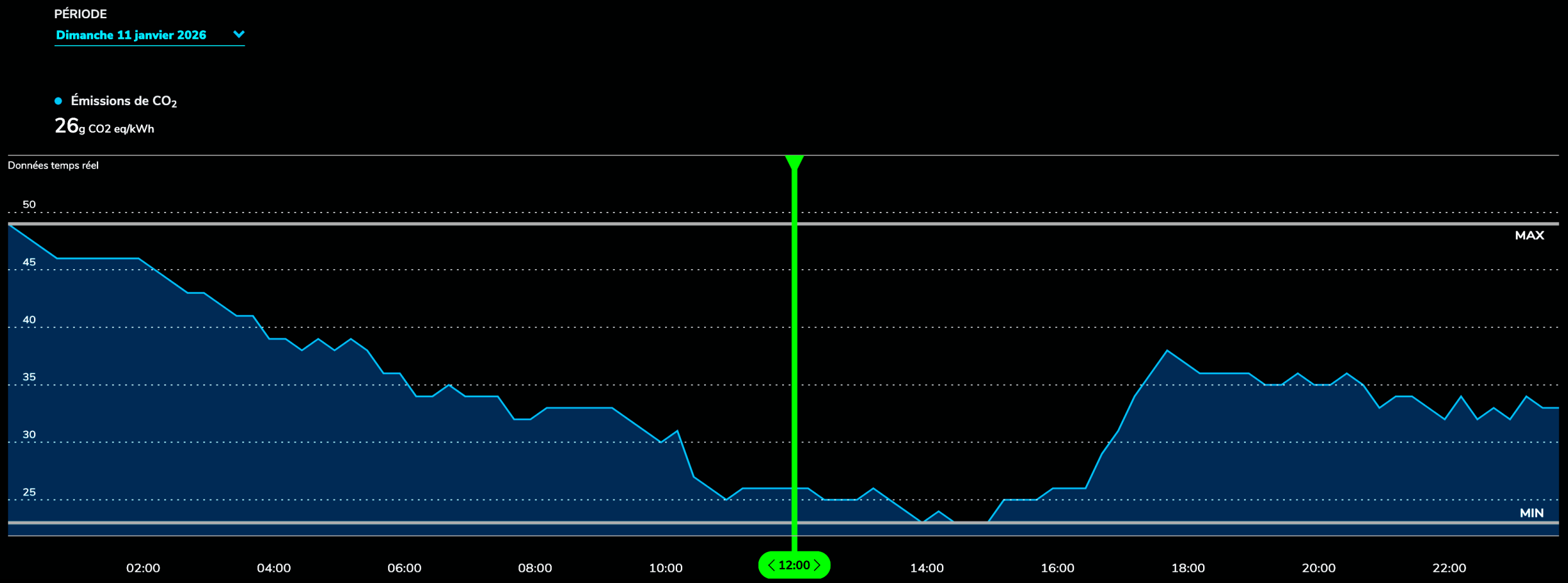
France : 18g CO_2eq/kWh
(60% nuclear, 39% renewable)

NY : 311g CO_2eq/kWh
(42% gas, 22% renewable)

Chine : 578g CO_2eq/kWh
(67% coal, 25% renewable)

https://app.electricitymaps.com/

© Constance Douwes

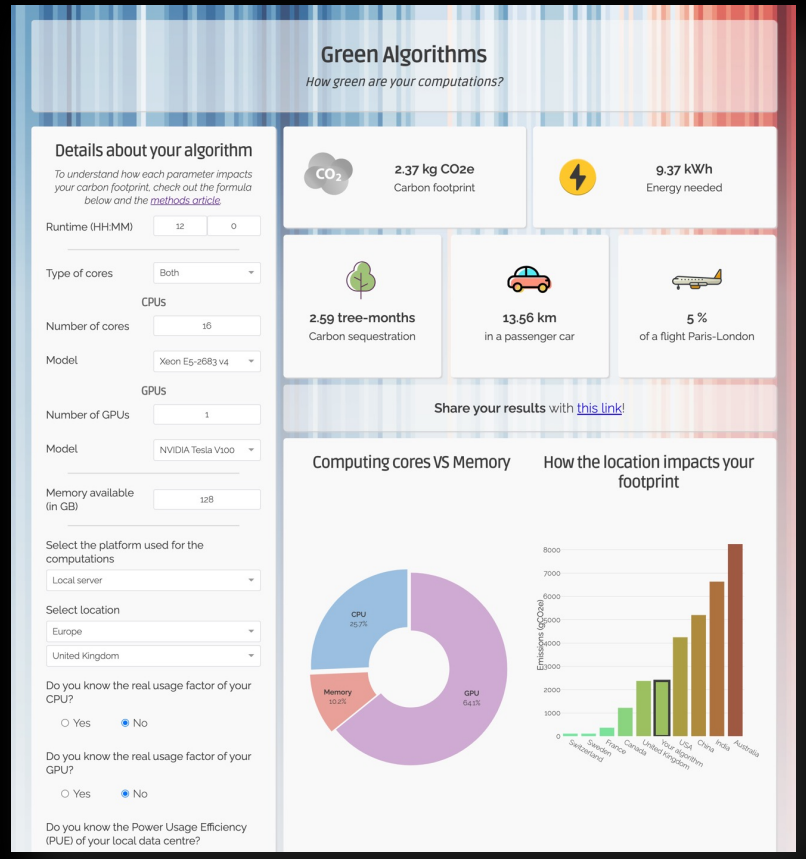


https://www.rte-france.com/donnees-publications/eco2mix-donnees-temps-reel/emissions-co2-par-kwh-produit-france

© Constance Douwes

Some tools

Online



https://calculator.green-algorithms.org/

Code level

CODE CARBON

lfwa/**carbontracker**

Track and predict the energy consumption and carbon footprint of training deep learning models.

Breakend/**experiment-impact-tracker**

Server level

```
    ######################################
    #                                    #
    #  Carbon footprint on <my_server>   #
    #         - user: <yourID> -         #
    #     (2024-01-01 / 2024-08-20)      #
    #                                    #
    ######################################


          --------------
         |   533 gCO2e   |
          --------------

...This is equivalent to:
     - 0.581 tree-months
     - driving 3.05 km
     - 0.01 flights between Paris and London

...13.3% of the jobs failed, these represent a waste of 33 gCO2e (0.035 tree-months).
...On average, the jobs request at least 1.1 times the memory needed.
   By only requesting the memory needed, 5 gCO2e (0.005 tree-months) could have been saved.

Energy used: 2.31 kWh
     - CPUs: 1.57 kWh (68.18%)
     - GPUs: 0.00 kWh (0.00%)
     - Memory: 0.43 kWh (18.77%)
     - Data centre overheads: 0.30 kWh (13.04%)
Carbon intensity used for the calculations: 231.12 gCO2e/kWh

Summary of usage:
- First/last job recorded on that period: 2024-01-29/2024-07-30
- Number of jobs: 15 (13 completed)
- Core hours used/charged: 332.1 (CPU), 0.0 (GPU), 332.1 (total).
- Total usage time (i.e. when cores were performing computations):
    - CPU: 9 days 05:26:47 (221 hours)
    - GPU: 0 days 00:00:00 (0 hours)
- Total wallclock time: 2 days 06:29:43
- Total memory requested: 351 GB
```

© Constance Douwes

Training cost

of Large Language Models

GPT-3 :

- 10000 NVIDIA Tesla V100 DGXS 32 GB
- 14,8 days
- Microsoft data center (PUE=1.125)
- Zone US

≈ 502 tCO_2eq

GPT-4 :

- 25000 NVIDIA A100 SXM4 40 GB
- 90-100 days
- Microsoft data center (PUE=1.125)
- Zone US

≈ 9700 tCO_2eq

© Constance Douwes

Training cost

of audio generative models

Model	GPU	TDP (W)	Hours	⚡Energy (kWh)	Carbon footprint (kgCO_2eq)
Jukebox	256 × Tesla V100	64 000	72	4 608	3 395
Diff-a-Riff	1 × RTX 3090	350	288	100	74
FloWaveNet	1 × V100	250	272	82	60
SING	4 × P100	1 000	52	52	38
SampleRNN	1 × Titan X	250	168	42	31
RAVE	1 × Titan V	250	168	42	31
GANSynth	1 × V100	250	108	32	24
WaveGAN	1 × P100	250	96	24	18

≈ **4 tCO_2eq**



AR ✈️
Paris - Tokyo

[Douwes 2023 & update 2025]

X PUE = 1.55
X IC = 0.43 kgCO2/kWh

© Constance Douwes

Training cost

of ISMIR conference

- Papers published at ISMIR 2017-2023
- Training details (hardware&time) – only 20% of papers
- Carbon intensities corresponding to affiliations

ISMIR = 8 tCO_2eq

[Holzapfel 2023]

