La face cachée de l'intelligence artificielle

consommation énergétique des algorithmes de IAg

Rémi Emonet – 2025-11-10 – UEOS-TEDS

UJM FST / Lab. Hubert Curien

cnrs
Inria
Laboratoire Hubert Curien
UMR ▪ CNRS ▪ 5516 ▪ Saint-Étienne
Université Jean Monnet Saint-Étienne

1. Impact environnemental du numérique ?
   … et autres activités
2. Qu'est-ce que l'intelligence artificielle (IA) ?
   Faire appendre une machine à partir d'exemples
3. Coût énergétique de l'IA générative ?
   Entraînement et utilisation
4. Réduire l'impact environnemental de l'IA générative ?
   Pistes et défis

Attention : la mesure précise est complexe

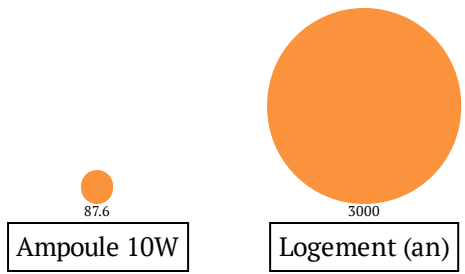
- les chiffres donnés sont des ordres de grandeur
- les sources varient sur leur date, leur méthodologie, leur précision
- les technologies évoluent rapidement

Impact environnemental du numérique ?

… et autres activités

Quelques Ordres de Grandeur (Énergie)

