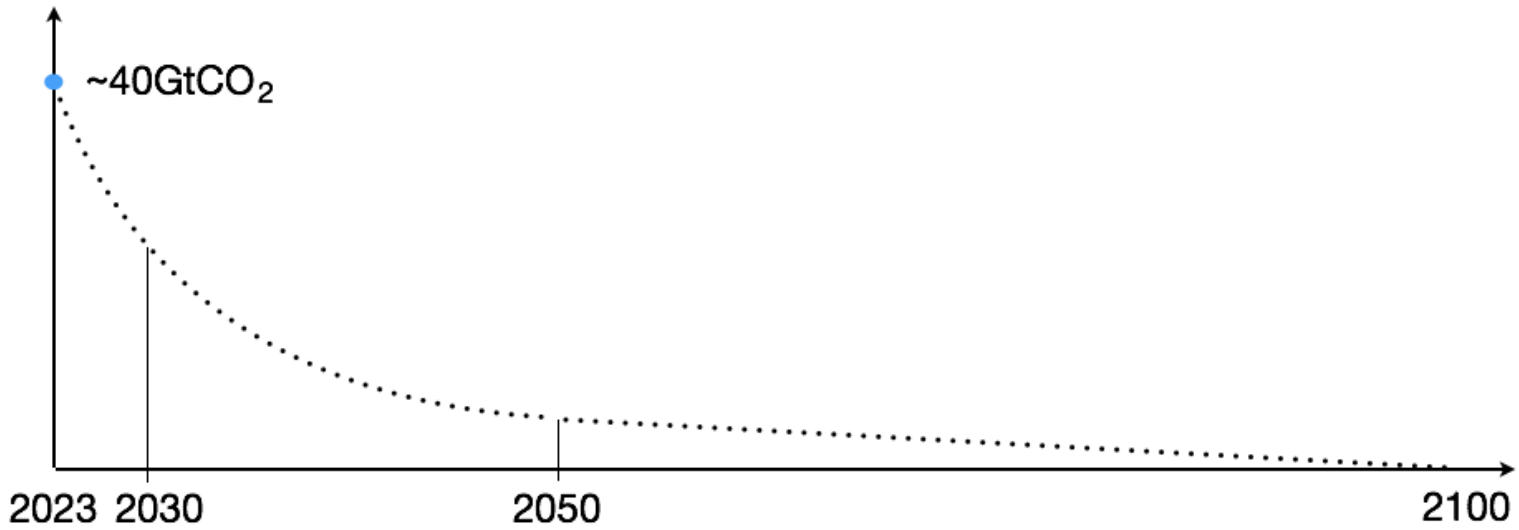
Focus on AI

Denis Trystram

MIAI, july 9, 2024

If we act now

- We emit around 40-45 Gt of CO_2 per year (worldwise).
- Maximum budget remaining to keep warming below 1.5 degrees : less than 1000 Gt of CO_2

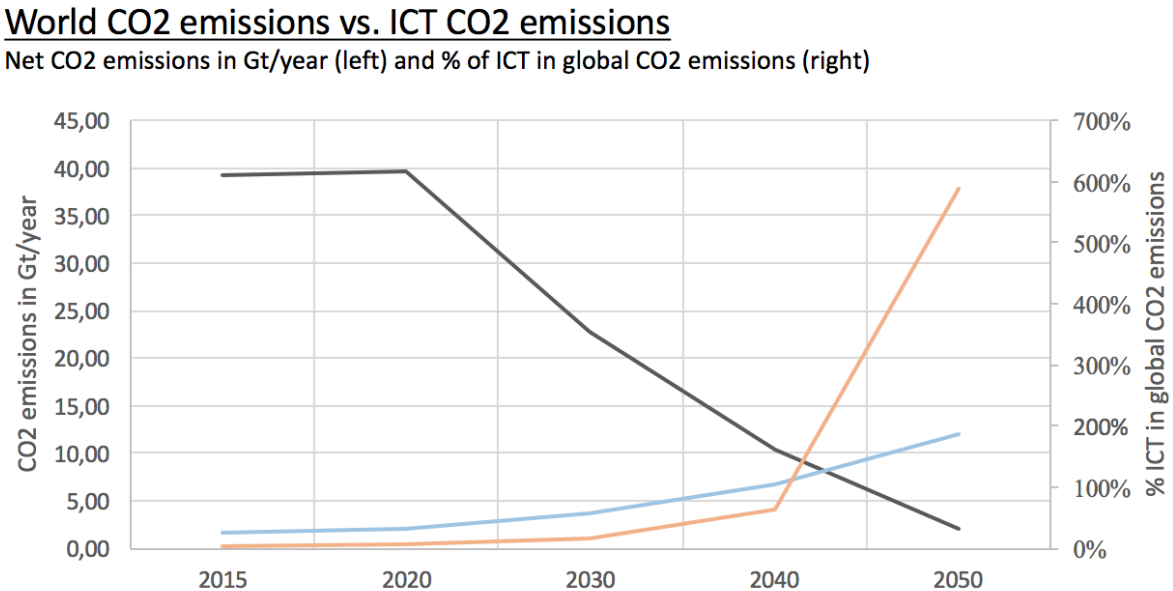


- The situations are very different in the world.
- We must reduce thee emissions targeting 2050, about 7 to 8% per year, even more if we delay...

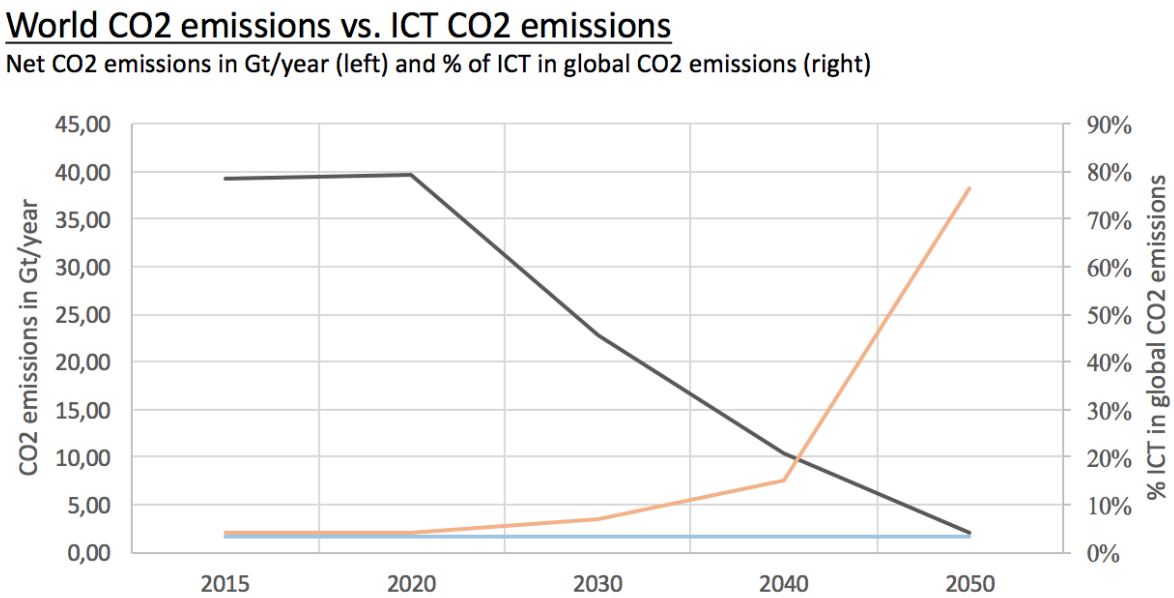
More precisely...

Simulator done with the help of Yannick Malot (PhD student at CEA-LIG) for comparing the SSP scenarios.

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



SSP1-1.9 et ICT constant



A complex reality

The previous example was a mental exercise...

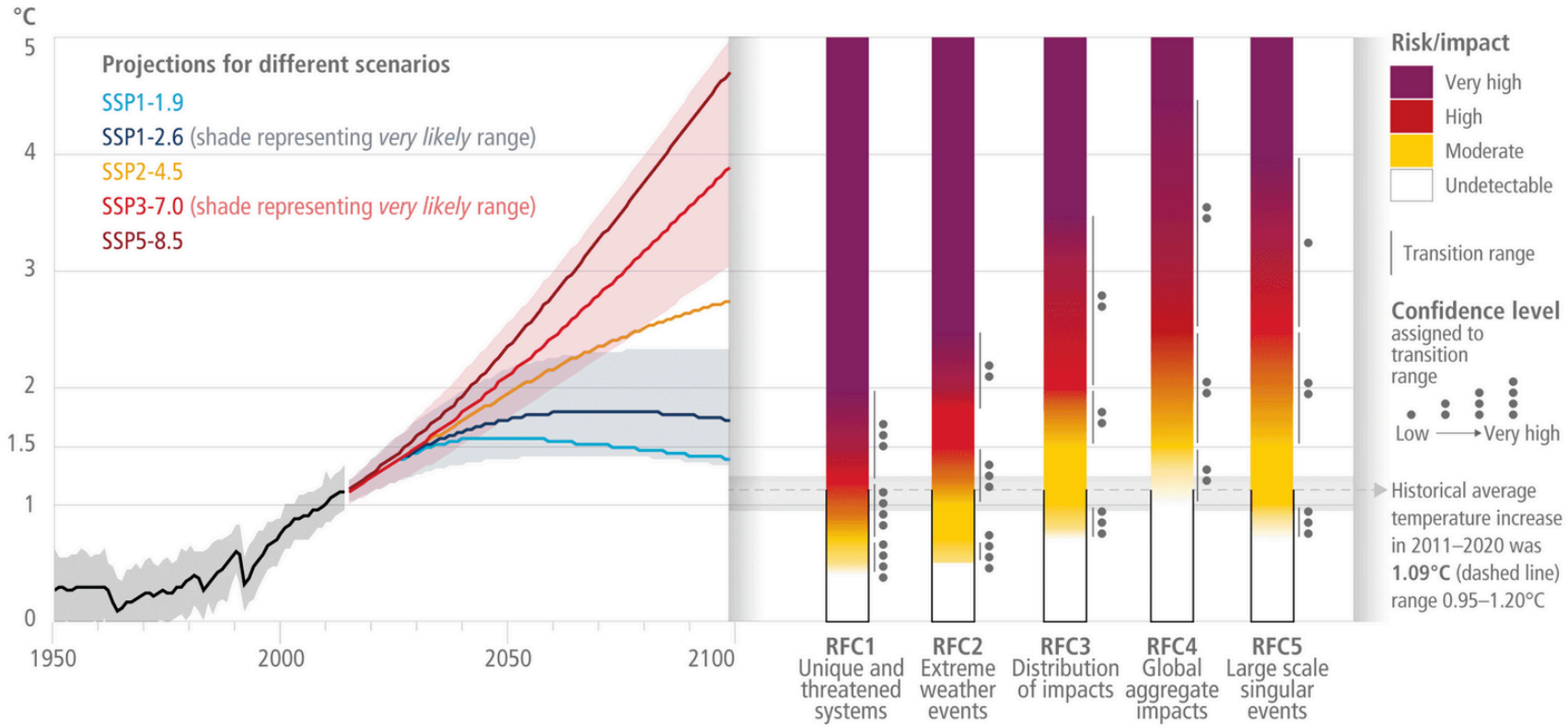
- Practically, the degrowth lies partially in technological advances.
- Electricty is always more decarbonized.

