Limiter les impacts environnementaux

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

pour les 20 ans du *groupe Calcul*
Paris-Jussieu, 3 juin 2024

Pourquoi ?

La crise environnementale – Etat des lieux

Les limites planétaires sont pour la plupart dépassées.

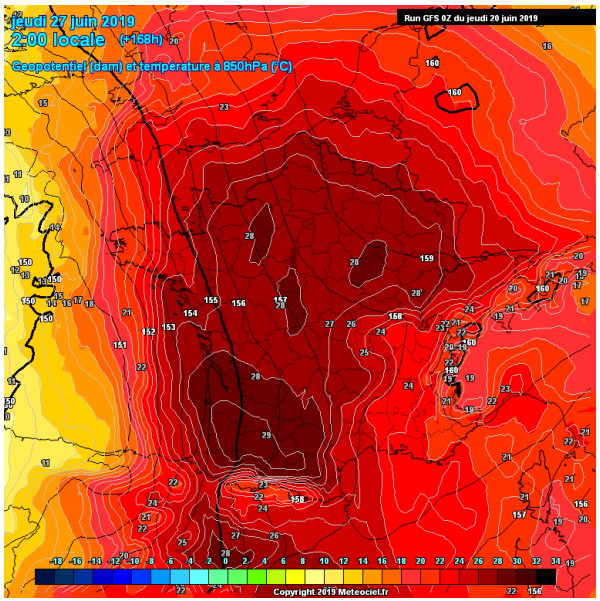
Il existe une prise de conscience, mais pas à la hauteur des enjeux.

Pourquoi ?

La crise environnementale – Etat des lieux

Les limites planétaires sont pour la plupart dépassées.
Il existe une prise de conscience, mais pas à la hauteur des enjeux.

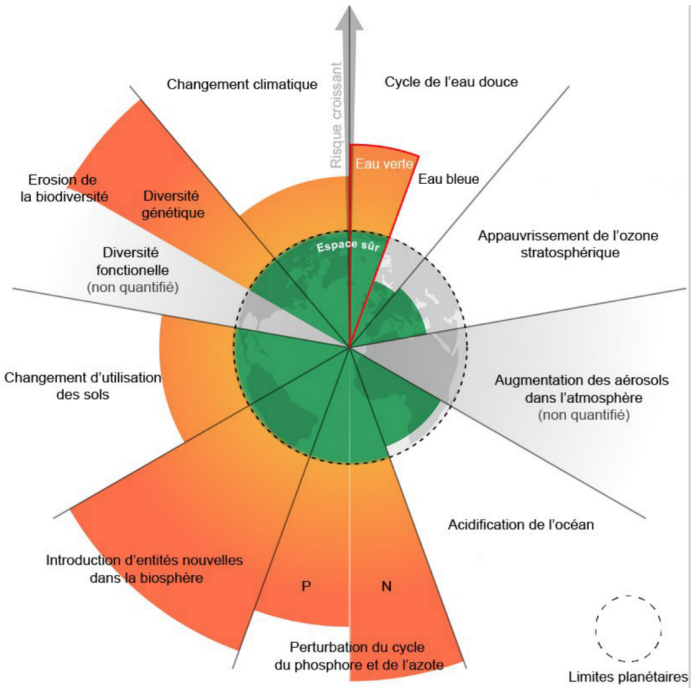
- La crise nous touche toutes et tous et s'exprime en particulier par le **dérèglement climatique**.
  - Des températures extrèmes
  - Des sécheresses
  - Des canicules
  - Des incendies
  - Des tempêtes
  - Des inondations
  - ...



Pour rester dans l'ambiance...

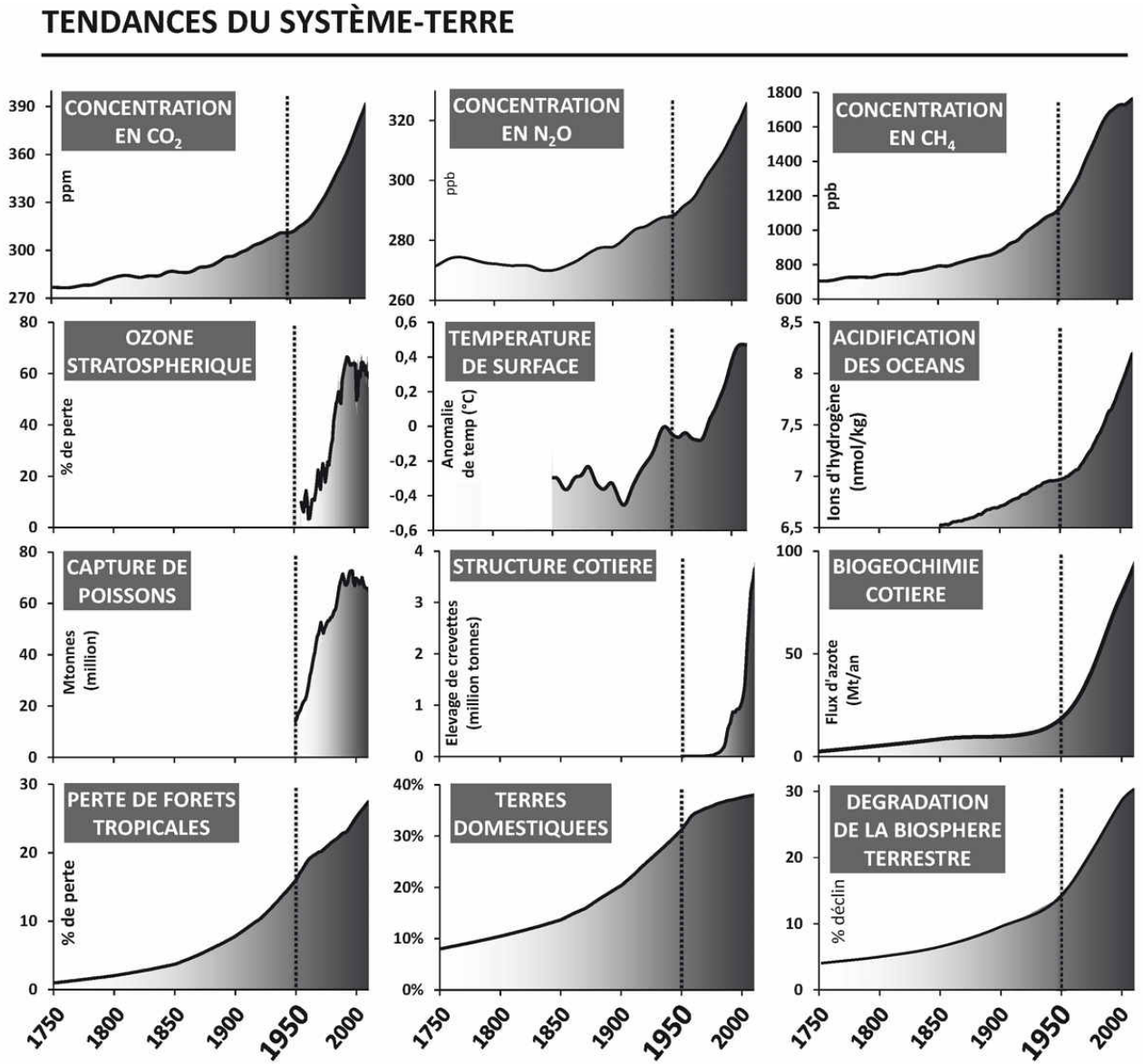
Il n'y a pas que le climat !

Violation des **limites planétaires** qui engendrent des grandes **menaces** civilisationnelles :



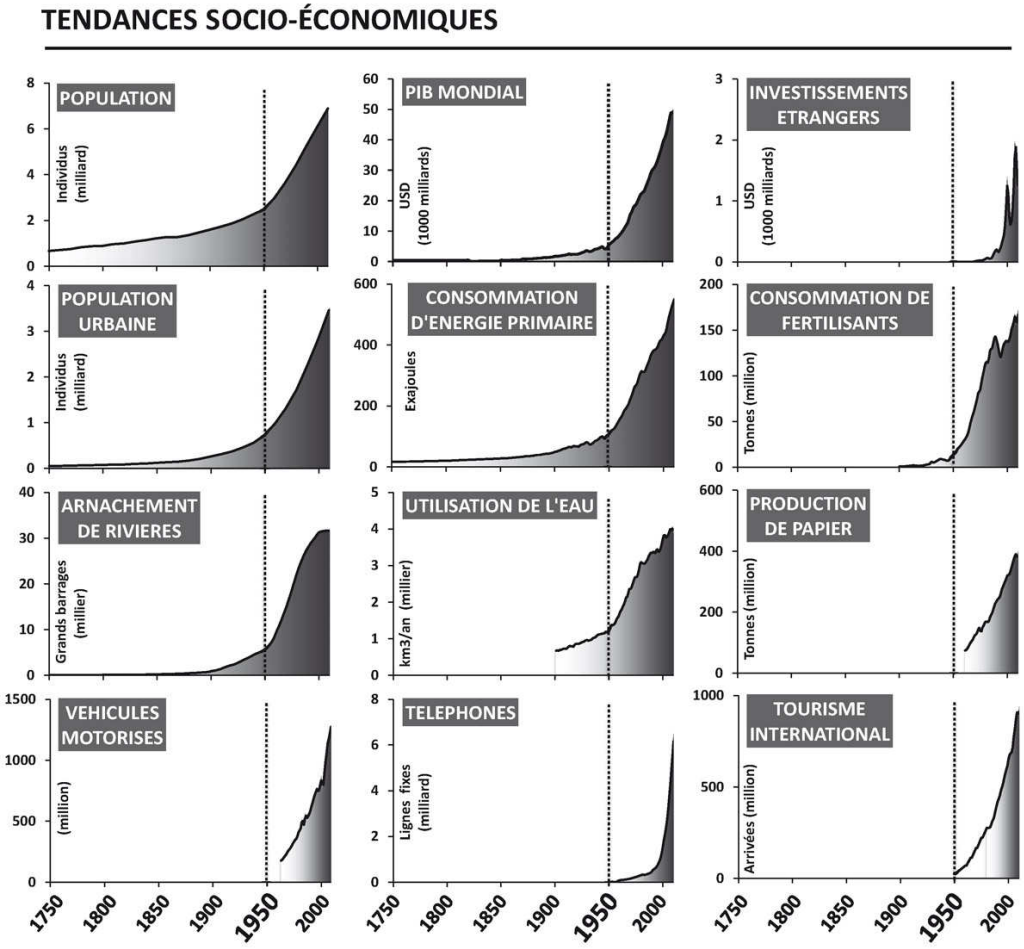
- Stress hydrique
  cycle de l'eau
  acidification des océans
- Epuisement
  des ressources abiotiques
- Erosion de la bio-diversité
  Extinctions des espèces
- Changement d'utilisation
  des sols

Etat de la planète



source : [Pablo Servine et Raphael Stevens 2015]

Origine de la crise



Clairement anthropique (attesté par le GIEC, confirmé par une étude fin 2023 avec un consensus de 97% de scientifiques).