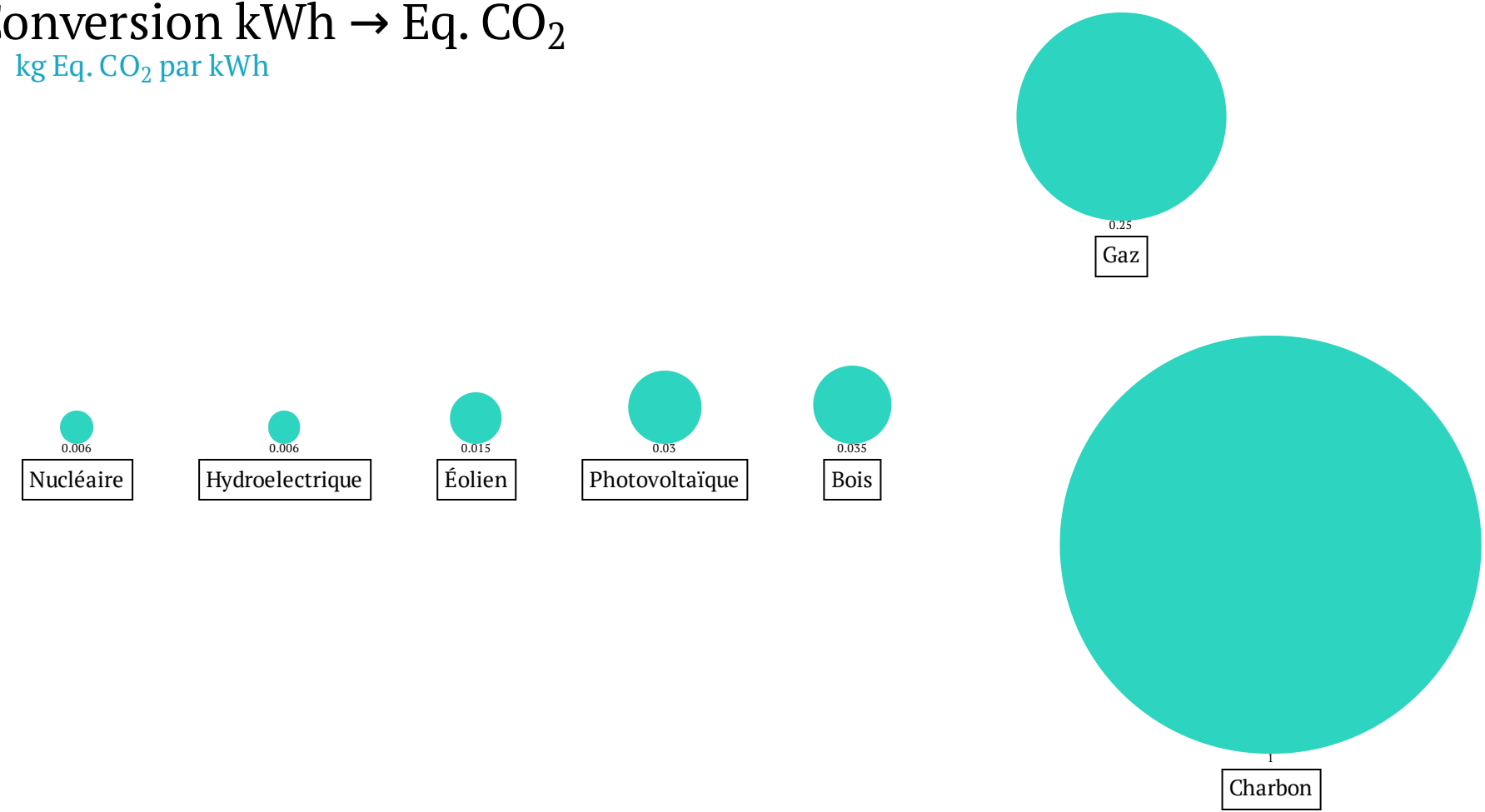
annuel, par personne, en kWh, en France



Production / consommation

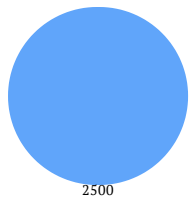
Conversion kWh → Eq. CO_2

kg Eq. CO_2 par kWh



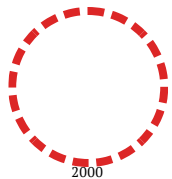
Quelques Ordres de Grandeur (Eq. CO_2)

annuel, en kg de CO_2, par personne, en France



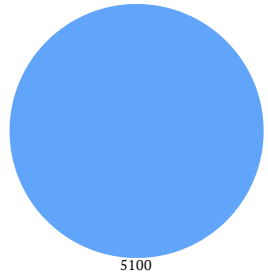
Logement (an)

Ampoule 10W



Objectif

Alimentation



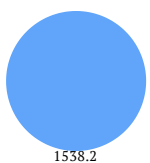
Alim. bœuf

Stockage cloud 50GB

Alim. poulet

Alim. végétarienne

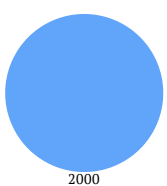
Nucléaire



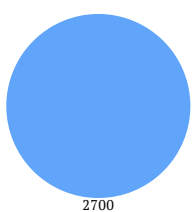
Carburant

3h de four par semaine

Transports



1 Vol Paris-New York (12k km)

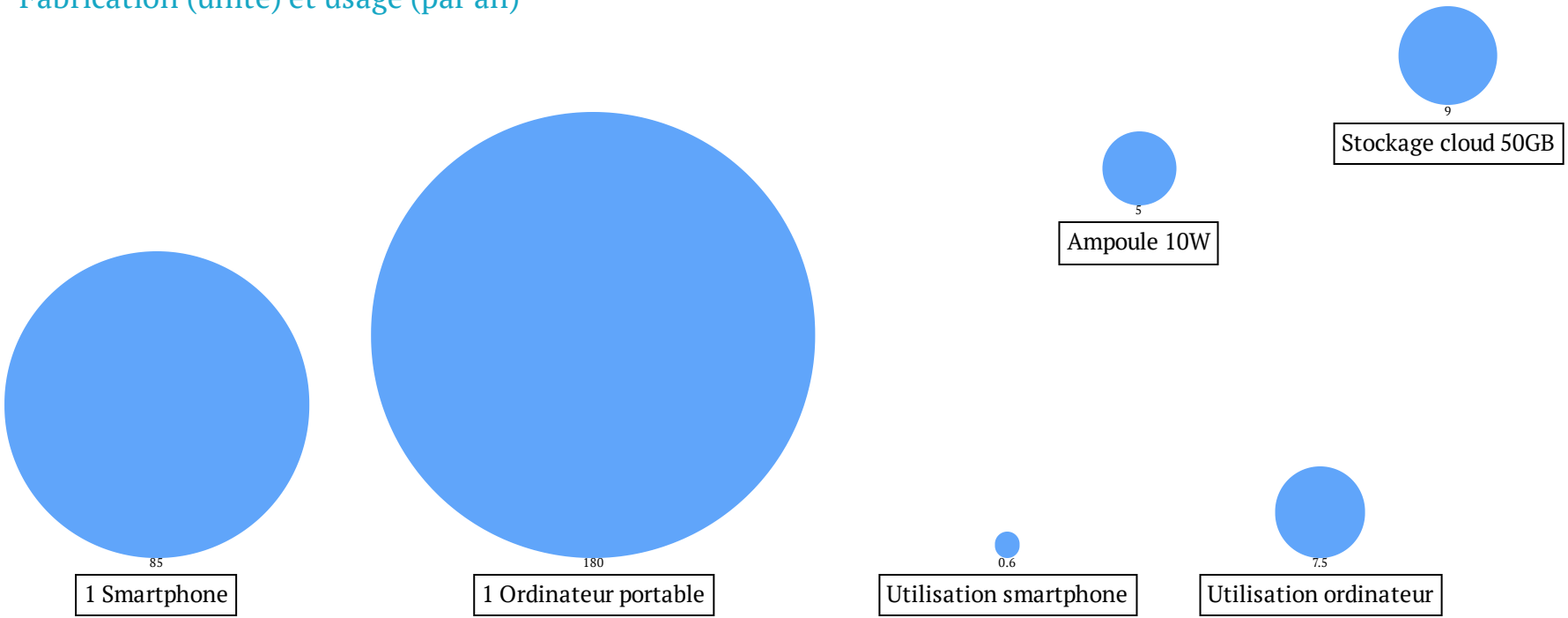


12k km voiture

12k km train

Impact du matériel numérique ?

