Impact environnemental de l'IA

Denis Trystram
Denis.Trystram@univ-grenoble-alpes.fr

24 octobre, 2024

A question to start

"How do you feel about the future of the earth?"

A question to start

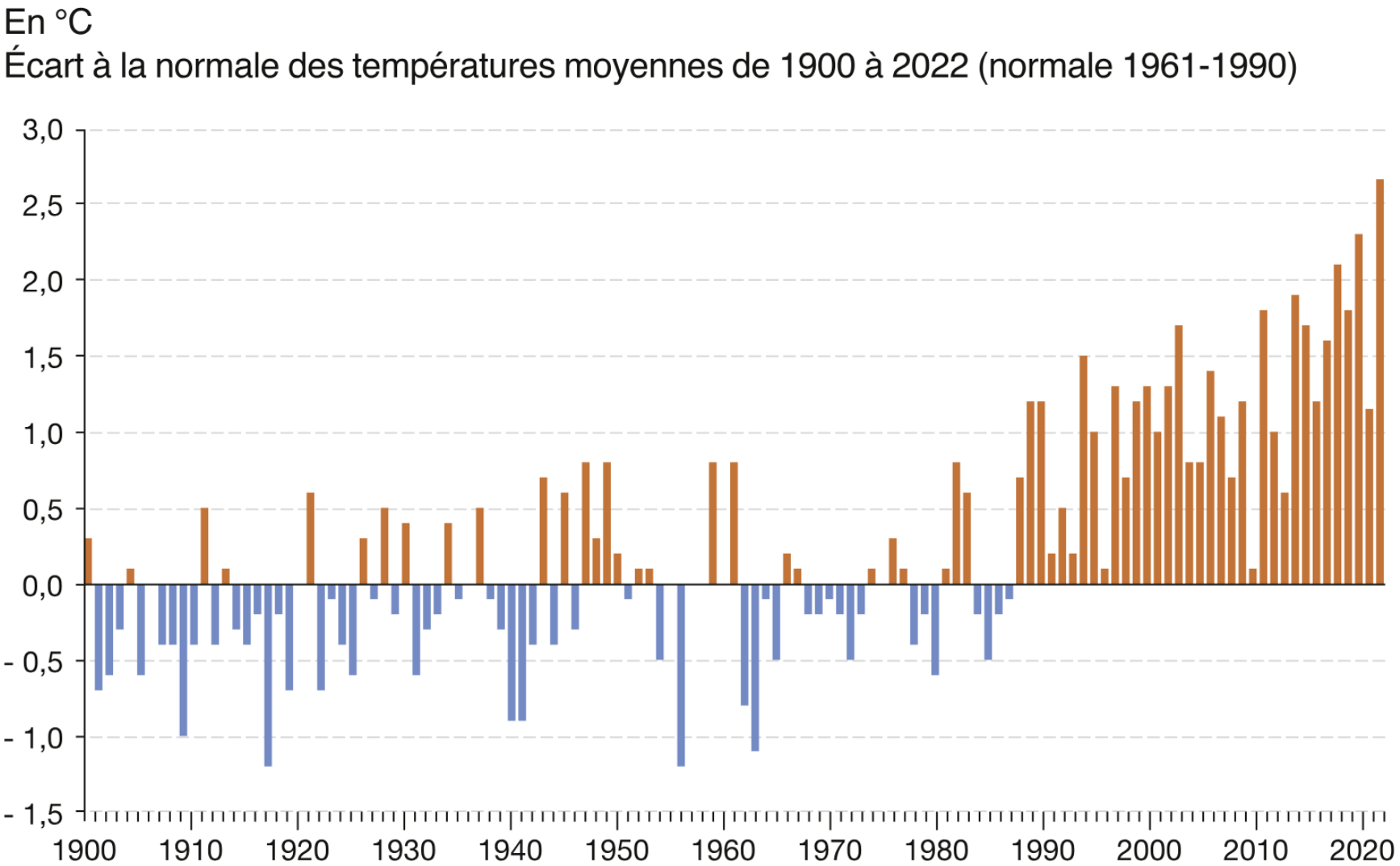
"How do you feel about the future of the earth?"

Anxious, Curious, Indifferent, Confident, Revolted, Militant, etc.

Some news about the world...

- more and more extreme weather events
  floods, heatwaves, tornadoes, drought, forest fires, etc.
- visible signs of a major environmental crisis
- partial awareness of people
- but, a possible shift to action

ÉVOLUTION DE LA TEMPÉRATURE MOYENNE ANNUELLE EN FRANCE MÉTROPOLITAINE DEPUIS 1900

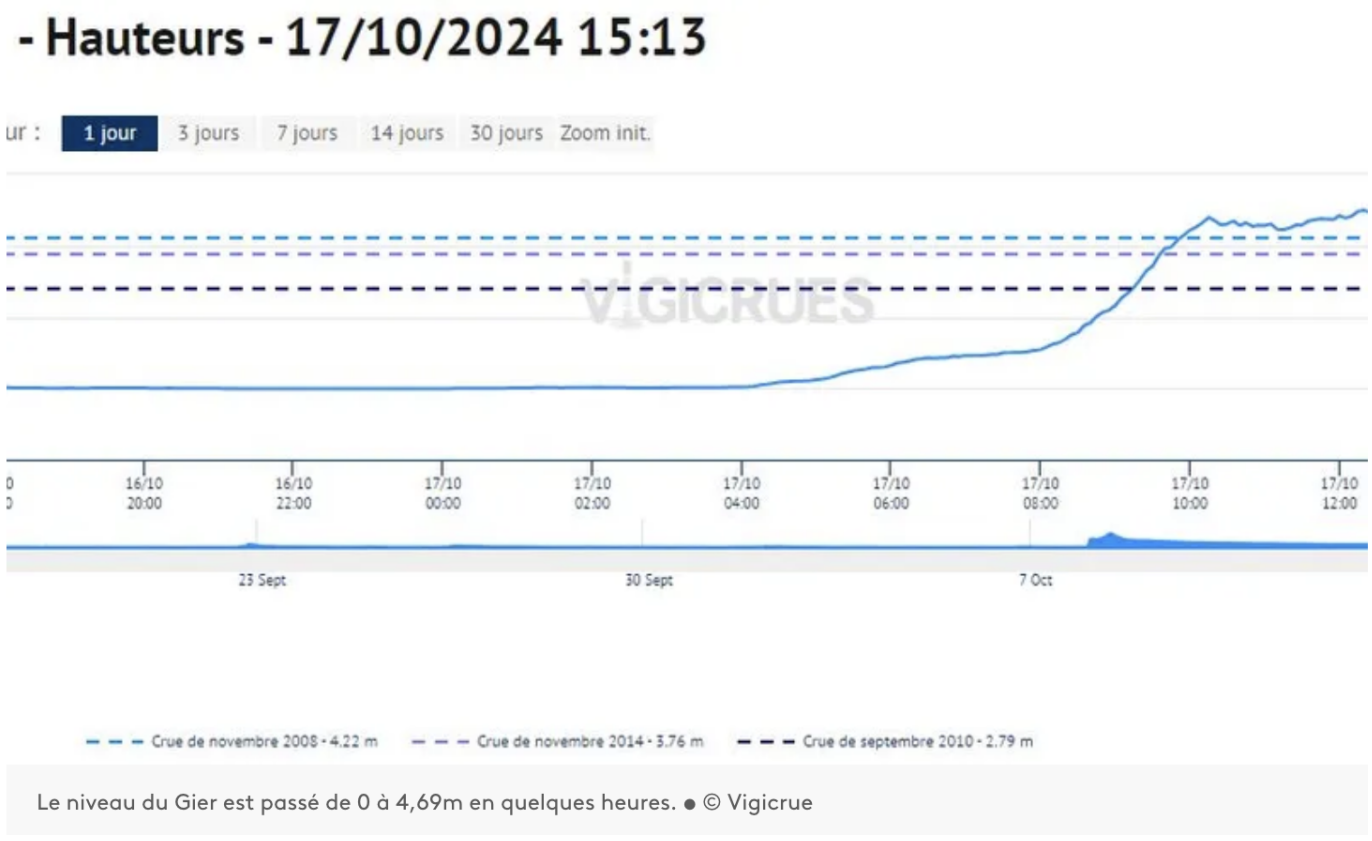


Source : Chiffres clés du climat 2023, ministère de la transition

Last friday in my region

Givors particulièrement touchée

A Givors, les niveaux des crues de 2008 et 2014 sont dépassés. Entre 04H00 et 10H00, le niveau est passé brusquement de 0 à 4,69m.



source: France TV info région AURA

Les images impressionnantes à Rive-de-Gier (Loire) et à Tence (Haute-Loire). © Radio France - Julien Frenoy / Aurélie Jacquand

A first observation

The environnemental crisis is a **reality**

...and its origin comes from **human activities**[1].

- The global warming comes from GHG emissions (mainly CO_2).

[1] as it is assessed by most scientists and IPCC reports

Objectives of this presentation

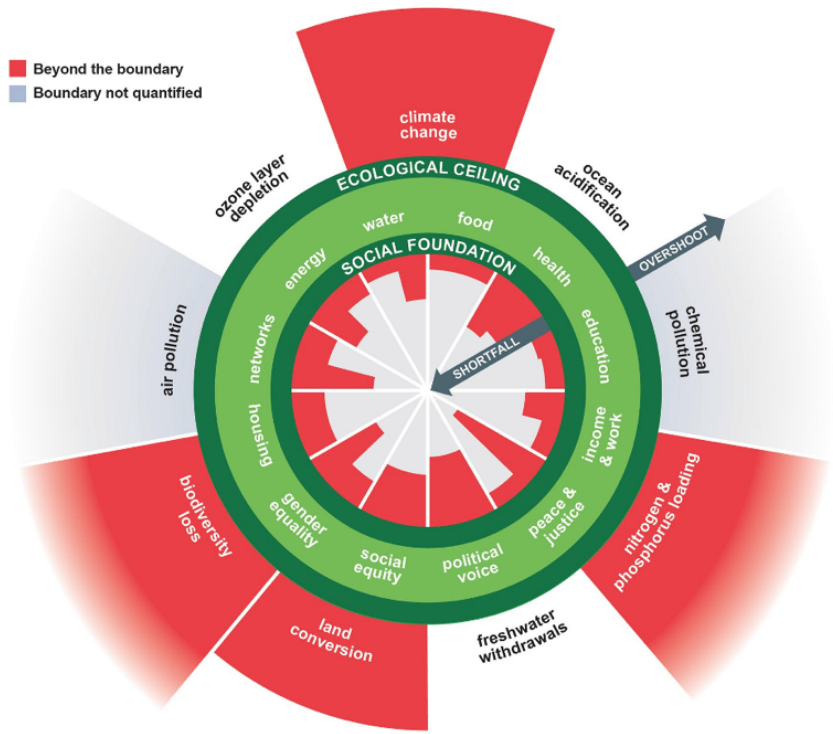
- **Provide a quick overview of the current state of the environmental crisis**
- Show the dynamics of AI in this crisis
- Give some elements to better understand/use machine learning.

The environmental question

From effects to causes?

The Donught theory

- Definition: The safe and just space for humanity lies between the environmental ceiling (9 earth limits) and the social floor (11 social objectives)