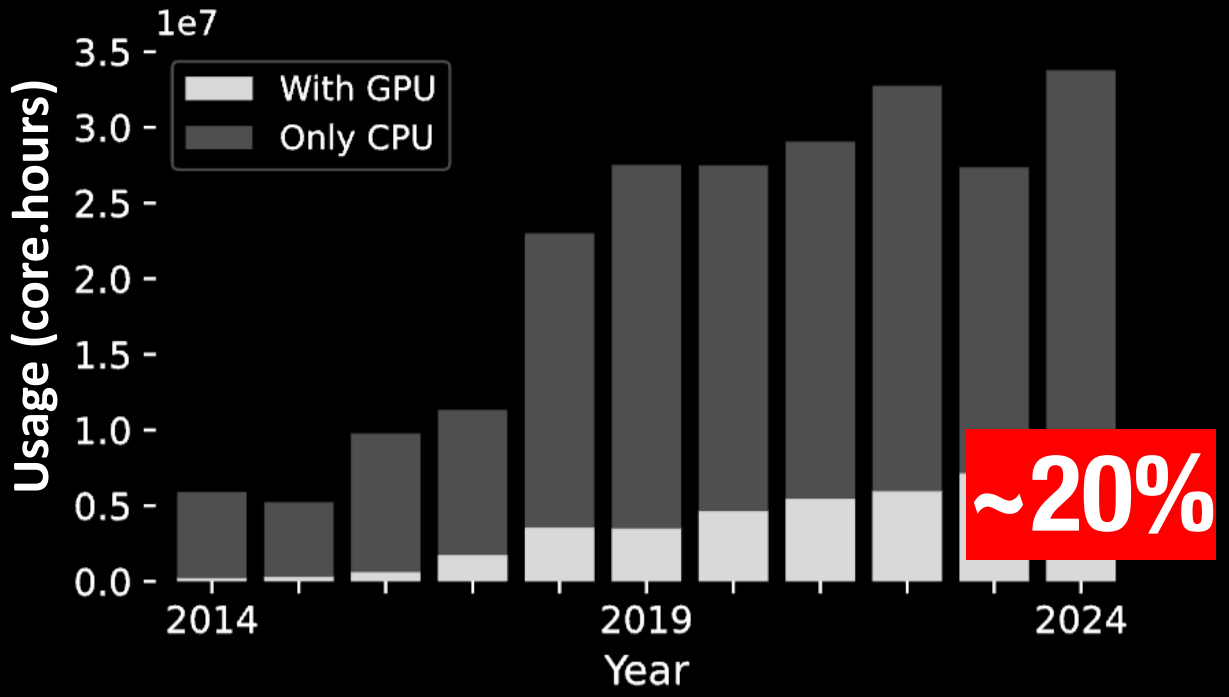
© Constance Douwes

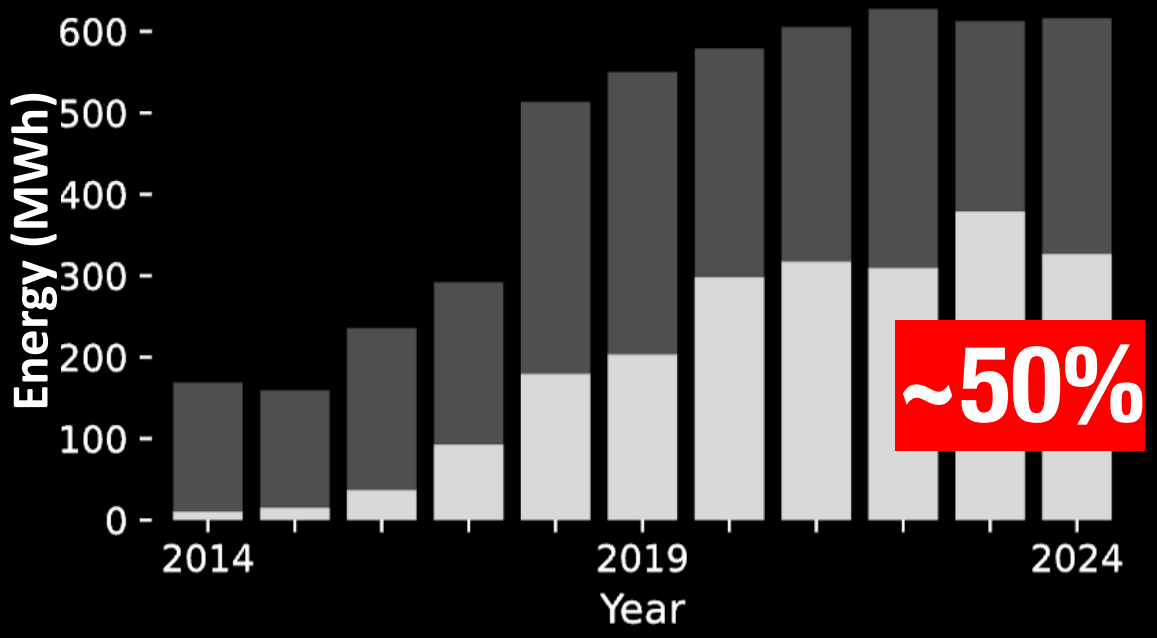
Development costs

of LORIA laboratory



INRIA Nancy (LORIA)



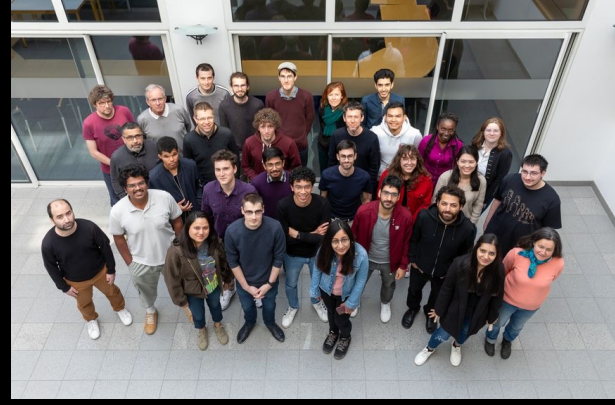


[Douwes 2027]

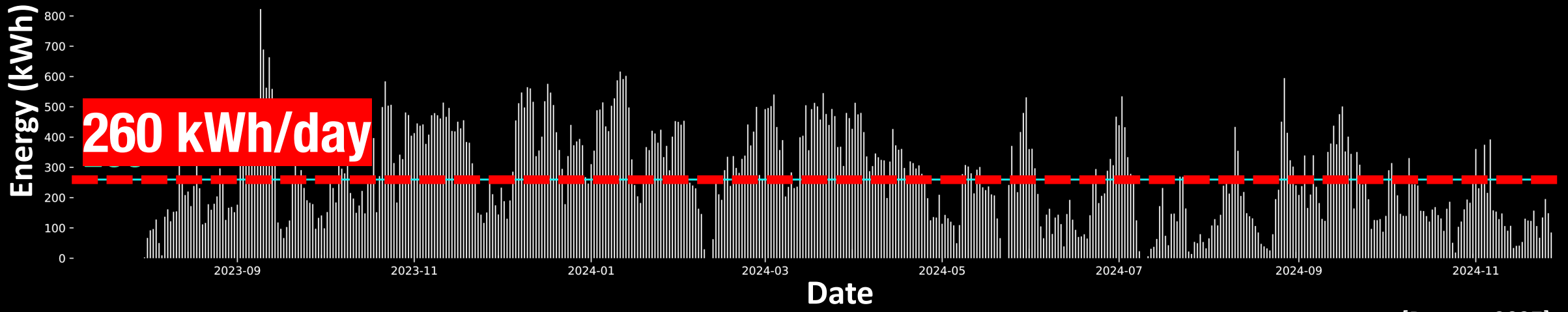
© Constance Douwes

Development costs

of Multispeech research team



Multispeech Team



[Douwes 2027]