Fabrication (unité) et usage (par an)



Ère du zettaoctet, 1E = 1,000,000,000,**000,000,000,000** (10^{21}) octets

⇒ importance du débit, dominé par le streaming vidéo

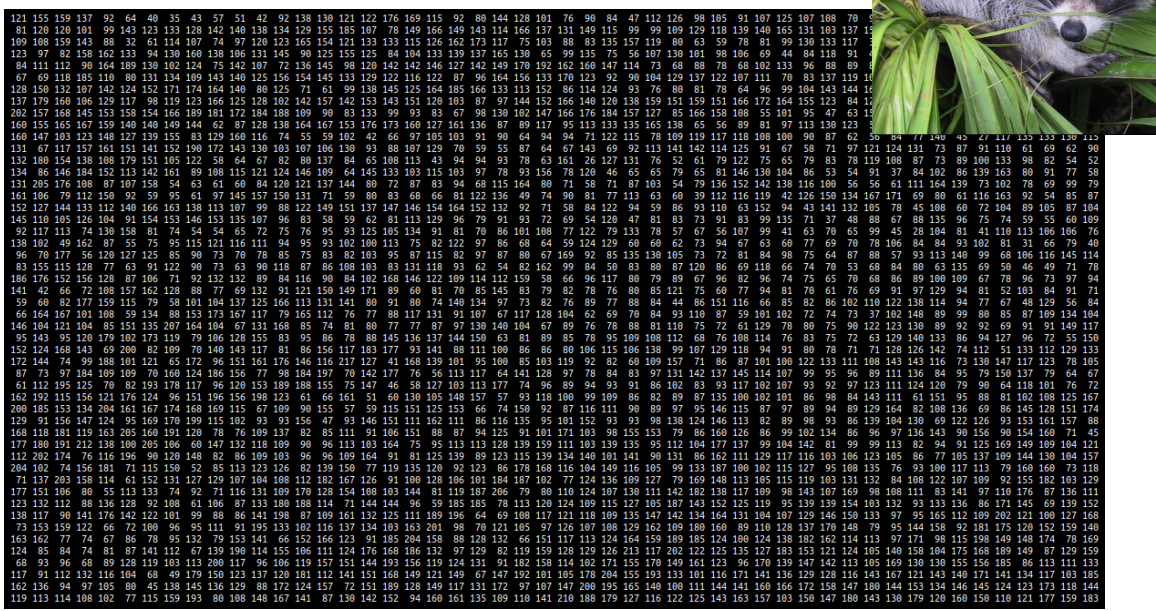
1. Impact environnemental du numérique ?
   … et autres activités
2. Qu'est-ce que l'intelligence artificielle (IA) ?
   Faire appendre une machine à partir d'exemples
3. Coût énergétique de l'IA générative ?
   Entraînement et utilisation
4. Réduire l'impact environnemental de l'IA générative ?
   Pistes et défis

Qu’est-ce que l’intelligence artificielle (IA) ?

Faire appendre une machine à partir d’exemples

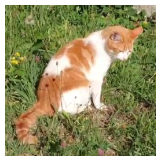
Limites de la programmation classique

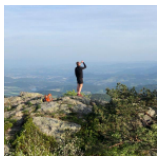
- complexité des programmes
- tâche impossible à décrire
- des « if » pour analyser une image ?

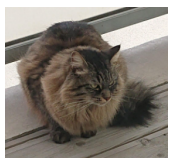


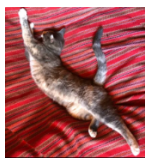
- ⇒ nécessité de programmer par l'exemple

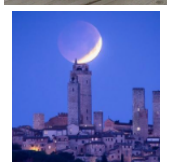
Création d'un jeu de données (pour la tâche Chat/NonChat)

