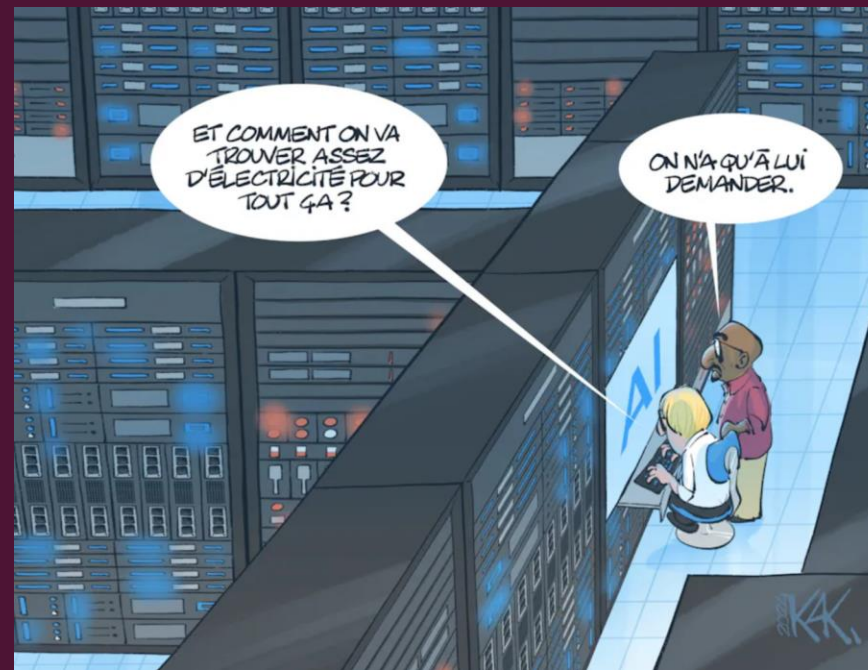


MISE EN PERSPECTIVE HISTORIQUE DE LA FRUGALITÉ EN IA ET STATISTIQUE

CHRISTOPHE BIERNACKI

WORKSHOP FRUGALIAS, 4 OCTOBRE 2024, PARIS





INTRODUCTION



DES DÉFINITIONS DE BASE MAIS RÉVÉLATRICES



■ Statistique

1. nom féminin Science et techniques d'interprétation mathématique de **données complexes et nombreuses**, permettant de faire des prévisions.

– Ensemble de données utilisables selon ces méthodes. *Statistiques économiques*

• abréviation, familier **stat**  .

■ Intelligence artificielle

3. Intelligence artificielle  (**IA** ) : ensemble des théories et des techniques développant des **programmes informatiques complexes** capables de simuler certains traits de l'intelligence humaine (raisonnement, apprentissage...).

– *Intelligence artificielle générative*, capable, à partir de **grands volumes de données** (textes, sons, images...), de dégager des modèles et d'en générer de nouveaux, ou d'améliorer les modèles existants.

} = stat + informatique
(ordinateur, algorithme)

■ Frugalité

3. Conçu de façon simple et raisonnée. Construction de bâtiments frugaux. Intelligence artificielle frugale, qui minimise son impact sur l'environnement.

Important mais
c'est plus général

UNE VISION PLUS PRÉCISE DES ENJEUX

- La création d'une image avec **l'IA générative consomme** autant d'énergie que la recharge d'un téléphone mobile
- IA frugale
 - c'est-à-dire **sobre énergétiquement** et/ou sur le plan des **données**
 - pour **répondre à des impératifs** environnementaux, mais aussi sociétaux et économiques

Ex : AI Act, RGPD

Power Hungry Processing: ⚡ Watts ⚡ Driving the Cost of AI Deployment?

Alexandra Sasha Luccioni
Yacine Jernite
sasha.luccioni@huggingface.co
Hugging Face
Canada/USA

Emma Strubell
Carnegie Mellon University, Allen Institute for AI
USA

arXiv:2311.16863v2 [cs.LG] 23 May 2024

UNE ACCÉLÉRATION DE LA NON FRUGALITÉ

THE CONVERSATION

L'expertise universitaire, l'exigence journalistique

L'IA peut-elle vraiment être frugale ?