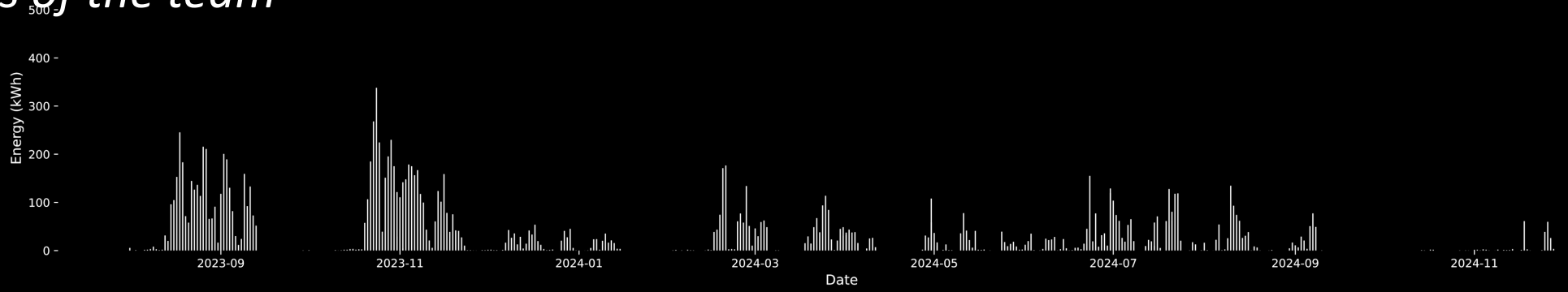
© Constance Douwes

Development costs

of members of the team

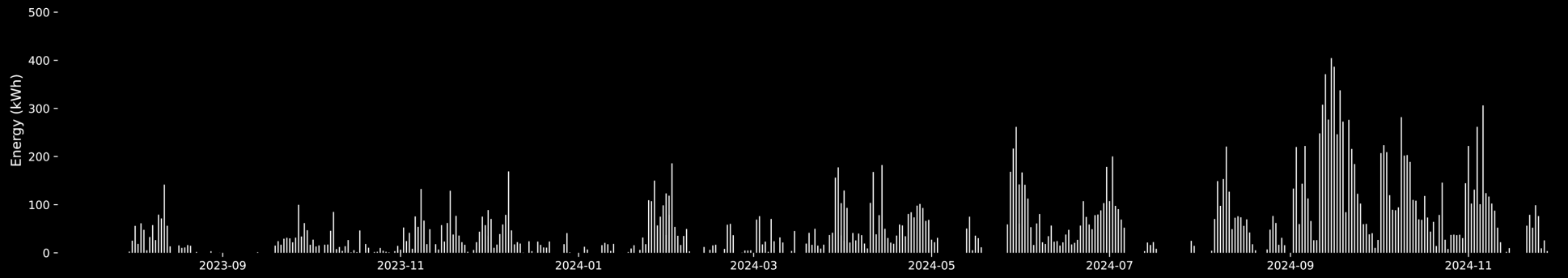


User 1



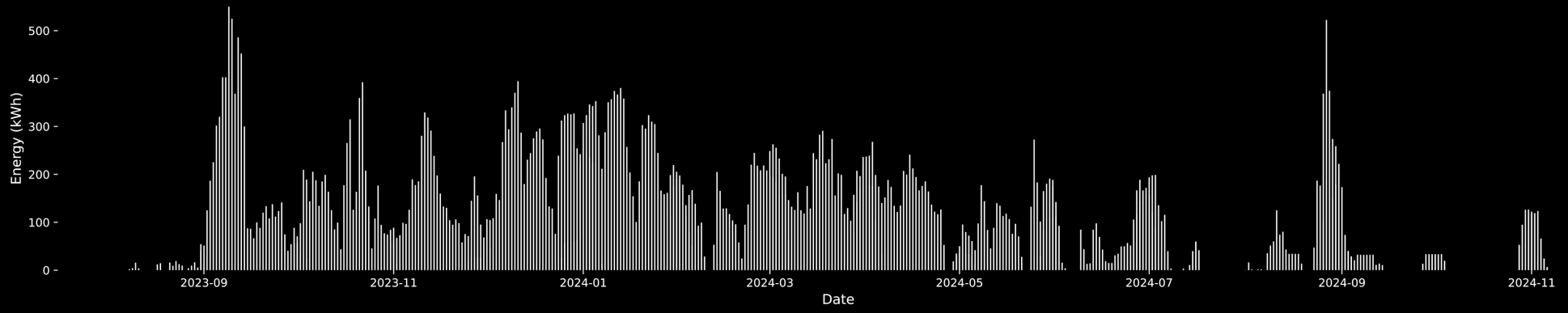


User 2





User 3



© Constance Douwes

[Douwes 2027]

Development costs

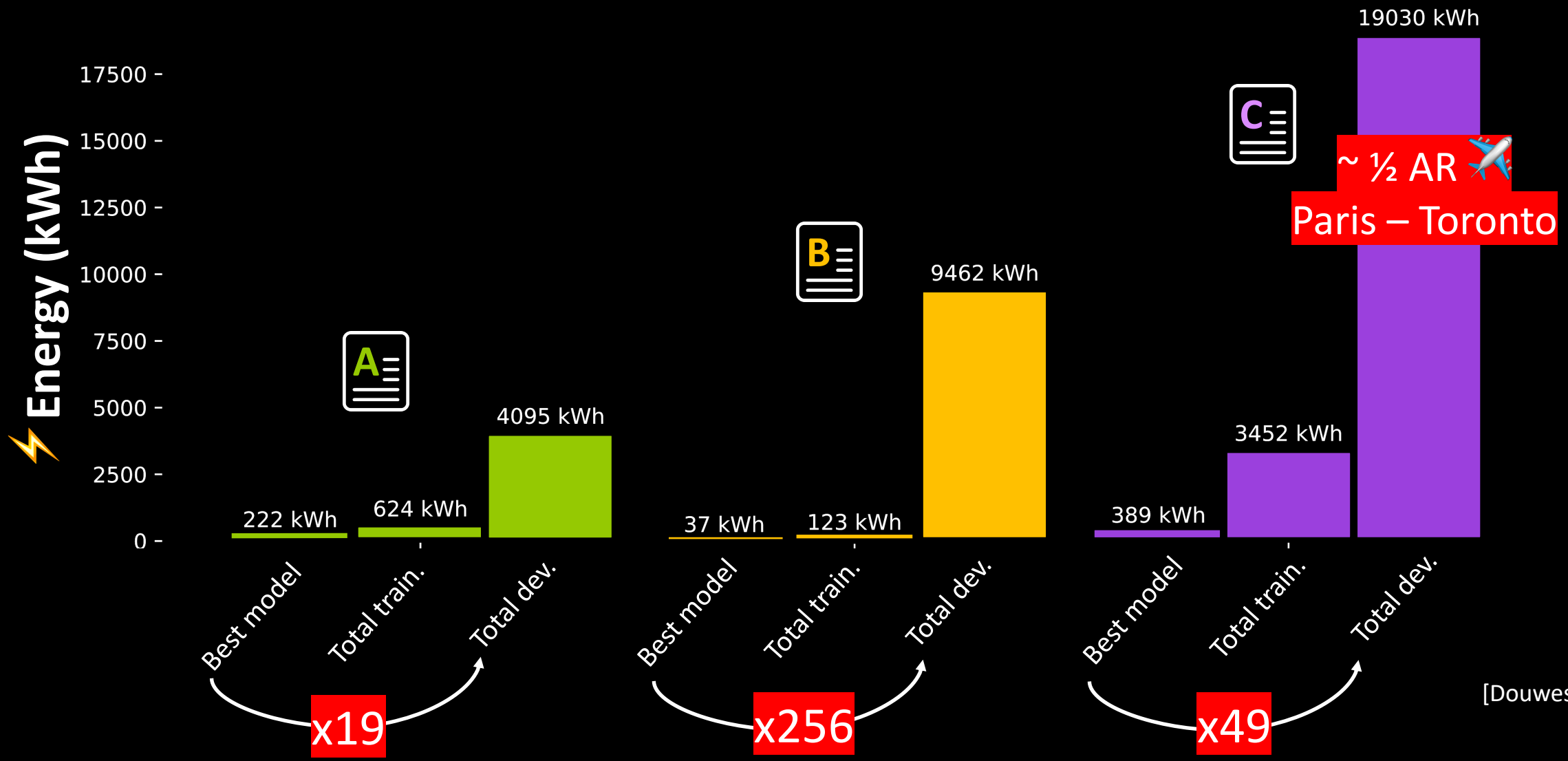
of members of the team



[Douwes 2027]

© Constance Douwes

Training vs. Development



[Douwes 2027]

© Constance Douwes

Inference costs

Navier Stokes

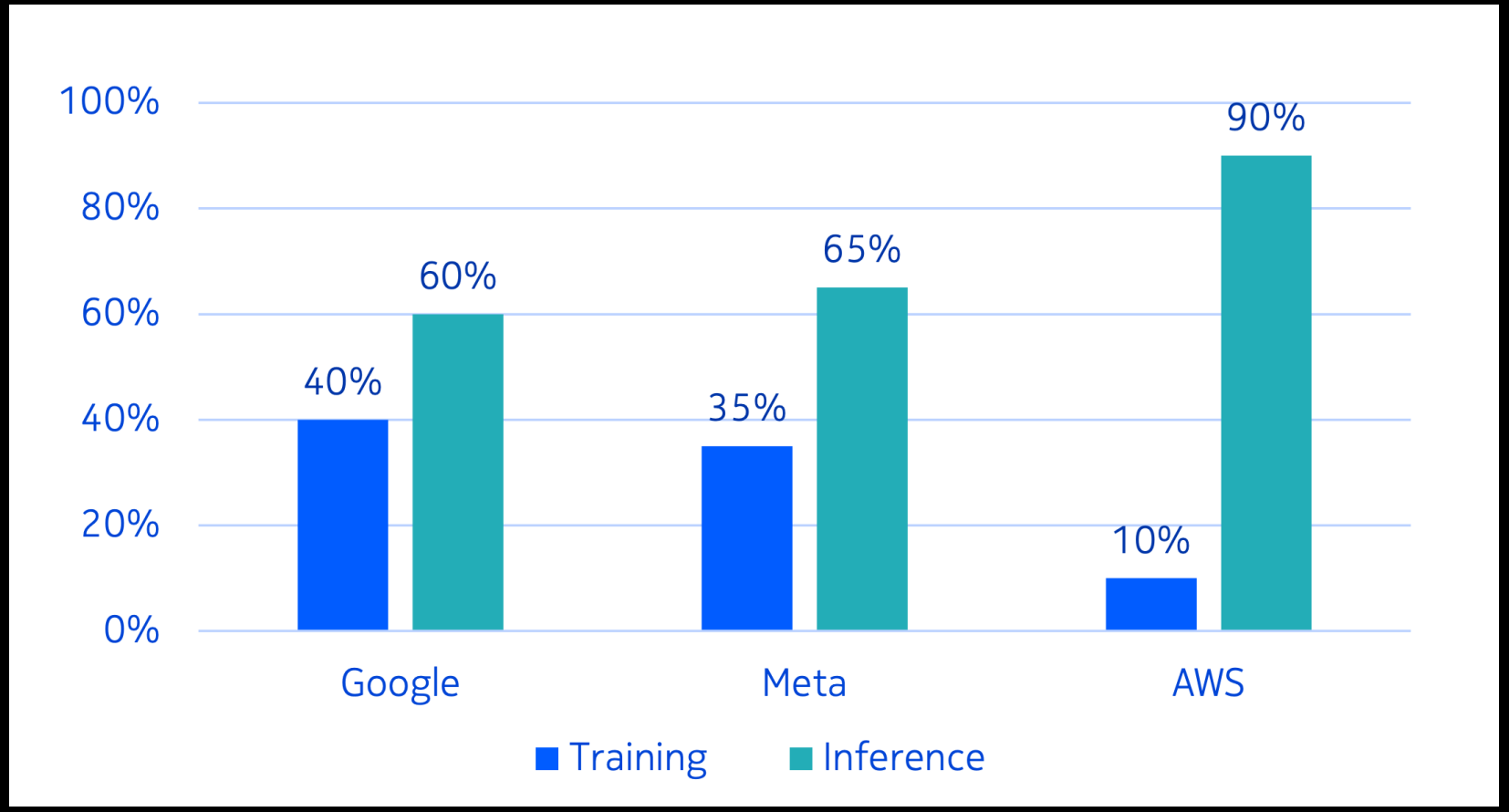
$$\frac{\partial \mathbf{u}}{\partial t} + (\mathbf{u} \cdot \nabla)\mathbf{u} = -\frac{1}{\rho}\nabla p + \mathbf{f} + \nu \Delta \mathbf{u}$$