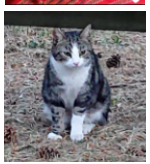

, Chat

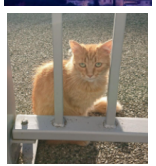

, NonChat


, Chat

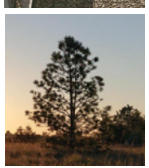

, Chat

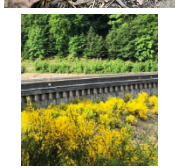

, NonChat


, Chat

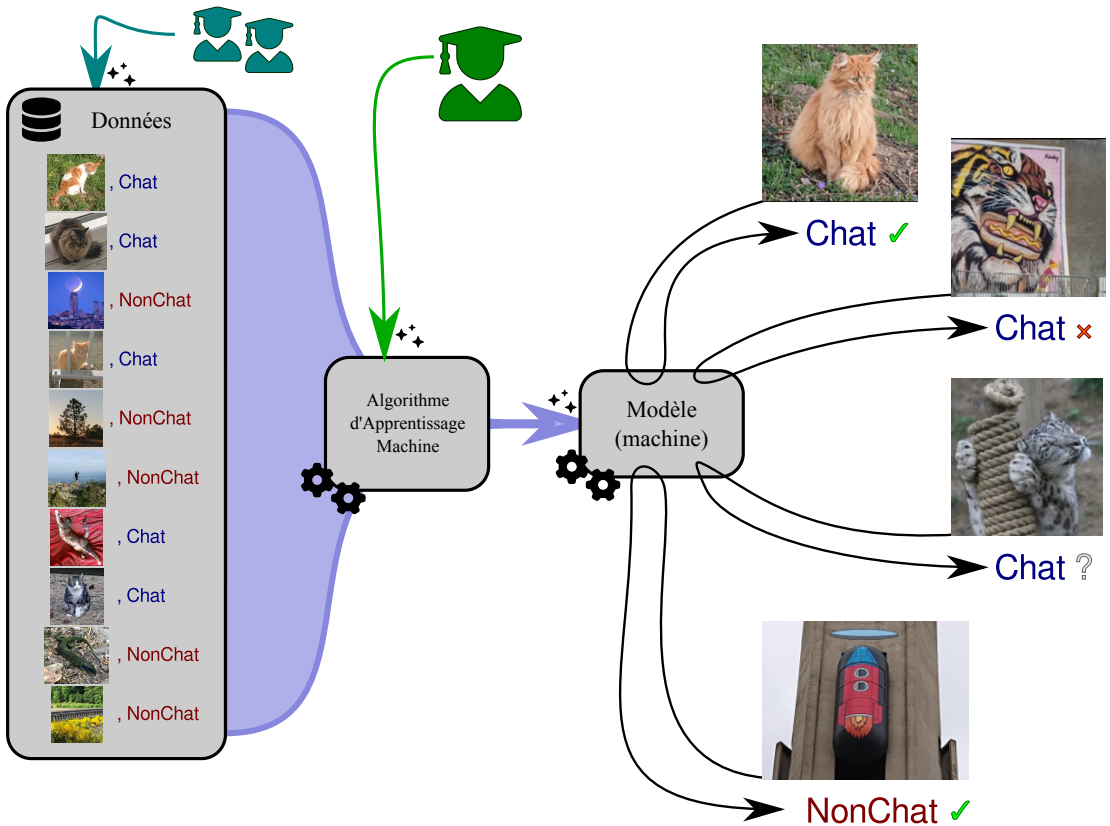

, Chat


, NonChat

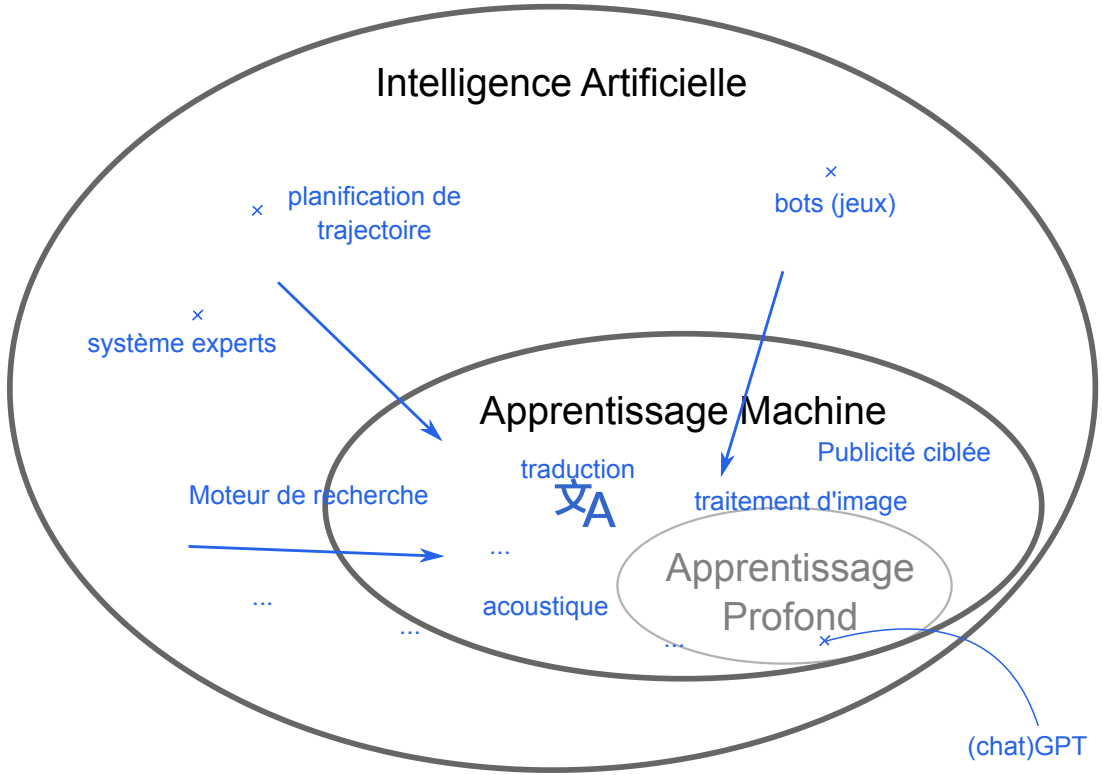

, NonChat


, NonChat

Apprentissage automatique : principe global

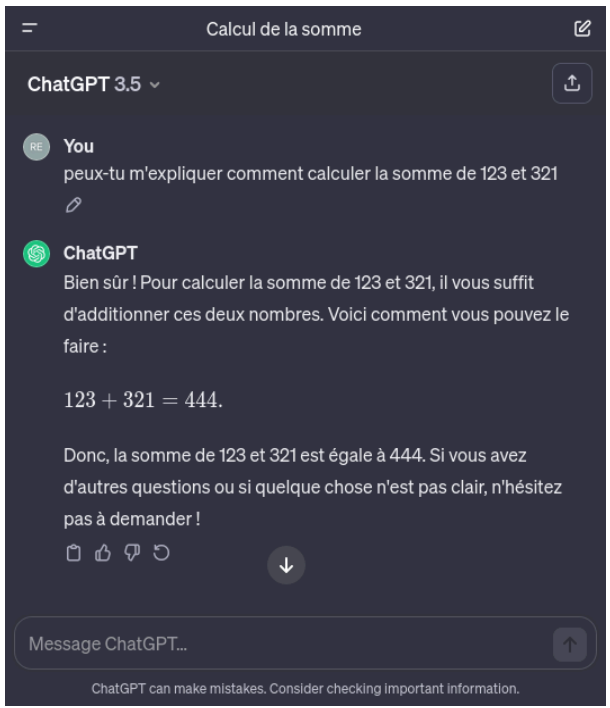


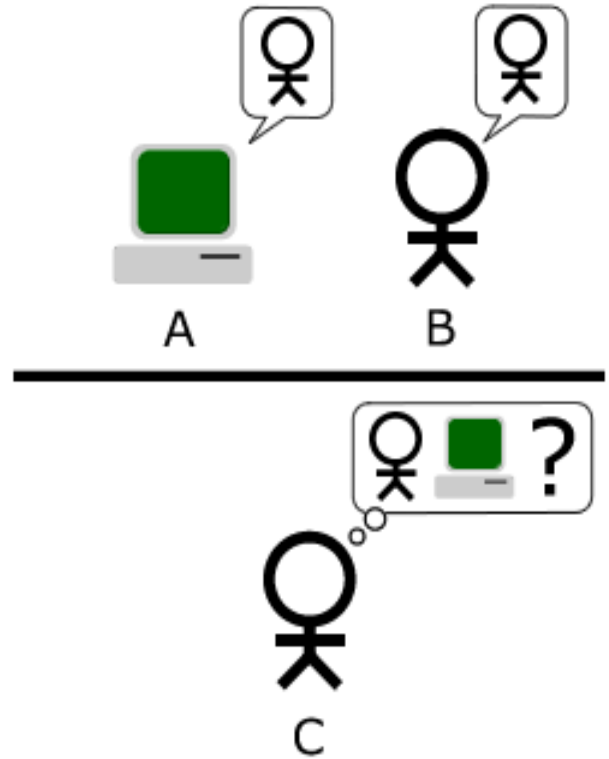
Terminologie



Intelligence artificielle ? Le test de Turing

- interaction textuelles entre humain et machine
- l'humain arrive-t-il à savoir qu'il interagit avec une machine ?





Intelligence ? Zoom sur chatGPT ?

- Modèle de language : apprendre à prédire le mot maquant

> Saint-Étienne appelée « Sainté » en langage familier, est une commune française située au sud-ouest de Lyon (60 km environ) et le quart-sud-est de la France, en région Auvergne-Rhône-Alpes. C'est le chef-lieu du département de la Loire. Avec 174 082 habitants en 20203, elle est la 13e commune la plus peuplée de France (2016) et la 2e commune d'Auvergne-Rhône-Alpes. Saint-Étienne Métropole constitue par sa population (400 813 habitants en 2020) la 3e métropole régionale après la métropole de Grenoble Alpes et la métropole de Lyon. La commune est par ailleurs au cœur d'une vaste aire urbaine de plus de 520 640 habitants en 2017, la dix-septième de France par sa population, regroupant 117 communes.