Objectif(s) et points abordés dans l'exposé

Constat

L'humanité fait face à des défis sans précédents.

L'humanité ne dispose ni d'une énergie bon marché, ni de ressources illimitées permettant au système technologique complexe actuel de perdurer.

Objectif(s) et points abordés dans l'exposé

Constat

L'humanité fait face à des défis sans précédents.

L'humanité ne dispose ni d'une énergie bon marché, ni de ressources illimitées permettant au système technologique complexe actuel de perdurer.

Les *solutions* envisagées, basées sur le numérique, pour faire face à la crise environnementale sont variées et souvent clivantes.

- Quelle est la place que peut prendre la communauté HPC ?
  - Se focaliser sur des applications au service du climat ?
  - Réduire les impacts négatifs du domaine ? et alors, comment ?
  - Réduire la voilure et /ou changer de paradigme de calcul ?

Agenda de cette présentation

- Une brève introduction sur les émissions Carbone.
- Impacts du HPC, de quoi parle-t-on au juste ?
- Dynamique d'emballement (Numérique et Calcul)
- Une brique de base : Mesurer les impacts
  Analyse de Cycle de Vie, effets indirects et rebonds
- Comment envisager un HPC compatible avec les limites planétaires ?

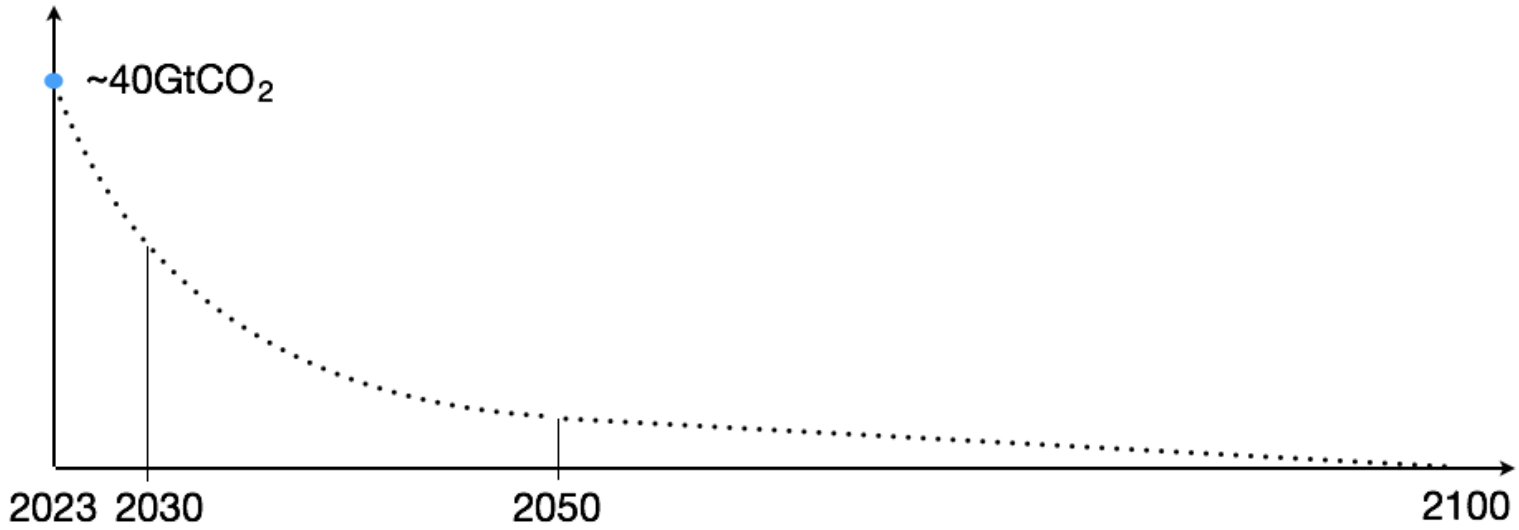
Préliminaires sur les émissions Carbone

Quiz en 1 slide pour revenir sur la dynamique du réchauffement avant de rentrer dans le vif du sujet.

- Gaz à effet de Serre, forçage radiatif, puits de Carbone, etc..
- Le CO_2 reste plus d'un siècle dans l'atmosphère.
- On émet de l'ordre de 40 Gt de CO_2 par an dans le monde.
- Il n'y a pas que le CO_2

Si l'on réagit dès maintenant

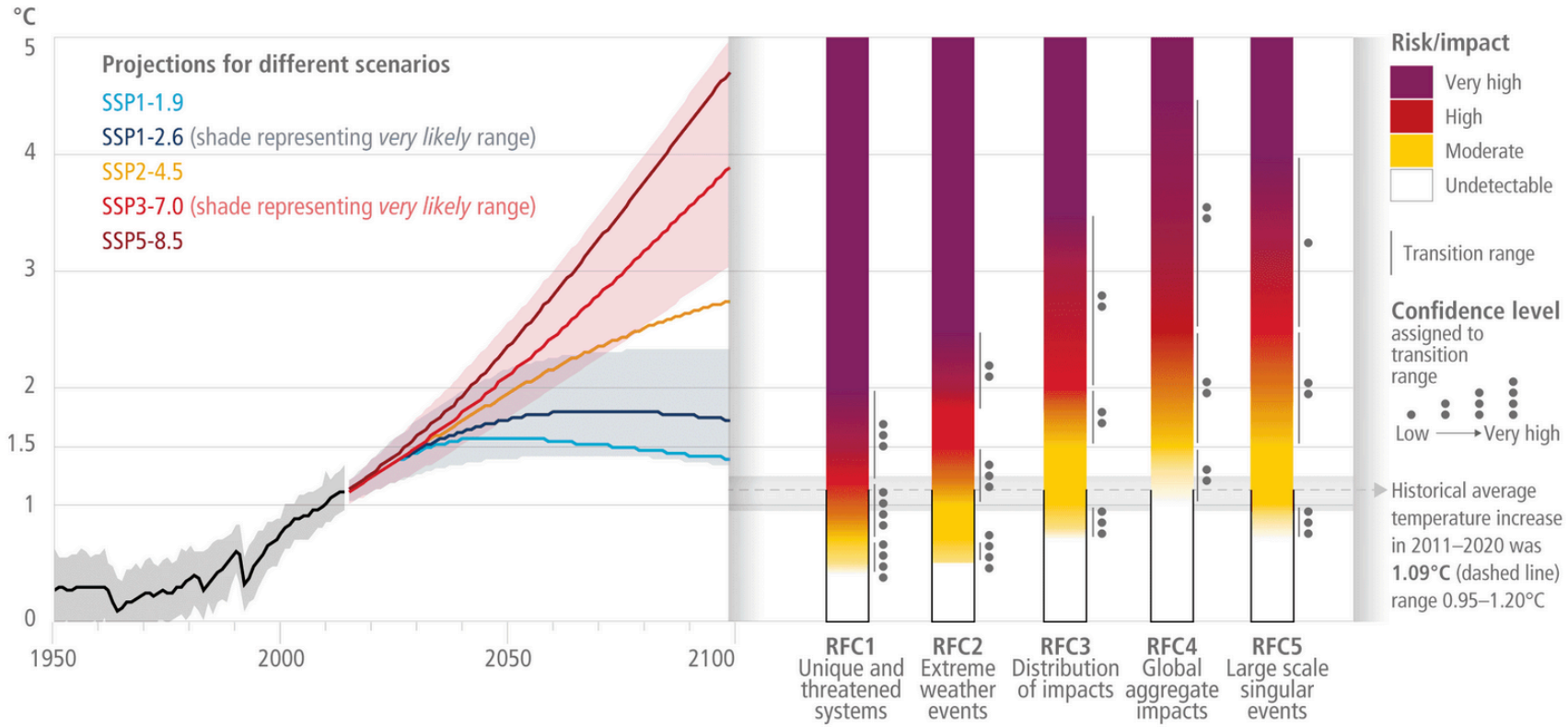
- Budget maximal restant pour maintenir le réchauffement en dessous de 1.5 degrés : moins de 1000 Gt de CO_2



- Les situations sont fortement différentes dans le monde.
- On doit réduire nos émissions d'ici 2050, de l'ordre de 7 à 8% par an d'ici 2050, voire plus si on tarde encore à réagir...