- 88 hours
- 17 hours of verification
- 10 000 IA agents
- $10 million
- **??? = tCO2e ?**

© Constance Douwes

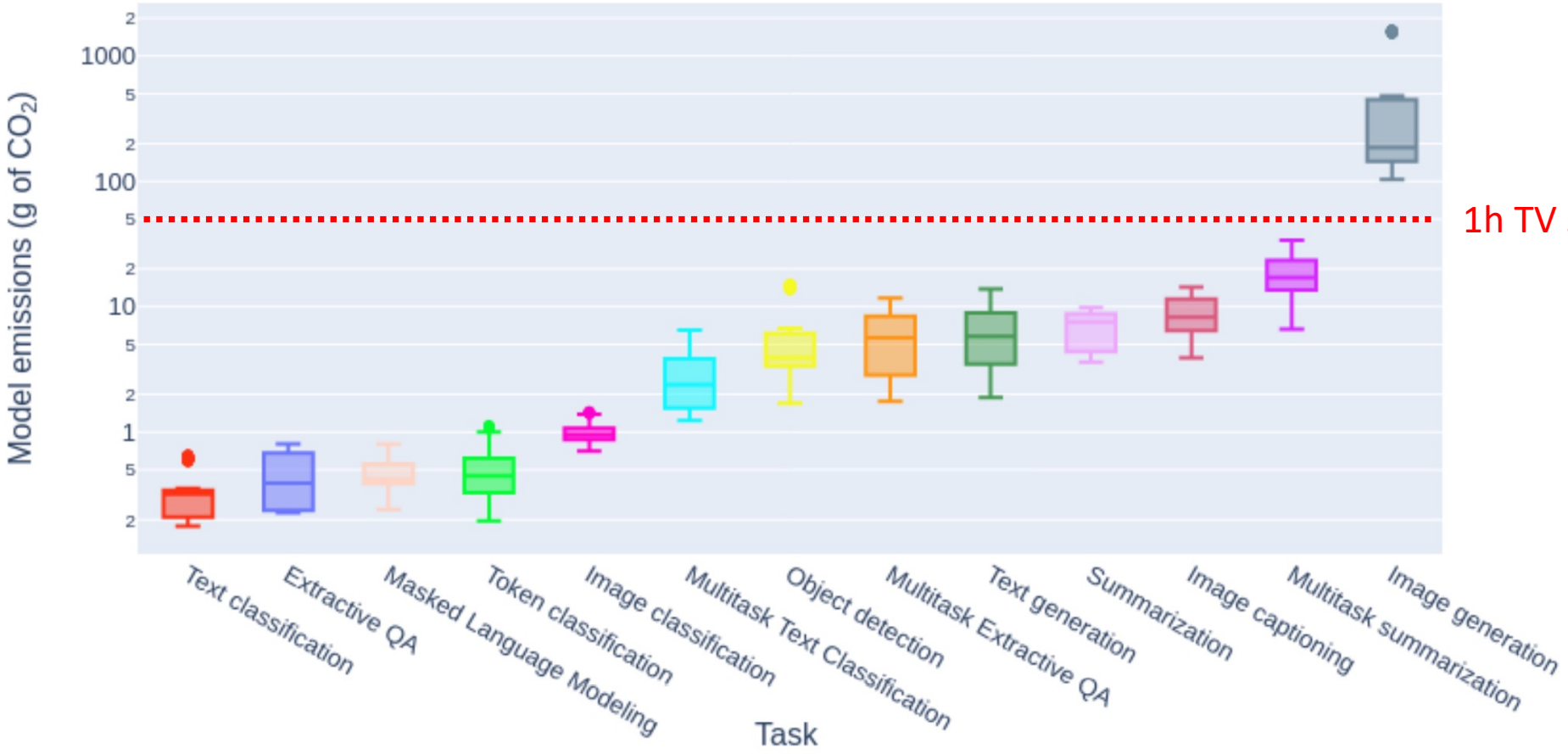
Training VS Inference



(from Sardi 2025; data from Patterson et al., 2022; Wu et al., 2022; De Chateauvieux et al., 2022)

© Constance Douwes

Emissions per task



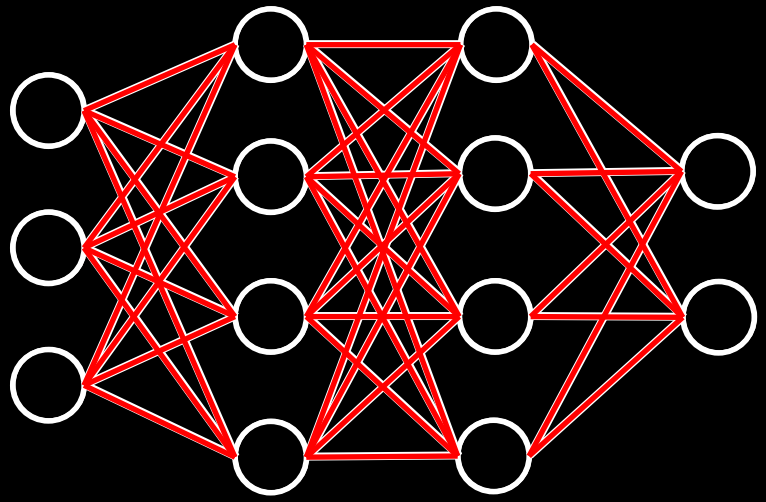
1h TV streaming ☺

[Luccioni 2023]

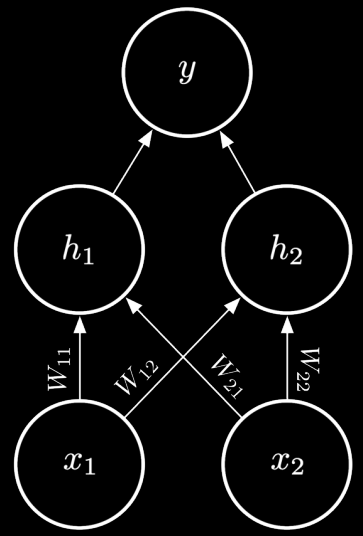
© Constance Douwes

Independant hardware metrics

Number of parameters



Number of operations (+,-,x, ÷)

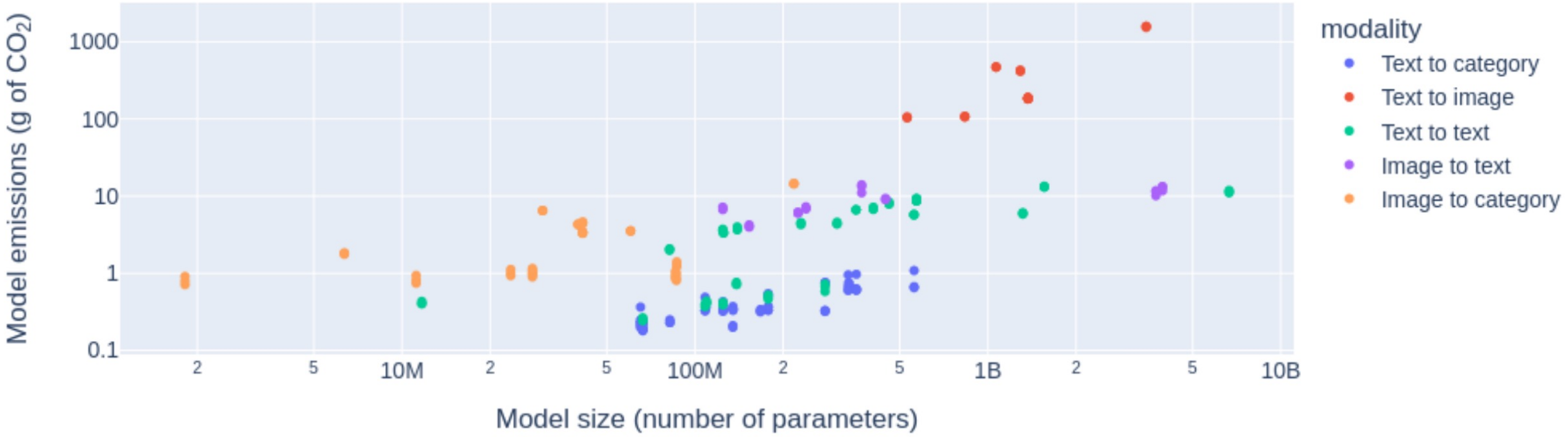


8 FPOs

$$y = x_1 \cdot W_{11} + x_1 \cdot W_{12} + x_2 \cdot W_{21} + x_2 \cdot W_{22} + b$$