  Mondial consumption in 2022 : 68 200 TeraWh.
  In 2022, la consommation mondiale d'électricité a augmenté de 2,5% par rapport à 2021, hausse proche de la croissance moyenne (+ 2,6% par an entre 2010 et 2021), mais l'intensité carbone de la production mondiale d'électricité a chuté à 436g CO_2 par kWh[1].

[1] Selon l'enquête récente de l'AIE

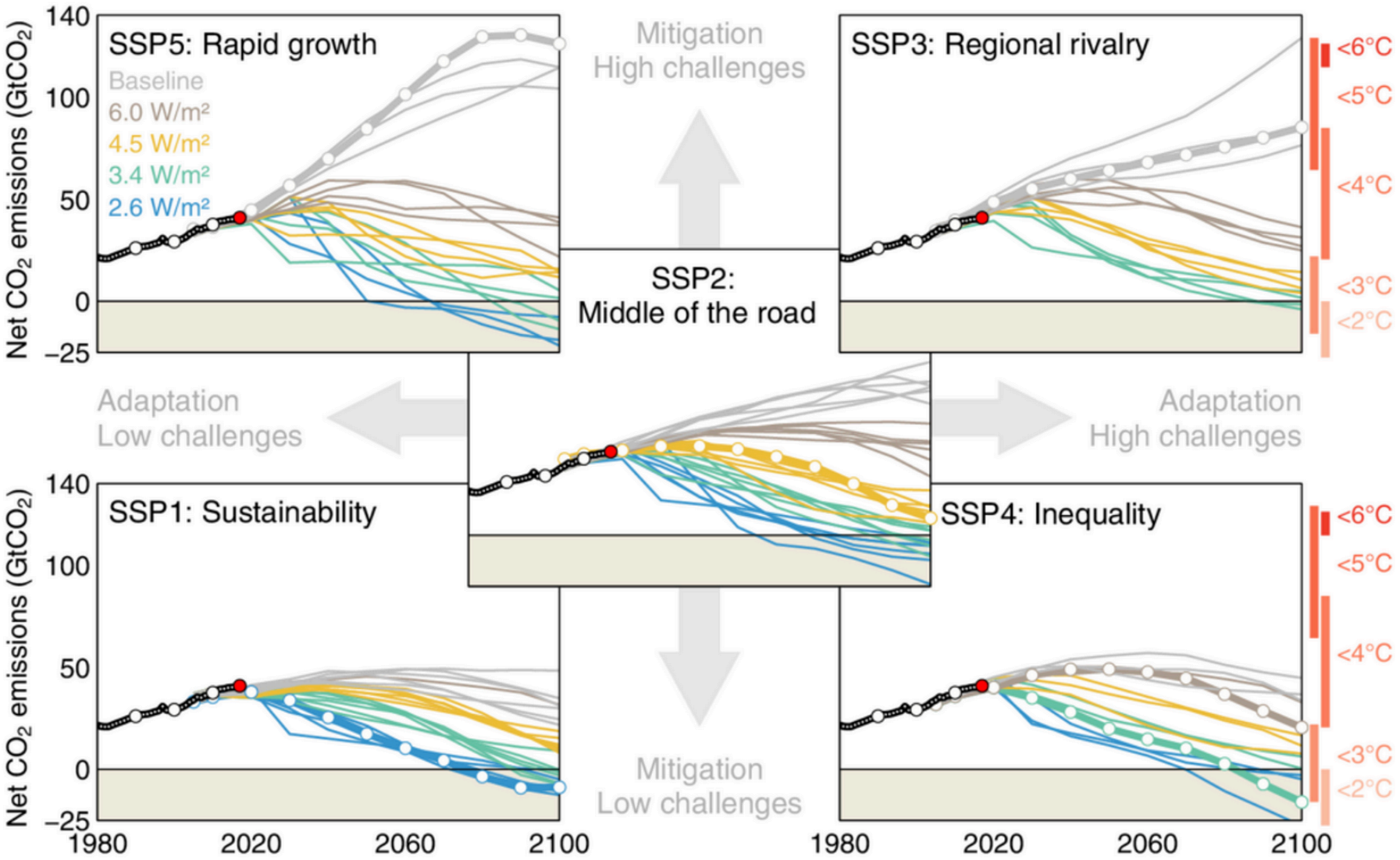
Trajectoire(s)

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



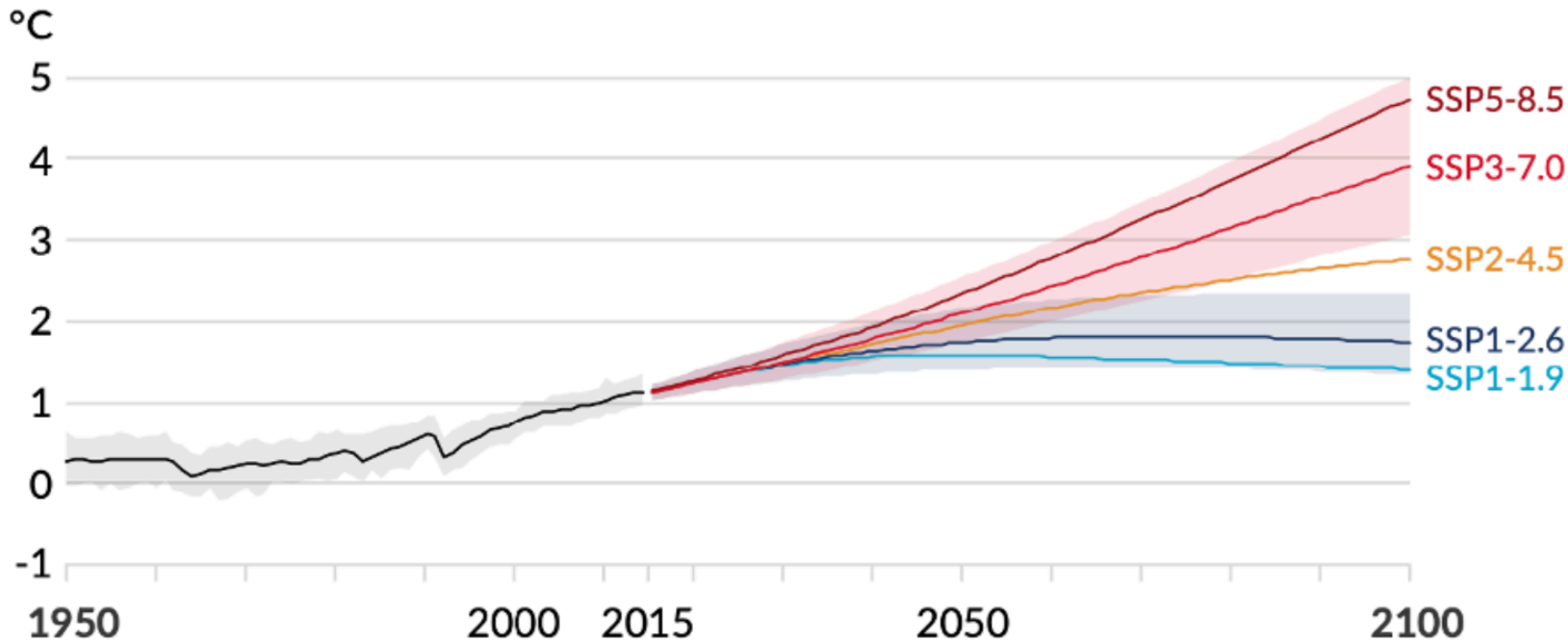
source : GIEC and Carbone4

Details of 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



The digital world

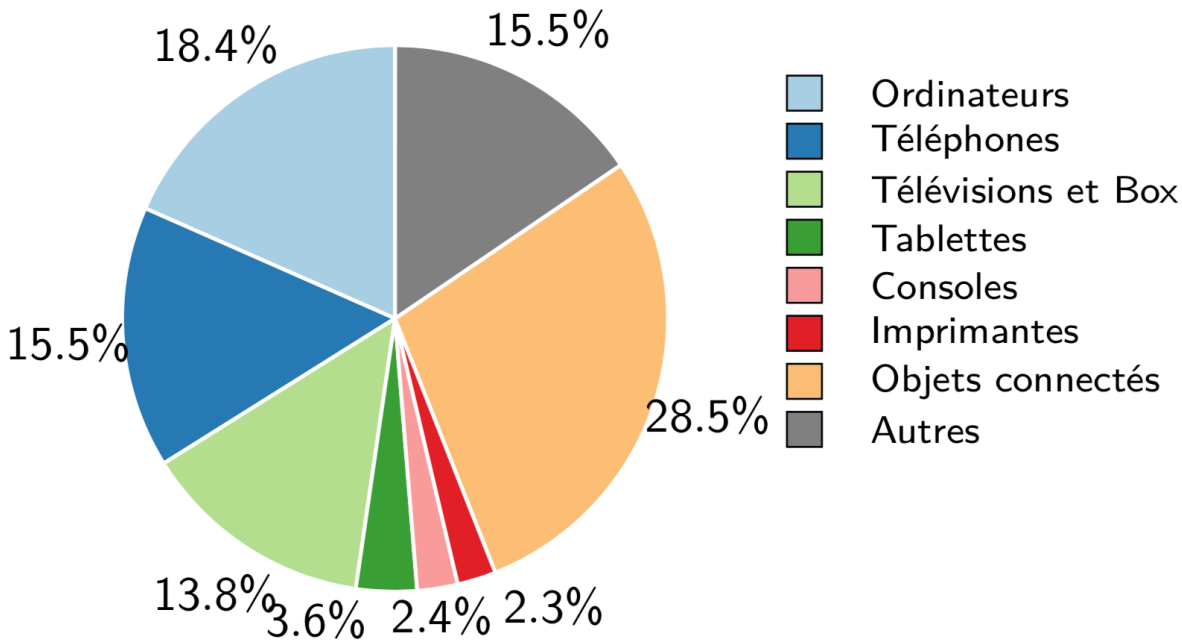
34 billions equipements for 4.1 billions people[2].
Vision classique **3 tiers**