Publié: 13 mai 2024, 16:51 CEST

Denis Trystram

Professeur des universités en informatique, Université Grenoble Alpes (UGA)

Thierry Ménissier

Professeur de philosophie politique, Université Grenoble Alpes (UGA)

ChatGPT3 : 175 milliards de paramètres
ChatGPT4 : beaucoup plus puissant...

Si on se focalise sur l'IA, on observe une rupture claire à partir de 2012. La croissance du secteur s'emballa alors avec un doublement des besoins en puissance de calcul tous les 5-6 mois au lieu de 24 mois, chiffre jusqu'alors stable de la classique loi empirique de Moore. Cette date correspond au développement des modèles d'IA reposant sur l'apprentissage profond, ou deep learning, rendus possibles par l'utilisation de processeurs graphiques (GPU) pour effectuer les calculs à la base de l'apprentissage profond et par le développement des données ouvertes sur Internet. Rappelons que l'IA n'est pas réduite à l'apprentissage par réseaux de neurones profonds, mais ce sont incontestablement ces derniers qui sont les plus gourmands.

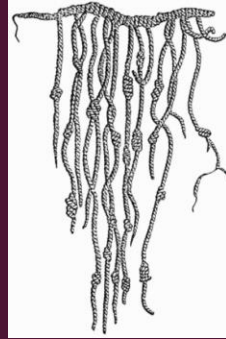
Un nouveau palier a été atteint en 2023, avec l'explosion des modèles génératifs comme l'agent conversationnel ChatGPT. Même s'il est difficile d'avancer des chiffres précis, étant donné que les « géants de la tech » comme OpenAI, Meta ou Microsoft qui sont à l'origine des plus gros modèles ne communiquent plus sur ces données, cette diffusion à large échelle est très inquiétante.



ANCRAGE HISTORIQUE DE CETTE TENDANCE

(DE NON FRUGALITÉ)





L'UNIVERSALITÉ DE LA DÉMARCHE STATISTIQUE

- Il y a toujours eu une connexion forte entre la collecte des données, leur stockage et leur traitement
- A l'origine du stockage de l'information on a trouvé depuis une centaine d'années un certain nombre d'objets gravés, principalement sur **des os ou des bois de rènes** au Paléolithique supérieur (environ -35 000 ans en Europe et -60 000 ans en Afrique)
- Les quipus des Incas étaient des systèmes fondés sur des **cordelettes, des nœuds et des couleurs** permettant d'avoir des statistiques sur les récoltes
- Plus récemment, nous ne devons pas sous-estimer l'importance **du stylo et du papier** comme forme d'informatique !

Jean-Claude Oriol. Formation à la statistique par la pratique d'enquêtes par questionnaires et la simulation : étude didactique d'une expérience d'enseignement dans un département d'IUT. Education. Université Lumière - Lyon II, 2007. Français. NNT: . tel-00191166

LA PUISSANCE STATISTIQUE COMME MOTEUR

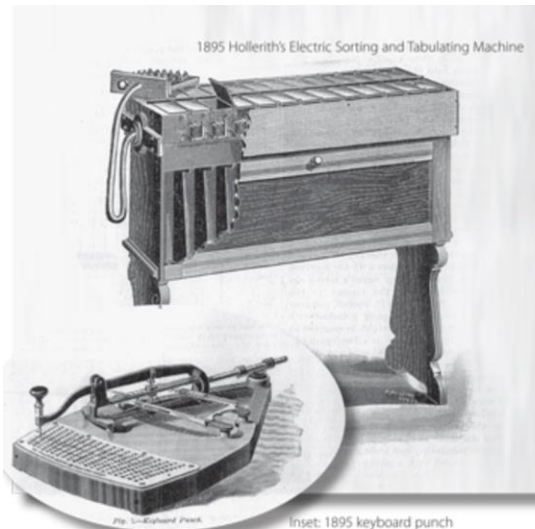
- La prévision d'un futur par essence imprévisible, va installer pour longtemps **respect et défiance** vis-à-vis du statistique
- **Staline déportait les statisticiens** qui avaient fait des sondages, seuls les recensements ayant un statut de preuve à ses yeux