© Constance Douwes

Emissions ≠ Size

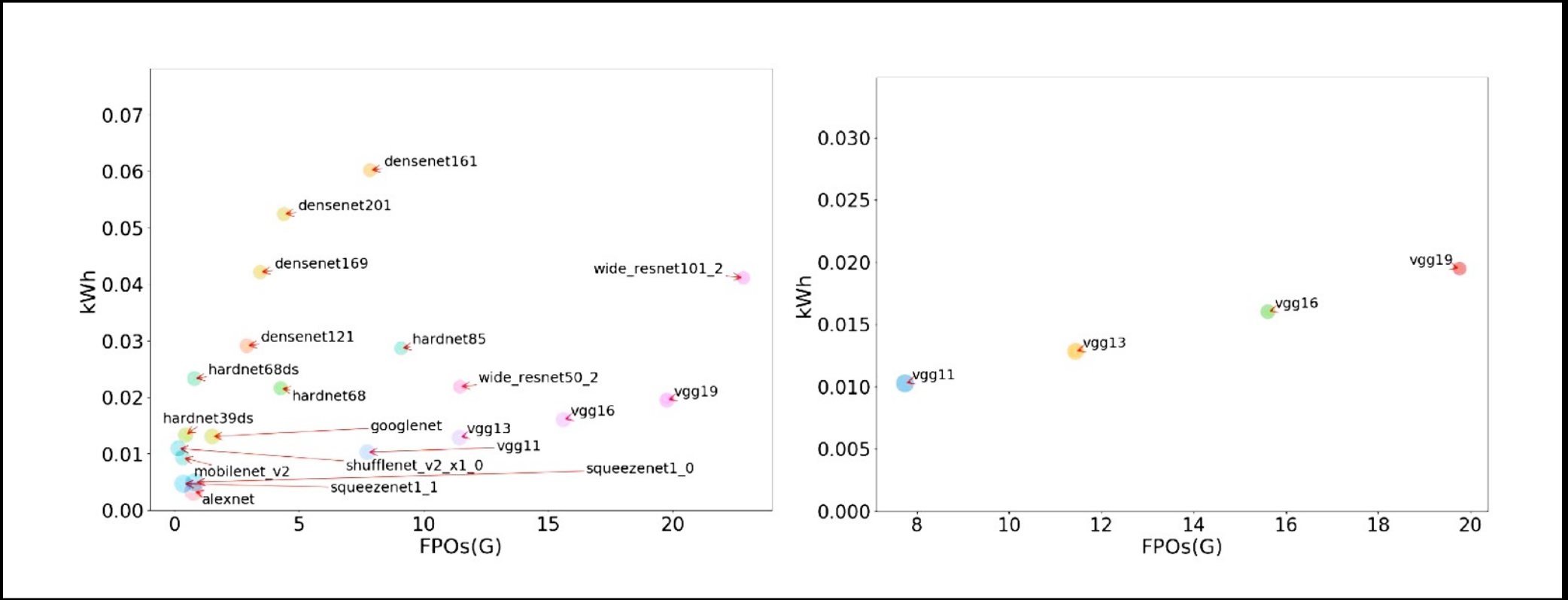


[Luccioni2024]

$$E_{network} \neq \alpha\ Size + \beta$$

© Constance Douwes

Energy ≠ FLOPs



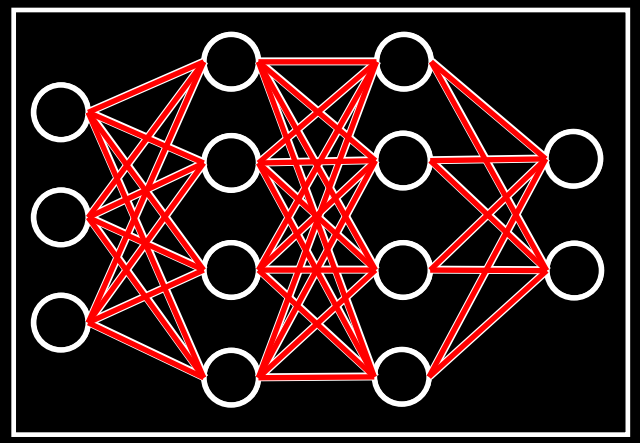
$$E_{network} \neq \alpha\, FPOs + \beta$$

© Constance Douwes

Per layer estimations

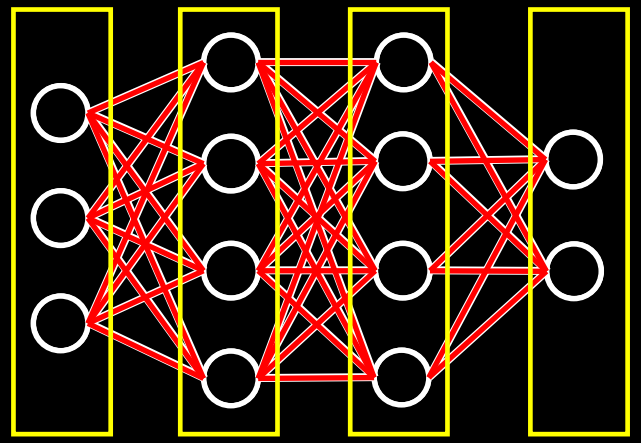
of any architecture

$E_{network}$



$\approx$

E_{l_1} E_{l_2} ... E_{l_N}



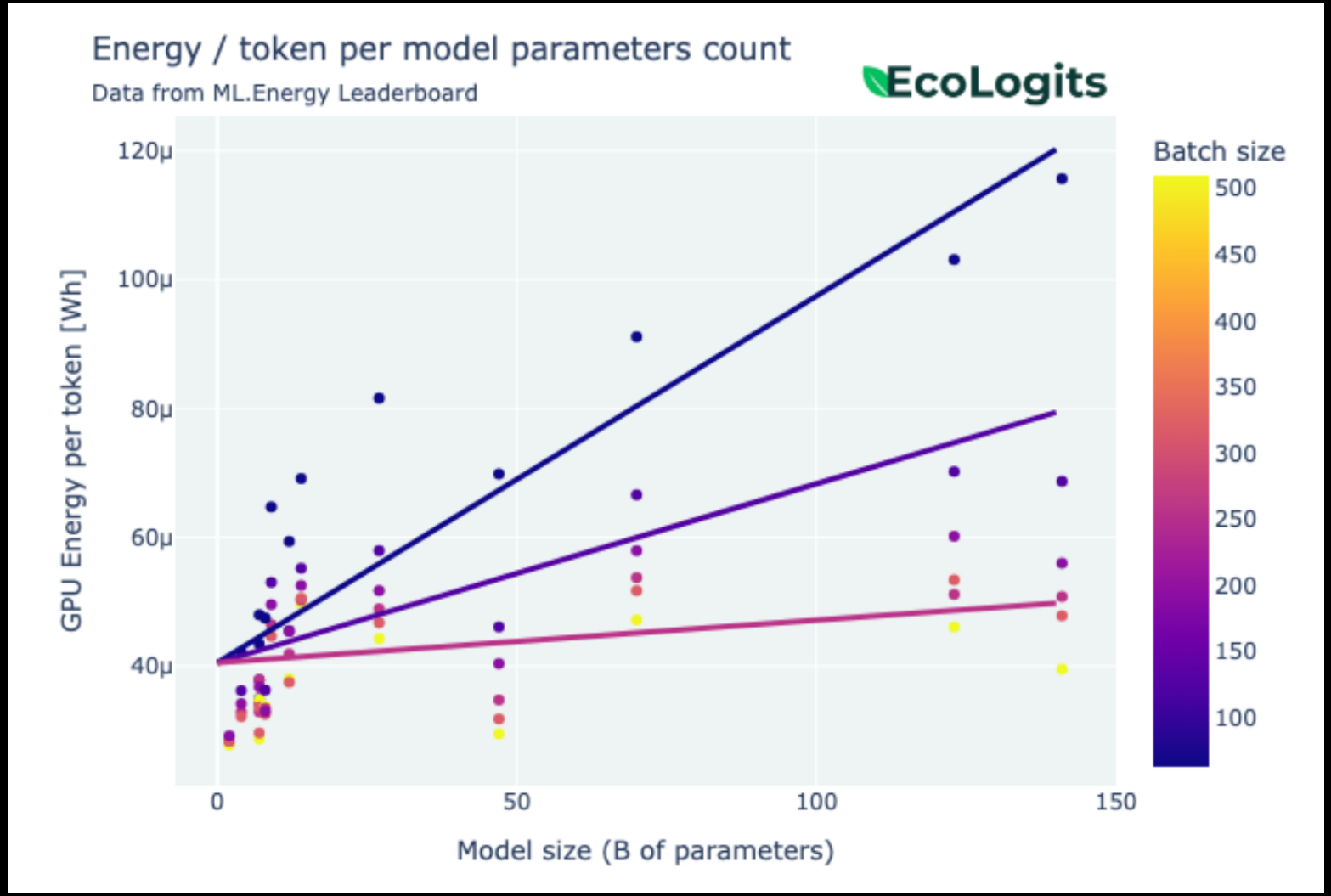
Additivity corrective term

$$= \alpha \sum_{i}^{N_{layers}} f_i(X_i)$$

Fitting function $f_i(X_i)$
Polynomial models
(i/o size, nb channels, FLOPs, etc…)