- **Data centers**
  67 millions servers
  1% de l'électricité mondiale
- **Terminals**
  3.5 billions de smartphones
  more than 3 billions screens
  entre 10 et 30 billions d'objets connectés
- **Networks**
  1 billion boxes
  10 millions d'antennas

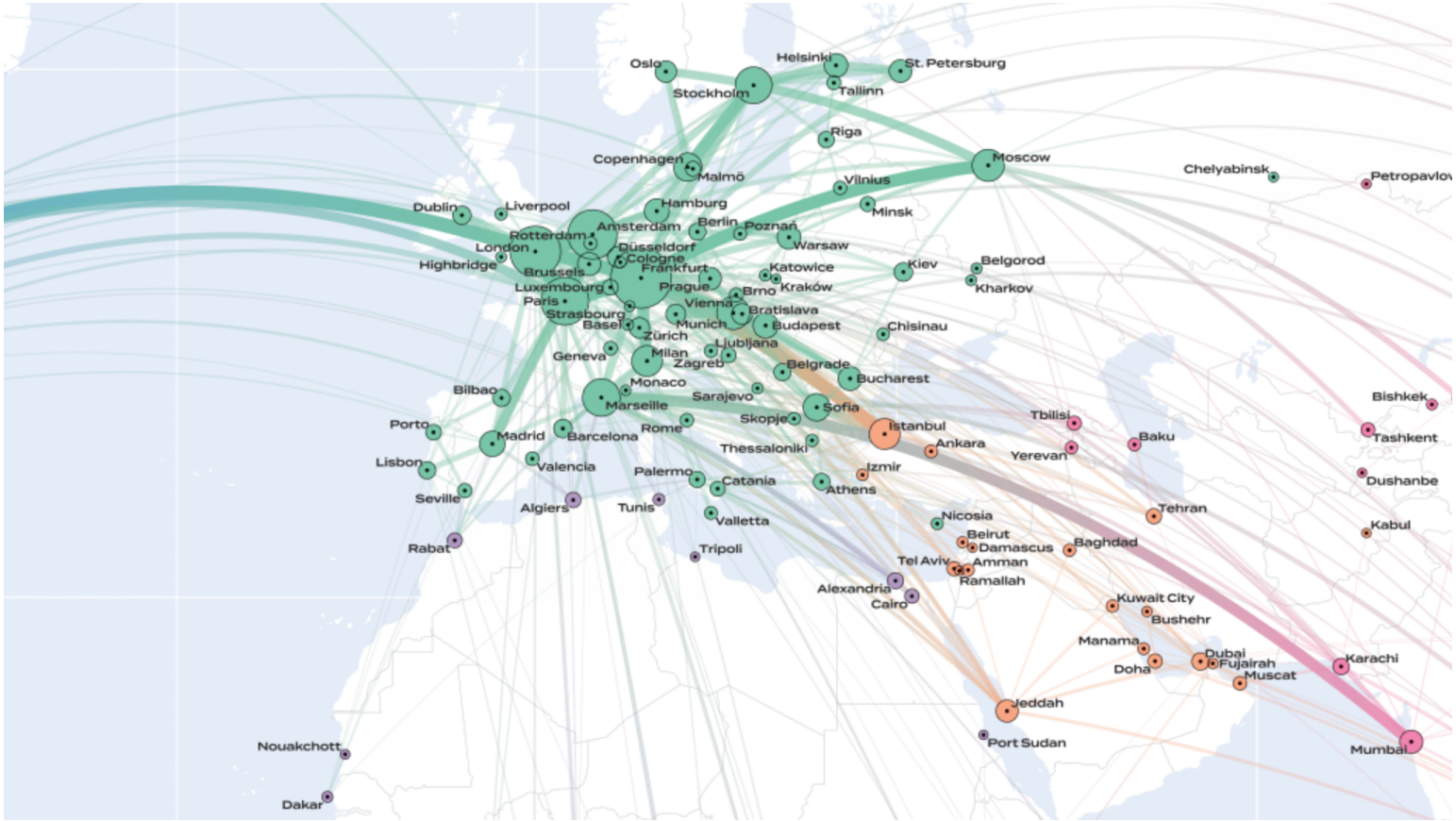
[2]Compilation de plusieurs sources de données dont GreenIT2019, Lean ICT 2018

In France

- 631 millions of equipements
- 58 millions of users (environ 11 composants par personne)

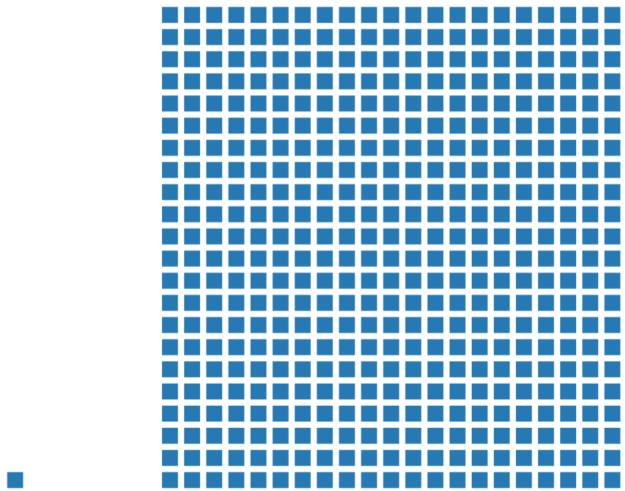


Internet traffic

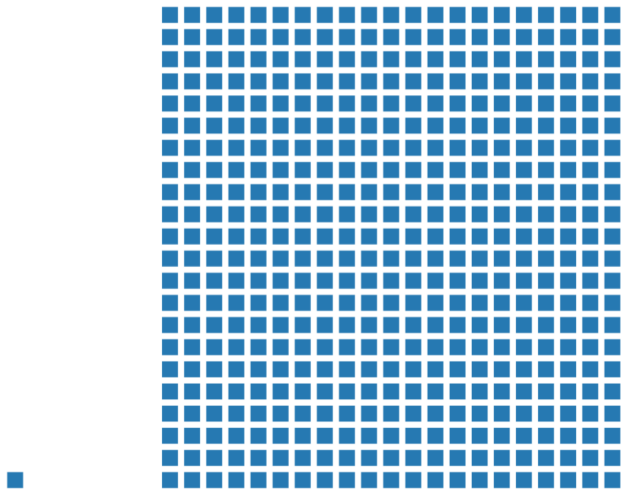


- 2002 : 3.2 exaBytes
- 2017 : 1.5 zettaBytes (10^{21})

- 500× more over 15 years !



- 500× more over 15 years !



Same order of evolution for Data.

entre 2010 et 2025 :

- GHG $\times$ 3,1
- Water $\times$ 2,4
- Abiotic resources $\times$ 2,1

WHY?

entre 2010 et 2025 :

- GHG $\times$ 3,1
- Water $\times$ 2,4
- Abiotic resources $\times$ 2,1

WHY?

- Doubling the size of the screens
- Slower energy gains
- Equipment for emerging countries
- Connected objects $\times$ 48
- Growth of digital services

- Digital is every where

- Digital is every where

- The Carbon cost of the digital is often hidden.

Impact of digital

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

[3] 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

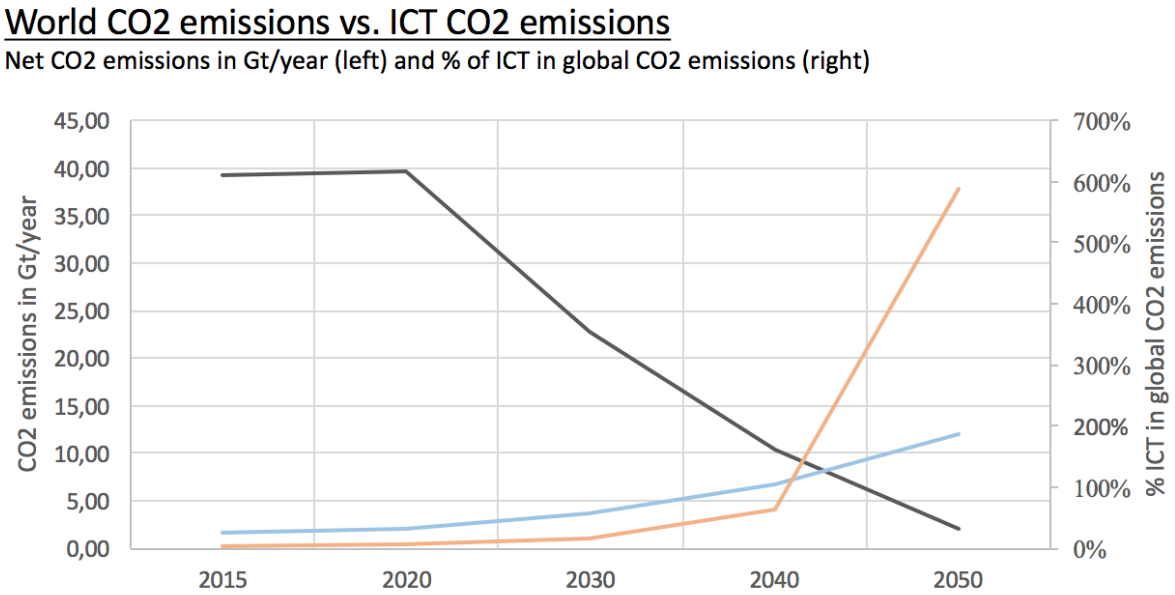
- Digital technology accounts for 5 to 6% of the world's primary energy consumption, roughly 4% of the total emissions [Lean ICT 19].
  Freitag et al. estimate the domain from 2.1 to 3.9 % of carbon emissions[3].
- Annual growth 6-9% (over 2015-2019).
- It is very difficult to quantified the part of AI (accelerator effect).