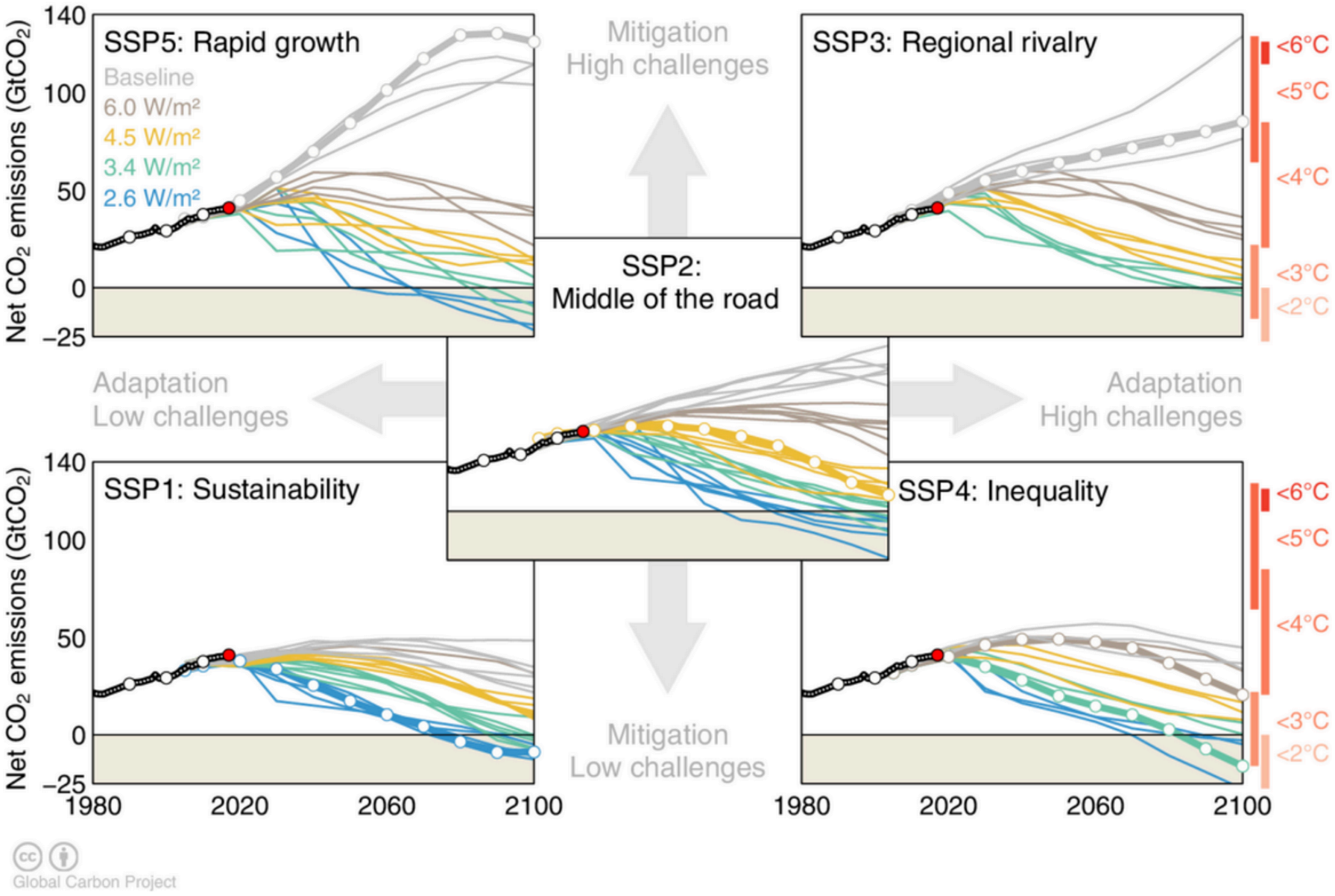
Quelles trajectoires

scénario de référence SSP x-y (Shared Socio-economic Pathways)
5 classes de scénarios
y : forçage radiatif à la fin du siècle (en W/m^2)



sources : GIEC et Carbone4

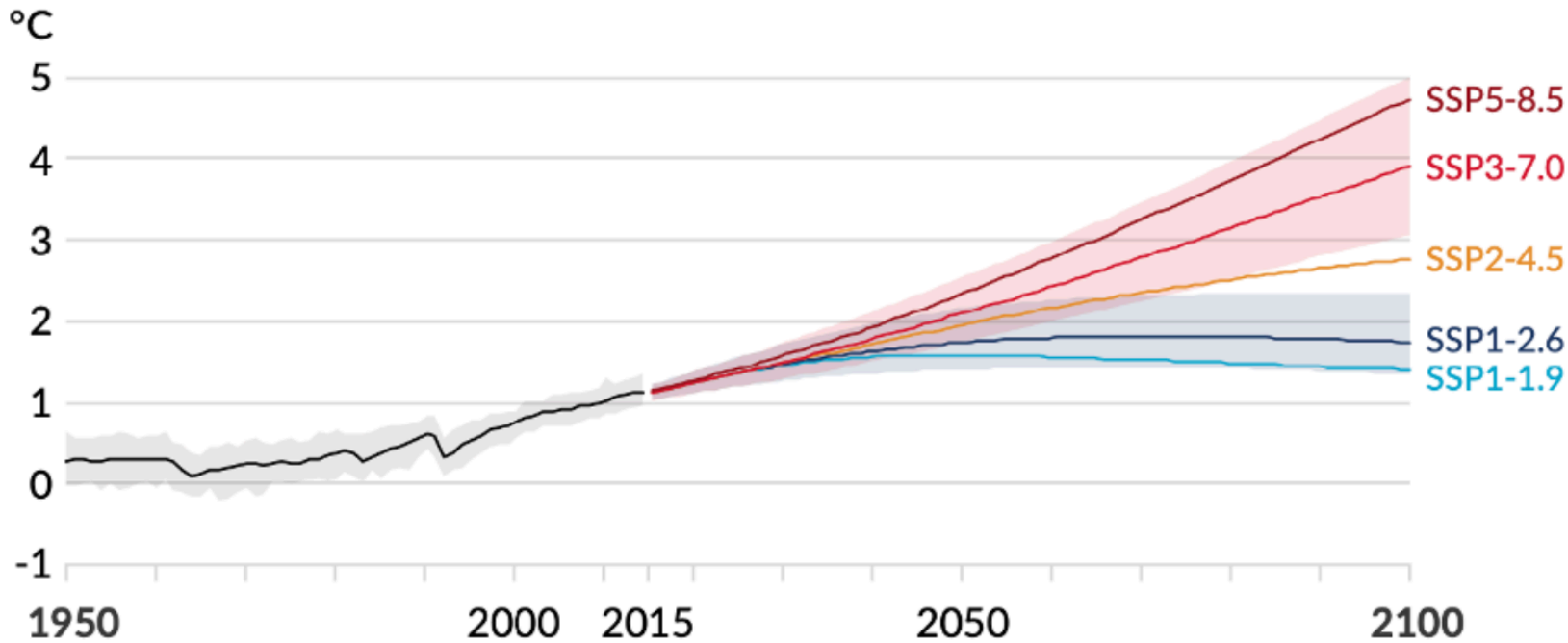
Détails des 5 classes de scénarios



Global Carbon Project

Projection des scénarios

	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



Le numérique :
Un domaine compliqué à cerner...

Distinguons les équipements numériques eux-mêmes de la *numérisation* qui ruisselle dans tous les domaines de la Société[1].

[1] classification OCDE et G. Roussihle: *situer le numérique* 2022

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

Le numérique : Un domaine compliqué à cerner...

Distinguons les équipements numériques eux-mêmes de la *numérisation* qui ruisselle dans tous les domaines de la Société[1].

Contribution aux émissions de CO_2

- le numérique représente de l'ordre de 5 à 6% de l'énergie primaire mondiale, soit en gros 4% des émissions modiales [Lean ICT 19].
  Freitag et al. estiment entre 2.1 et 3.9 d'émissions carbone[2].
- Croissance annuelle de 6-9% (sur 2015-2019).
- Il est très difficile de quantifier la part du HPC et encore plus pour l'IA (effet accélérateur).

---

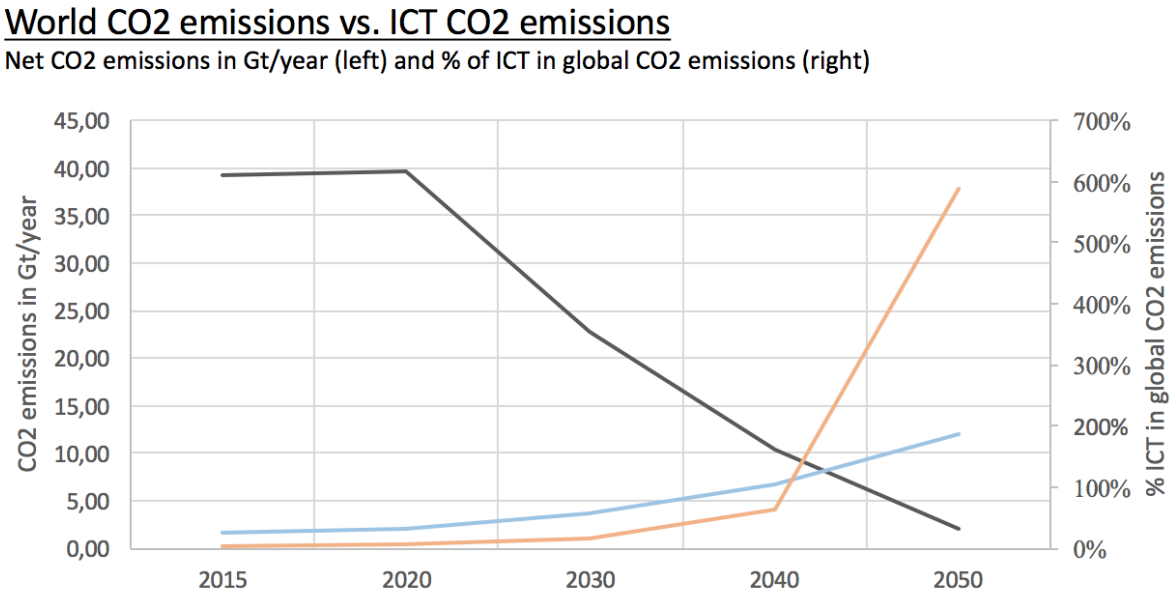
[1] classification OCDE et G. Roussihle: *situer le numérique* 2022

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

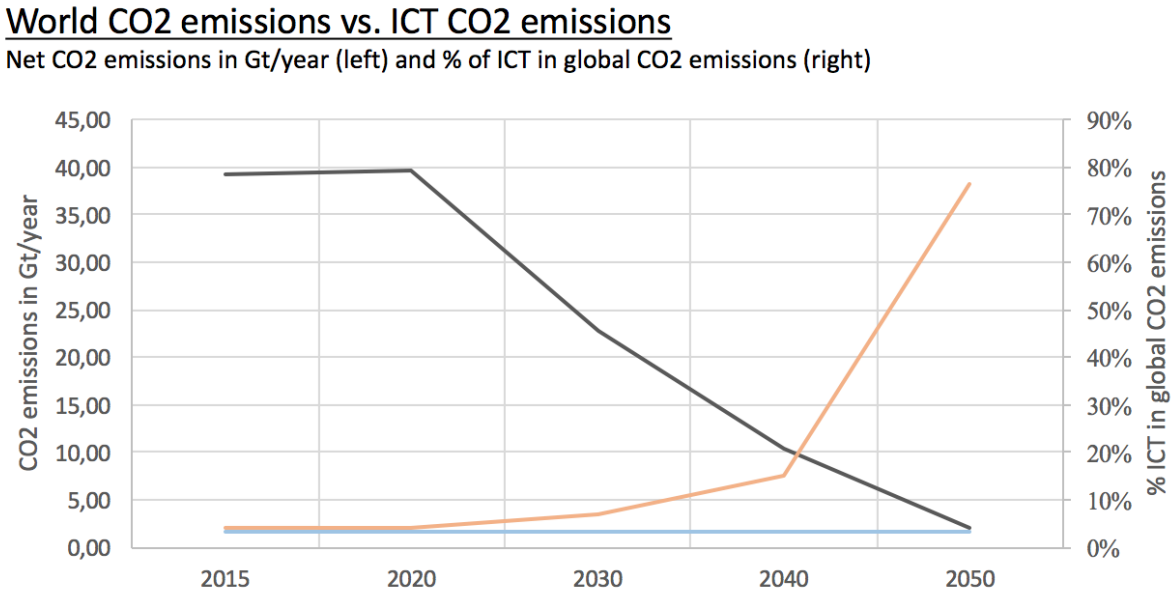
Plus précisément...