© Constance Douwes

Energy per token



© Constance Douwes

Manufacture cost





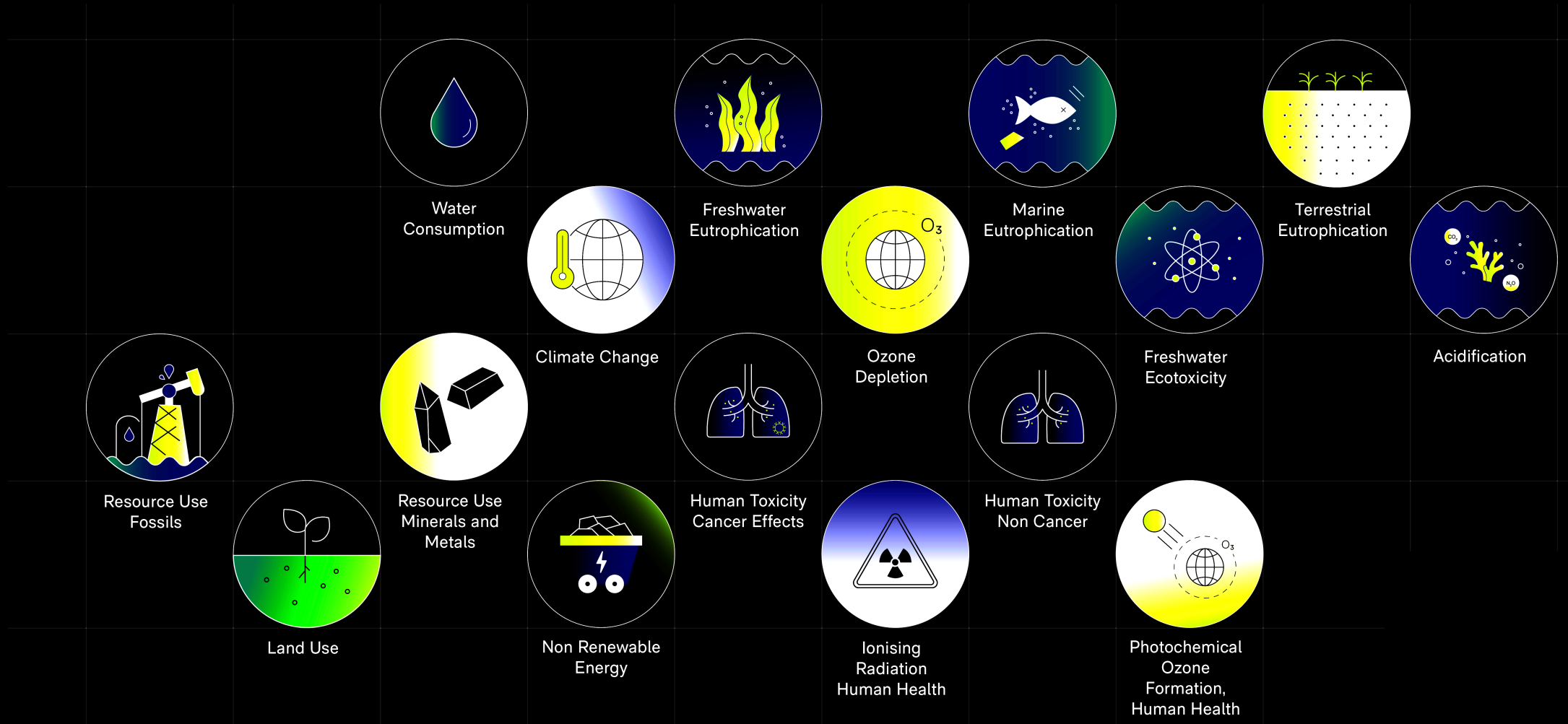
$$\frac{\text{Duration use}}{\text{Lifespan of the GPU}} \times CO_2\text{eq Production}$$

Allocation factor

- Meta : 70% Usage / 30% Manufacture
- BLOOM : 78% Usage / 22% Manufacture [Luccioni 2022]
- GenAI : 87% Usage / 13% Manufacture [Falk 2025]

© Constance Douwes

Impacts categories



© Constance Douwes

Text generation

Company mode ⓘ Expert mode ⓘ

Task: ✍️ Write a 5-page report

Provider: OpenAI

Model: Gpt 4.1

✍️ 5,000 output tokens · ⚠️ The model architecture has not been released and the model architecture is multimodal, expect lower precision.

Environmental impacts

⚡ Electricity consumption	🌍 Carbon Footprint	🚰 Water consumption	🪨 Metals & Minerals
9.22 Wh	3.85 gCO2eq	33.3 mL	18.2 µgSbeq
between 7.28-11.2 Wh	between 3.11-4.6 gCO2eq	between 26.3-40.3 mL	between 18-18.4 µgSbeq

https://calculator.ecologits.ai/

© Constance Douwes

Video generation

Company mode Expert mode

Task: 🎬 Generate a 4K video

Provider: OpenAI

Model: Sora 2 pro

🎬 3840x2160, 8s with audio

Environmental impacts

⚡ Electricity consumption	🌍 Carbon Footprint	🚰 Water consumption	🪨 Metals & Minerals
6.08 kWh	3.04 kgCO2eq	26.6 L	14.1 mgSbeq
between 4.5-7.65 kWh	between 2.32-3.76 kgCO2eq	between 19.7-33.5 L	between 14.1-14.2 mgSbeq

https://calculator.ecologits.ai/

© Constance Douwes

LCA example

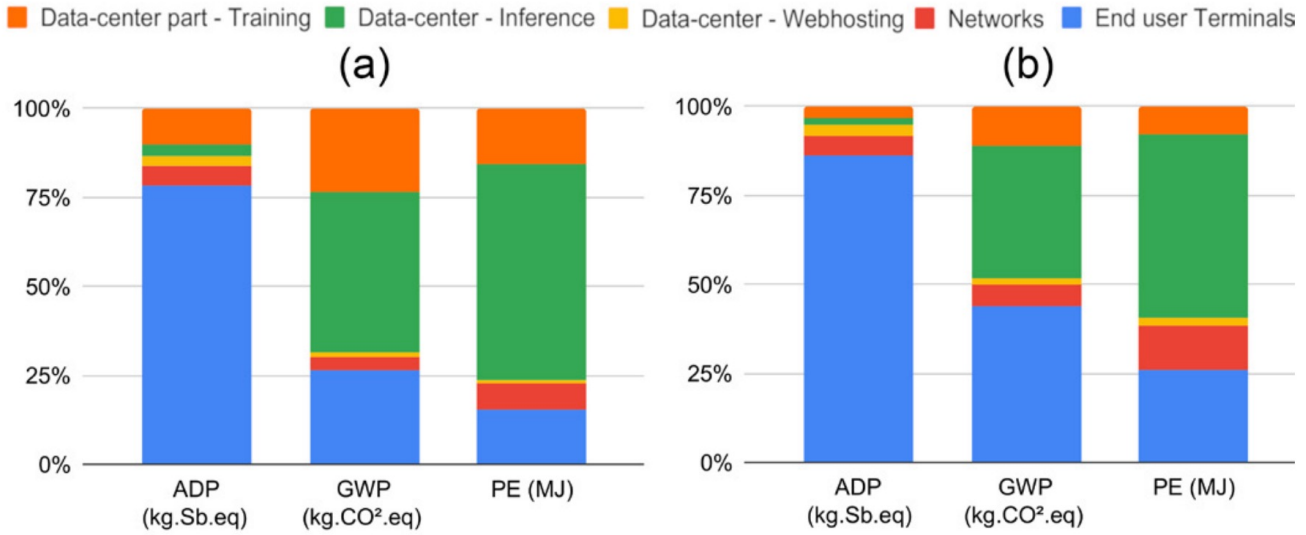


Stable diffusion

https://stablediffusionweb.com/fr

Table 2. Environmental impact of Stable Diffusion for FU1 and FU2

FU	Abiotic Depletion Potential (kgSb eq)	Warming Potential (kgCO² eq)	Primary energy (MJ)
FU1 - Single use of service	$6.72e^{-08}$	$7.84e^{-03}$	$2.02e^{-01}$
FU2 - A year of service	$4.64e^{+00}$	$3.60e^{+05}$	$8.93e^{+06}$

