L'IA générative (de l'entraînement à l'utilisation) Quel coût pour l'environnement ?

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

Journée Capsule
Sorbonne Université, 20 juin 2024

Quelques repères

- 2000 – Stratégie de Lisbonne : axe de politique économique de l'UE pour numériser la société
  (numérique à l'école, dématérialiser les services publics, faciliter l'accès à un Internet massif et peu cher, etc.).
- 2013 – La révolution GPUs et l'apprentissage profond
- 2022 – chatGPT et l'IA générative
- Stratégie française : La relance H2030 par des start-ups innovantes.
  L'IA doit diffuser dans tous les secteurs.

Quelques repères

- 2000 – Stratégie de Lisbonne : axe de politique économique de l'UE pour numériser la société
  (numérique à l'école, dématérialiser les services publics, faciliter l'accès à un Internet massif et peu cher, etc.).
- 2013 – La révolution GPUs et l'apprentissage profond
- 2022 – chatGPT et l'IA générative
- Stratégie française : La relance H2030 par des start-ups innovantes.
  L'IA doit diffuser dans tous les secteurs.

On peut en tirer deux enseignements :

l'IA est partout et en croissance exponentielle

Développement à rythme très rapide, inédit.

Objectif(s)

Constat

L'humanité fait face à des défis environnementaux sans précédents.

On ne dispose ni d'une énergie facile à utiliser et bon marché, ni de ressources illimitées permettant au système technologique complexe actuel de perdurer.

Objectif(s)

Constat

L'humanité fait face à des défis environnementaux sans précédents.

On ne dispose ni d'une énergie facile à utiliser et bon marché, ni de ressources illimitées permettant au système technologique complexe actuel de perdurer.

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

- Quelle place peut/doit prendre la communauté ESR sur la question de l'IA générative ?

Points abordés dans l'exposé

- Opportunité ou désastre ? Bien entendu, c'est un peu des deux...
  Il s'agit dans cet exposé de partager quelques faits et reflexions sur le volet environnemental

Points abordés dans l'exposé

- Opportunité ou désastre ? Bien entendu, c'est un peu des deux...
  Il s'agit dans cet exposé de partager quelques faits et reflexions sur le volet environnemental

- Réduire les impacts négatifs du domaine ? Est-ce suffisant et pour les technoptimistes : comment ?
- Réduire la voilure et /ou changer de paradigme ?

Des avantages certains

Dans le meilleur des mondes, l'IA permet (entre autre) de :

- développer un apprentissage personnalisé
- réduire les inégalités (Nord/Sud)
- diminuer la charge des enseignants (mais attention ici à l'acceptance/acceptabilité –des collègues– et à l'éthique)
- ...

Agenda de cette présentation

- Une brève introduction sur la dynamique *émissions Carbone / IA*.
- Impacts du numérique et de l'IA, de quoi parle-t-on au juste ?
- Comprendre la dynamique d'emballement.
- Une brique de base : Mesurer les impacts.
  Analyse de Cycle de Vie / effets indirects et rebonds
- Comment envisager une IA générative compatible avec les limites planétaires ?

Préliminaires sur les émissions Carbone

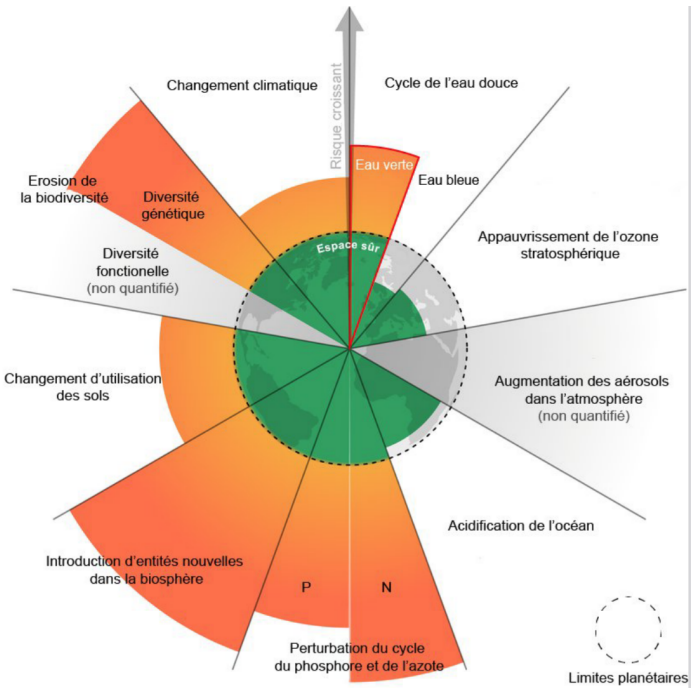
Pour rappeler quelques notions importantes

- Le CO_2 est un des 7 Gaz à effet de serre identifiés au protocole de Kyoto.
- 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
- On a un budget maximal restant pour maintenir le réchauffement en dessous de 1.5 degrés en 2100

Attention, il n'y a pas que les émissions Carbone : stress hydrique et épuisement des ressources abiotiques (métaux)

Attention, il n'y a pas que le climat !