- Des quantités de données
  - tout **wikipedia**
  - **et 100× plus** avec le web, des livres, etc
- La partie *chat* : faire un agent conversationnel
  - InstructGPT
  - étiquetage manuel d'interactions (~100k)

Défis et difficultés de l'apprentissage automatique

- concevoir des algorithmes d'apprentissage
- prouver que ces algorithmes marchent
- évaluer ces algorithmes
- défis divers
  - biais des données
  - interprétabilité/explicabilité
  - adaptation/transfert
  - attaques et manipulation
  - *efficacité en données*
  - *efficacité en calcul/énergie*
- choix sociétaux et legislation

Et la génération d'image ?

- Données
  - des paires (image, description textuelle)
  - e.g. LAION-5B: **5 milliards d'images**
- Modèles
  - e.g. DALL·E 2:
    modèle propriétaire d'OpenAI
  - e.g. Midjourney:
    modèle propriétaire
  - e.g. Stable Diffusion:
    modèle open-source très populaire

1. Impact environnemental du numérique ?
   … et autres activités
2. Qu'est-ce que l'intelligence artificielle (IA) ?
   Faire appendre une machine à partir d'exemples
3. Coût énergétique de l'IA générative ?
   Entraînement et utilisation
4. Réduire l'impact environnemental de l'IA générative ?
   Pistes et défis

Coût énergétique de l'IA générative ?

Entraînement et utilisation

Entraînement et Inférence

Entraînement

- création du modèle
- phase d'apprentissage
- données massives

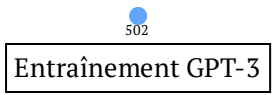
Inférence

- utilisation du modèle
- phase de génération
- utilisateurs finaux
- volume d'utilisation