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

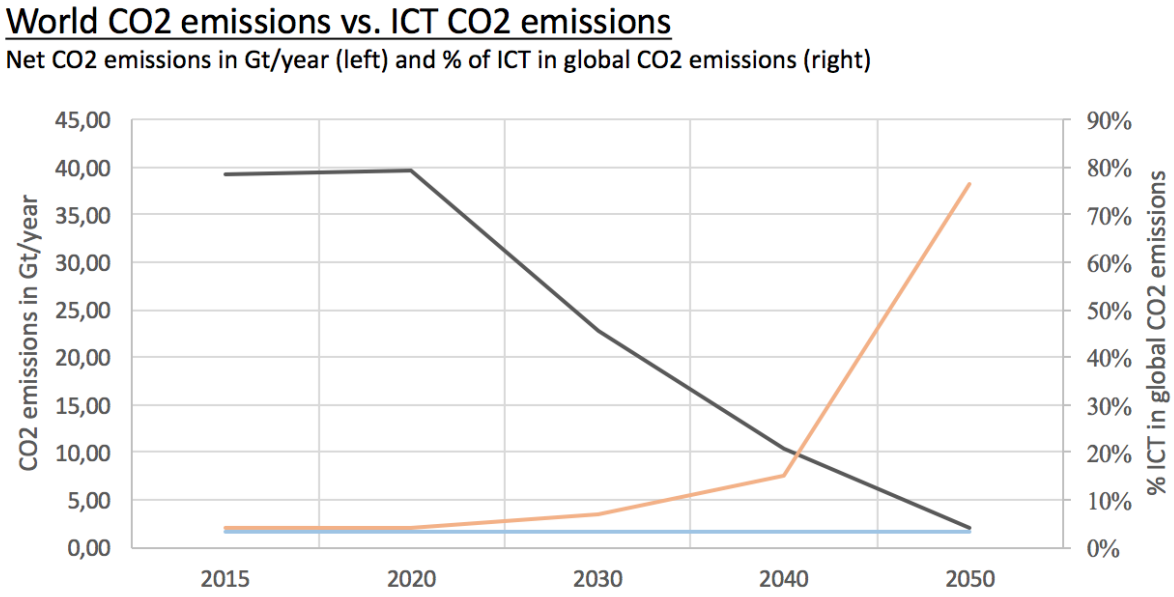
Plus précisément...

Simulateur réalisé avec l'aide de Yannick Malot (Doctorant CEA-LIG) pour comparer les scénarios SSP.

- Scénario le plus favorable SSP 1-1.9 avec ICT base de croissance minimale (6%)



SSP1-1.9 et ICT constant



- Ceci est une vue de l'esprit comme base de discussion pour alimenter le débat "le Numérique peut nous sortir de la crise"...

Focus sur le domaine HPC

- Le monde du numérique est très large
  il englobe le domaine du *calcul*
- Il existe des données sur la face visible de l'iceberg du calcul :
  le TOP500

TOP500

- Depuis 1993
- Classe les systèmes HPC les plus puissants (parmi plus de 10,000 supercomputers issus de 2,800 organisations, la plupart academiques).

- Il fournit des attribus sur les architectures, performances and location (nombre de cores, CPU/GPU, capacités mémoire, constructeurs, etc.).
- R_{max}: maximum Linpack perf achieved
- R_{peak}: theoretical performances

Green500

- Initié en 2008, créé officiellement en 2013.
- Il utilise les mêmes benchmarks que le TOP500
- Basé sur les mêmes attribus.
- Non renseigné pour plusieurs systèmes

Analyse critique du TOP500

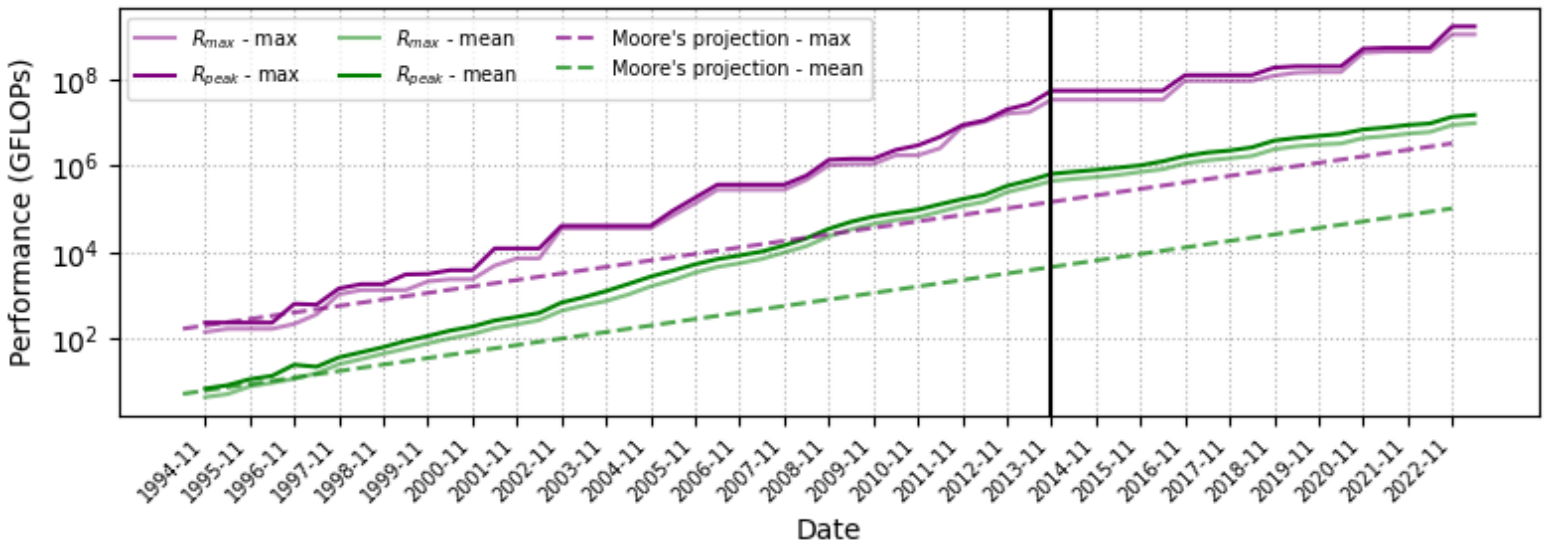
- Caractère déclaratif :
  base volontaire, non contractuelle.
  Il manque de gros industriels et certains pays
- Gros biais sur les très gros systèmes HPC, en particulier pour le Green, des systèmes plus petits pouvant être beaucoup plus efficaces relativement.

Empiric laws

Some macroscopic indicators

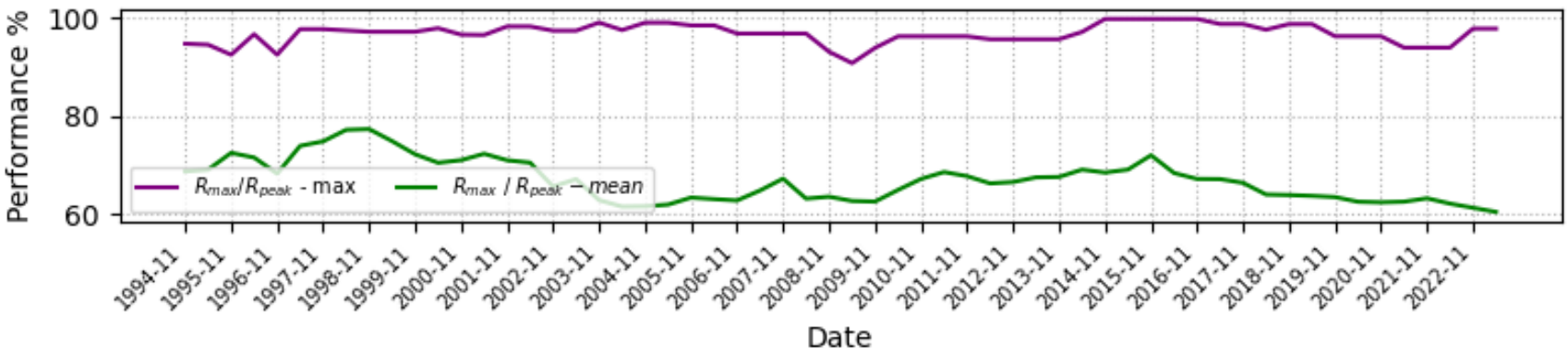
- **Moore** : Performances d'un système
  Le nombre de transistors des circuits intégrés double tous les 2 ans.
  Extension aux systèmes parallèles.
- **Koomey** : Similaire mais cible l'efficacité énergétique
  Nombre de calculs élémentaires par Joule d'énergie dissipée.
  Double tous les 18 mois, avant 2010. Aujourd'hui, tous les 2 ans et quelques mois.

Performance



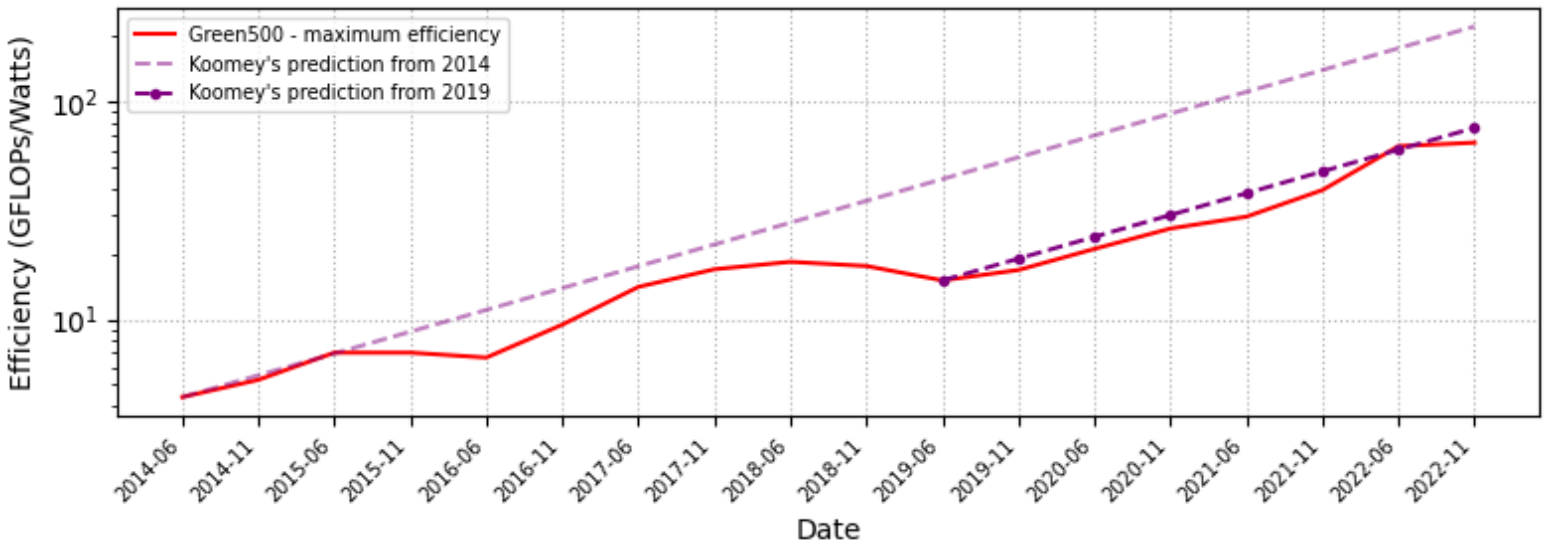
- Clairement une rupture autour de 2013-2014