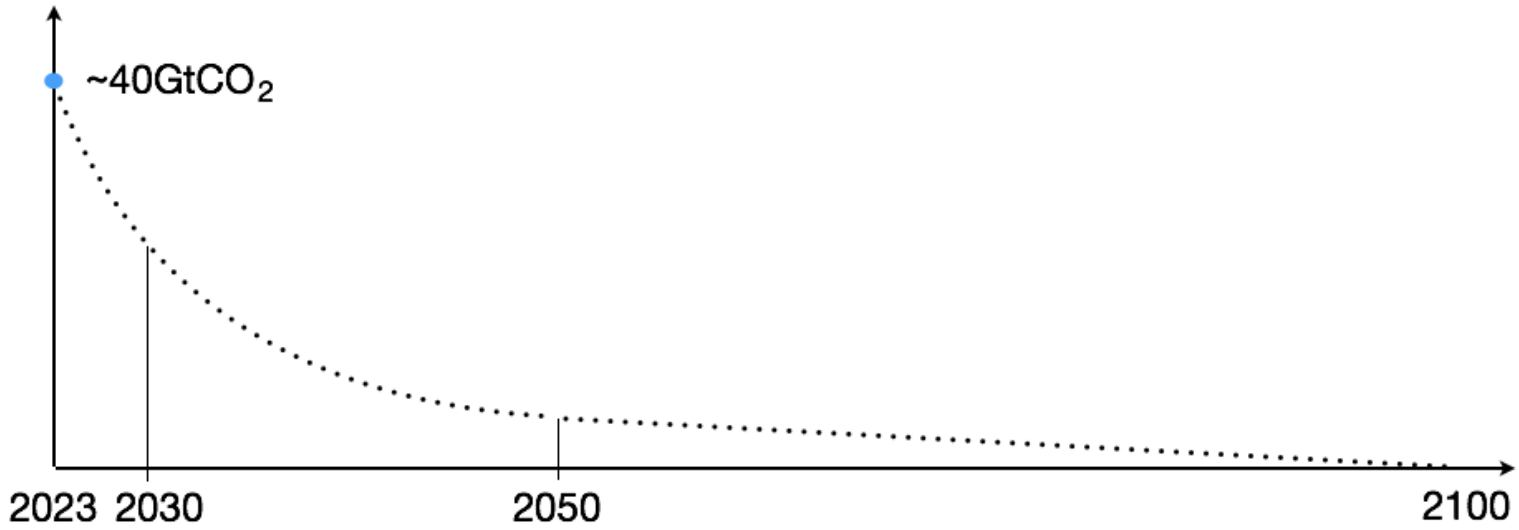
Violation des **limites planétaires** qui engendrent des grandes **menaces** civilisationnelles :



- 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

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...

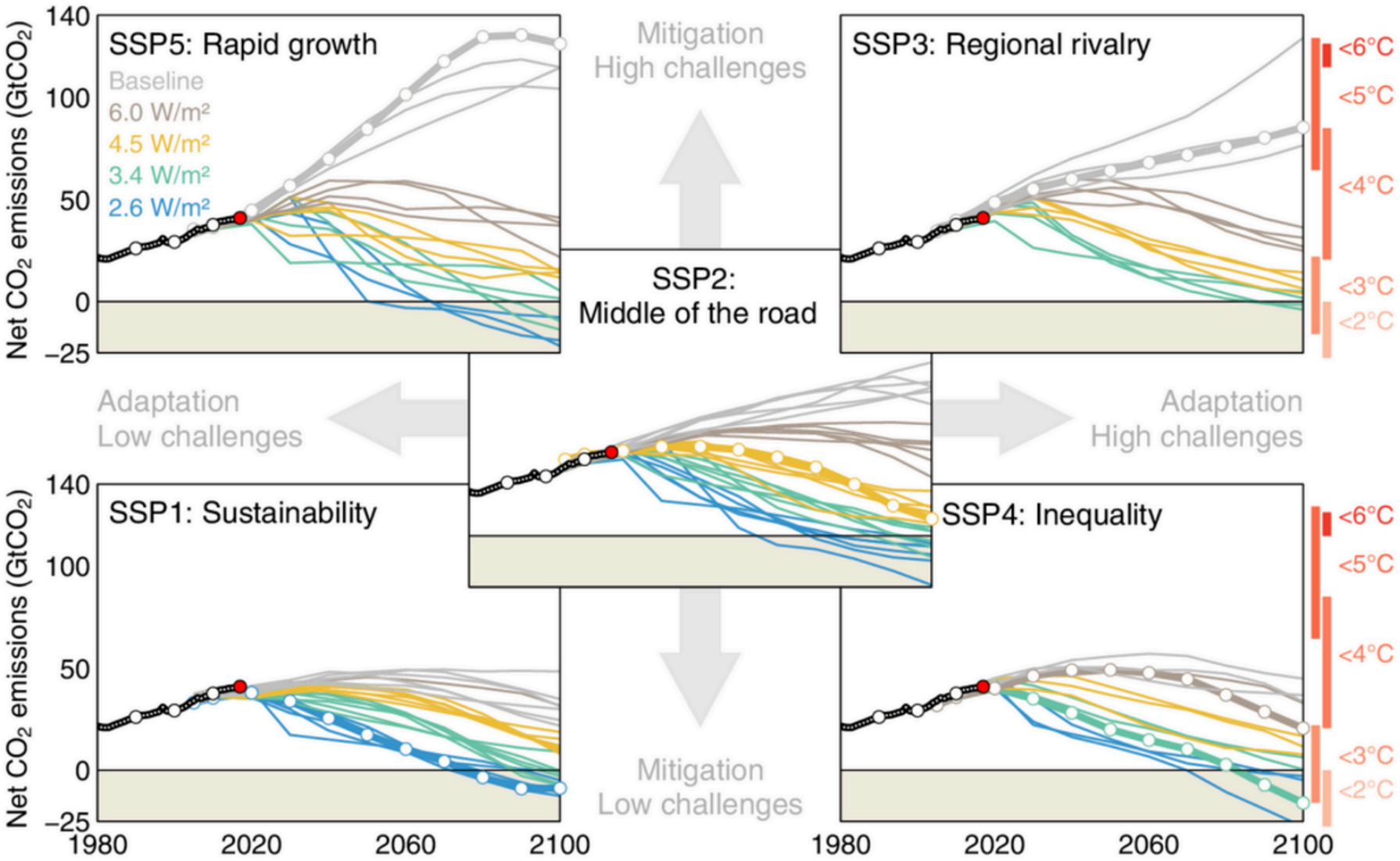
Quelles trajectoires pour y arriver ?