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

Bref rappel sur le 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 systèmes et benchmarks que le TOP500
- Basé sur les mêmes attributs.
- Non renseigné pour plusieurs systèmes

Analyse critique du TOP500

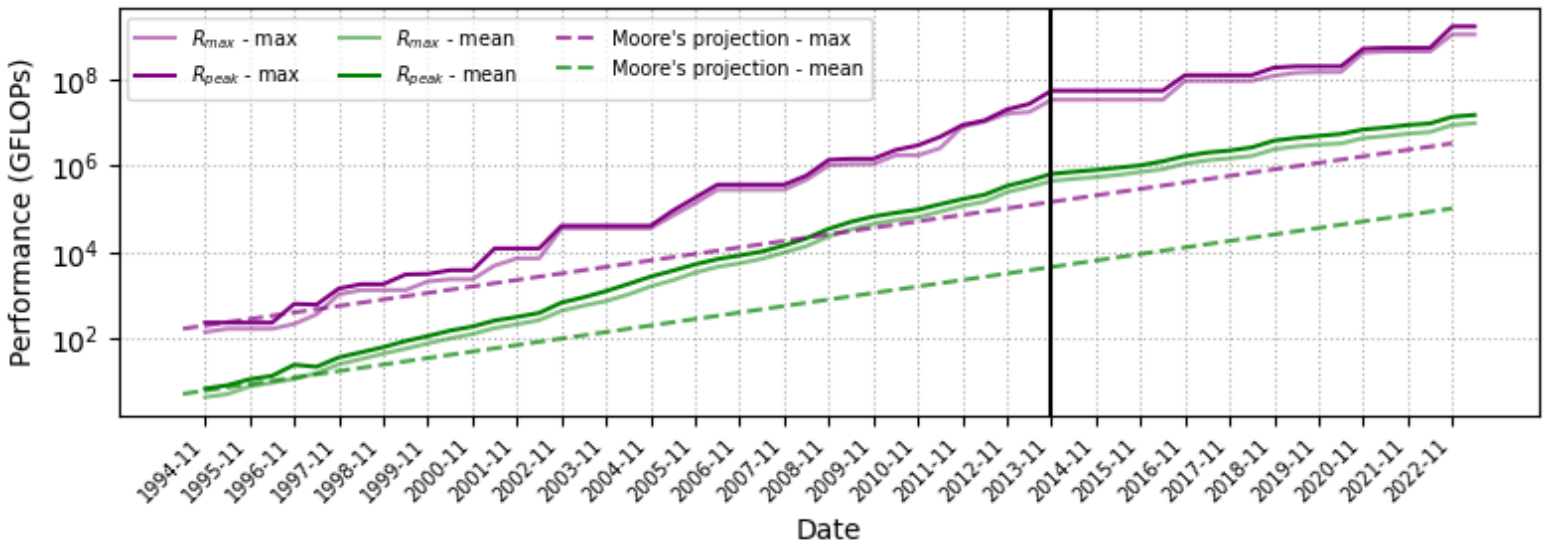
- Caractère déclaratif :
  base volontaire, non contractuelle.
  Il manque de gros industriels et certains pays
- Gros biais sur les très gros systèmes HPC, en particulier pour le Green, des systèmes plus petits pouvant être beaucoup plus efficaces relativement.

Les lois empiriques

Capturer quelques indicateurs macroscopiques

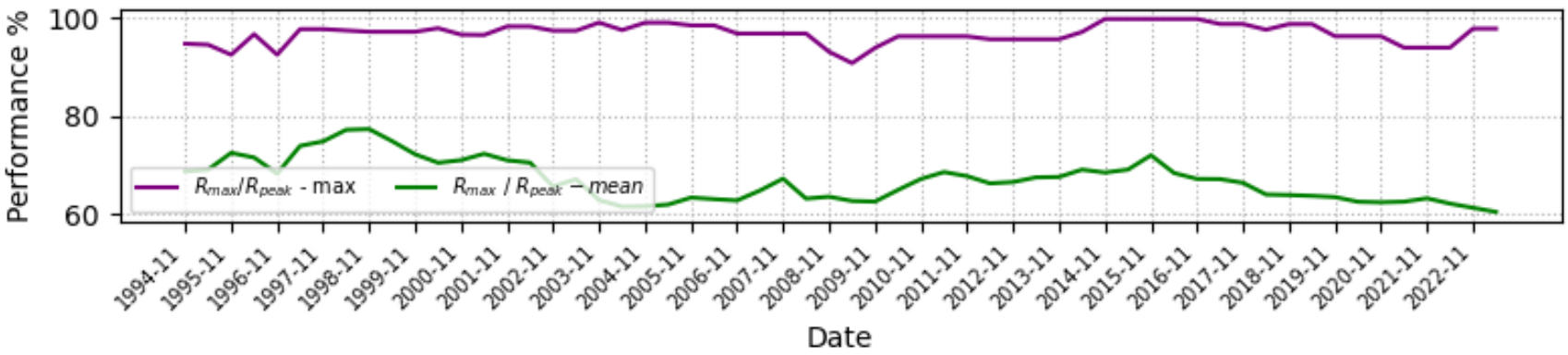
- **Moore** : Performances d'un système
  Le nombre de transistors des circuits intégrés double tous les 2 ans.
  Extension aux systèmes parallèles.
- **Koomey** : Similaire mais cible l'efficacité énergétique
  Nombre de calculs élémentaires par Joule d'énergie dissipée.
  Double tous les 18 mois, avant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

Performance



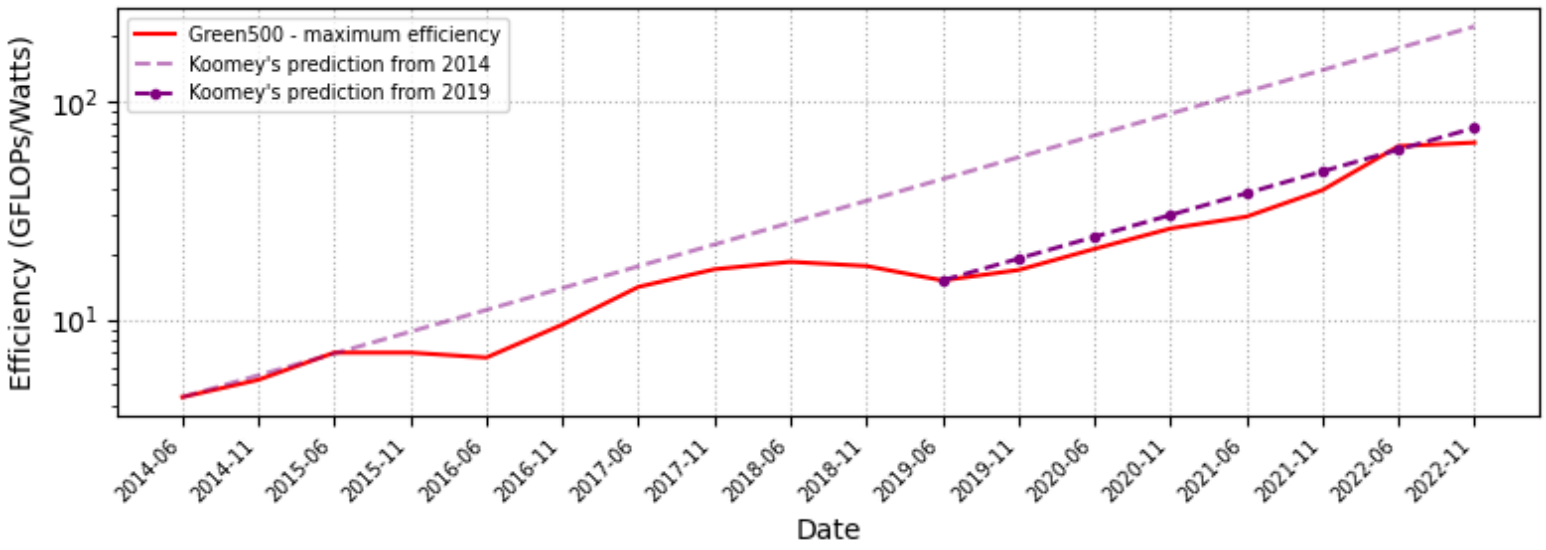
- Clairement une rupture autour de 2013-2014

Des systèmes toujours plus complexes

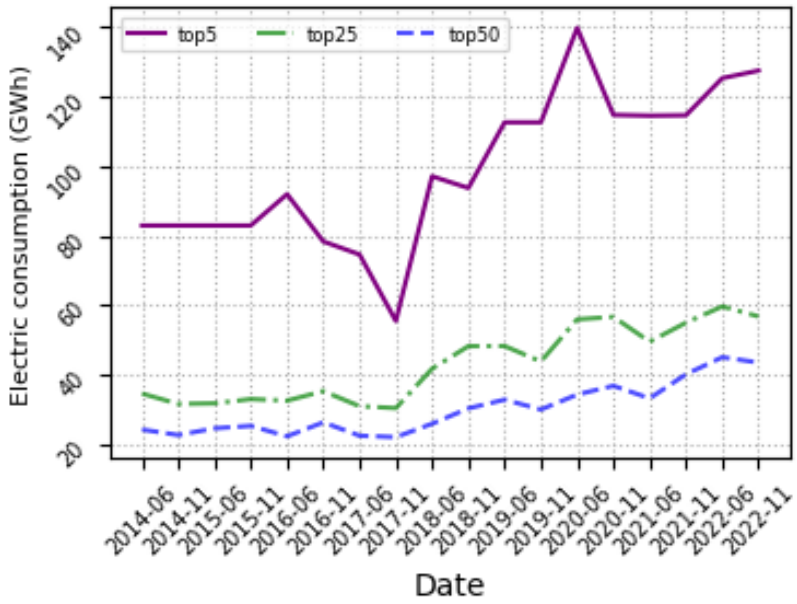


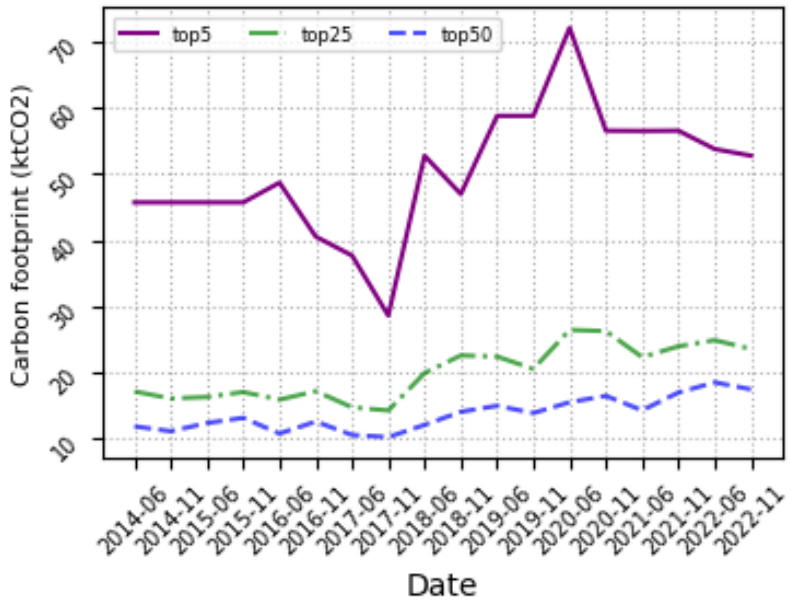
- C'est encore pire sur les applications réelles !