Des systèmes toujours plus complexes

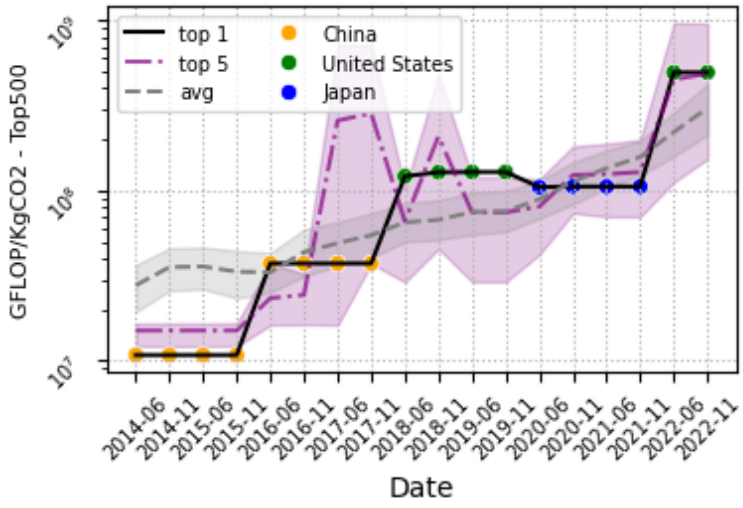


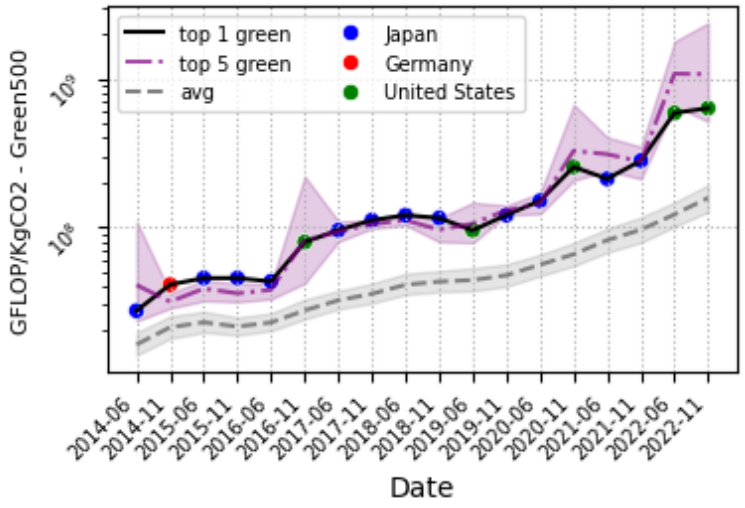
- C'est encore pire sur les applications réelles !

Efficacité énergétique



Outil prospectif

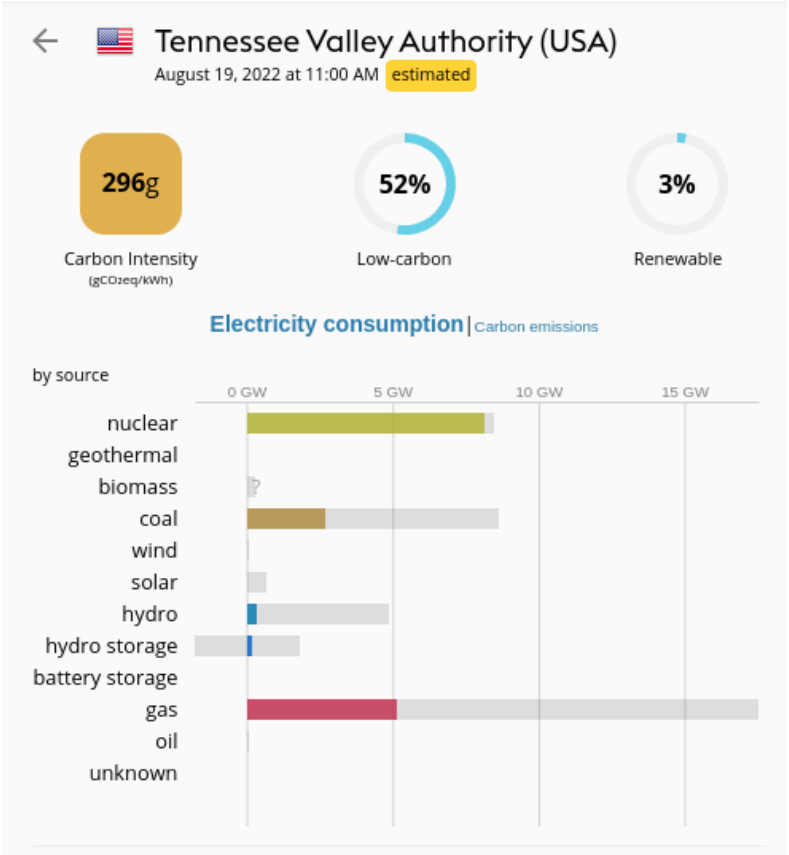




- Cibler 2030 : est-ce soutenable ?
- Question sous-jacente :
  Que veut dire un HPC au service de l'écologie ?

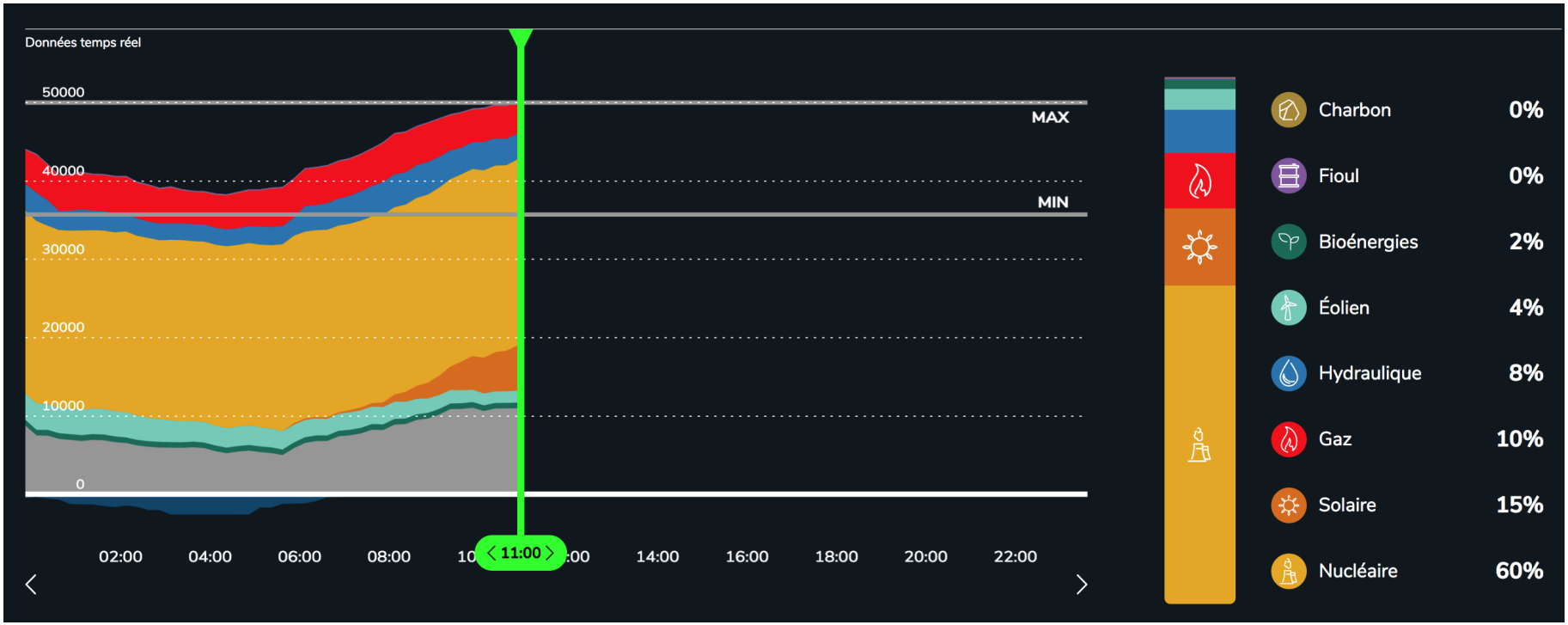
From KWh to CO2eq

Exemple de Frontier

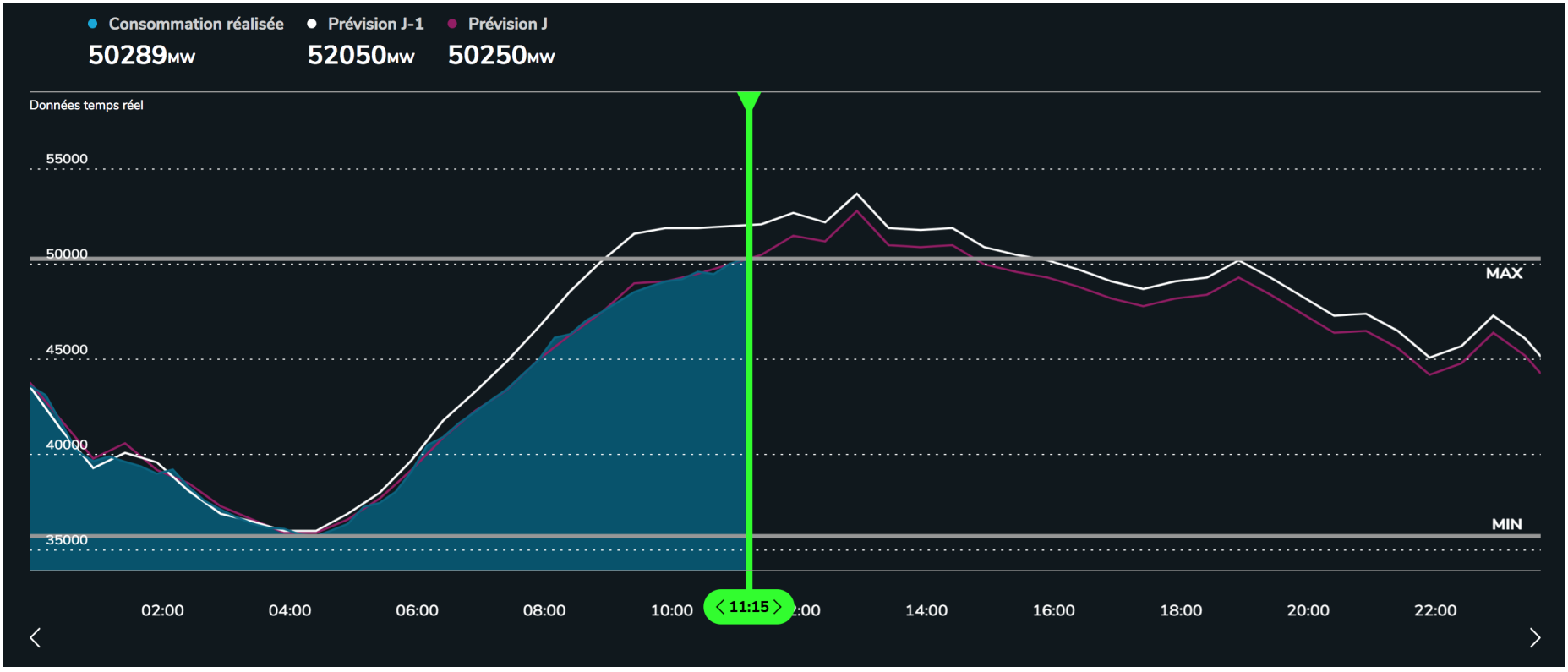


En France

Mixte énergétique



Prévisions


Consommation réalisée
50289MW
Prévision J-1
52050MW
Prévision J
50250MW
Données temps réel
55000
50000
45000
40000
35000
MAX
MIN
02:00
04:00
06:00
08:00
10:00
11:15
14:00
16:00
18:00
20:00
22:00