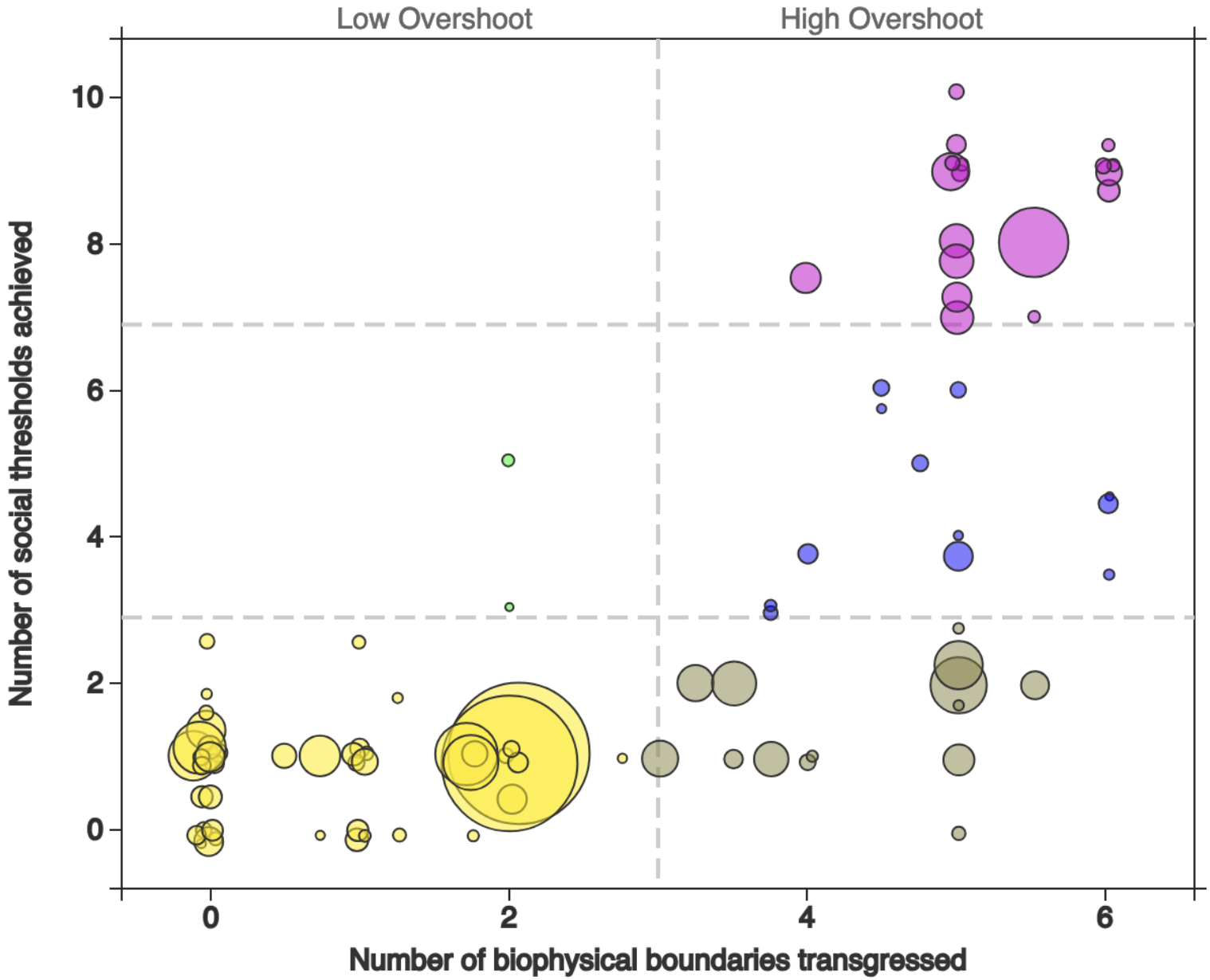
source: Kate Raworth, 2017

Let us play together

Guess who is where

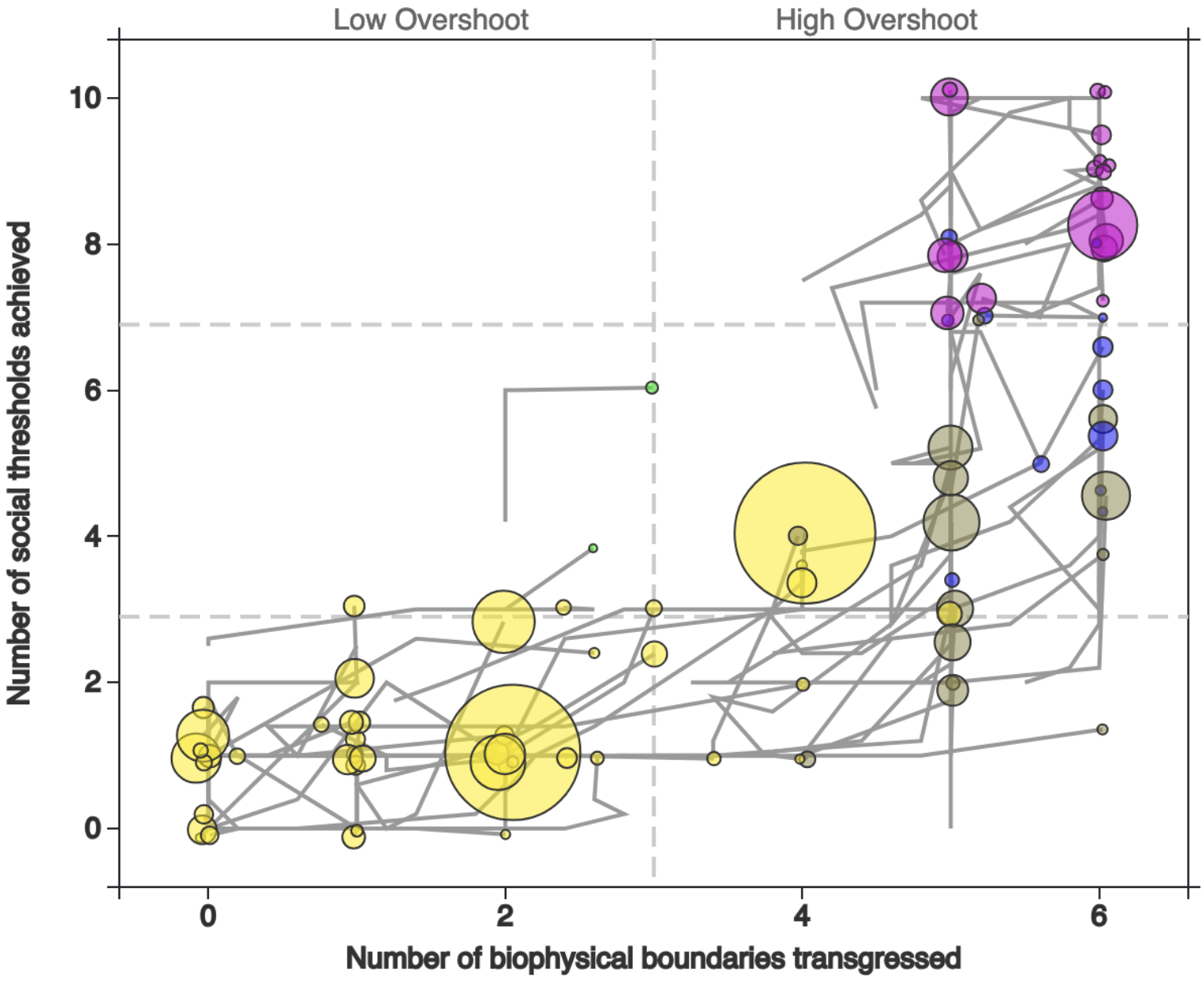
Table 1. Country performance with respect to per capita biophysical boundaries

Biophysical Indicator	*N*	Planetary Boundary	Per Capita Boundary	Countries Within Boundary (%)
CO_2 Emissions	145	2 °C warming	1.61 t CO_2 y^{-1}	34
Phosphorus	144	6.2 Tg P y^{-1}	0.89 kg P y^{-1}	44
Nitrogen	144	62 Tg N y^{-1}	8.9 kg N y^{-1}	45
Blue Water	141	4000 km^3 y^{-1}	574 m^3 y^{-1}	84
eHANPP	150	18.2 Gt C y^{-1}	2.62 t C y^{-1}	44
Ecological Footprint	149		1.72 gha y^{-1}	43
Material Footprint	144		7.2 t y^{-1}	44



O'Neill et al. A good life for all within planetary boundaries. Nat Sustain (2018).

https://goodlife.leeds.ac.uk/national-trends/pathways/



The energy paradox

- The steam engine in the late 18th century (coal)
- One century later, the oil revolution $\rightarrow$ What was the real advantages?

source: Jean-Baptiste Fressoz

All activities need energy.

Energy is always a superposition of the successive energy types!

Two examples

- In 1900, England was gobbling up 4.5 million m^3 of wood a year for use as props in mine galleries.
  In the 1750s, the English burned 3.6 million m^3.
  Thus, just to extract coal, the English used more wood in 1900 than they had burnt in 1750!

All activities need energy.

Energy is always a superposition of the successive energy types!

Two examples

- In 1900, England was gobbling up 4.5 million m^3 of wood a year for use as props in mine galleries.
  In the 1750s, the English burned 3.6 million m^3.
  Thus, just to extract coal, the English used more wood in 1900 than they had burnt in 1750!
- Oil is used to run cars.
  Back in the 1930s, it took around 7 tons of coal to make a car, i.e. as much coal by weight as the oil it burned during its lifetime.

Focus on electricity

- The electricity fairy
  The usage of digital devices is mostly based on electricity.
  The building needs a lot of water and metals (around 70 to make a smartphone).
- According to AIE, the proportion of decarbonized electricity may reach 42% in 2030[2].

[2] However, renewable energies are not really decarbonized

Focus on electricity

- The electricity fairy
  The usage of digital devices is mostly based on electricity.
  The building needs a lot of water and metals (around 70 to make a smartphone).
- According to AIE, the proportion of decarbonized electricity may reach 42% in 2030[2].

Let us take some more time for understanding the global landscape...

[2] However, renewable energies are not really decarbonized

Global warming

GHG and Carbon cycle

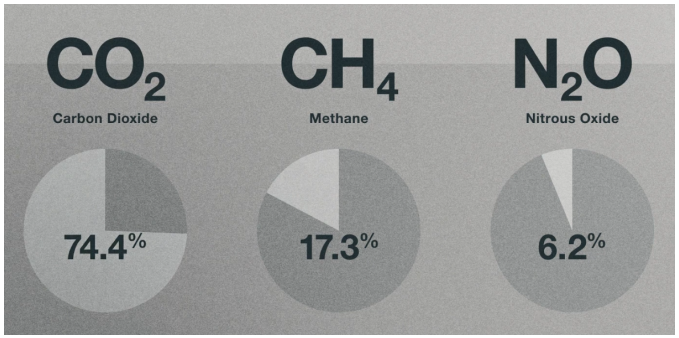
A gas in the atmosphere that intercepts infrared radiation emitted by the earth's surface.

GHG naturally exist, the Earth system was perfectly well-balanced for years, giving humans the condition for life.
(750 GTons/year emitted and absorbed by carbon sinks).

- we know: H_2O and CO_2
- we know less CH_4, N_2O and O_3
- The three CO_2 CH_4 and N_2O cover more than 96 % of the seven GES of the Kyoto protocol.
- GHGs remain in the atmosphere for a long time: more than 100 years for CO_2 !

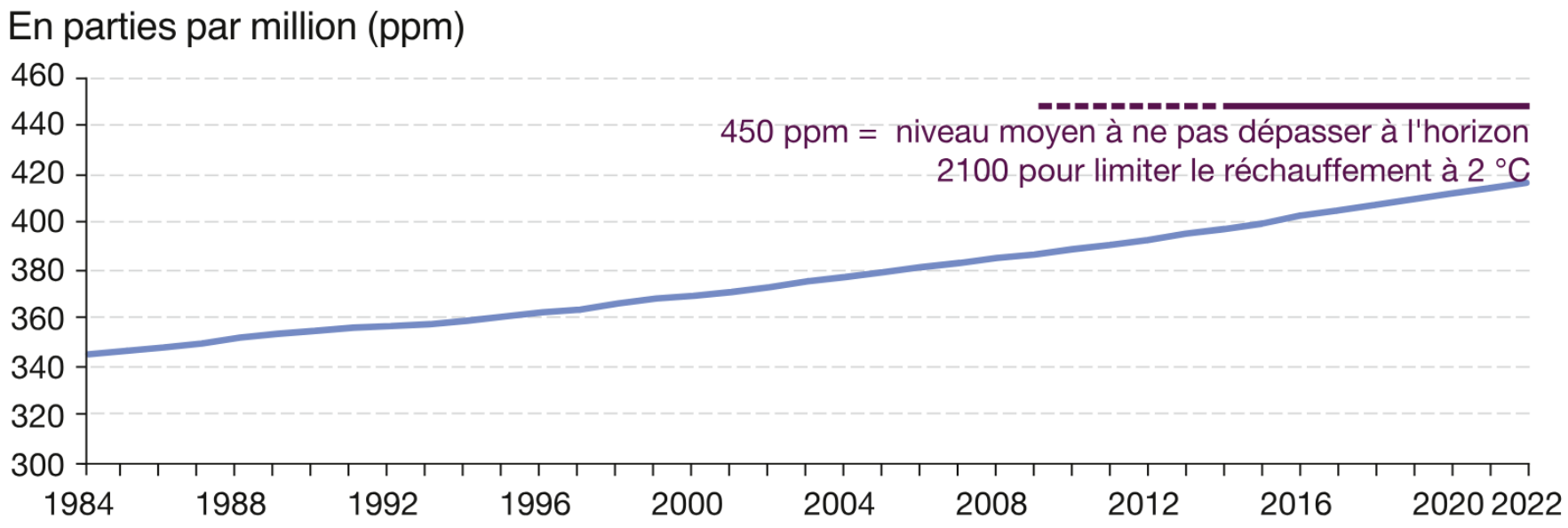
Why studying CO_2 emissions?

- the most of importance



- well-studied, with clear and reliable evaluation parameters (even with some uncertainties).
- visible and understandable to everyone.

Concentration in CO_2



Source : *National Oceanic and Atmospheric Administration (NOAA), USA, 2023*

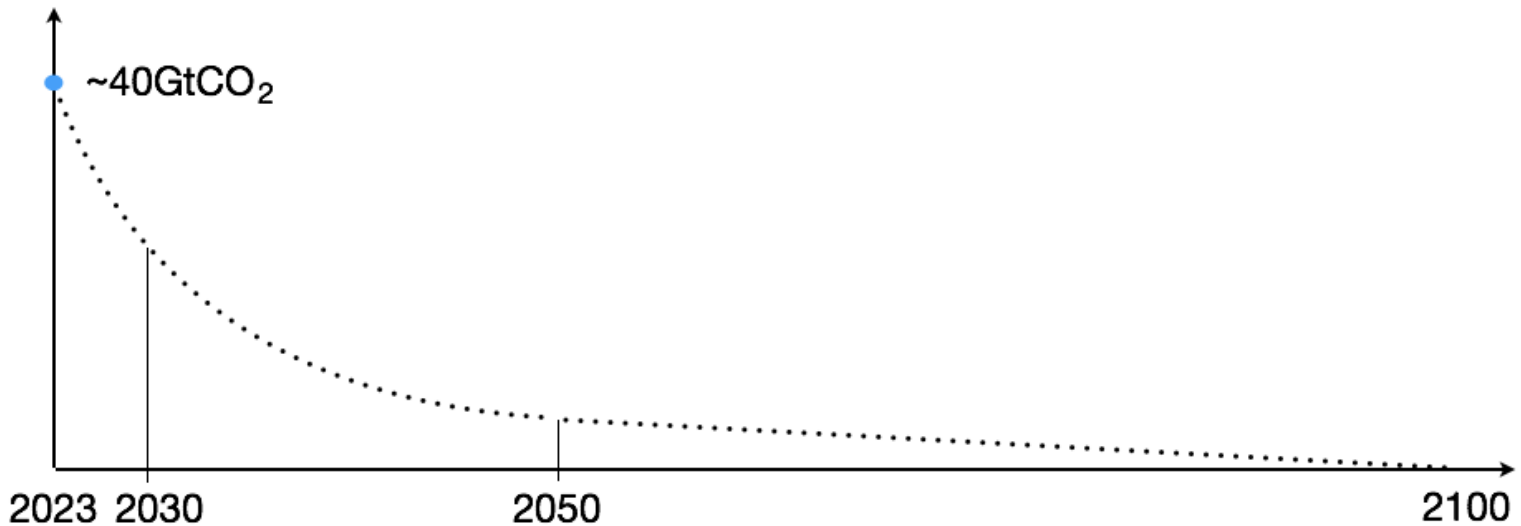
Concept of remaining budget

- We emit 40 about GigaTons of CO_2 per year (worldwise)[3].
- Maximum budget remaining to limit the warming
  - below 1.5 degrees: 325 Gt of CO_2
  - below 2 degrees: 1075 Gt of CO_2

[3] with 10% uncertainty. Never fight with numbers...