Impact de l’entraînement

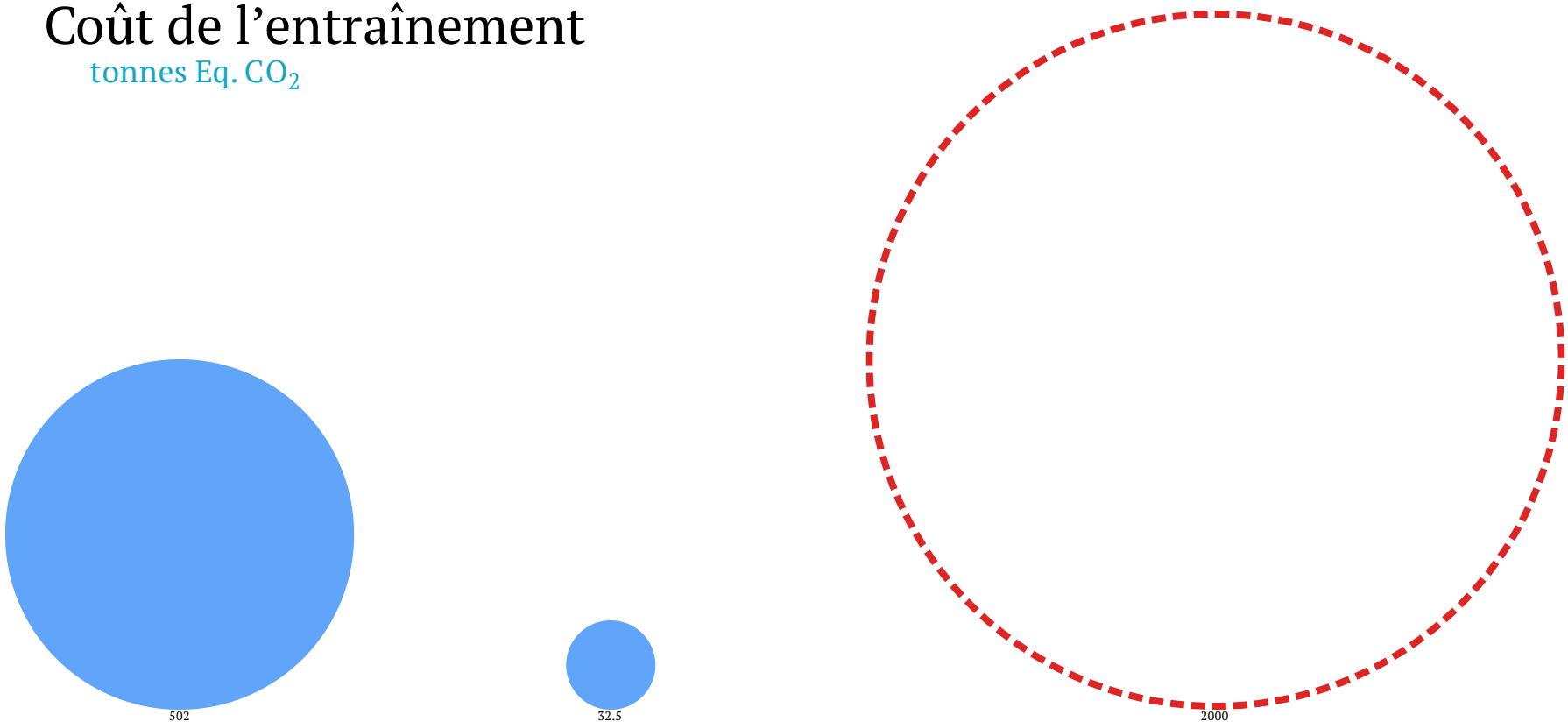
tonnes Eq. CO_2

- Entraînement GPT-3 ≈ 500 t Eq. CO_2



Coût de l’entraînement

tonnes Eq. CO_2



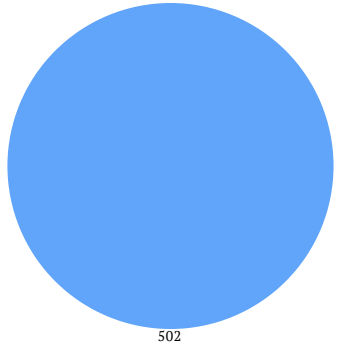
Entraînement GPT-3

Électricité annuelle 1000 personnes

Objectif 1000 personnes

L’exemple de Llama

transparence sur les coûts, deux version 8B et 70B

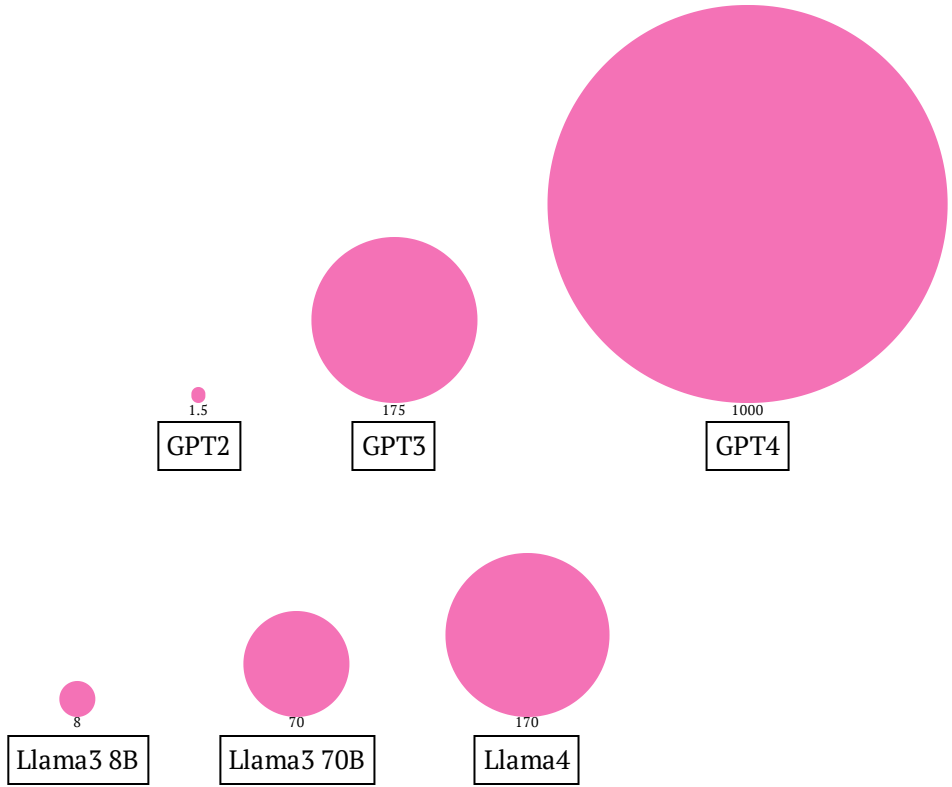


Entraînement GPT-3

- 8 milliards de paramètres (8B) ou 70 milliards (70B)
- Jeu de données
  - 15T "tokens" (parties de mots)
  - soit 15×10^{12}
- coût (en t Eq CO_2)
  - matériel (GPU)
  - calcul
  - **climatisation**