GIEC – scénario de référence SSP x-y (Shared Socio-economic Pathways)

- 5 classes de scénarios (x)
- y : forçage radiatif à la fin du siècle (en W/m^2)

Rapport ADEME

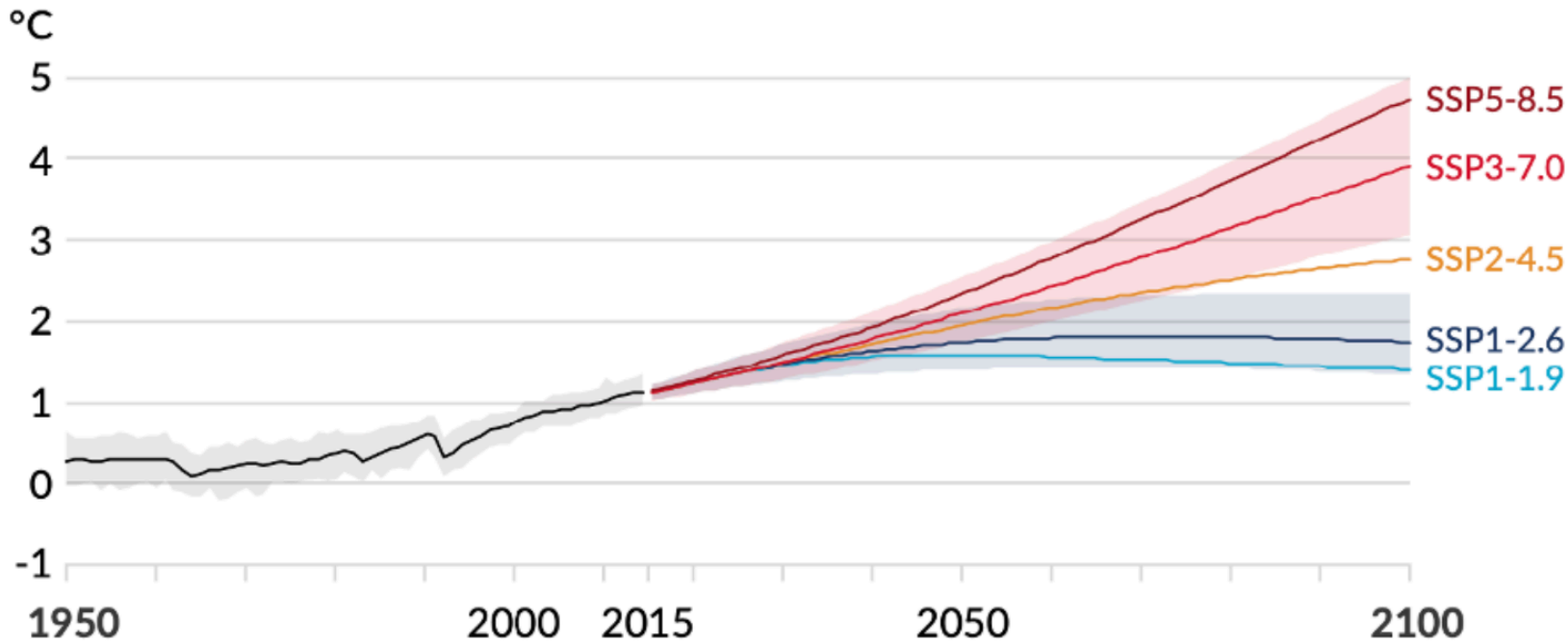
Une vue rapide des 5 classes de scénarios



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



Impact du numérique

Selon le baromètre du numérique, en France, c'est 10% de l'électricité consommée.

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

Impact du numérique

Selon le baromètre du numérique, en France, c'est 10% de l'électricité consommée.