Concept of remaining budget

- We emit 40 about GigaTons of CO_2 per year (worldwise)[3].
- Maximum budget remaining to limit the warming
  - below 1.5 degrees: 325 Gt of CO_2
  - below 2 degrees: 1075 Gt of CO_2

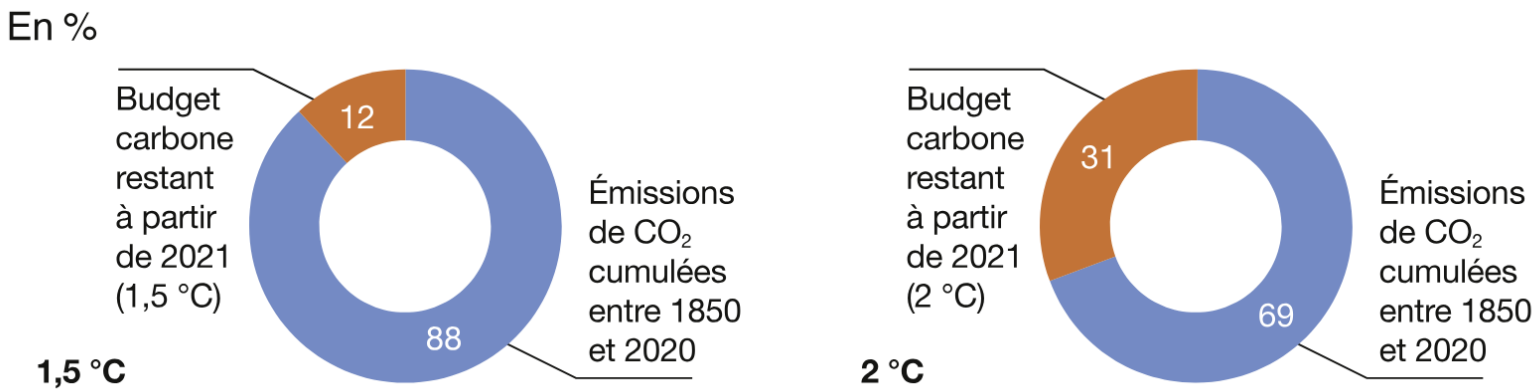


[3] with 10% uncertainty. Never fight with numbers...

- The situations are very different all over the world.
- Targeting 2050, we must reduce the emissions about 7 to 8% per year, for reaching the neutrality in 2100.

Prospective scenarios (global)

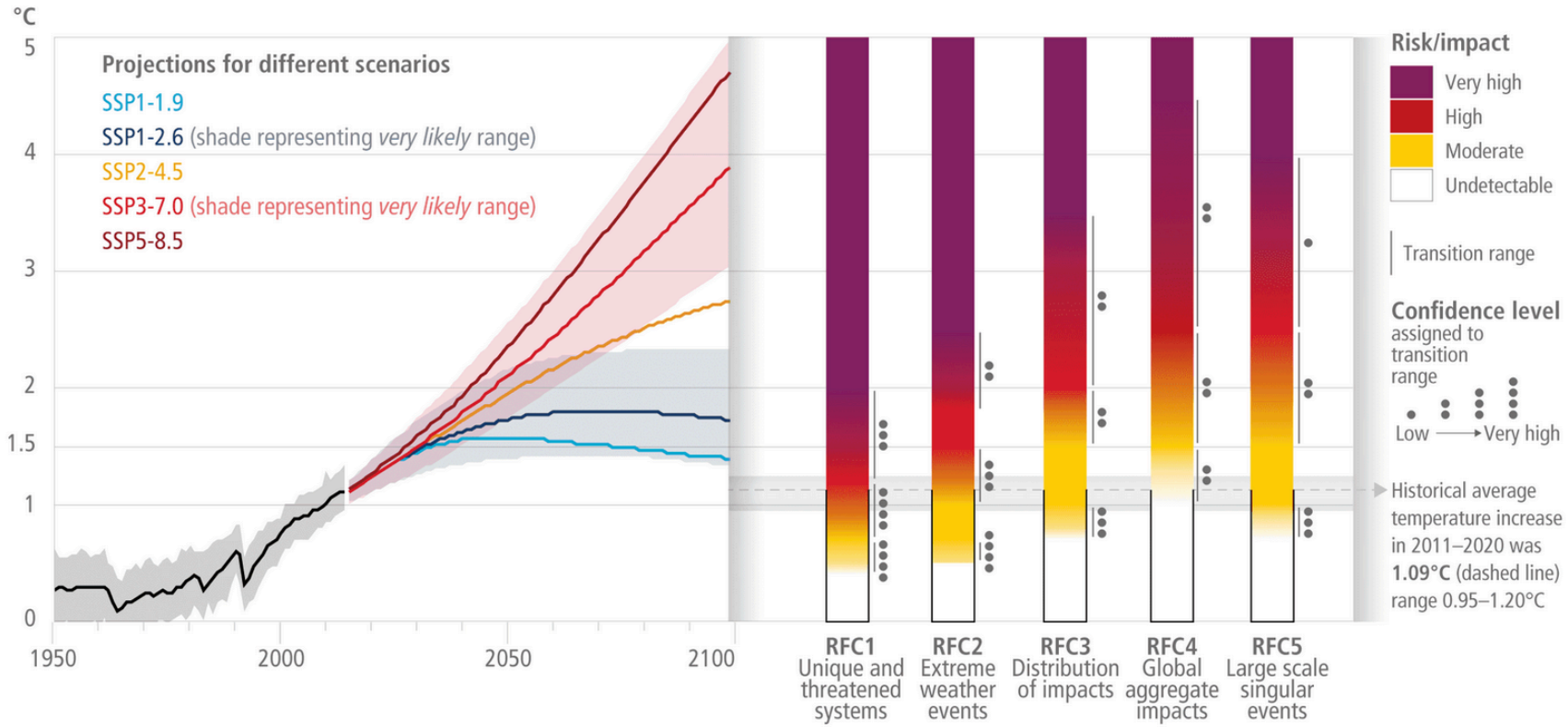
The Paris agreements established in 2015 targeted SSP1 with 1.5 degrees
Today, the trajectory is SSP2-4.5 with 2 degrees



Source: ADEME/ARCEP et chiffre clés 2023 du ministère

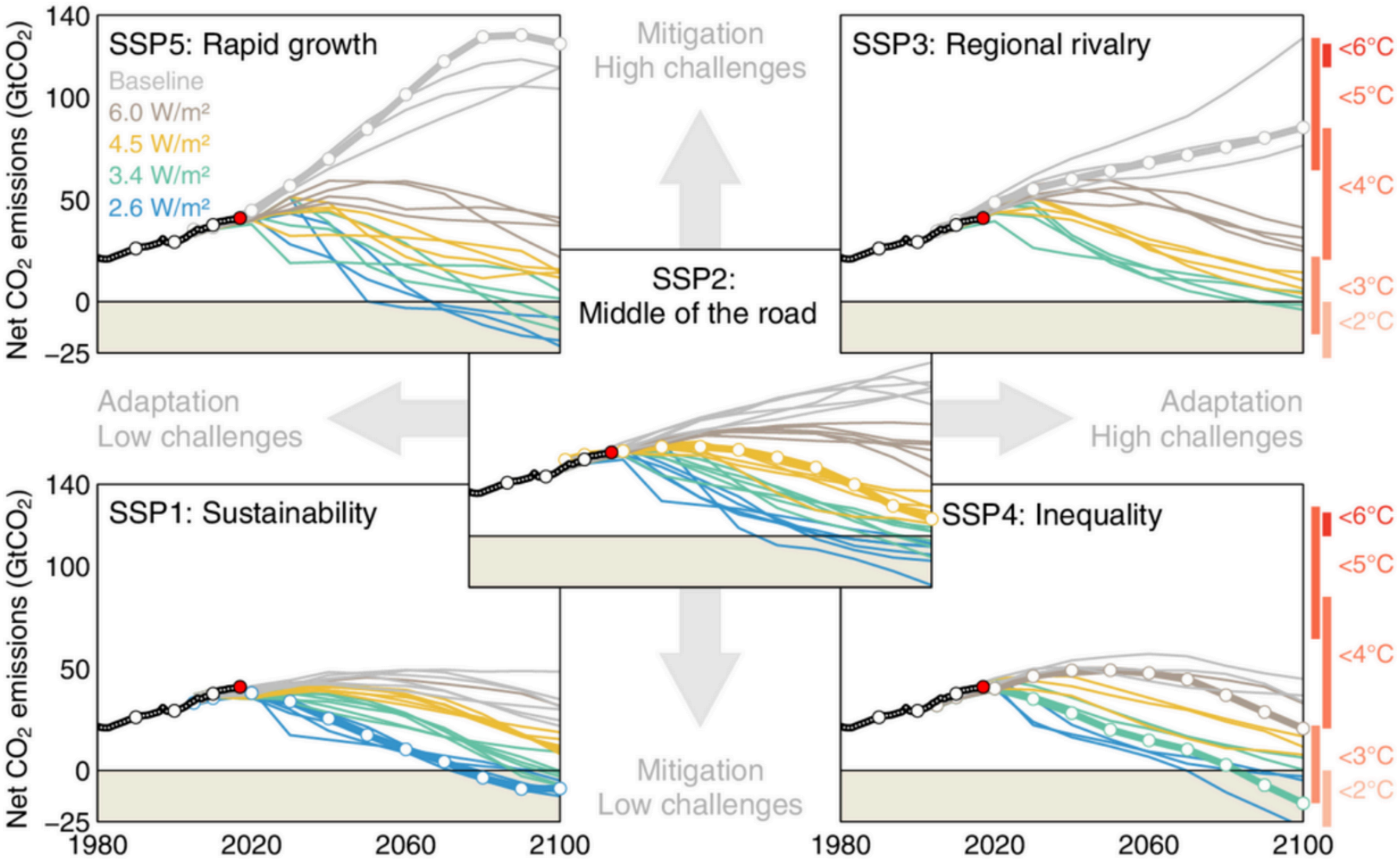
Trajectories

- Reference SSP x-y.z (Shared Socio-economic Pathways)
- 5 classes of scenarios
  y.z : radiative forcing at the end of century (in W/m2)



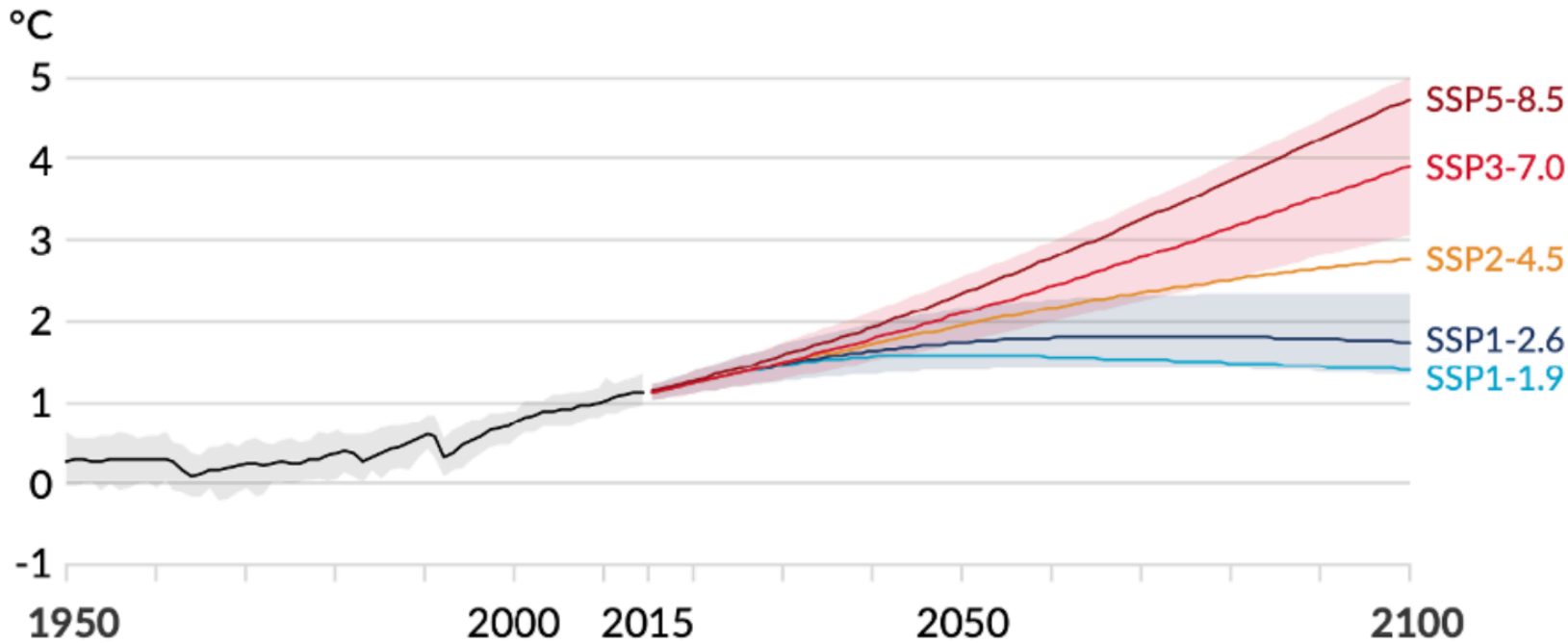
Source: IPCC reports synthetized by Carbone4

A quick look at the 5 classes of scenarios



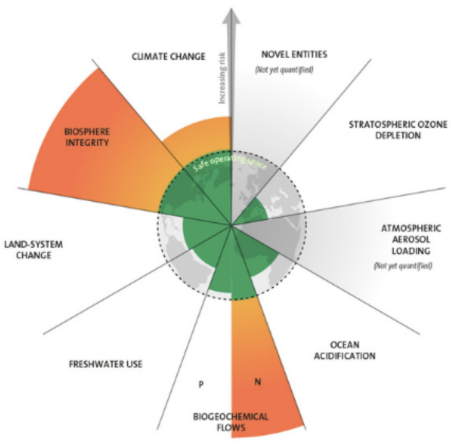
Projection

	Court terme : 2021-2040	Moyen terme : 2041-2060	Long terme : 2081-2100
SSP1-1.9	1,5	1,6	1,4
SSP1-2.6	1,5	1,7	1,8
SSP2-4.5	1,5	2,0	2,7
SSP3-7.0	1,5	2,1	3,6
SSP5-8.5	1,6	2,4	4,4