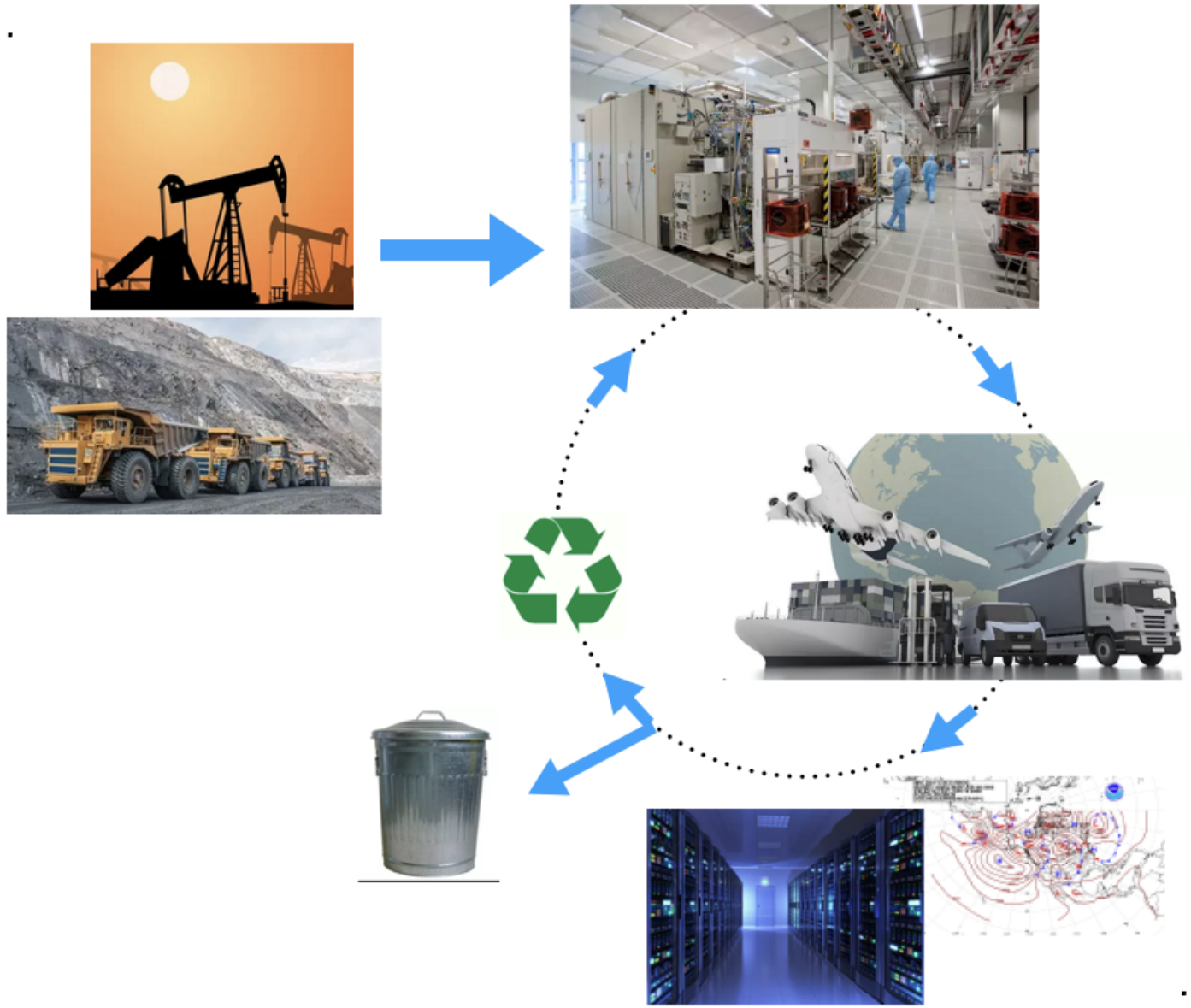
Evaluer le coût d'une application HPC

- C'est indispensable !
- Il existe des méthodologies sur les différentes phases du cycle de vie.
  Il faut tout compter
  Au premier ordre, i.e. dans le périmètre de l'application déployée.

Analyse de cycle de vie

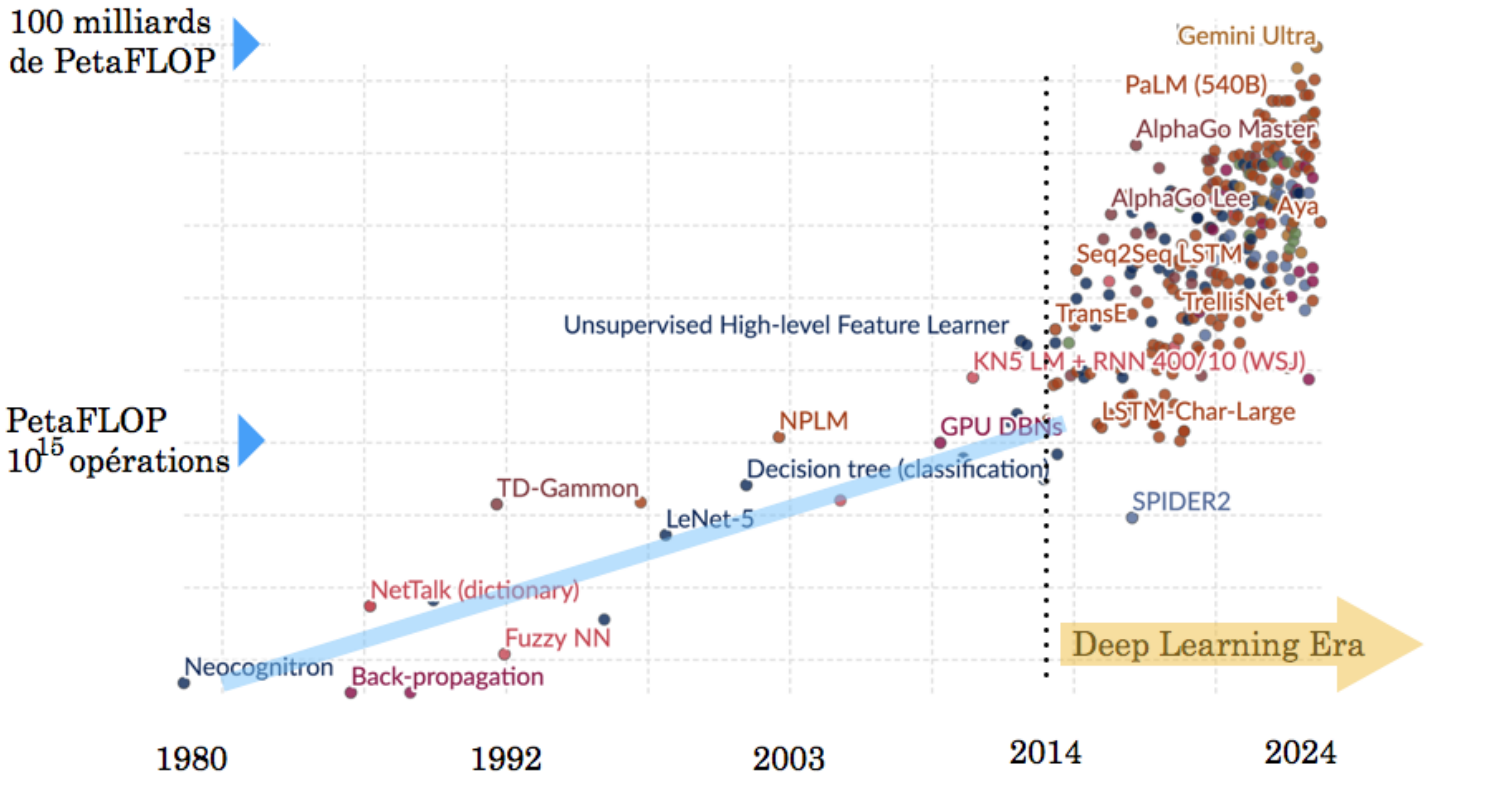
- Une ACV cible essentiellement les *effets directs*.
- Il faut aussi prendre en compte les *effets indirects* et rebonds.
  Ce qui n'est pas compté dans le périmètre initial.