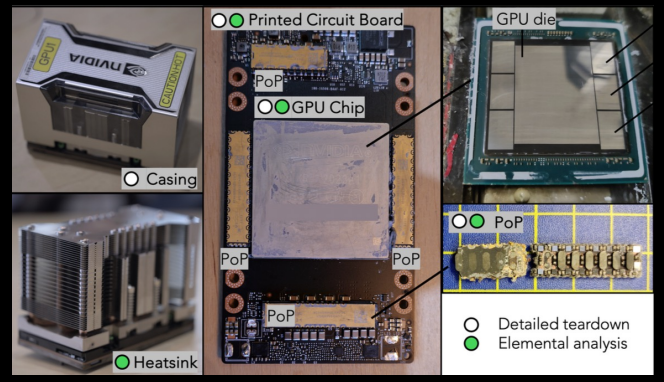


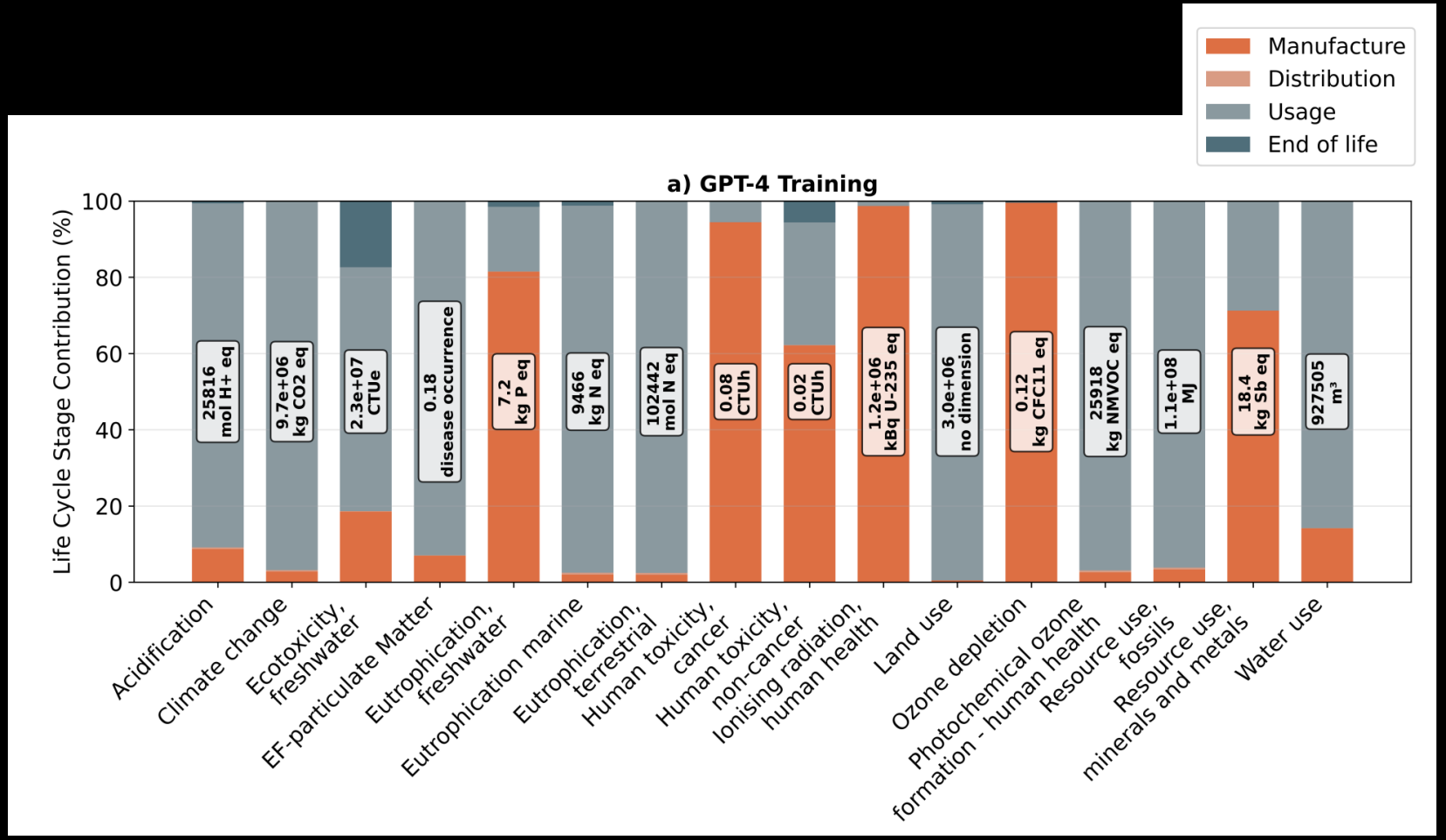
Source: Berthelot et al, 2024

© Constance Douwes

LCA example

Train GPT-4 model on Nvidia A100 SXM 40GB

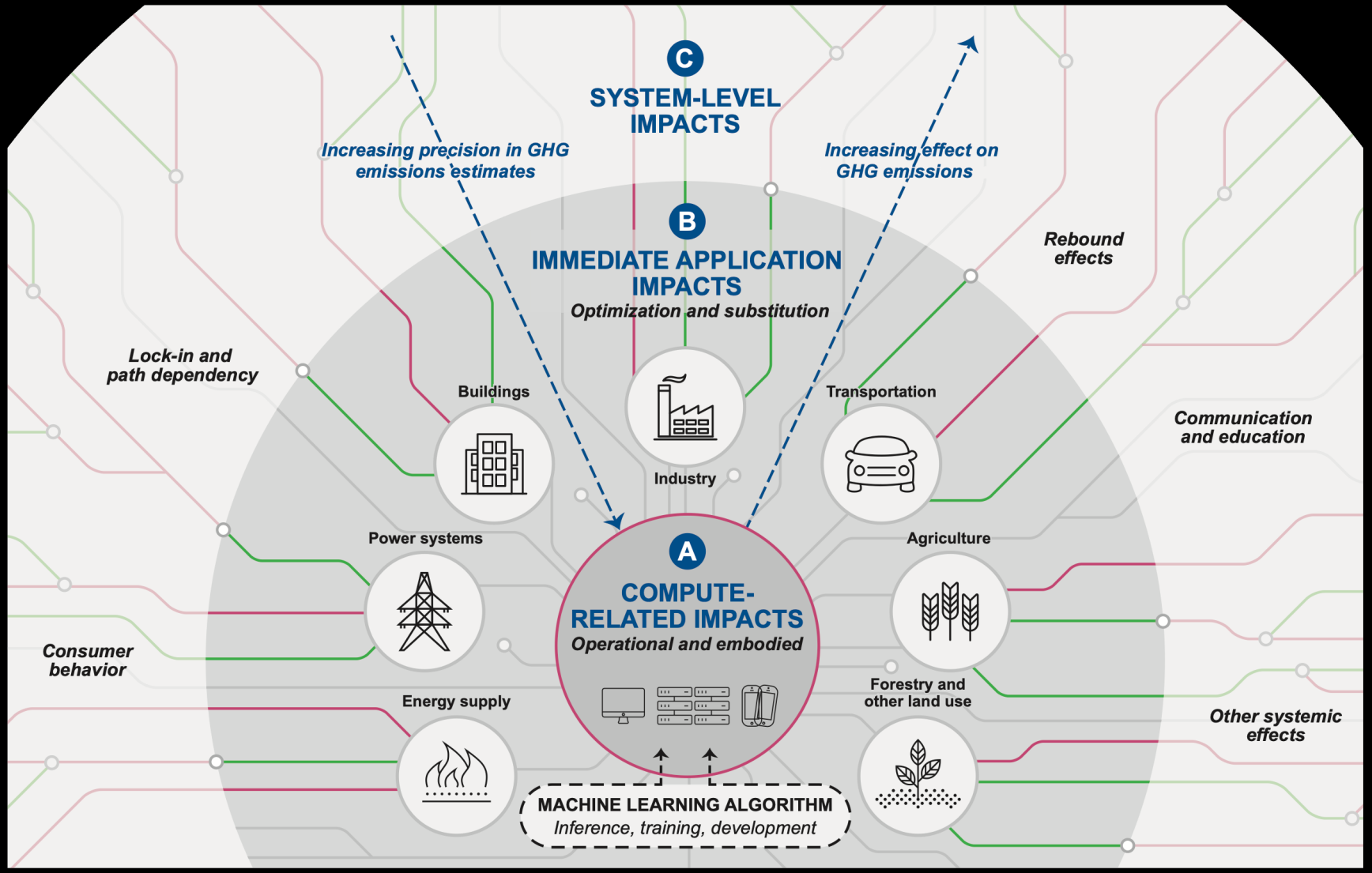




Falk, Ekchajzer, et al.. More than Carbon: Cradle-to-Grave environmental impacts of GenAI training on the Nvidia A100 GPU, 2025