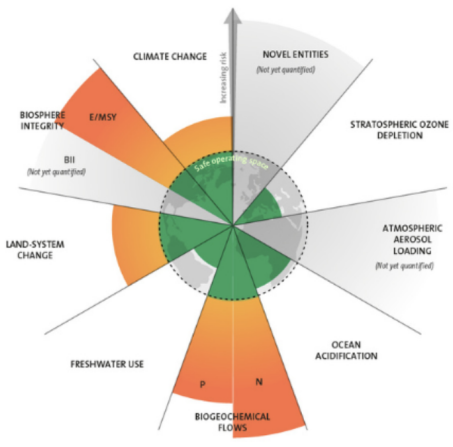
Not only CO_2
Boundaries limits

2009



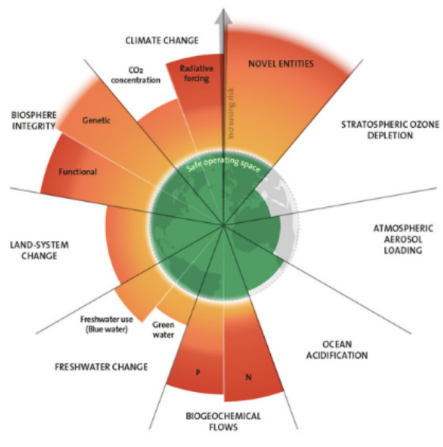
7 boundaries assessed,
3 crossed

2015



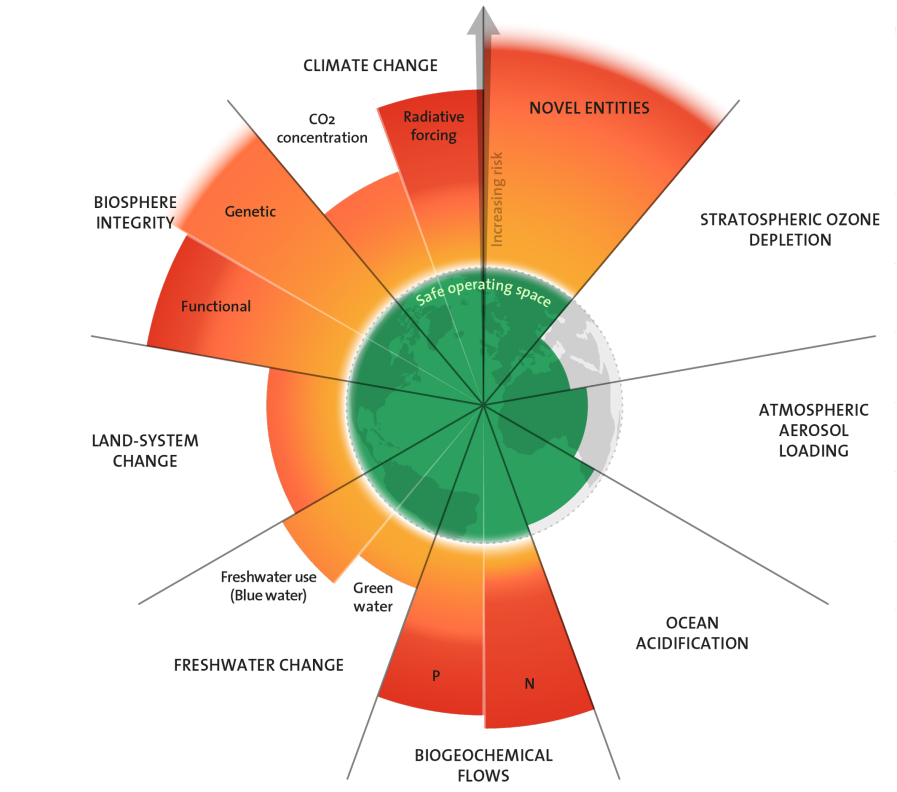
7 boundaries assessed,
4 crossed

2023



9 boundaries assessed,
6 crossed

Focus today



A first assessment

- We are on a trajectory that calls into question the very future of humanity.
- It is too late on most indicators (we can not turn back the clock).
  - What can we do about it?
  - Can we still react?
  - Limit the damage / get out of it?

AI part of the solution?

- Can AI help to mitigate the crisis?
- A least partially...

AI part of the solution?

- Can AI help to mitigate the crisis?
- A least partially...
- However, how much it contributes to the emissions?

It is also part of the problem

Let us open the second part of this talk:
Impact of digital (and AI/GenAI)

First of all, what are we talking about?

What is digital technology (ICT)?

eco-system that groups together all the devices used to manipulate information in electronic form.

Two ways of looking at digital

- The digital objects themselves, structured in three tiers: Data centers / users-terminals / networks.
- The digitization of society
  Agriculture, buildings, transport, industry, health, etc.
  These are the effects of all the changes brought about by digital technology.

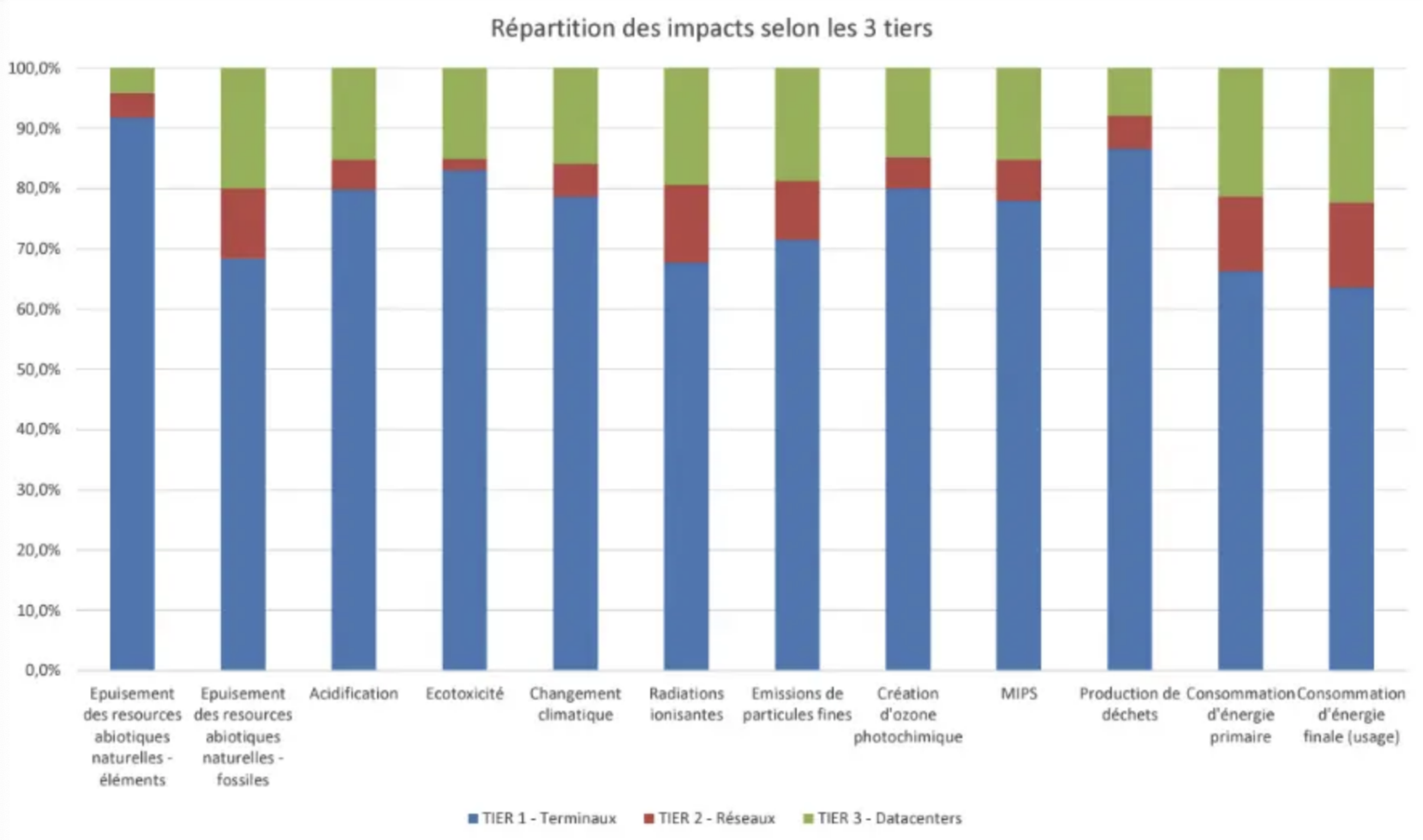
- **It is an ever-increasing snowball.**
- ML and GenAI are accelerators of digital technologies.

More precisely

The classical vision of the field is the **3 tier**

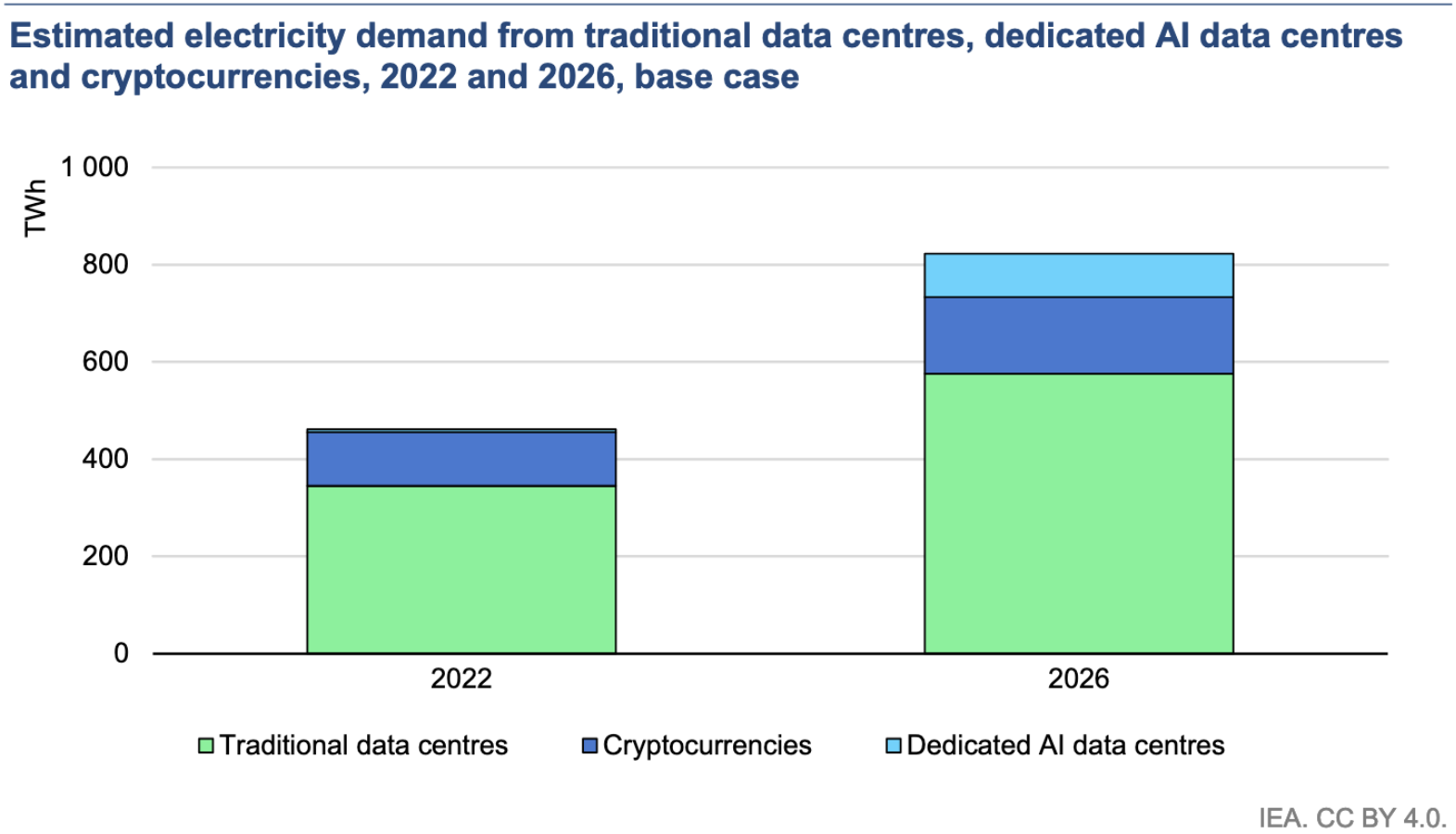
- **Data centers**
  more than 1% of the electricity (in the world)
- **Terminals**
- **Networks**

Repartition (France)



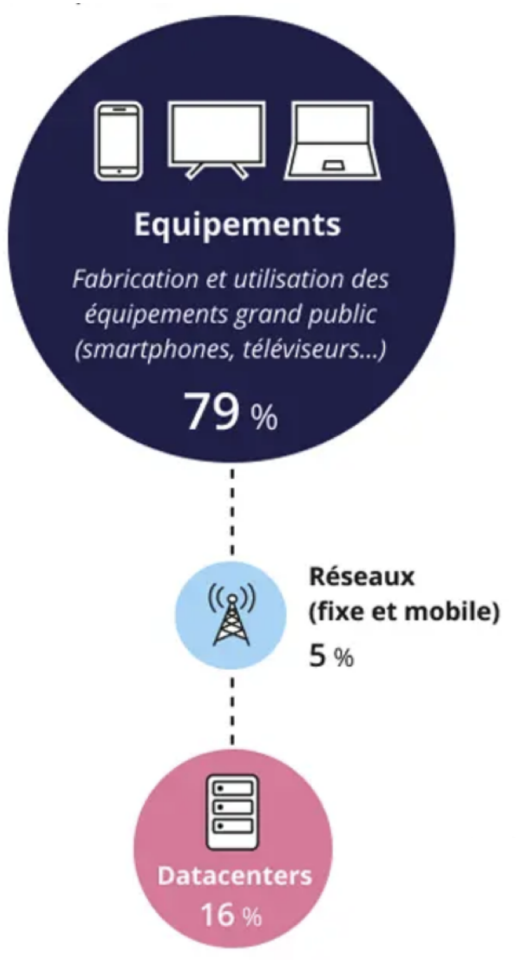
Focus on data Centers

Most AI training is done in large scale data centers.