Contribution aux émissions de CO_2

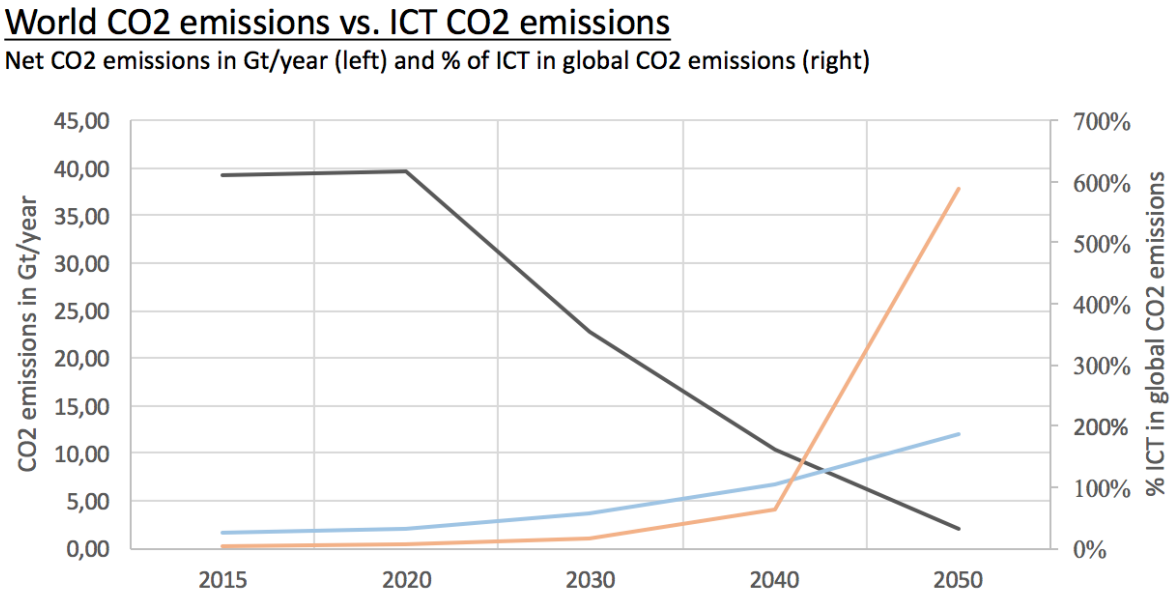
- le numérique représente de l'ordre de 5 à 6% de l'énergie primaire mondiale, soit en gros 4% des émissions mondiales [Lean ICT 19].
  Freitag et al. estiment entre 2.1 et 3.9 d'émissions carbone[1].
- Croissance annuelle de 6-9% (sur 2015-2019).
- Il est très difficile de quantifier la part de l'IA (effet accélérateur).

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

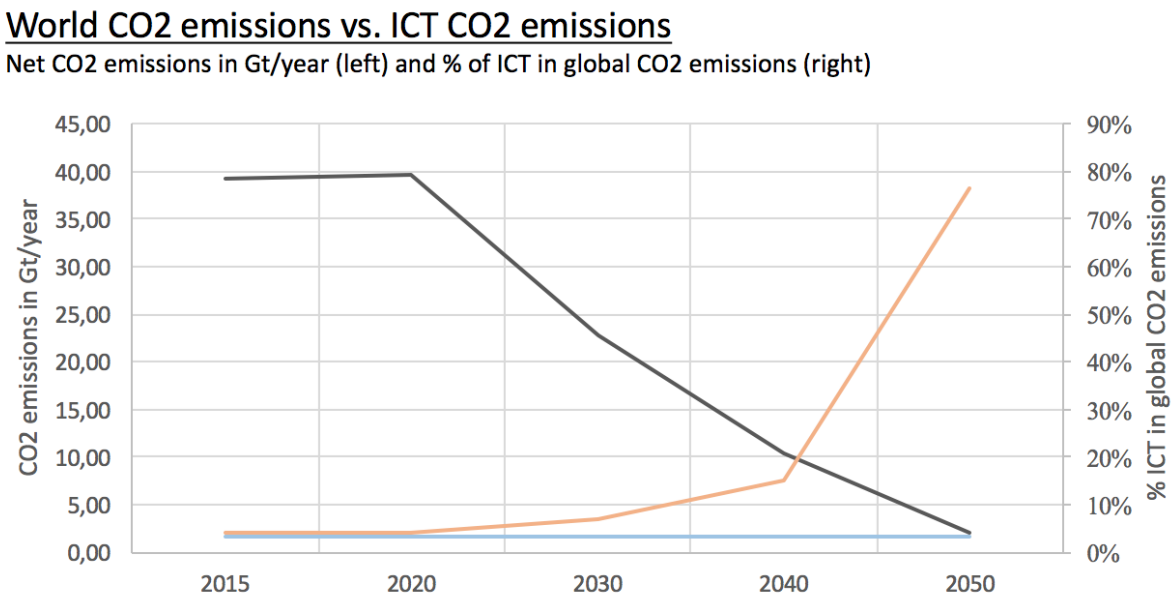
Plus précisément...

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



Une réalité complexe

L'exemple précédent était un exercice mental...

- En pratique, la décroissance tient compte en parti d'avancées technologiques.
- L'électricté se décaborne significativement.

  Consommation mondiale en 2022 : 68 200 TeraWh.
  En 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[2].

[2]Selon l'enquête récente de l'AIE

Un 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.