Stockage des modèles

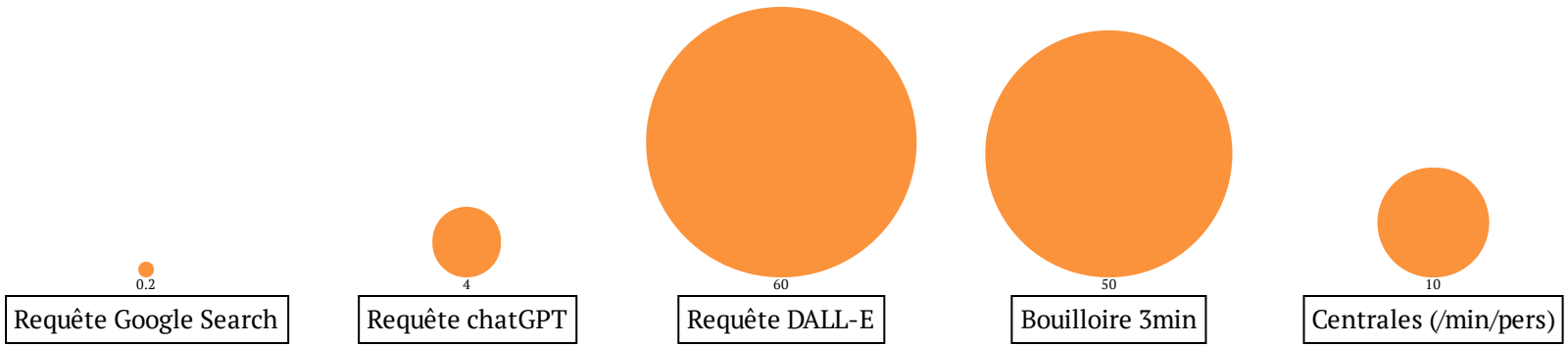
- Dépend du nombre de paramètres
  - GPT-2 : 1,5 milliards de paramètres (1.5B)
  - GPT-3 : 175 milliards de paramètres (175B)
  - GPT-4 : 1 trillion de paramètres (1000B)
  - GPT-5 : …
- Exemple de Llama3
  - estimation de taille
    - 8B : ~60 Go
    - 70B : ~500 Go
  - compression possible
    - 8B : 5Go, voir 2Go
    - 70B : similaire



Coût énergétique de l'inférence

Wh par requête

- 1 requête chat GPT = 10 requêtes Google Search



1. Impact environnemental du numérique ?
   … et autres activités
2. Qu'est-ce que l'intelligence artificielle (IA) ?
   Faire appendre une machine à partir d'exemples
3. Coût énergétique de l'IA générative ?
   Entraînement et utilisation
4. Réduire l'impact environnemental de l'IA générative ?
   Pistes et défis

Réduire l’impact environnemental de l’IA générative ?

Pistes et défis

Pistes pour réduire l'impact

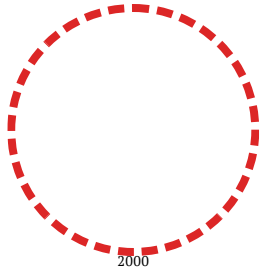
- Coté modèles / algorithmes / entraînement :
  - Optimiser les algorithmes d'entraînement
  - Réutiliser des modèles pré-entraînés
  - Réduire la taille des modèles
  - Réduire la quantité de données d'entraînement
- Coté infrastructure :
  - Utiliser des sources d'énergie renouvelable
  - Utiliser le matériel de l'utilisateur
  - Améliorer l'efficacité matérielle
- Coté usage :
  - Réduire le nombre de requêtes
  - Réduire la taille des réponses
- Exemples :
  - Mélange d'experts
  - Distillation de modèles
  - Quantification et compression

1. Impact environnemental du numérique ?
   … et autres activités
2. Qu'est-ce que l'intelligence artificielle (IA) ?
   Faire appendre une machine à partir d'exemples
3. Coût énergétique de l'IA générative ?
   Entraînement et utilisation
4. Réduire l'impact environnemental de l'IA générative ?
   Pistes et défis

Conclusions

Conclusions

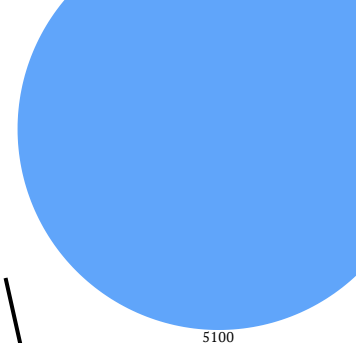
(par an et/ou personne)



Objectif

1 Ordinateur portable

Utilisation ordinateur

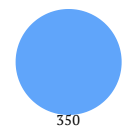


Alim. bœuf

Entraînement GPT-4 (pour 1M personne)

10 Requêtes ChatGPT par jour

Merci !

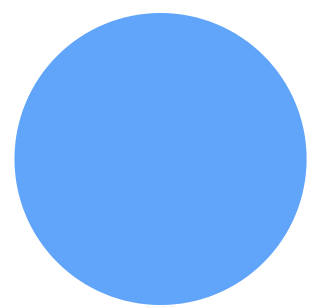


Alim. végétarienne

!!! Augmentation en continue !!!

1000 Requêtes ChatGPT par jour

1 Vol Paris-New York (12k km)



12k km voiture

12k km train