Following the current evolution, the part of electricity in irish data centers (2030) is expected to reach 30%...

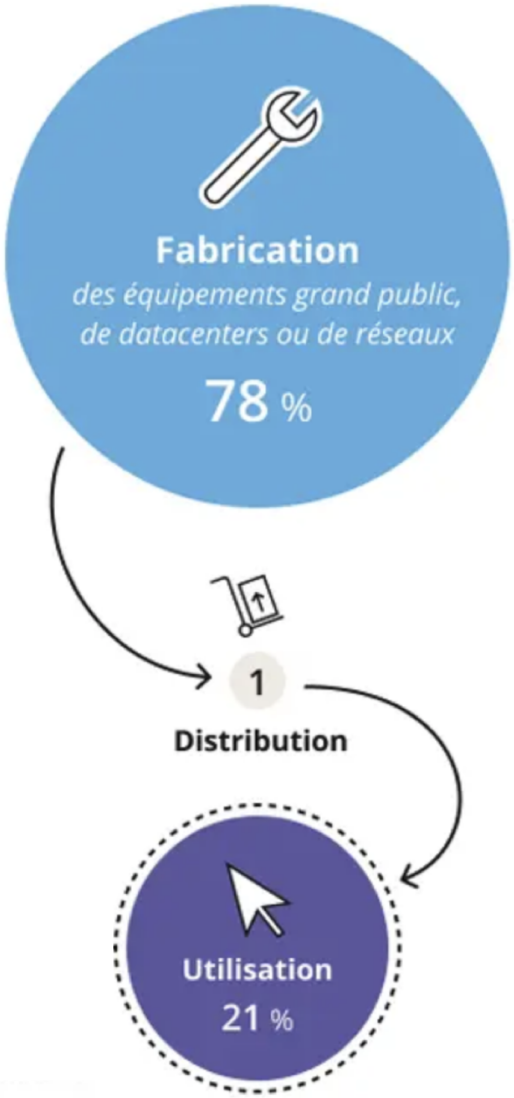
In France



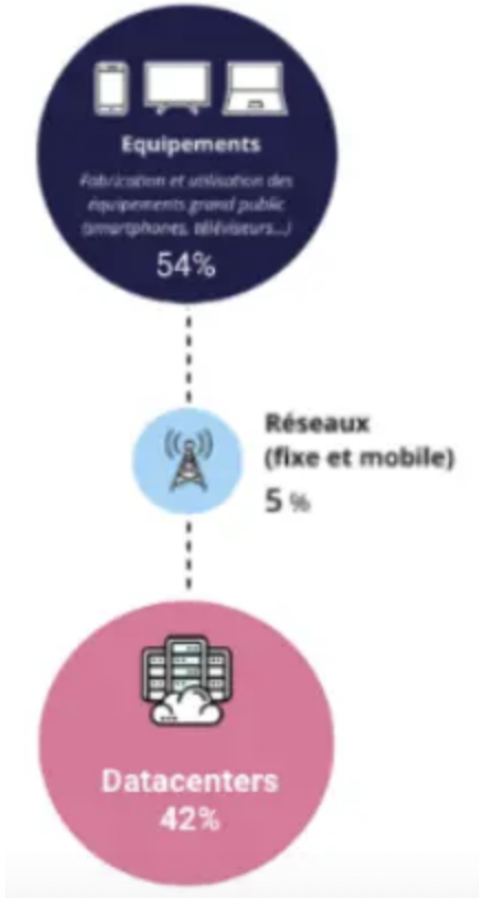
Take care: our energy mix is decarbonized... The situation is more balanced elsewhere.

source: Ademe-ARCEP

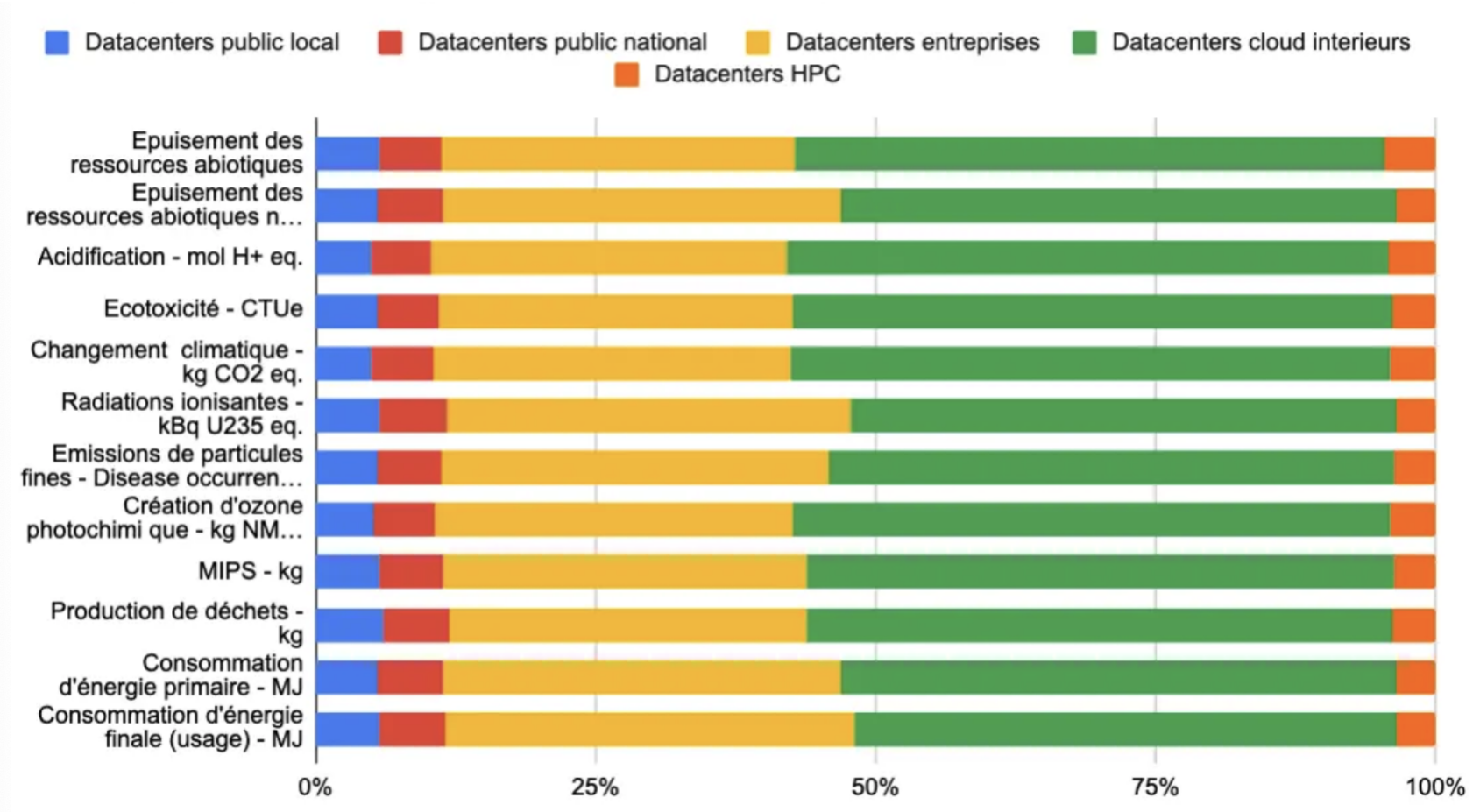
Same but from the ACV perspective



Revisited by taking into account what comes from abroad



https://hubblo.org/fr/blog/datacenters-imported-impacts/


Datacenters public local
Datacenters public national
Datacenters entreprises
Datacenters cloud interieurs
Datacenters HPC
Epuisement des ressources abiotiques
Epuisement des ressources abiotiques n…
Acidification - mol H+ eq.
Ecotoxicité - CTUe
Changement climatique - kg CO2 eq.
Radiations ionisantes - kBq U235 eq.
Emissions de particules fines - Disease occurren…
Création d'ozone photochimi que - kg NM…
MIPS - kg
Production de déchets - kg
Consommation d'énergie primaire - MJ
Consommation d'énergie finale (usage) - MJ
0%
25%
50%
75%
100%

Message

- Digital is everywhere.
- The Carbon cost of the digital is most often hidden.
  - Digital is not virtual.
  - We are invaded by very real digital objects.
    Their numbers are increasing rapidly (exponential growth)

Impact of digital

In France, this is 10% of the whole consumed electricity.

[4] ShiftProject: Lean ICT 2019

[5] The real climate and transformative impact of ICT: A critique of estimates, trends, and regulations, 2021

Impact of digital

In France, this is 10% of the whole consumed electricity.

Contribution to the Carbon emissions

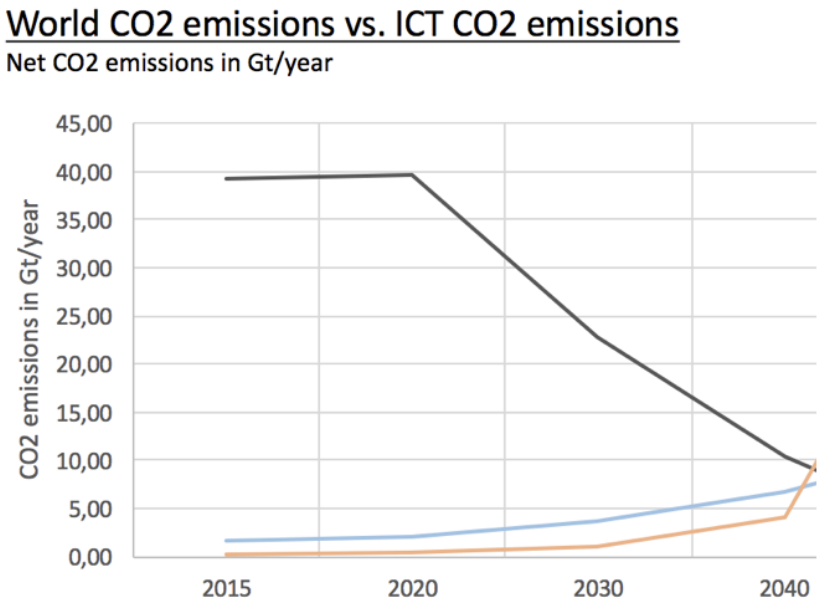
- Digital technology accounts for 5 to 6% of the world's primary energy consumption, roughly 4% of the total emissions[4]. Freitag et al. estimate the domain from 2.1 to 3.9 % of carbon emissions[5].
- Annual growth 6-9% (over 2015-2019).
- It is very difficult to quantified the part of AI (accelerator effect).

[4] ShiftProject: Lean ICT 2019

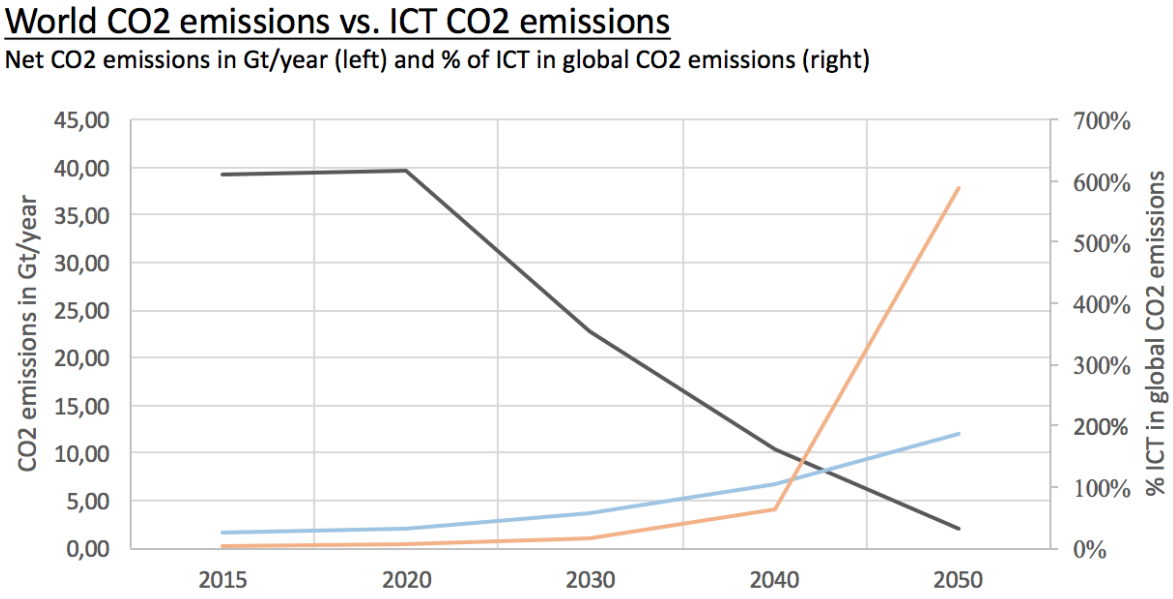
[5] The real climate and transformative impact of ICT: A critique of estimates, trends, and regulations, 2021

Simulator done with the help of Yannick Malot and Guillaume Raffin (at LIG-Inria) for comparing the SSP scenarios.

- Most favorable SSP 1-1.9 with ICT basis minimal growth (6%)



More precisely...



- This is, of course, just a thought as a basis for discussion in the "Digital can get us out of the crisis" debate...
- Are you ready in 10 years to leave in a world where half of the electricity is dedicated to ICT?

Mitigation: A complex reality

The previous example was a mental exercise...

- Practically, the degrowth lies partially in technological advances.
- Electricty is always more decarbonized.

  Mondial consumption in 2022 : 68 200 TeraWh.
  In 2022, global electricity consumption increased by 2.5% compared to 2021, close to the average growth rate (+2.6% per year between 2010 and 2021), but the carbon intensity of global electricity production dropped to 436g CO_2 per kWh[6] because of the development of renewable energies in the mix.