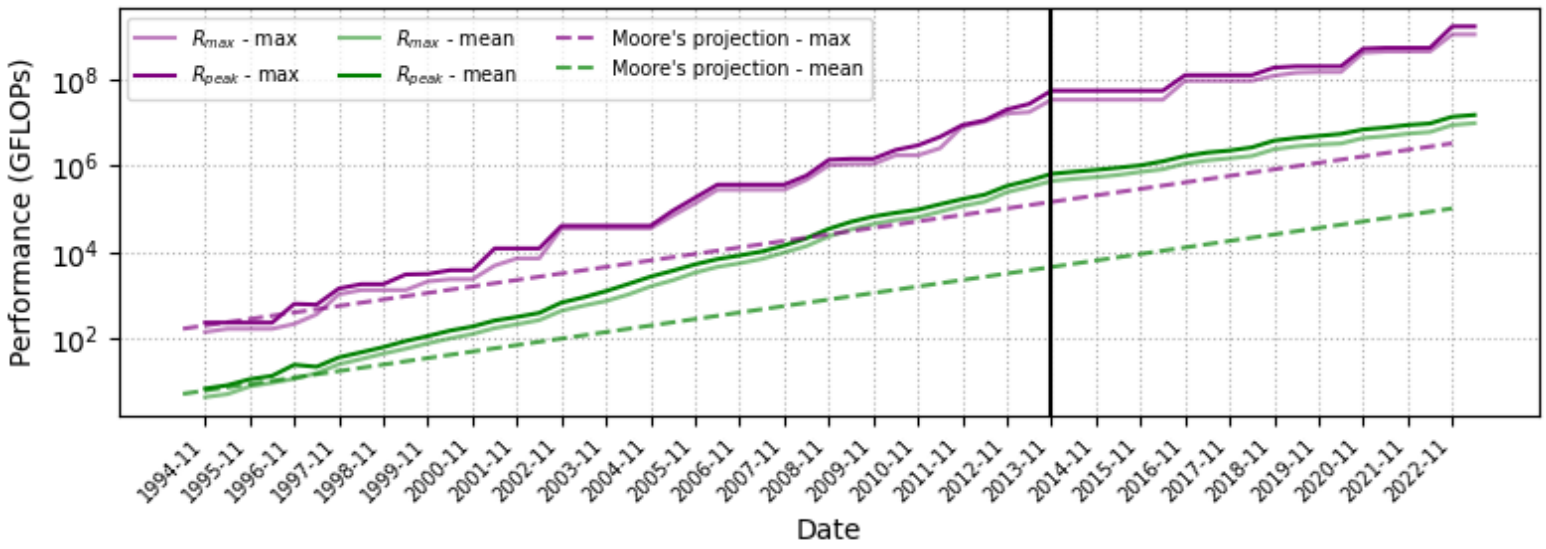
Les lois empiriques

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.

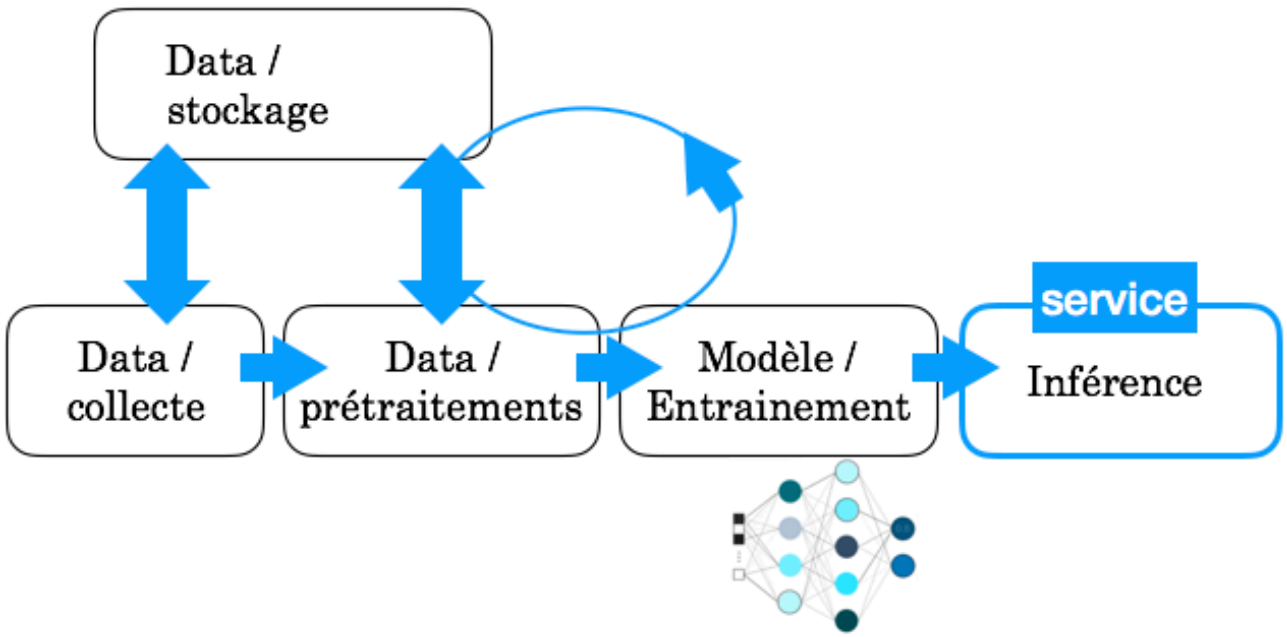
Performance des grandes plates-formes (TOP500)