ACV



Effet accélérateur de l'IA

Computation used to train notable artificial intelligence systems



Data source: Epoch (2024)

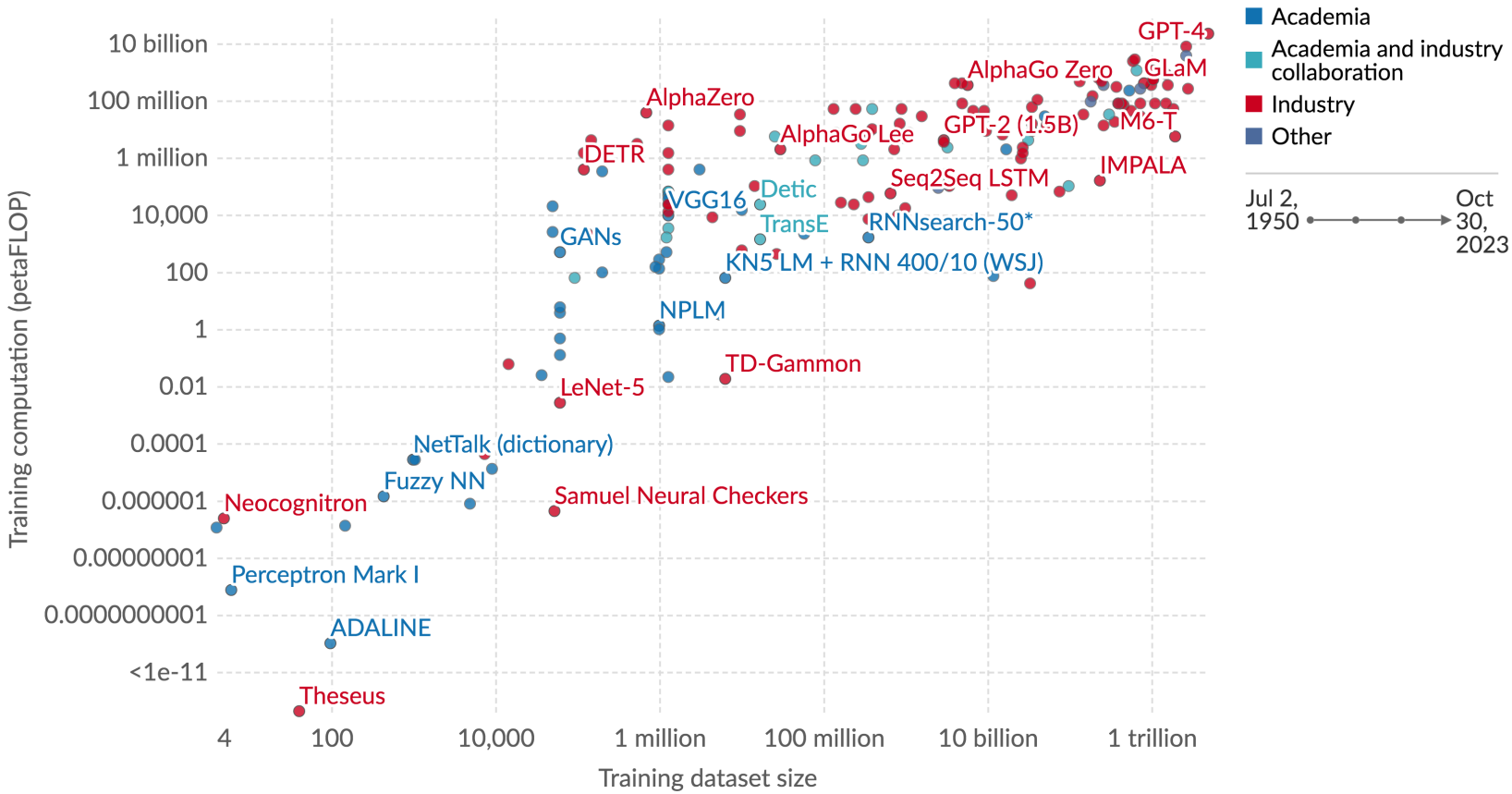
OurWorldInData.org/artificial-intelligence

Et les données dans tout cela ?

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.

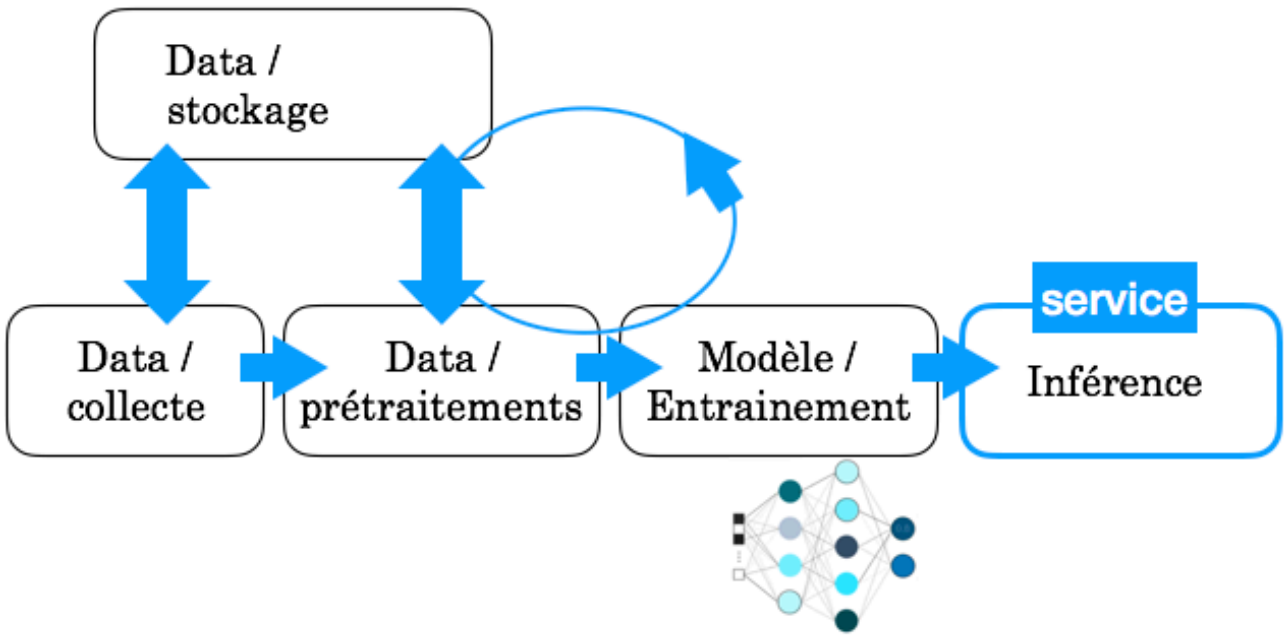


Data source: Epoch (2024)

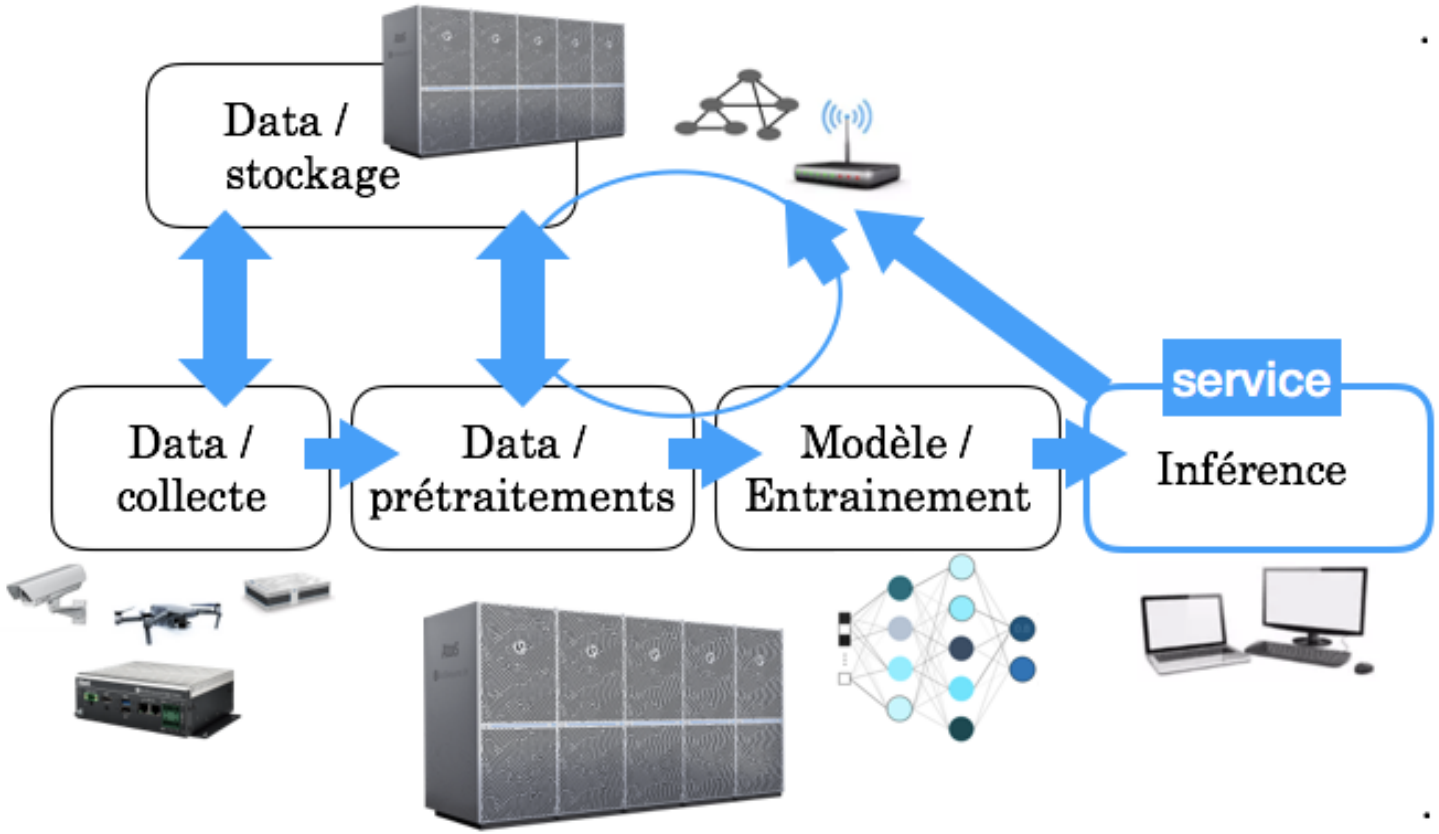
OurWorldInData.org/artificial-intelligence | CC BY

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.

ACV d'un service numérique (IA)



Il faut tout compter !



Mesurer

Pourquoi ?

- Quantifier l'ordre de grandeur d'un équipement-service numérique
- pour comparer ? Comme base du Politique (éclairer les décideurs) ?
- Casser l'illusion de la dématérialisation
- Remise en cause potentielle sur une base bénéfice/risque

Comment ?

- Voir Mise en situation avec Guillaume.

Un premier impératif : évaluer/mesurer

Sans mesure : Pas de Science !