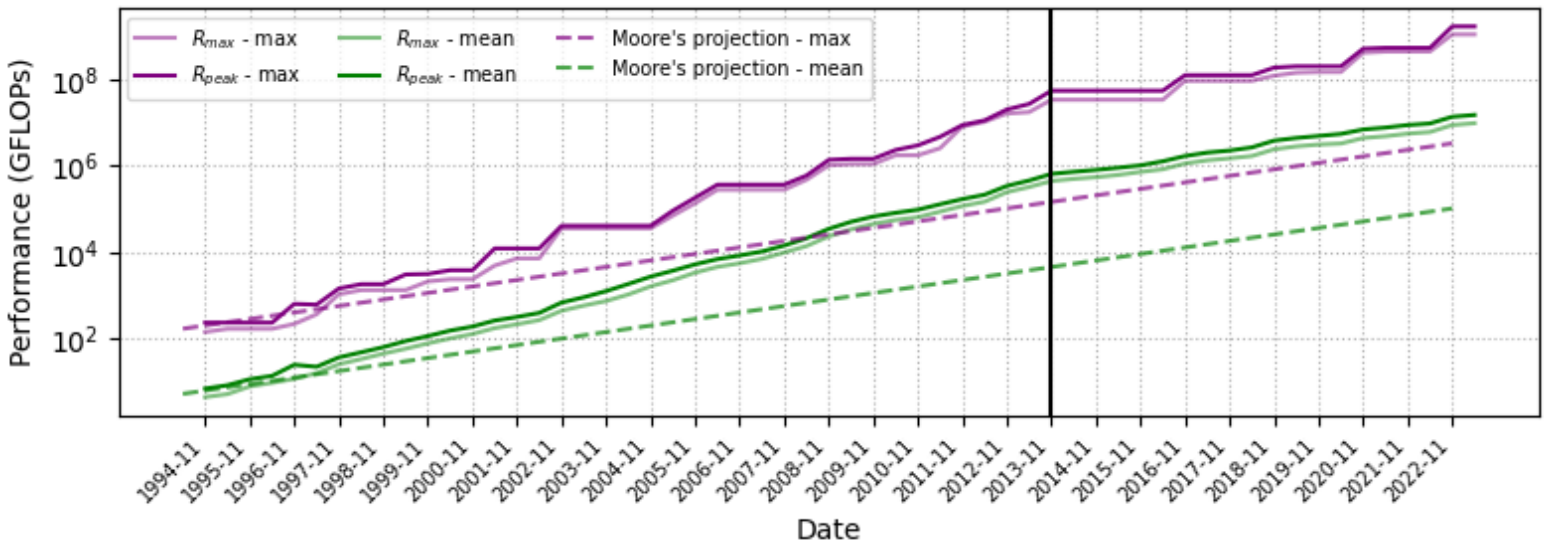
[6] source: IEA 2023

Two useful empiric laws

Macroscopic indicators

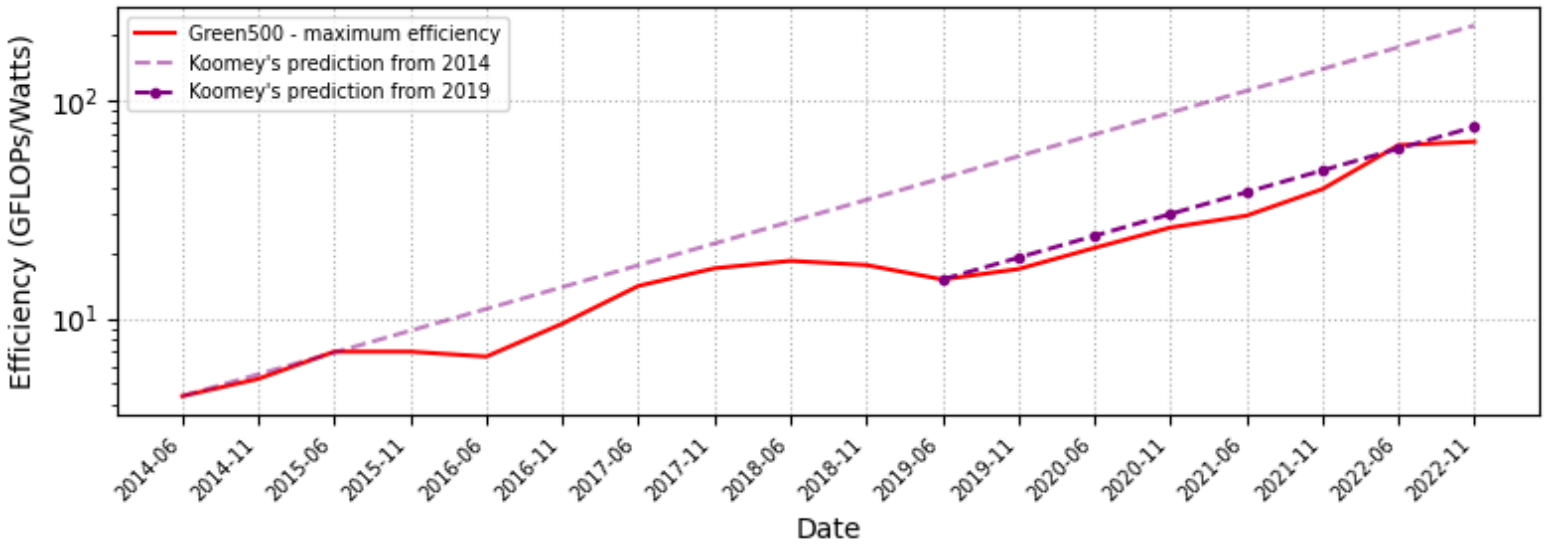
- **Moore**: Performance
  The number of transistors in integrated circuits intégrés doubles every 2 years, and by extension, so it is for global performance of computing systems.
  Extension to parallel clusters.
- **Koomey**: Similar but it targets energy efficiency (number of Operations per Watt/Joule)
  Double every 18 months before 2010, today, double every 26-27 months.

Performance (TOP500)



- Clear break around 2013-14
- It corresponds to general use of GPUs in parallel
- This is the reason of development of deep learning

Energy efficiency

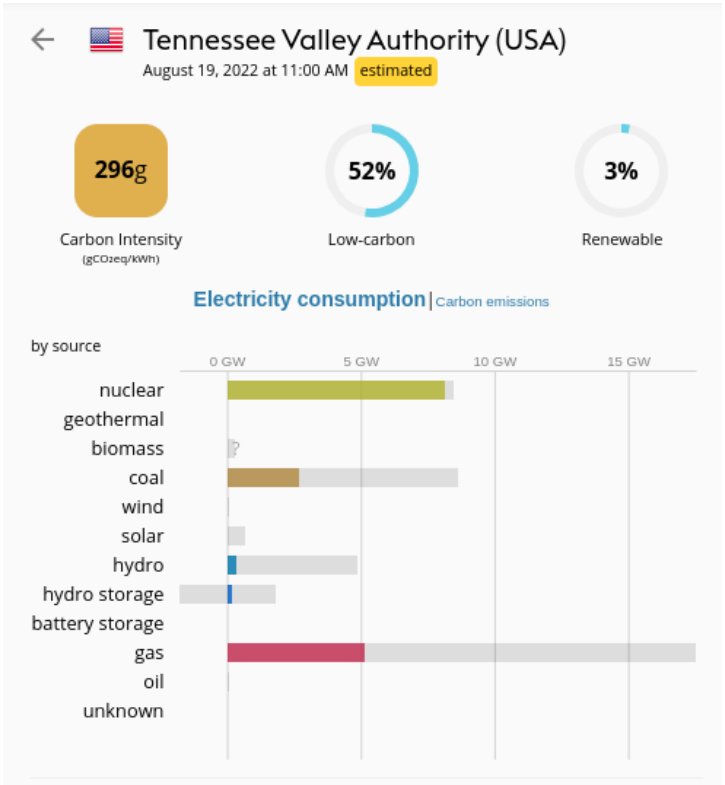


- The energy efficiency gains are less than the performance growth.

From KWh to CO2eq

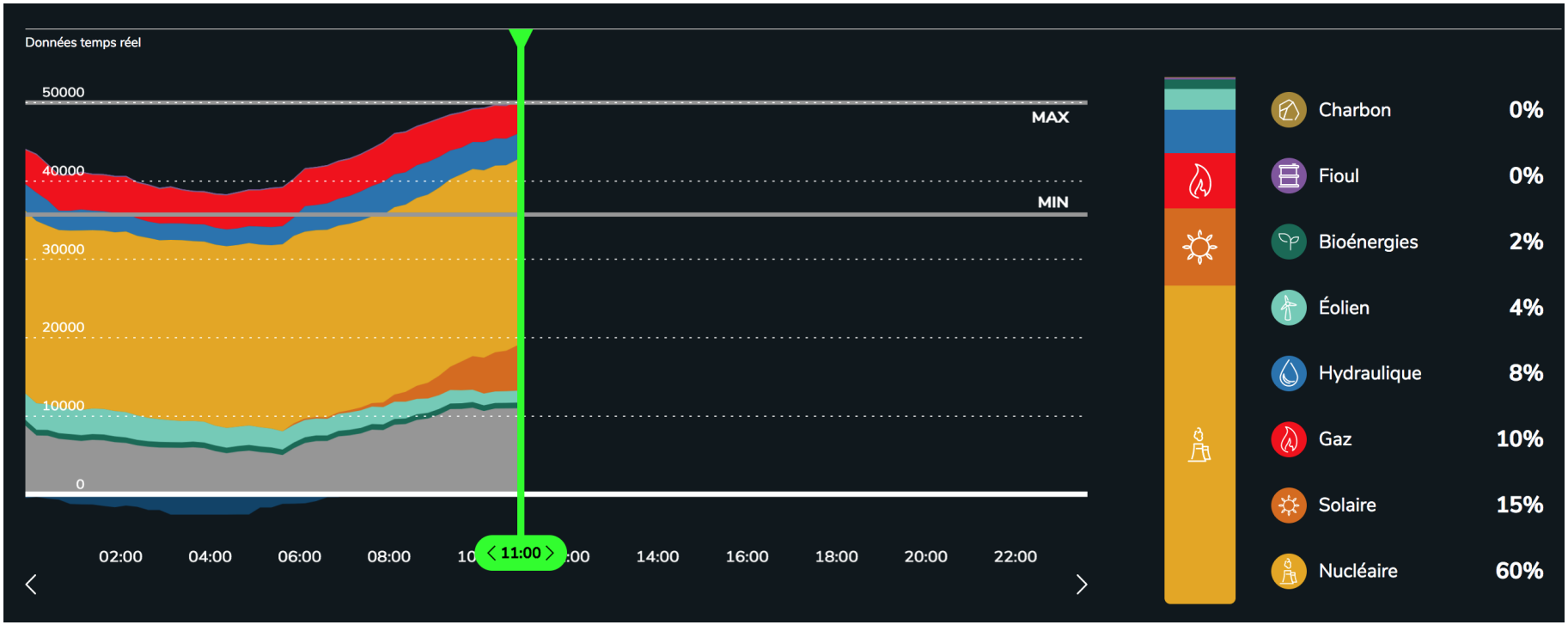
- It was the first exascale platform.
- its Power is more than 30MWatts.

Example of *Frontier* supercomputer

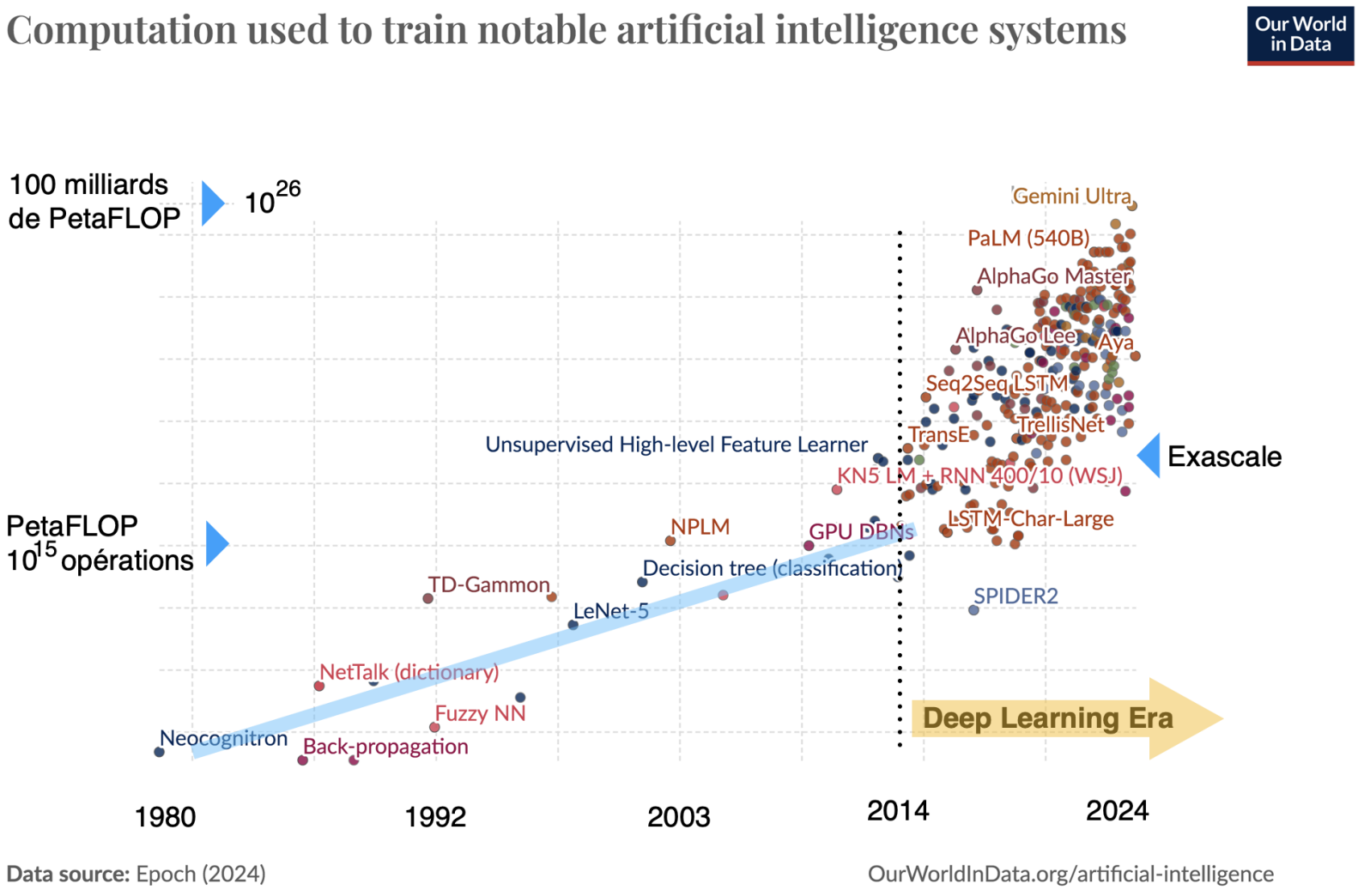


In France (RTE)

Energy mix



What about AI?