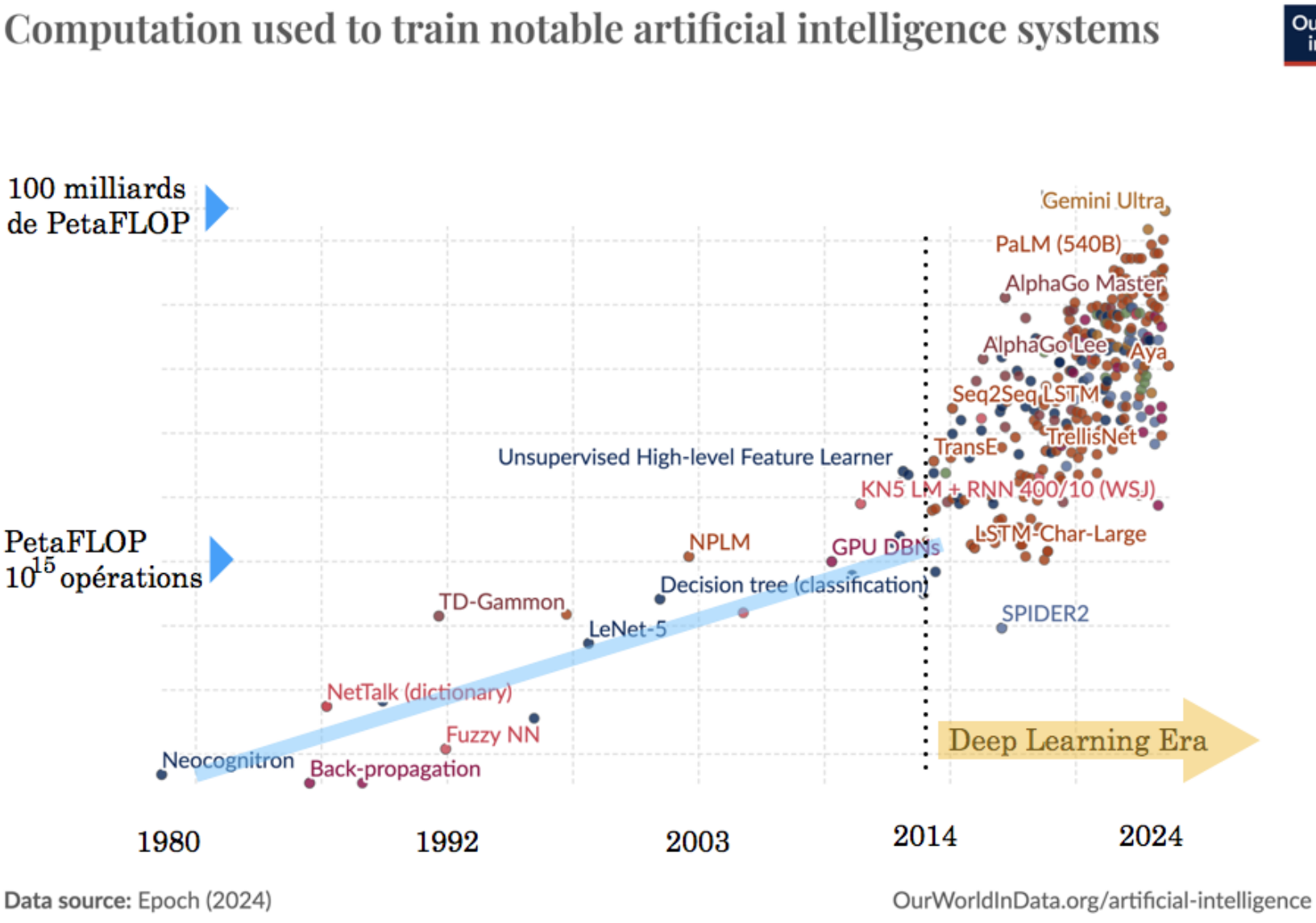
- Clairement une rupture autour de 2013-2014

On peut mettre cette courbe au regard du développement des grands modèles d'apprentissage...