Efficacité énergétique



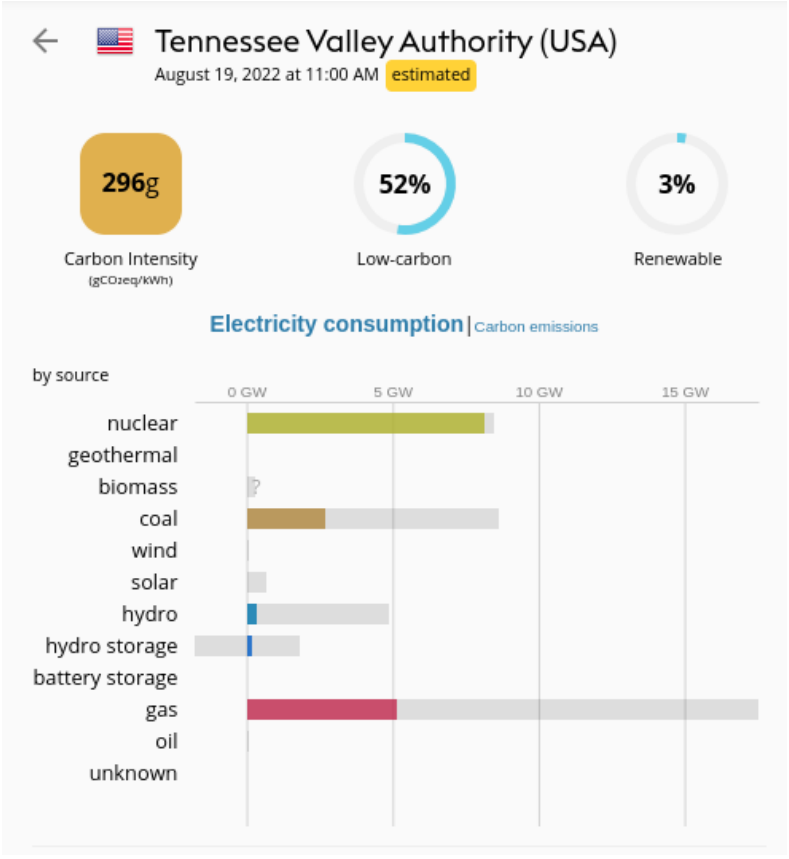
Et si on ramène en émissions carbone





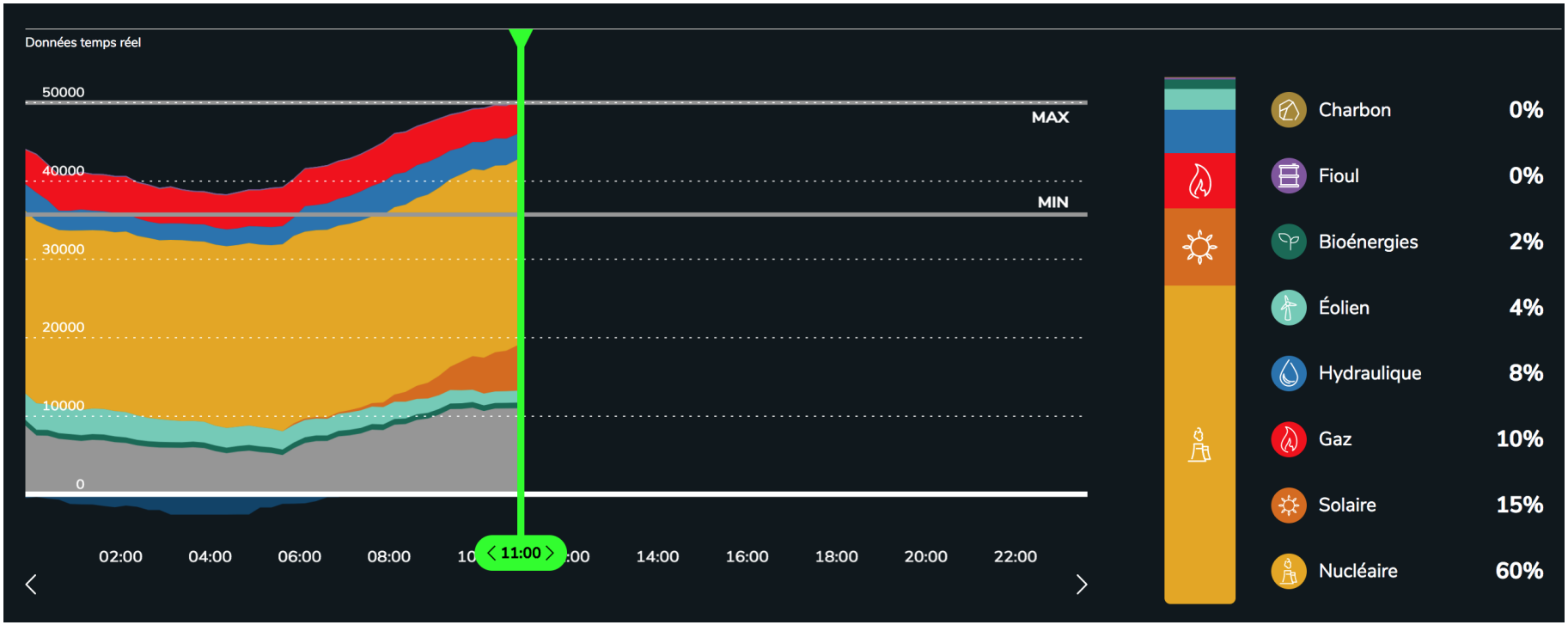
Passer des KWh au CO2eq

Exemple de Frontier

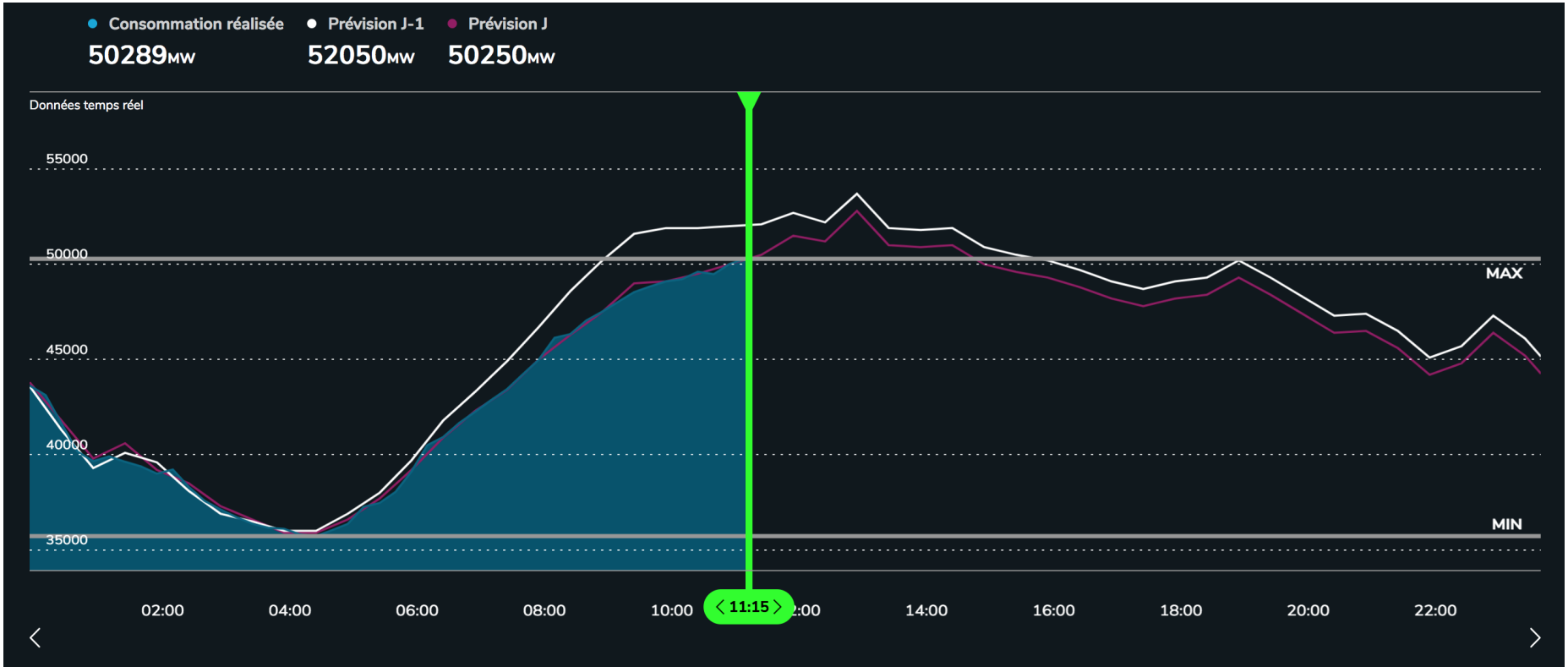


En France avec RTE

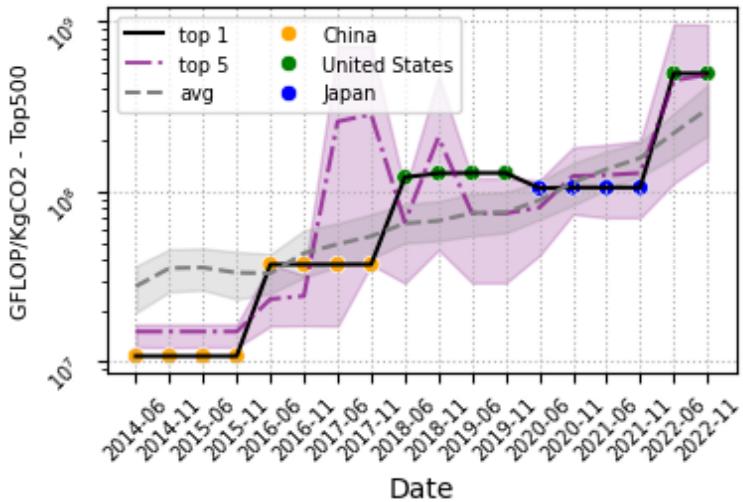
Mixte énergétique

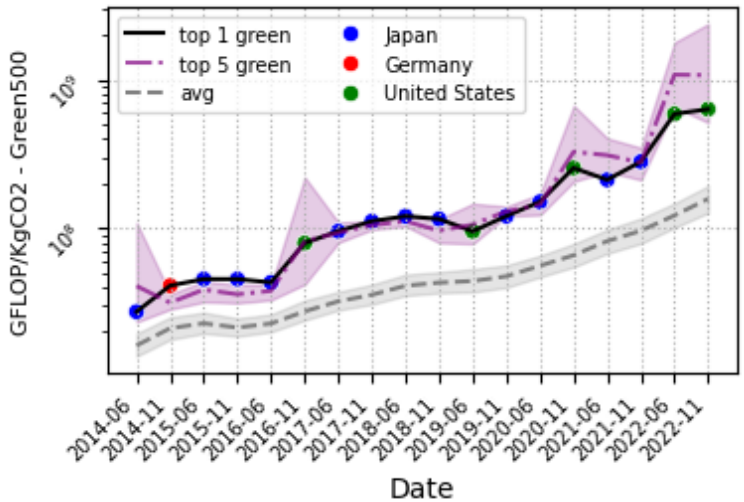


Prévisions


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

Outil prospectif





- Cibler 2030 : est-ce soutenable ?
- Question sous-jacente :
  Que veut dire un HPC au service de l'écologie ?[3]

[3] j'y reviendrai en conclusion...