Jean-Claude Oriol. Formation à la statistique par la pratique d'enquêtes par questionnaires et la simulation : étude didactique d'une expérience d'enseignement dans un département d'IUT. Education. Université Lumière - Lyon II, 2007. Français. NNT : . tel-00191166

DE LA STATISTIQUE À LA STATISTIQUE CALCULATOIRE

[ASA 175th Anniversary](#)

[Celebrate Our Past, Energize Our Future](#)



- **Les fondateurs** du domaine des statistiques **se sont appuyés sur les mathématiques** et les approximations asymptotiques dans leur développement de méthodologie statistique
- **L'informatique statistique est devenue un domaine d'étude populaire dans les années 1920 et 1930**, lorsque les universités et les laboratoires de recherche ont commencé à acquérir les premières tabulatrices mécaniques à cartes perforées d'IBM. Ces machines étaient utilisées pour tabuler et calculer des statistiques descriptives et pour ajuster des modèles statistiques plus complexes, tels que les analyses de variance et les régressions linéaires.
- **Sans eux, la méthodologie statistique moderne aurait pu dépérir** en tant que théorie intéressante, car seulement utile pour les petits problèmes

DE L'ORDINATEUR AUX LOGICIELS DÉDIÉS À LA STAT

- **Les développements les plus fondamentaux de la statistique au cours des 60 dernières années sont dus aux technologies de l'information**
- À partir du **début des années 1950**, nous sommes entrés dans l'ère de l'informatique
- **Pas simple à utiliser** : toute la programmation devait être en code machine avec les instructions et les données sur un disque rotatif avec une longueur de mot de 32 bits
- **Pas frugal en ressources** : En 1963, la dernière année où les ordinateurs Elliott 401 et Elliott 402 furent utilisés, les statisticiens de Rothamsted Research analysèrent 14357 variables de données, ce qui leur prit 4731 heures pour terminer le travail. Il est difficile d'imaginer la consommation d'énergie ainsi que la quantité de bande de papier utilisée pour la programmation. **La bande de papier (collée ensemble) serait probablement assez longue pour faire le tour de l'équateur.**
- Le développement des **logiciels statistiques** comme **autre booster de la statistique** (~70') : Genstat, SAS, SPSS



L'ESSORT DE LA STATISTIQUE CALCULATOIRE

- Carlo Lauro (ancien président de l'Association internationale pour l'informatique statistique, 1996) définit « les statistiques computationnelles » comme « visant à la conception d'algorithmes pour la **mise en œuvre de méthodes statistiques sur des ordinateurs, y compris celles impensables avant l'ère informatique** (par exemple, le **bootstrap**, la **simulation**), ainsi que pour faire face à des problèmes analytiquement insolubles »
- Le terme « statistiques computationnelles » peut également être utilisé pour désigner des méthodes statistiques nécessitant beaucoup de calcul, notamment les méthodes de **rééchantillonnage**, les méthodes de Monte Carlo par chaîne de Markov, la régression locale, l'estimation de la densité du noyau, les **réseaux neuronaux artificiels** et les modèles additifs généralisés



WIKIPEDIA
The Free Encyclopedia

CALCUL BAYÉSIEN : DU FRUGAL AU CALCUL INTENSIF

Les lois *a priori* avant les MCMC

- Calcul analytique de la loi *a posteriori* limité aux lois conjuguées
- Lois conjuguées disponibles uniquement sur quelques modèles simples
- Information *a priori* pas forcément facile à modéliser sous la forme d'une loi conjuguée

Les lois *a priori* à l'ère des MCMC

L'utilisation des MCMC (*Markov Chain Monte Carlo*) pour estimer par échantillonnage la loi *a posteriori* a donné un **nouvel élan à la statistique bayésienne**

- Algorithme de Metropolis-Hastings : 1970
- Échantillonneur de Gibbs : 1984
- ...