It is very hard to figure out what does it represent!

- 10000000000000000000000000 operations for training.
- Several months of full utilization while converted on the most powerful platforms.
- Usage (inference) is an order of magnitude larger for the most popular GenAI (chat-GPT / smart search engines)

What is the part of AI in ICT?

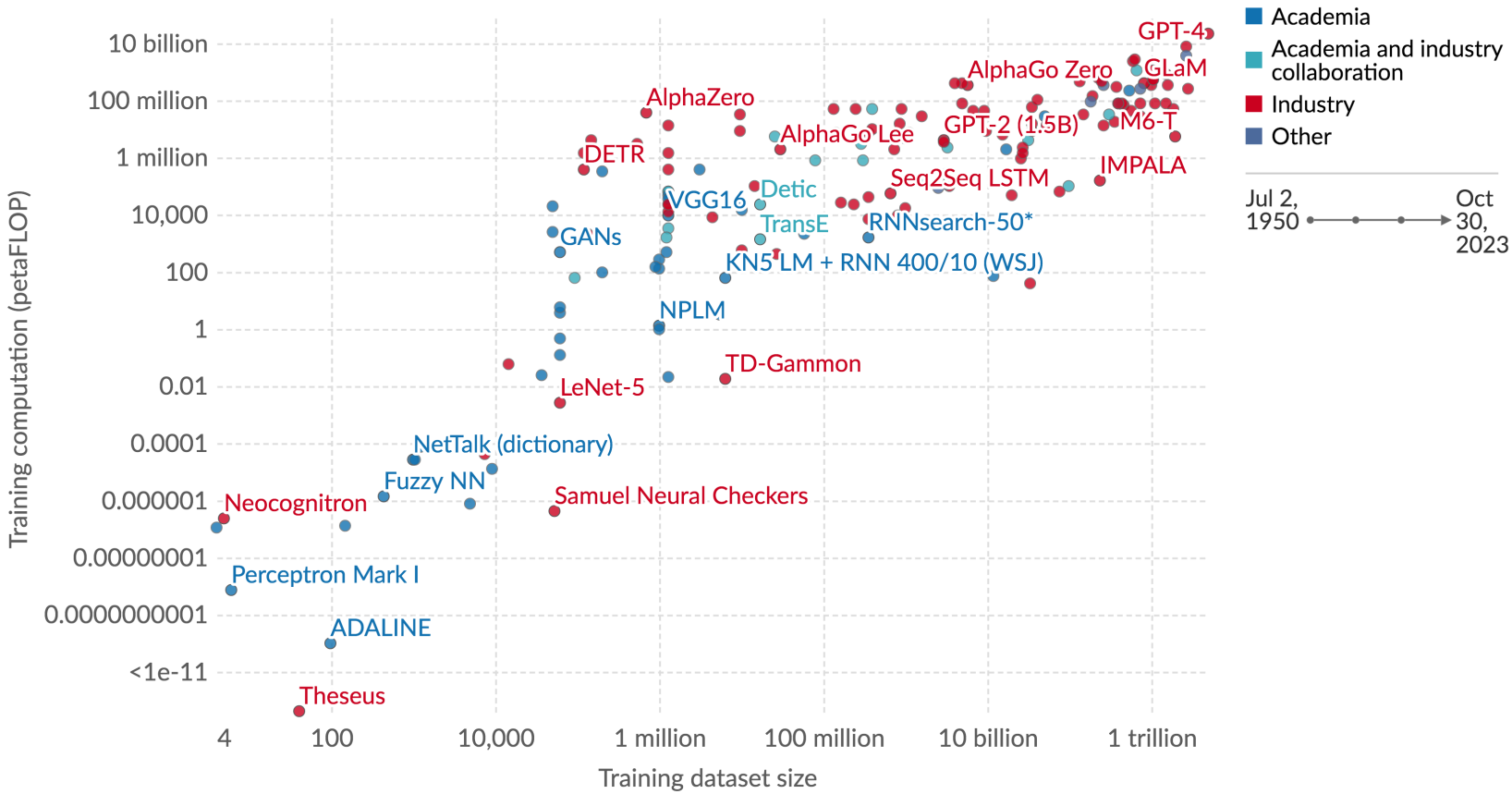
- Very hard to estimate since GenAI is spread in all domains.
- Global view: all Deep Learning and GenAI models run with accelerators and Nvidia produces 90% of this market.
- Another possibility is to develop methodology for estimating/measuring the Carbon cost of ML services.

And data?

Training computation vs. dataset size in notable AI systems, by researcher affiliation

Our World in Data

Computation is measured in total petaFLOP, which is 10^{15} floating-point operations[1] estimated from AI literature, albeit with some uncertainty. Training dataset size refers to the volume of text that is employed to train a model effectively.



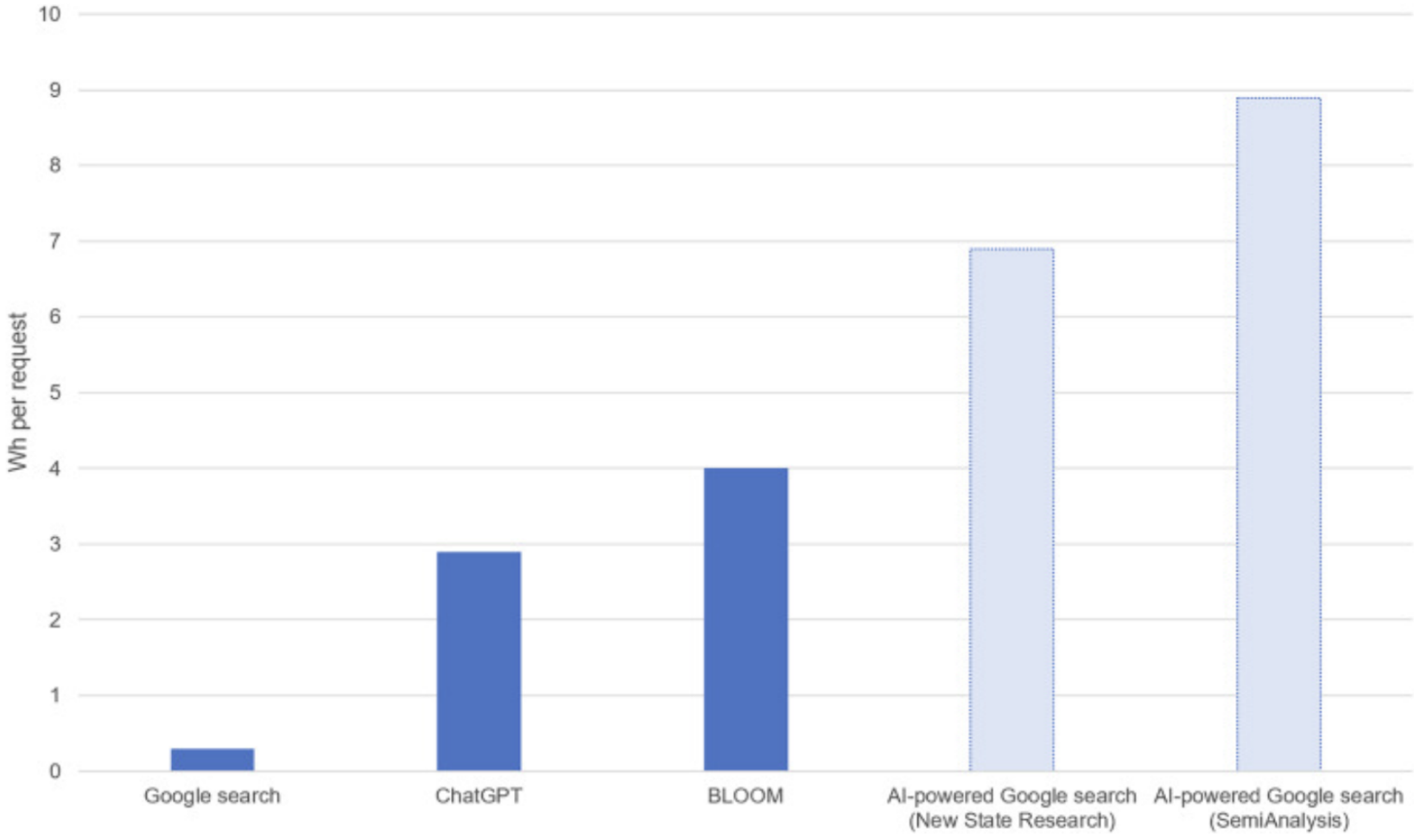
Data source: Epoch (2024)

OurWorldInData.org/artificial-intelligence | CC BY

1. Floating-point operation: A floating-point operation (FLOP) is a type of computer operation. One FLOP represents a single arithmetic operation involving floating-point numbers, such as addition, subtraction, multiplication, or division.

An example of impact

Smart Internet search.



source: A. de Vries, 2023

Comparison

- 9 billions of requests daily.
- no decrease since 2009.
- If integrated into Google tools, the demand will need hundreds of thousands GPUs.

source: https://googleblog.blogspot.com/2009/

Measuring

Measure is a pillar of Science.
We wish to evaluate the cost of what we use.

Why?

- for quantifying the order of magnitude of a ML service and the materiality of equipments it uses.
- for comparing several services as a basis for policy (enlightening decision-makers)?
- for breaking the illusion of dematerialization
- for questioning the ratio potential benefits/costs

What is a "good" measure?

- The purpose of measurement should be well-defined
- Consensual and practical method:
  acceptable (in a given time)
- Accuracy
- Clear hypotheses

Estimate the carbon cost of applications

- There exist methodologies for all the phases of the life cycle of equipments.

Life Cycle Analysis

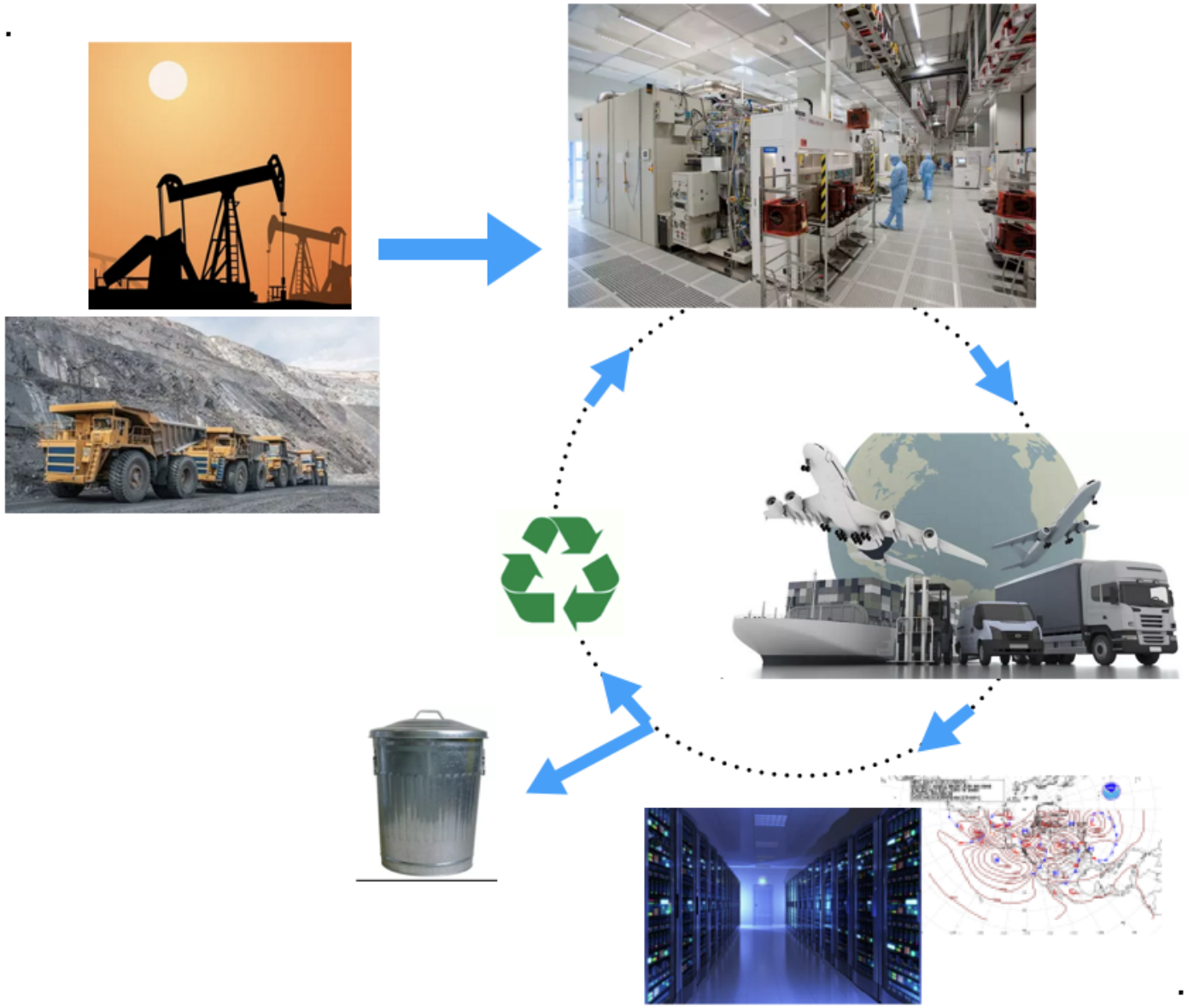
- It targets mainly the *direct effects*.
- we must also include *indirect effects* et rebound effects.
  That are all what is not accounted into the initial perimeter.