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
  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*[4].
  Ce qui n'est pas compté dans le périmètre initial.

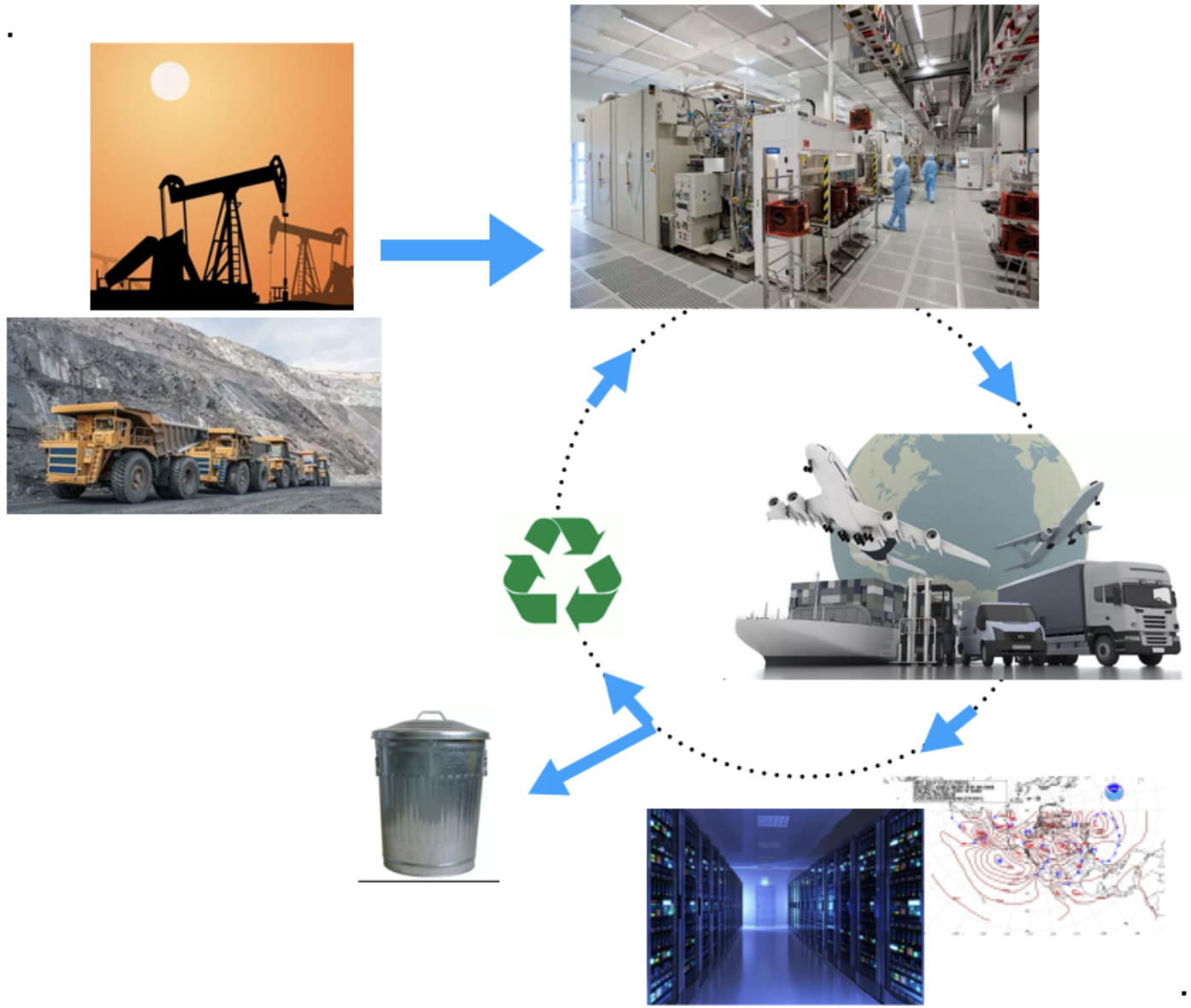
[4] Question à un champion pour le HPC...

Rebond

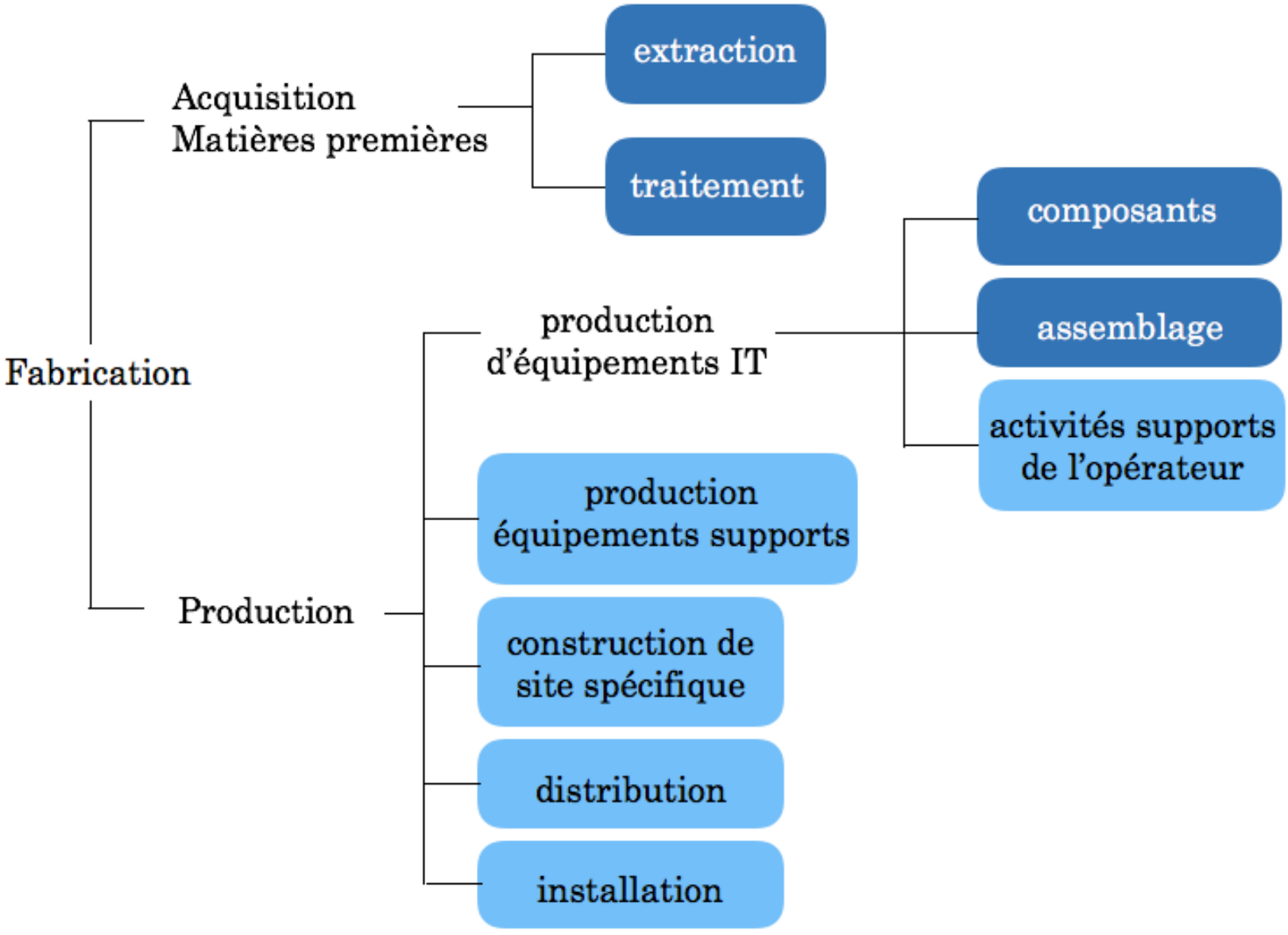
- Direct :
  Une technologie plus efficace augmente les usages.
- On a aussi un effet rebond indirect lorsque des gains réalisés dans un domaine génèrent de la consommation dans un autre.