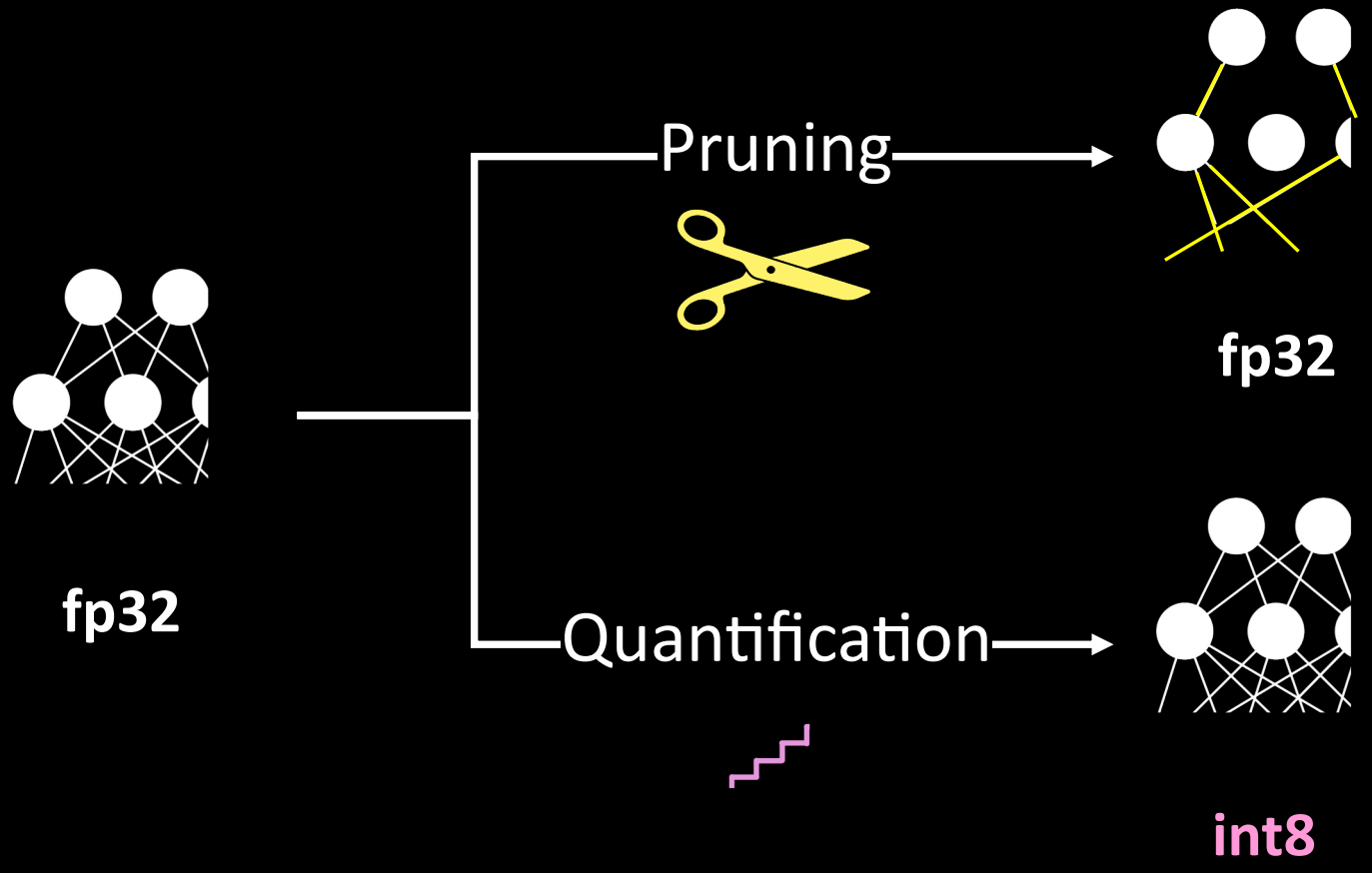
© Constance Douwes

1st, 2nd, 3rd order

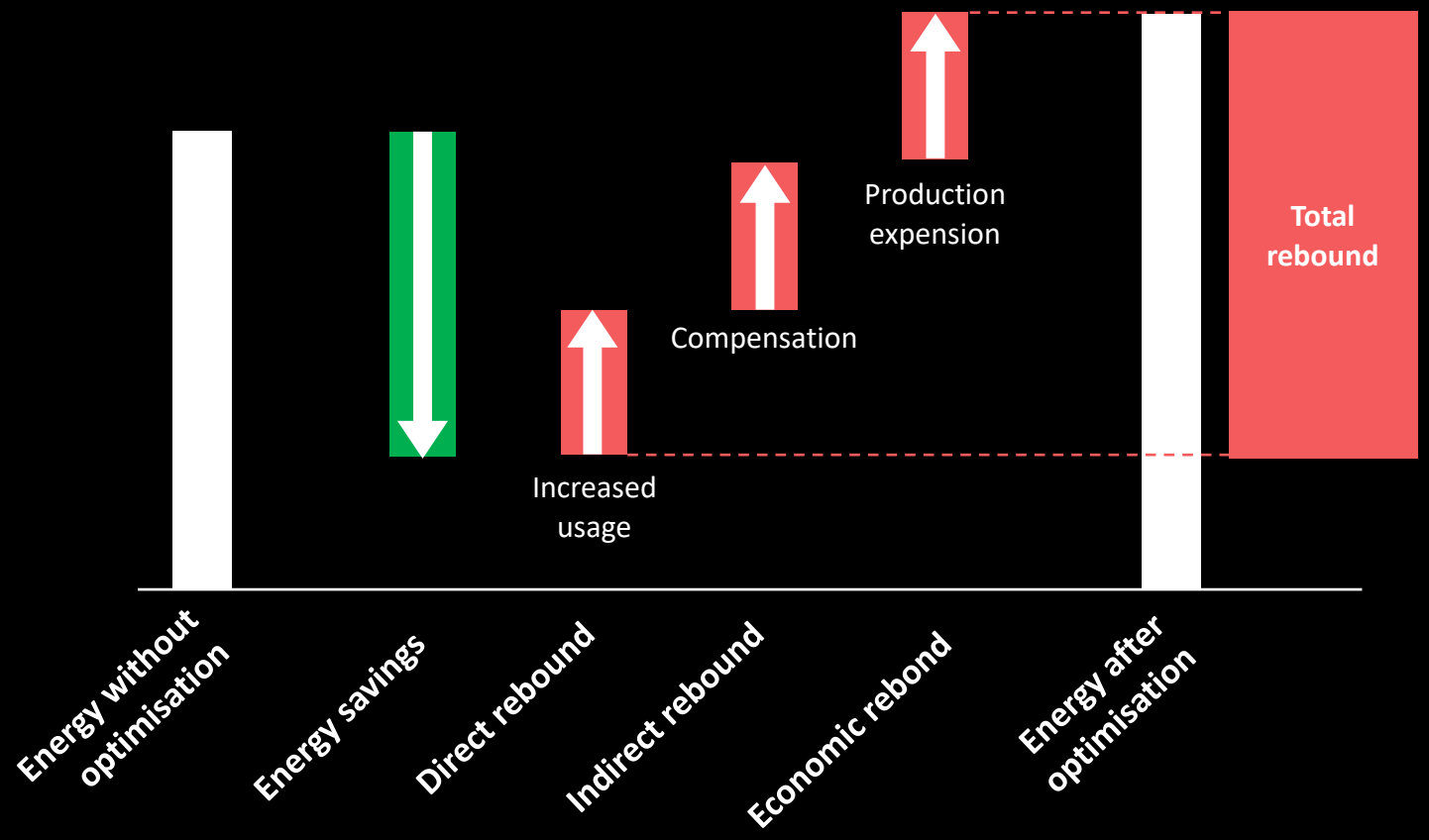


© Constance Douwes

Efficient AI



Rebound effect

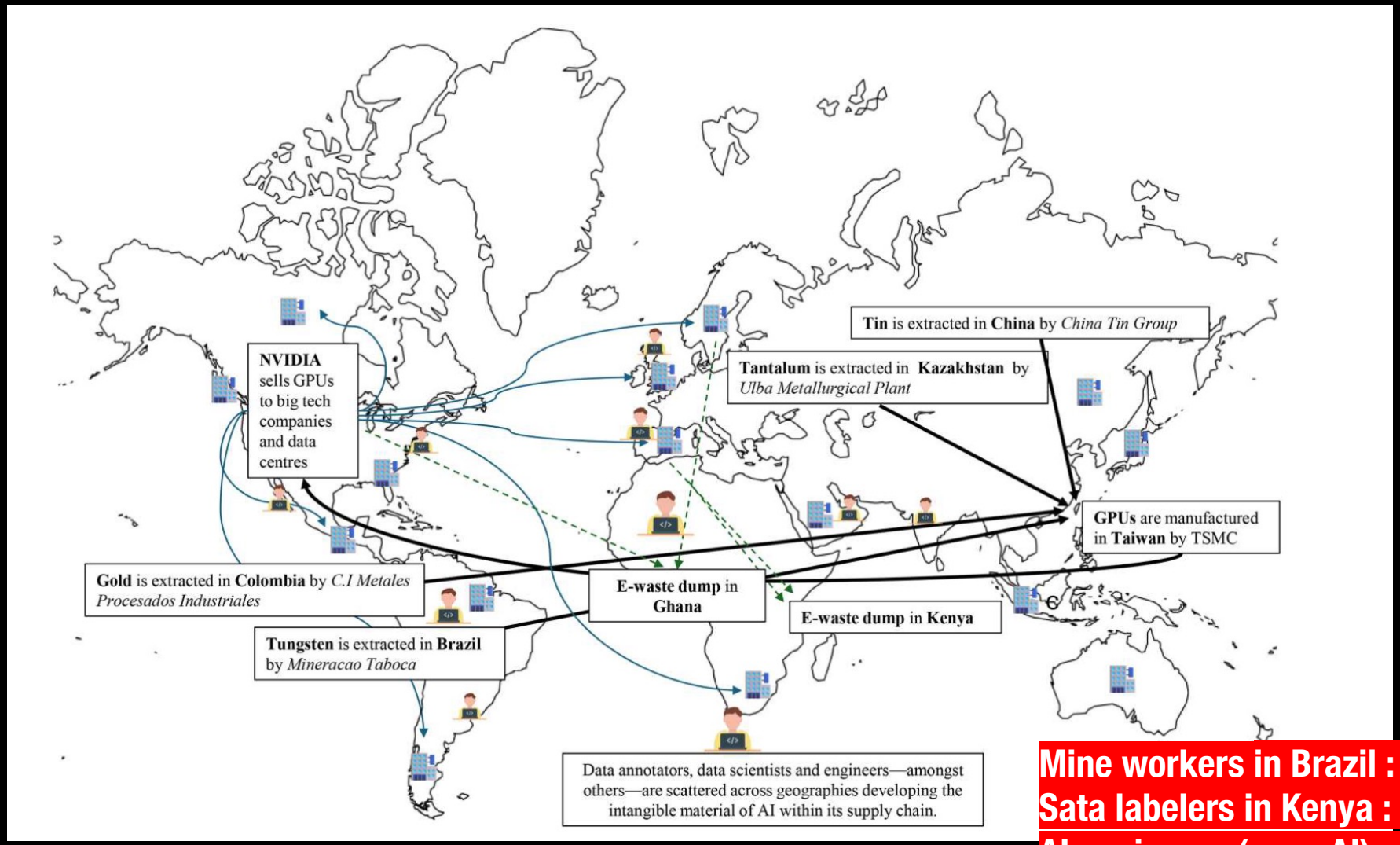


Frugality vs Efficiency

	Définition	Notions connexes	Raisonnement	Approche	Précisions
Efficience	Aptitude à optimiser les moyens alloués pour atteindre un résultat défini	Efficacité, optimisation	En relatif/par unité d'usage Le besoin prime : optimisation d'une solution jugée celle répondant le mieux au besoin	Recherche d'un optimum local ou d'un compromis sur un niveau de résultat fortement contraint	Prise en compte des effets de premier ordre pour les minimiser Prise en compte des parties prenantes de l'IA
Frugalité	Aptitude à se contenter d'un niveau de résultat jugé suffisant en redéfinissant les usages et les besoins	Sobriété (ou *Sufficiency* [10] en anglais)	En global La contrainte sur les ressources prime : recherche de la solution utilisant le moins de ressources possible et apportant une réponse satisfaisante au besoin	Recherche d'un optimum global ou d'un compromis large sur un niveau de résultat, ce qui nécessite d'élargir ou d'assouplir le besoin	Prise en compte des effets de premier ordre et de second ordre pour minimiser les impacts environnementaux négatifs Prise en compte de tous les acteurs au-delà des seules parties prenantes de l'IA

Référentiel général pour l'IA Frugale, AFNOR, 2025

Social impacts



[Valdivia 2025]

Mine workers in Brazil : 5EUR/hour
Sata labelers in Kenya : 2EUR/hour
AI enginners (open AI) : 75 EUR/hour

© Constance Douwes

Thanks !

Constance Douwes
MCF au LIS / Centrale Méditerranée
constance.douwes@lis-lab.fr