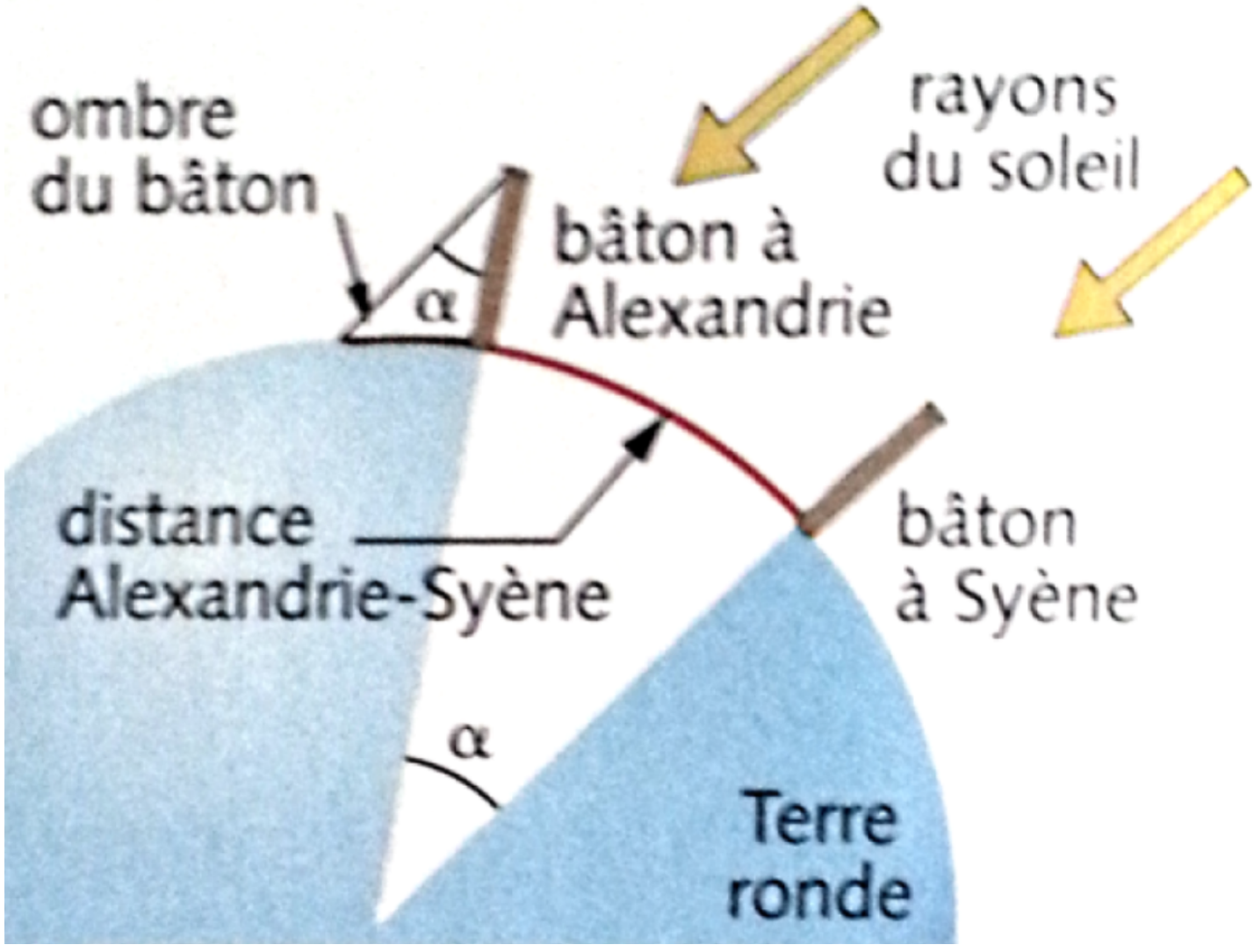
- Quantifie
- Objectivise
- Relativise (rend comparable)

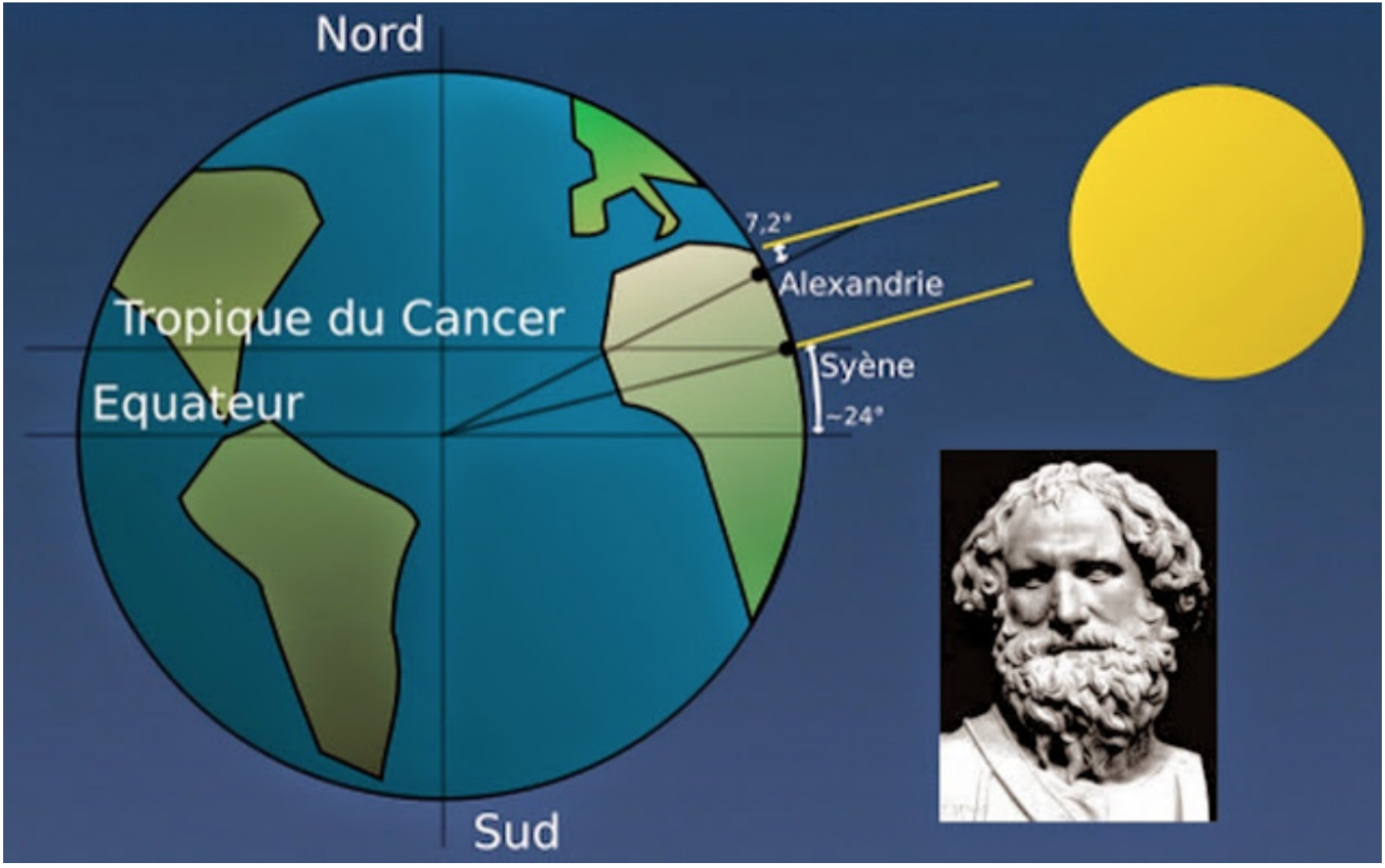
Qu'est-ce qu'une "bonne" mesure ?

- Objet de la mesure : bien défini
- Méthode consensuelle et pratique :
  acceptable pour l'époque
- Unité de mesure : longueur d'un "stade"
  Attention : pas très bien défini (olympique, égyptien)[4]
- Remise en cause potentielle sur une base bénéfice/risque
- incertitude élevée
- Hypothèses
  terre parfaitement ronde
  Alexandrie et Syène alignées sur le même méridien
  Approximation des petits angles

[4] Archimède, Ptolémée

Principe


ombre
du bâton
rayons
du soleil
bâton à
Alexandrie
α
distance
Alexandrie-Syène
bâton
à Syène
α
Terre
ronde


Nord
7,2°
Alexandrie
Tropique du Cancer
Syène
Equateur
~24°
Sud

Etait-ce une "bonne" mesure ?

Si l'on se fie au résultat :

- Syène-Alexandrie : 5000 stades
- Angle à Alexandrie : $\frac{1}{50}$ de cercle, soit environ 7.2 degrés

Circonférence $= 50 \times 5000$ stades
Donc, 39375 km
Impressionnant !
Le chiffre actuel est 40 075,017 km à l'équateur [wikipedia], soit, moins de 10^{-3} en erreur relative !

Oui, mais, c'est par chance !

Lesson learned

Comment mesurer ?

Le CPU, la mémoire, les réseaux, tout est électricité !
que l'on transforme en eqCO2.

Rebond

- Direct :
  Une technologie plus efficace augmente les usages.
- On a aussi un effet rebond indirect lorsque des gains réalisés dans un domaine génèrent de la consommation dans un autre.

Ainsi une démarche de sobriété peut aussi être source d'effets rebond. du fait des économies réalisées qui sont réinvesties (qu'elles soient monétaires ou temporelles), ou du fait de déculpabilisation sur la consommation d'autres produits

L'accélération et les bouleversements induits

La technologie s'améliore très vite

Mais la société (y compris les chercheurs) ne peuvent suivre le rythme :

- Evolution du public pour adopter les nouvelles technologies.
- Incompréhension des mécanismes profonds.
- La construction scientifique nécessite un temps long.
- Problèmes éthiques : à prendre en compte en amont en évitant le mécanisme d'action-réaction.

Evaluer le coût d'un service

Distinguer entrainement versus inférence
Pour l'IA générative, l'inférence est bien supérieure !
Mais ceci est très dur à déterminer...

- Il existe des méthodologies sur les différentes phases du cycle de vie.
- Il faut tout compter
  Relativement bien renseigné pour le premier ordre, i.e. dans le périmètre de l'application déployée.

Analyse de cycle de vie

- Une ACV cible essentiellement les *effets directs*.
- Il faut aussi prendre en compte les *effets indirects* et *rebonds*.
  Ce qui n'est pas compté dans le périmètre initial.

Rebond

- Direct :
  Une technologie plus efficace augmente les usages.
- On a aussi un effet rebond indirect lorsque des gains réalisés dans un domaine génèrent de la consommation dans un autre.

Ainsi une démarche de sobriété peut aussi être source d'effets rebond du fait des économies réalisées qui sont réinvesties (qu'elles soient monétaires ou temporelles), ou du fait de déculpabilisation sur la consommation d'autres produits

Comment garantir que le bilan est vraiment positif ?

Remettre la question du sens au centre de nos sujets.

L'analyse critique nous impose une nouvelle manière d'évaluer le rapport de l'IA aux questions environnementales.

- mesurer. On ne remet rien en cause, on observe sur des bases scientifiques
- améliorer à partir de ce que l'on a mesuré
- remettre en question les usages d'un nouvel algo/outil/usage avant de le déployer

"Efficiency" or "Sufficiency" ?

- The community has realized that it needs to react.
- The main way forward is to optimize platforms and applications from an energy standpoint.
  This is eco-efficiency: reducing the intensity of environmental impacts or resource use per unit of economic value produced.
- We can also switch to renewable energy for decarbonized computing.