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

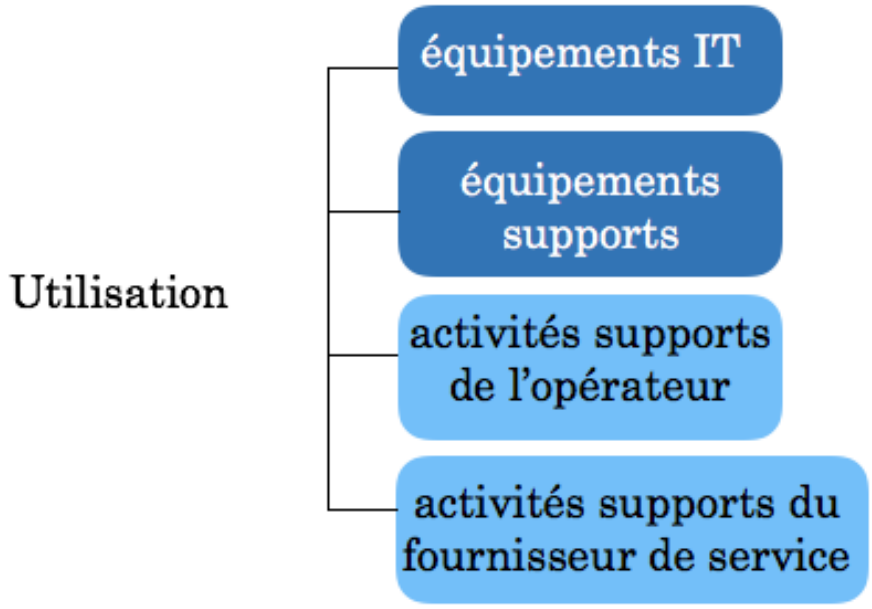
ACV



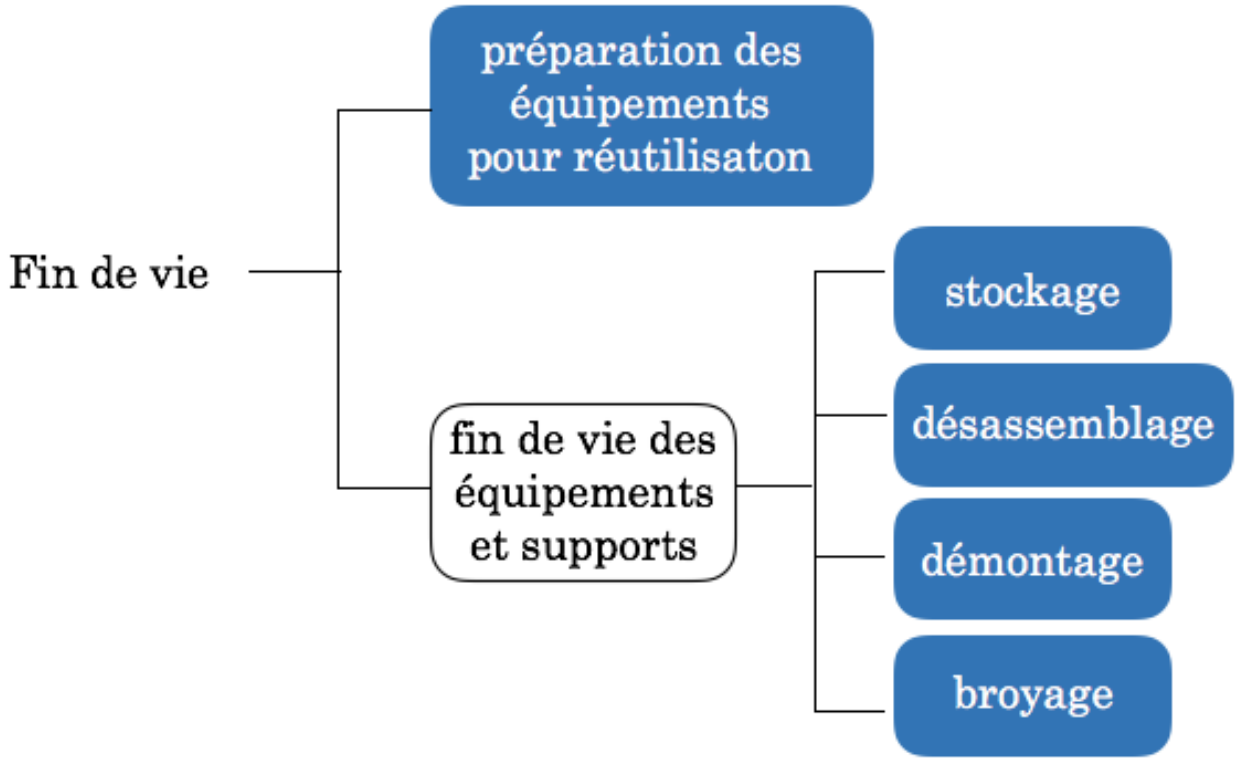
ACV d'un service numérique (fab)



ACV d'un service numérique (usage)



ACV d'un service numérique (fin de vie)



Un premier impératif : évaluer/mesurer

Pourquoi ? Car "sans mesure : Pas de Science"

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

Comment ?

- Quantitatif et qualitatif.
- Il existe pas mal d'outils pour les mesures de programmes et des ACV des matériels.

Un retour d'expérience sur Jean Zay

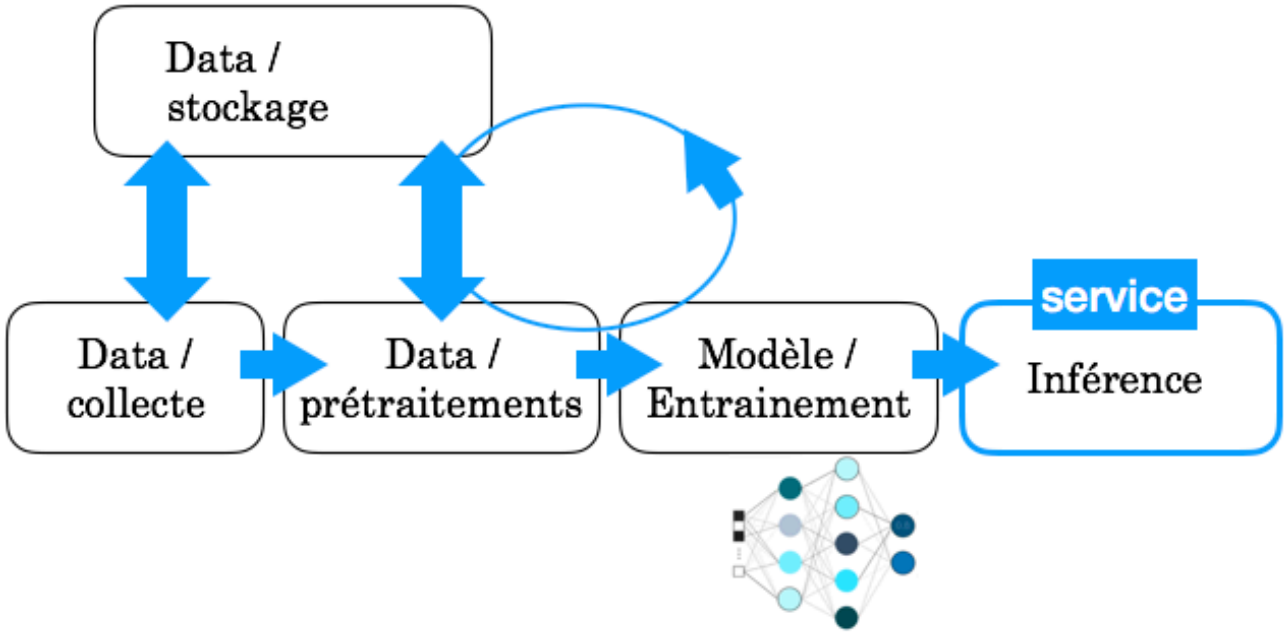
- La méthodologie de calcul prend en compte aussi bien les GES liés à la construction que l'utilisation avec toute son infrastructure technique (refroidissement, électricité).
  - Le GES d'une heure.coeur de calcul sur Jean Zay-CPU est de : 1,39 gCO2e
  - Le GES d'une heure.GPU V100 sur Jean Zay-GPU est de : 41,26 gCO2e

Le facteur d'émission électricité qui a été pris pour la modélisation est : 0,06 kgCO2/kWh (Base Carbone V19, 2020).
Pour la partition CPU, 50% du GES est dû à la fabrication et environ 50% à l'usage.
Pour la partition GPU, on passe à 1/3 pour la fabrication et 2/3 pour l'usage.

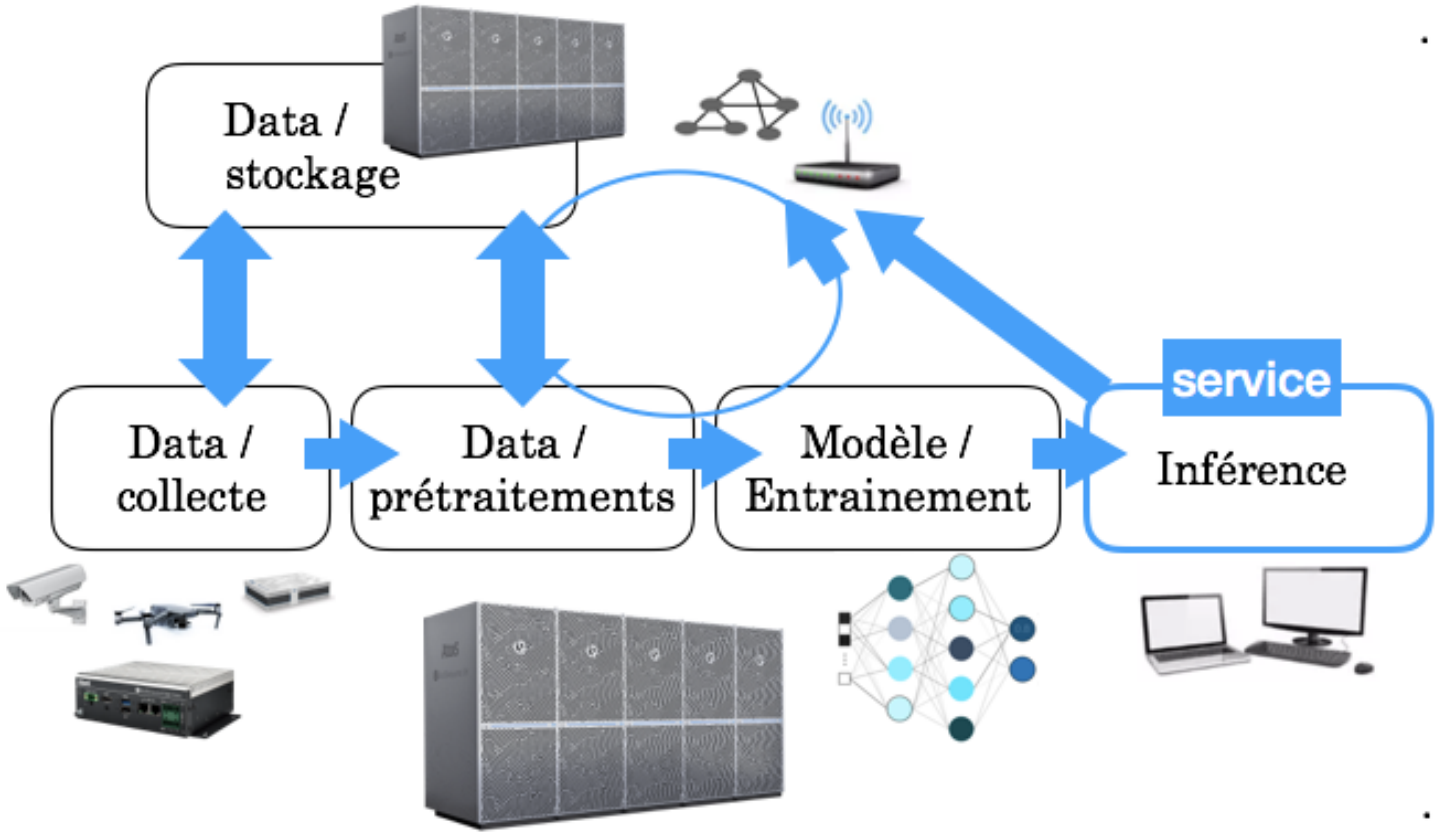
Si je résume

- La crise environnementale est avérée et ses conséquences sont sans précédent.
  La croissance économique est la base de nos sociétés (occidentales) et le Numérique (IA) est un des principaux moteurs de la croissance.
- Il peut jouer un rôle positif dans la crise.
- Mais, c'est aussi un domaine coûteux dans un contexte où il nous faut réduire.
- Comment garantir que le bilan est vraiment positif ?