Focus sur l'IA : cycle de vie d'un service d'IA



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

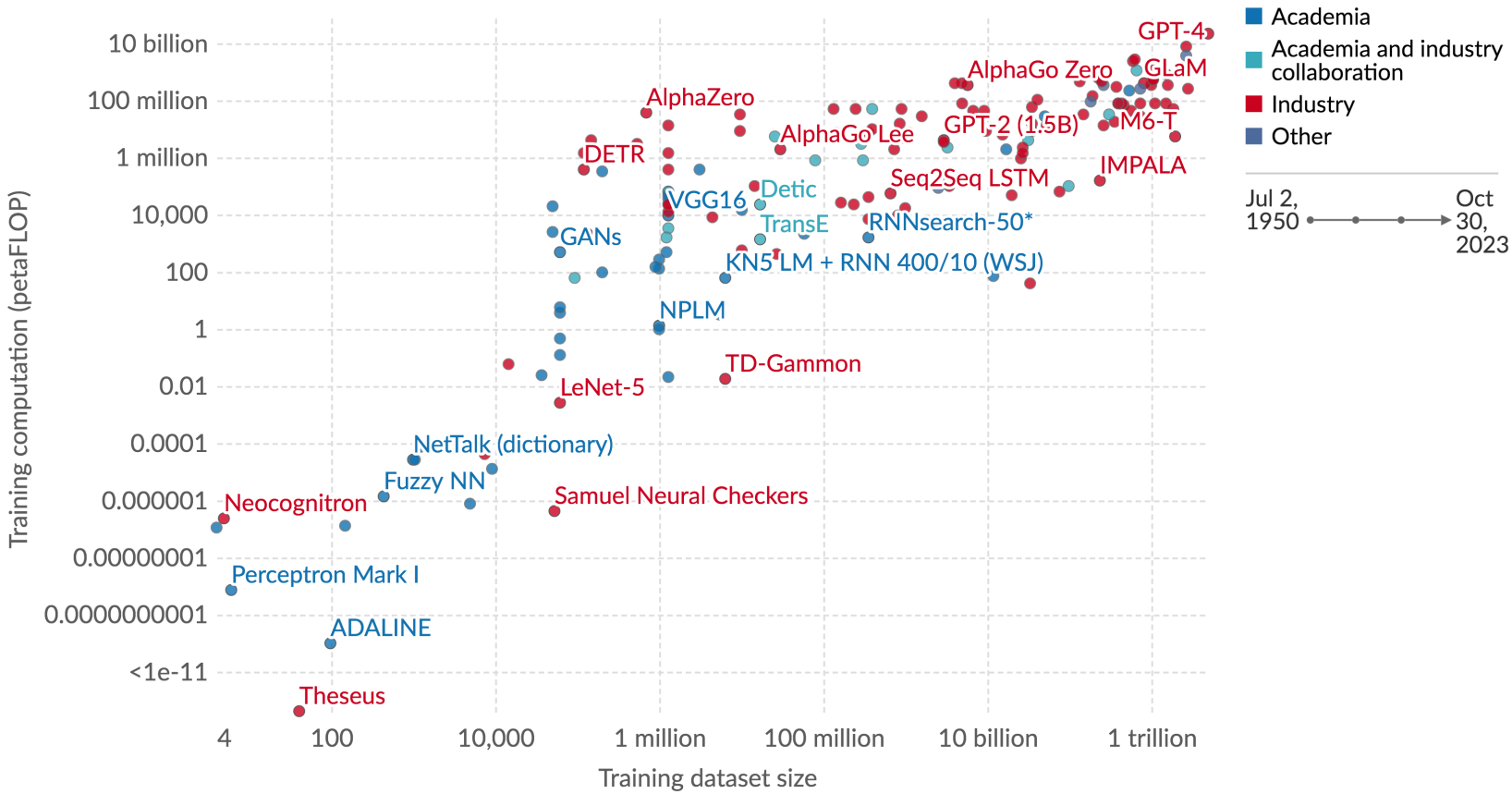


L'effet rebond des données

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.