Rebound effect

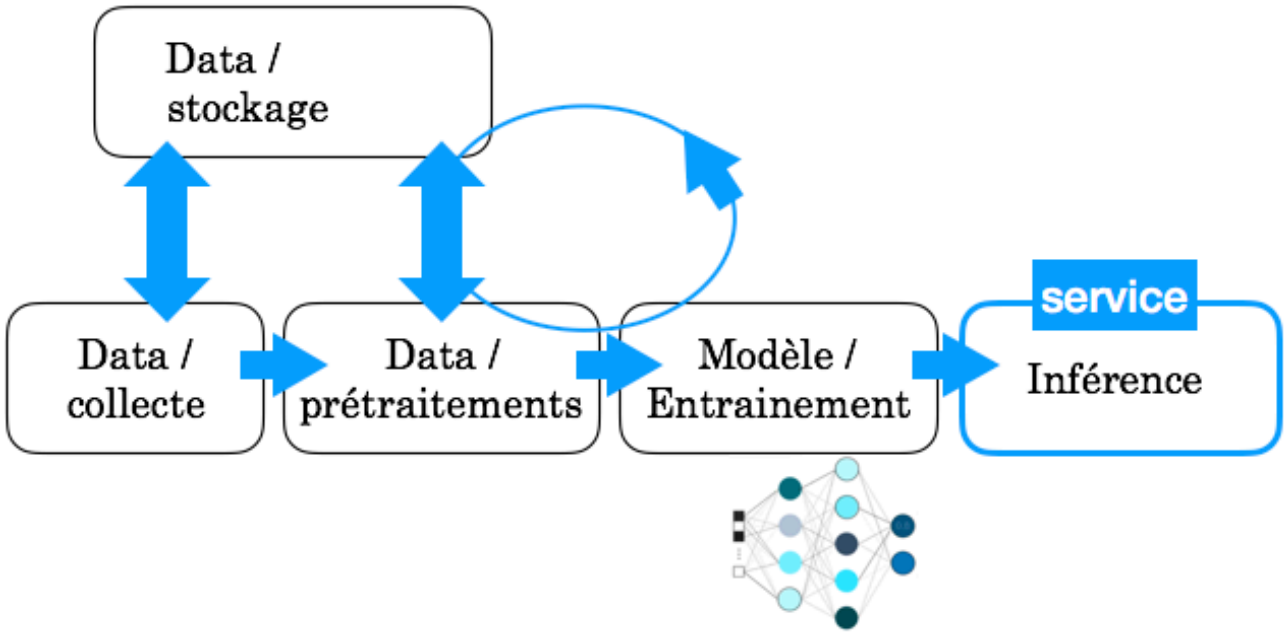
- Direct:
  Technological progress improves a device, making it more efficient and thus, increasing the number of users.
- There is also an indirect rebound effect when gains in one area generate consumption in another.

  In this way, a sobriety approach can also be a source of rebound effects, as the savings made are reinvested (whether in money or time), or as the consumption of other products is relieved of guilt.

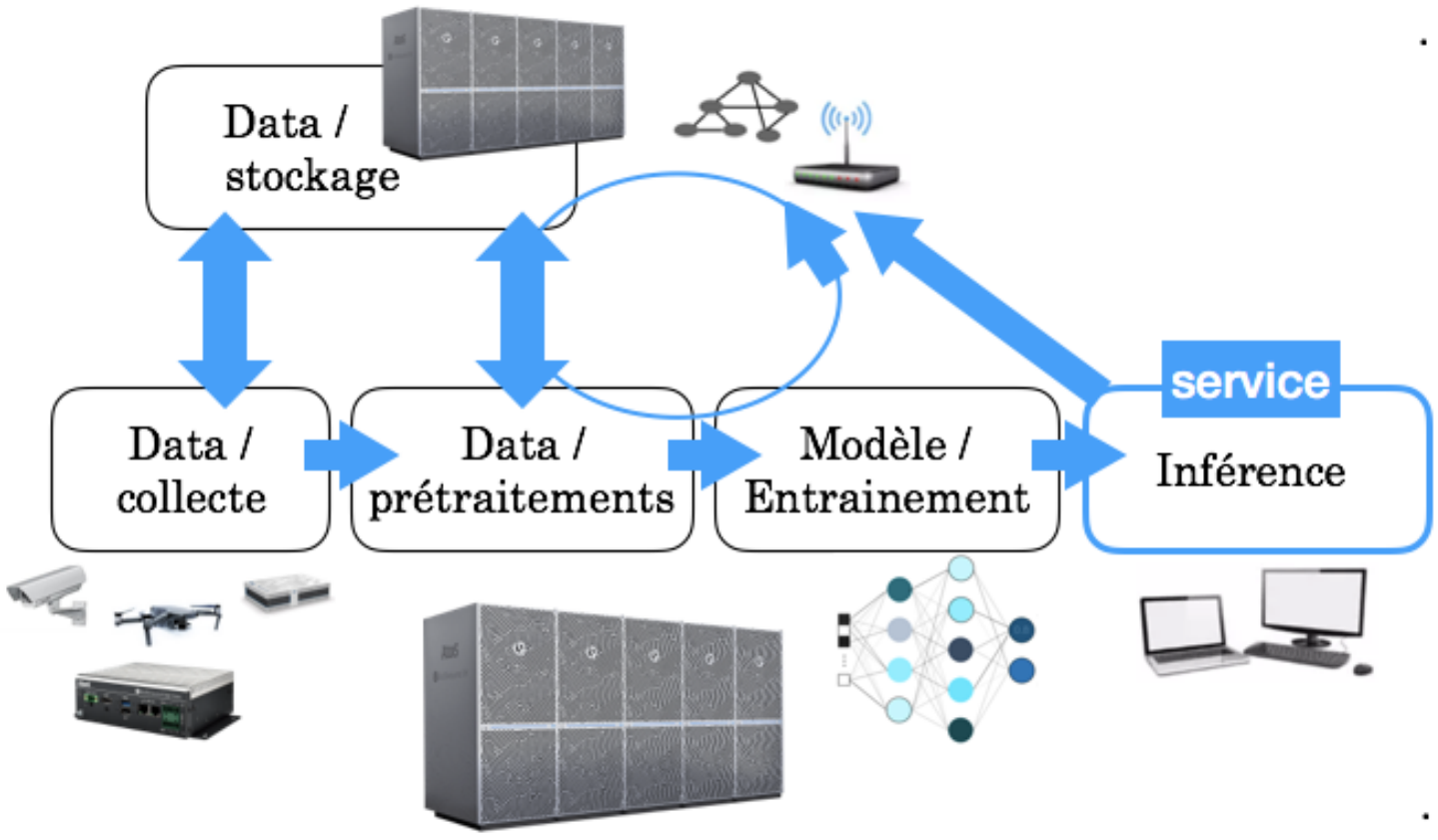
LCA for a digital device



LCA of a digital service (AI):
Example, control temperature system

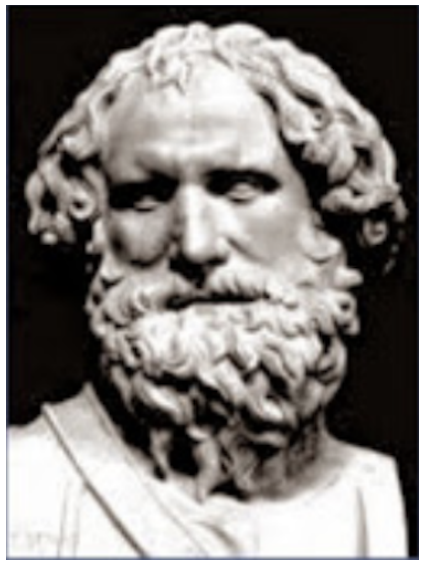


Counting everything: materiality matters



I invite you to digress for a moment

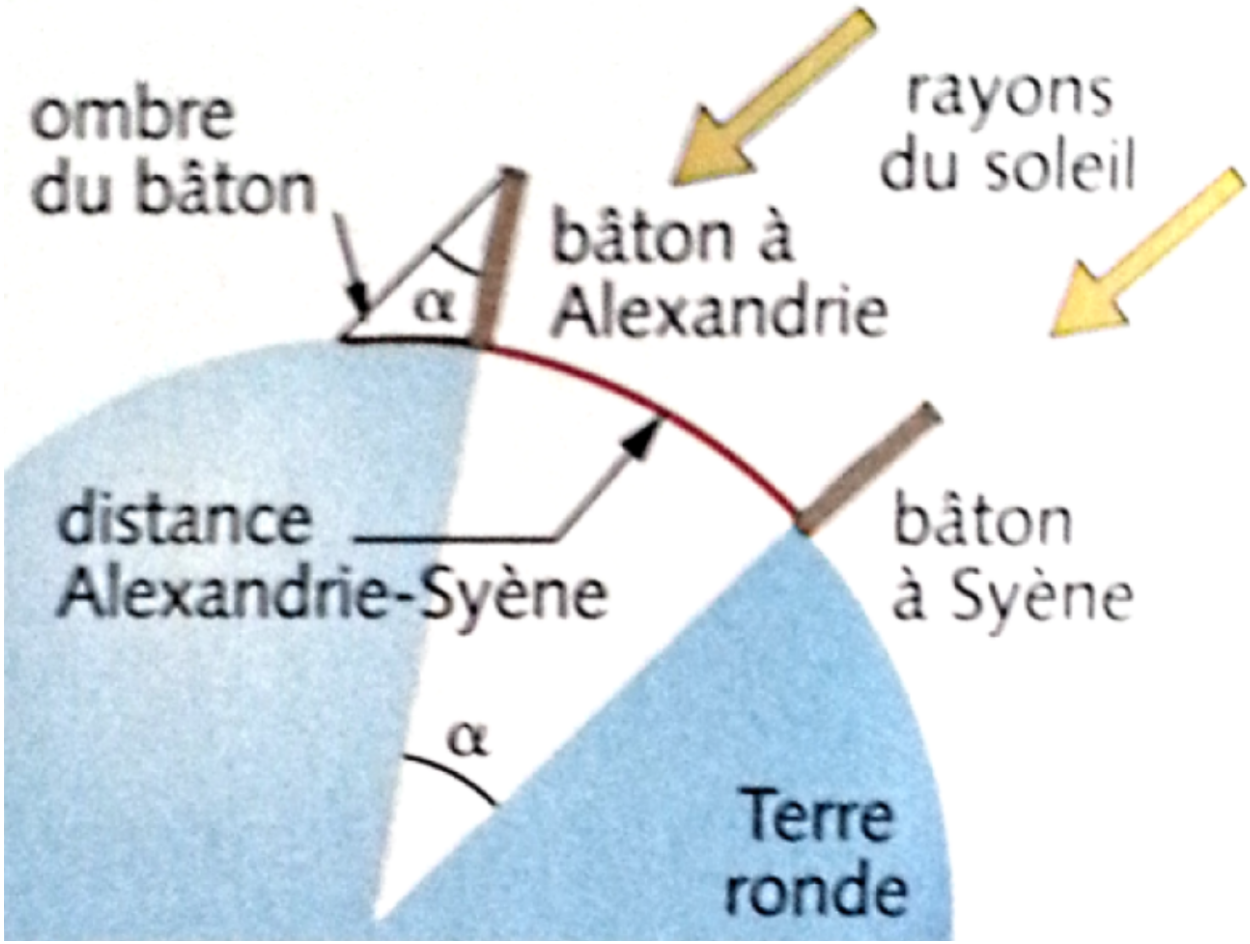
Let us go back a few millennia for measuring the circumference of the earth (as did Eratosthene)



Hypotheses

- perfectly round earth
- The two selected cities Alexandria and Syene aligned on the same meridian
- Approximation of small angles

Principle



source: Wikipedia

Do you think it was a "good measure"?

Yes, if we look at the result:

- Syène-Alexandria 5000 stadium
- Angle : $\frac{1}{50}$ of circle, that is roughly 7.2 degrees

circumference $= 50 \times 5000$ stadium
Thus, 39375 km
Impressive!

Do you think it was a "good measure"?

Yes, if we look at the result:

- Syène-Alexandria 5000 stadium
- Angle : $\frac{1}{50}$ of circle, that is roughly 7.2 degrees

circumference $= 50 \times 5000$ stadium
Thus, 39375 km
Impressive!

We know today it is 40 075,017 km at the equator [wikipedia], thus, less than 10^{-3} relative error!

Okay, but it was by chance!

A tool for measuring electricity in computing systems

Here is an example of designing a software according to the ideas of conviviality of I. Illitch:

- ALUMET delivers the consumption of memory+CPU for any device, no need to specified in advance.
- It was developed in RUST, (relatively) easy to use, light-weight and open-sourced.

Acceleration and induced upheaval

Technology is improving fast

But society (including researchers) canot follow:

- Evolution of the public to adopt new technologies.
- Lack of understanding of underlying mechanisms.
- Scientific constructions take time.
- Ethical issues: to be taken into account upstream avoiding the action-reaction mechanism.

Another message

- The environmental crisis is here to stay, and its consequences are unprecedented.
  Economic growth is the foundation of our (Western) societies and digital technology (AI) is one of the main engines of growth.
- It can play a positive role in the crisis.
- But it is also a costly area in a context where we need to cut back.
- How can we be sure that the balance is truly positive?
  If not, how to decide if we develop a new app or not?

"Efficiency"

- The computing (and AI) community has realized that it needs to react.
- There are fruitful proposals in many places.
- However, the main way forward is to optimize hardware/platforms and applications from the energy point of view.

  This is **eco-efficiency**:
  reducing the intensity of environmental impacts or resource use per unit of economic value produced.
  switching to renewable energy for decarbonized computing.
- However, can we believe in eco-efficiency?

- How can we guarantee that the balance sheet is truly positive?
- Putting the question of meaning back at the heart of our issues.
- Critical analysis imposes us to imagine new ways for the relationships of ML in regard to environmental issues.

More and more voices are being raised to react and rethink Machine Learning services by developing only applications based on the needs and usefulness of services **this is sufficiency**.

In conclusion

- Efficiency" or "Sufficiency"?

Thank you, I am open to discuss!