Focus sur l'IA : ACV d'un service d'IA



Il faut tout compter !

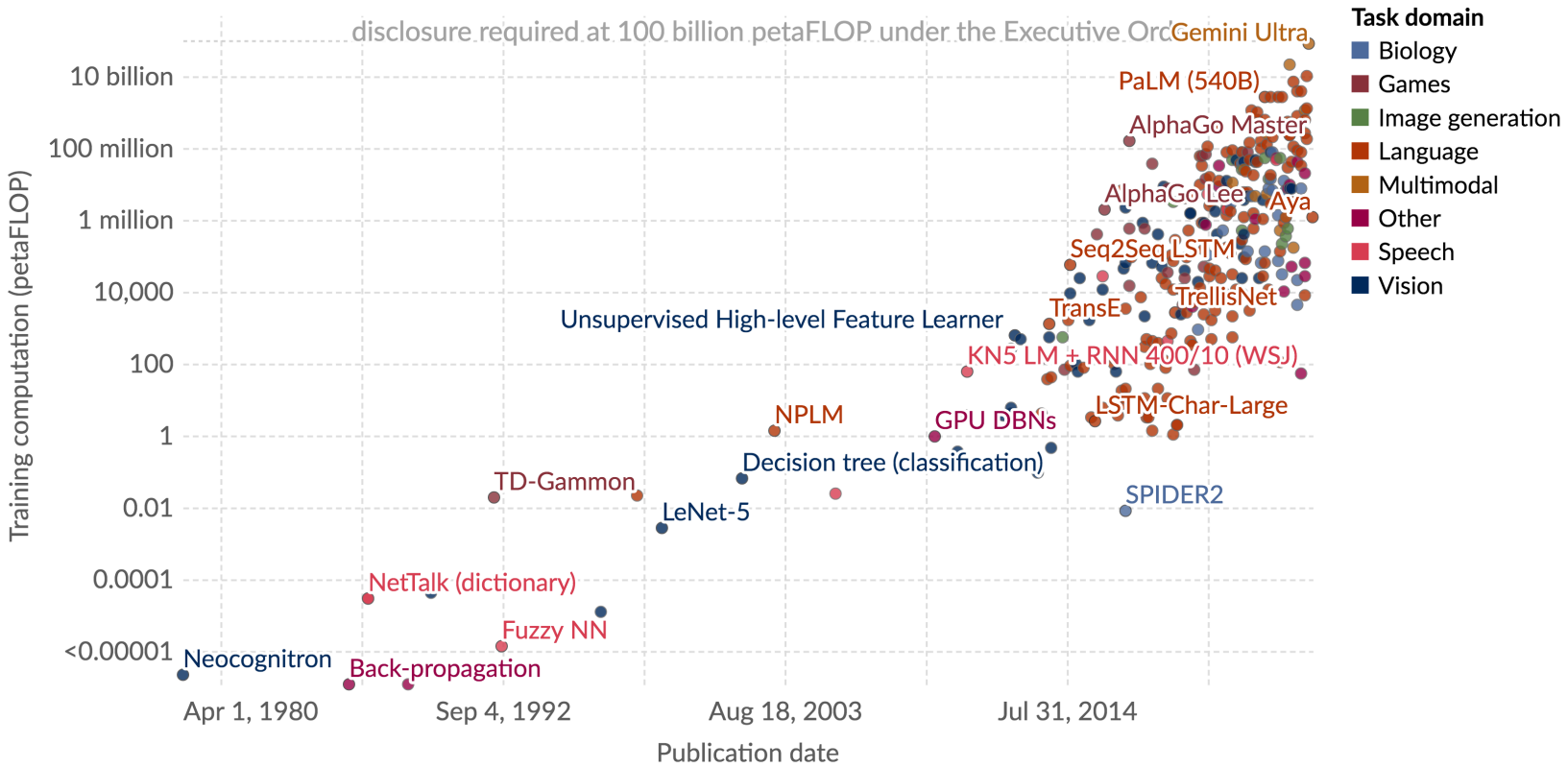


L'effet accélérateur du Big Data et de l'IA

Computation used to train notable artificial intelligence systems

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. Estimates are expected to be accurate within a factor of 2, or a factor of 5 for recent undisclosed models like GPT-4.



Data source: Epoch (2024)

OurWorldInData.org/artificial-intelligence | CC BY

Note: The Executive Order on AI refers to a directive issued by President Biden on October 30, 2023, aimed at establishing guidelines and standards for the responsible development and use of artificial intelligence within the United States.

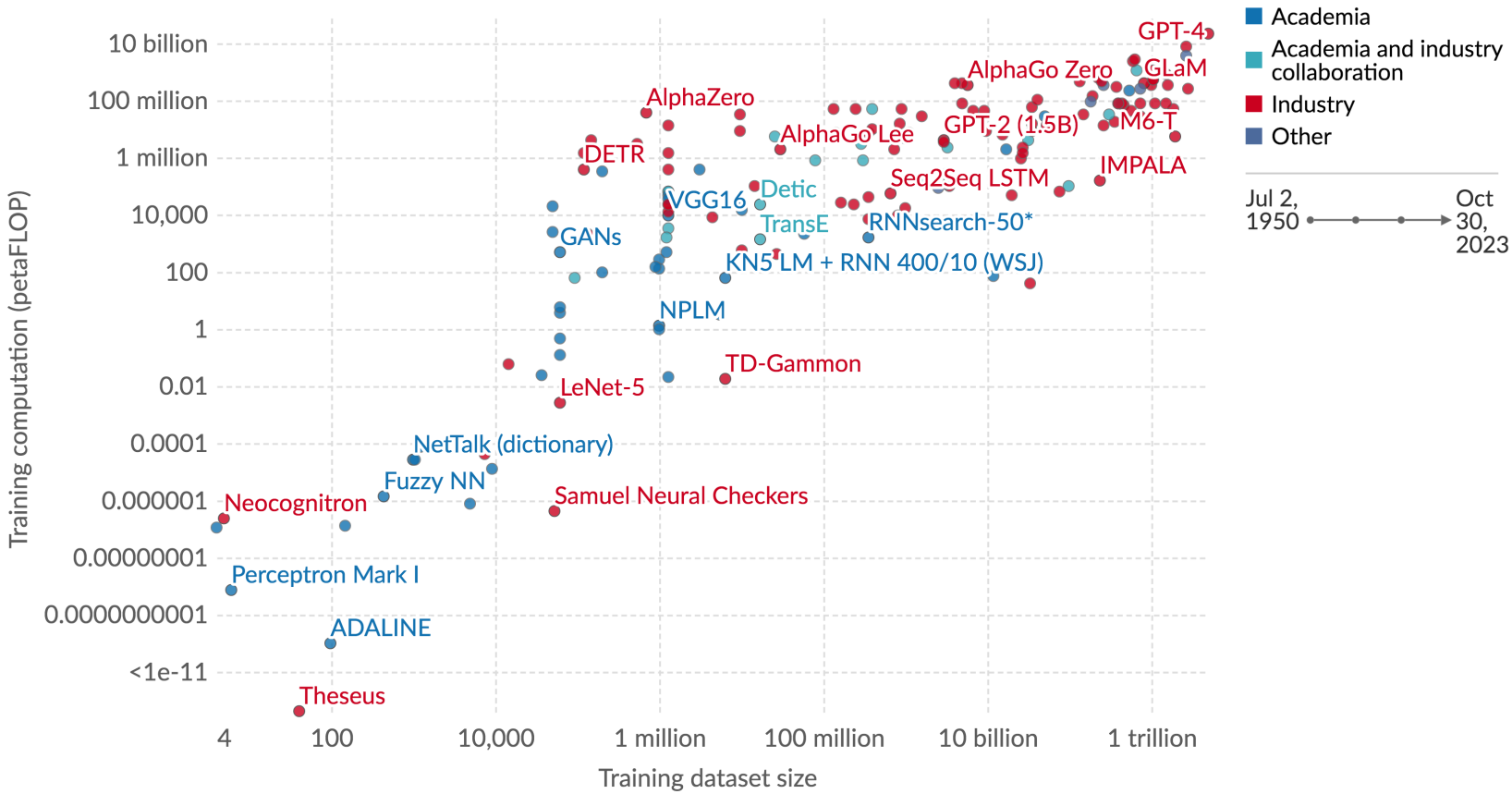
1. Floating-point operation: A floating-point operation (FLOP) is a type of computer operation. One FLOP represents a single arithmetic operation involving floating-point numbers, such as addition, subtraction, multiplication, or division.

L'effet rebond des données

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.

"Efficiency" ou "Sufficiency" ?

- La communauté a pris conscience qu'il faut réagir.
- La voie principale consiste à optimiser les plates-formes et les applications du point de vue énergétique.
  C'est l'éco-efficacité : réduction de l'intensité des impacts environnementaux ou de l'usage de ressources par unité de valeur économique produite.
- On peut aussi passer au renouvelable pour des calculs décarbonés.

"Efficiency" ou "Sufficiency" ?

- La communauté a pris conscience qu'il faut réagir.
- La voie principale consiste à optimiser les plates-formes et les applications du point de vue énergétique.
  C'est l'éco-efficacité : réduction de l'intensité des impacts environnementaux ou de l'usage de ressources par unité de valeur économique produite.
- On peut aussi passer au renouvelable pour des calculs décarbonés.
- L'autre voie est de se poser les questions sur les applications a priori, quitte à renoncer aux calculs.

Politique et Environnement

Le développement du HPC et de l'IA n'est que la continuité des politiques publiques de la numérisation des sociétés.

- Stratégie de Lisbonne : axe de politique économique de l'Union Européenne actée en 2000 pour numériser la société (numérique à l'école, dématérialiser les services publics, faciliter l'accès à un Internet massif et peu cher, etc.).
- Stratégie française : La relance par les start-ups innovantes. L'IA doit diffuser dans tous les secteurs.

Merci pour votre attention