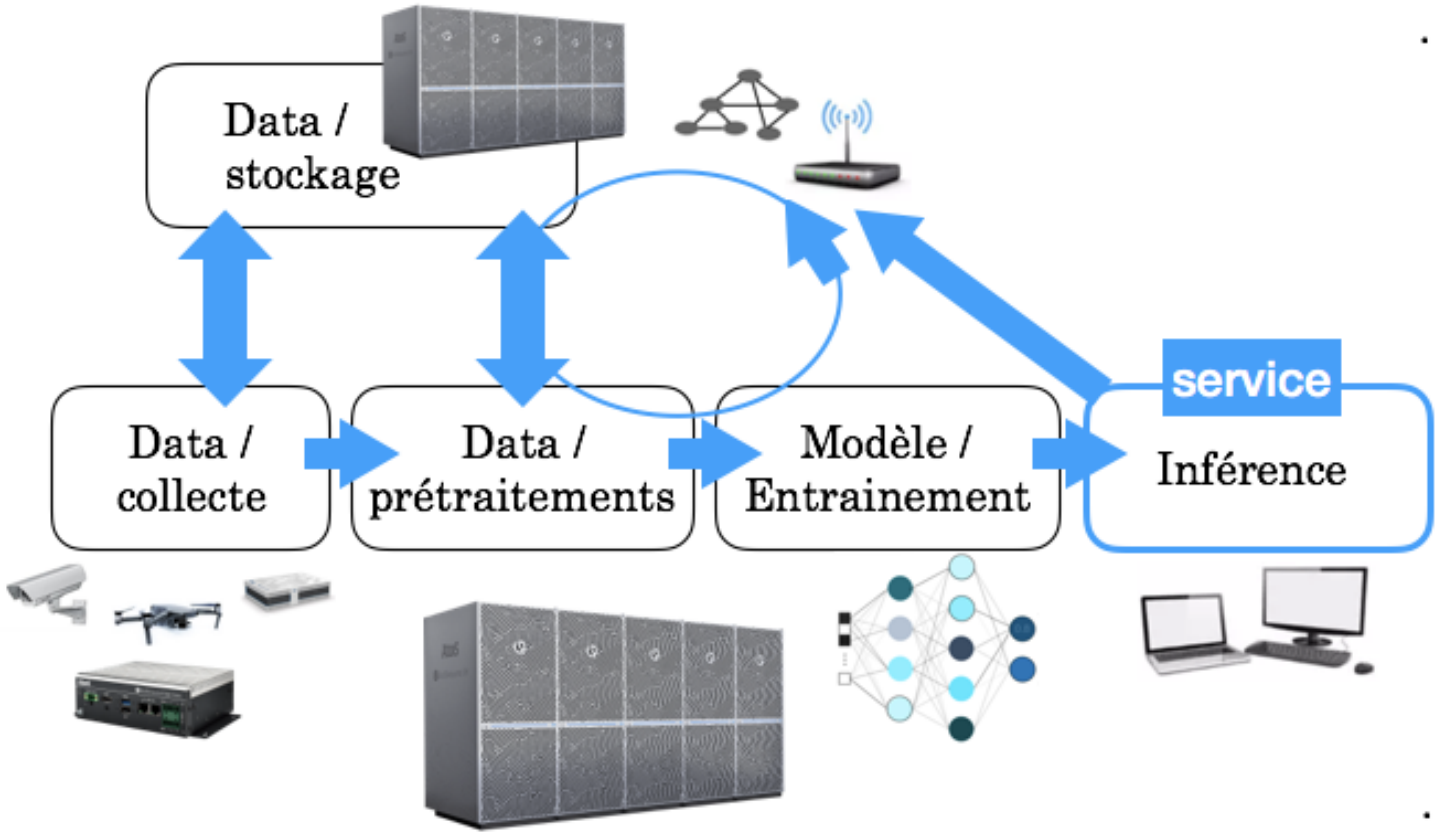


Data source: Epoch (2024)

OurWorldInData.org/artificial-intelligence | CC BY

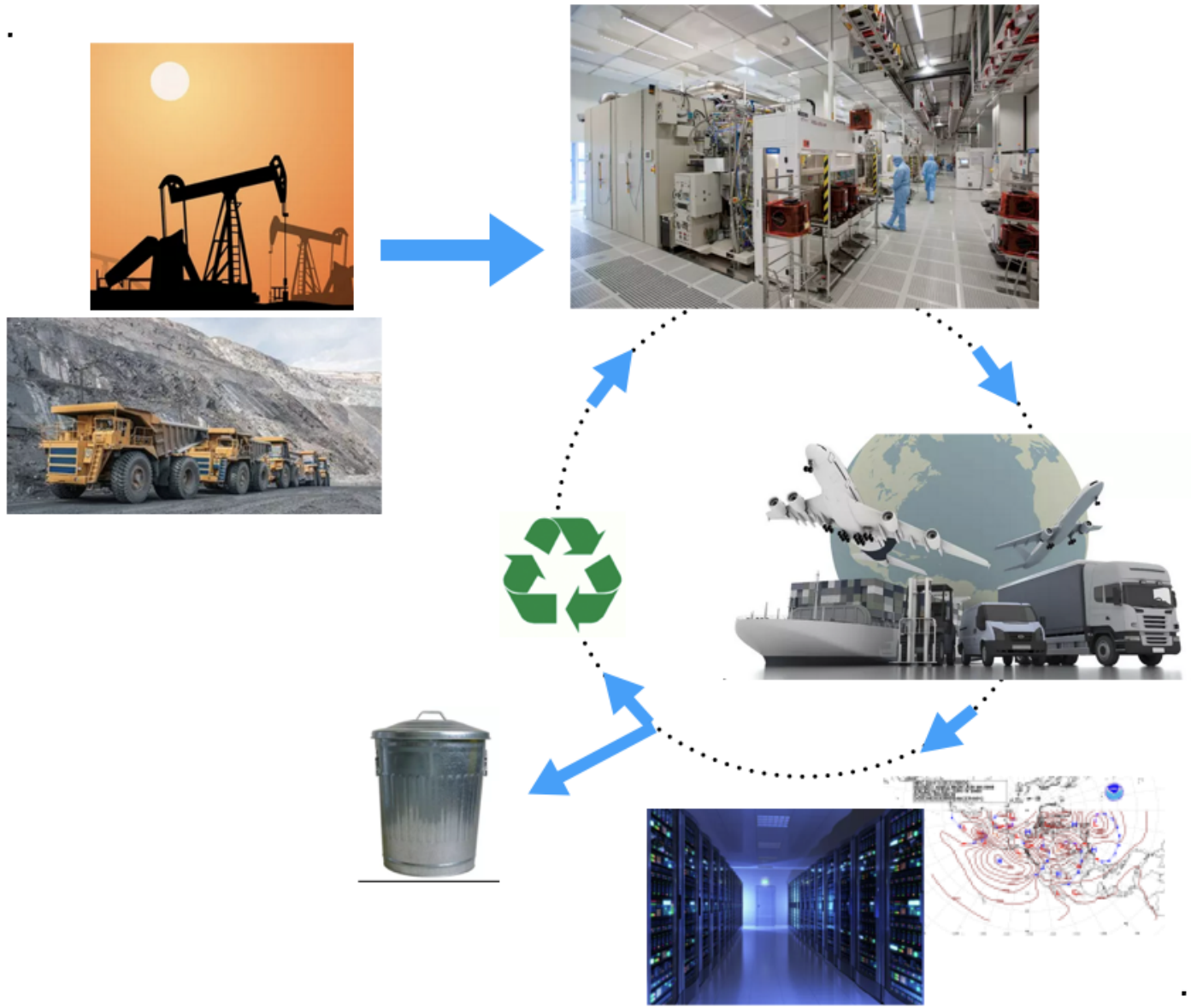
1. Floating-point operation: A floating-point operation (FLOP) is a type of computer operation. One FLOP represents a single arithmetic operation involving floating-point numbers, such as addition, subtraction, multiplication, or division.

Il faut tout compter !



- L'IA générative se mesure par la fabrication des équipements, la phase d'usage (l'électricité) et la fin de vie.

ACV "produit"



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" 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.

Merci pour votre attention