LES DIFFÉRENTS FACTEURS EN STAT CALCULATOIRE

- L'histoire générale des ordinateurs couvre une trentaine d'années (Davis 1977). Des changements technologiques substantiels ont été constants tout au long de cette période. Une grande partie de ces changements spectaculaires ont concerné le matériel informatique, les périphériques qui composent l'ordinateur et leur organisation physique. Le matériel informatique peut être divisé en quatre composants : **les processeurs**, qui contiennent des équipements permettant de manipuler logiquement ou numériquement les informations ; **la mémoire** programme, qui stocke le programme et les données qu'il traite actuellement ; **le stockage de données en masse**, qui stocke toutes les informations que l'ordinateur peut être appelé à traiter (y compris les programmes des utilisateurs) ; et la **communication, l'échange d'informations** entre l'ordinateur et ses utilisateurs ou entre les sous-systèmes de l'ordinateur lui-même.
- L'utilisation des ordinateurs dans les statistiques et ailleurs sera fortement affectée par deux évolutions liées : l'évolution rapide du matériel que nous venons de mentionner et **l'importance croissante des logiciels**, qui entraînera une crise dans la disponibilité de bons programmes (« The Oregon Report » 1978).

Statistical Computing: History and Trends

Author(s): John M. Chambers

Source: *The American Statistician*, Nov., 1980, Vol. 34, No. 4 (Nov., 1980), pp. 238-243

Published by: Taylor & Francis, Ltd. on behalf of the American Statistical Association

LA FRUGALITÉ SE DÉCLINE SOUS DIFFÉRENTES FORMES

- **La frugalité des entrées** met l'accent sur le coût associé aux données, en particulier à l'acquisition des données d'apprentissage, à l'exploitation des caractéristiques descriptives, ou aux deux. Les entrées frugales peuvent impliquer moins de données d'apprentissage ou moins de caractéristiques que nécessaire pour la meilleure qualité de prédiction réalisable dans un contexte non frugal. La frugalité des entrées peut être **motivée par des contraintes de ressources et par des contraintes de confidentialité**.
- **La frugalité du processus d'apprentissage** met l'accent sur le coût associé au processus d'apprentissage, en particulier les ressources de calcul et de mémoire. L'apprentissage frugal peut produire un modèle avec une qualité de prédiction inférieure à celle réalisable dans un contexte non frugal, mais le faire beaucoup plus efficacement. La frugalité du processus d'apprentissage est principalement motivée par des **contraintes de ressources**, notamment une puissance de calcul limitée et une capacité de batterie limitée.
- **La frugalité du modèle** met l'accent sur le coût **associé au stockage ou à l'utilisation d'un modèle d'apprentissage** automatique, tel qu'un classificateur ou un modèle de régression. Pour l'apprentissage supervisé, les modèles frugaux peuvent nécessiter moins de mémoire et produire des prédictions avec moins d'effort de calcul que nécessaire pour une qualité de prédiction optimale. La frugalité du modèle est principalement motivée par des **contraintes de ressources** telles qu'une mémoire limitée ou des capacités de traitement limitées.

LE MUR DE L'INSOUTABILITÉ CALCULATOIRE (DÈS 2015...)

- **Des charges de calcul « colossales »**. De nombreux modèles et méthodes d'analyse de modèles modernes nécessitent des ressources de calcul de la taille d'un Goliath. Par exemple, les méthodes d'analyse de modèles par échantillonnage de Sobol et par Monte Carlo par chaîne de Markov (**MCMC**) **nécessitent généralement des milliers, des dizaines de milliers ou plus d'exécutions de modèles** pour fournir une exploration approfondie de l'espace des paramètres du modèle (par exemple, Razavi et al. 2010 ; Herman et al. 2013a).
- Confrontés à **des tâches d'analyse de modèles nécessitant des calculs écrasants**, les modélisateurs sont souvent obligés de : (1) **simplifier** — et peut-être même de simplifier à l'excès ! — les modèles simplement pour réduire les temps d'exécution, et/ou (2) effectuer des analyses avec **moins d'exécutions de modèles** que nécessaire pour obtenir des résultats fiables.

Practical Use of Computationally Frugal Model Analysis Methods

Article *in* Ground Water · March 2015

DOI: 10.1111/gwat.12330



PISTES DE SOLUTIONS

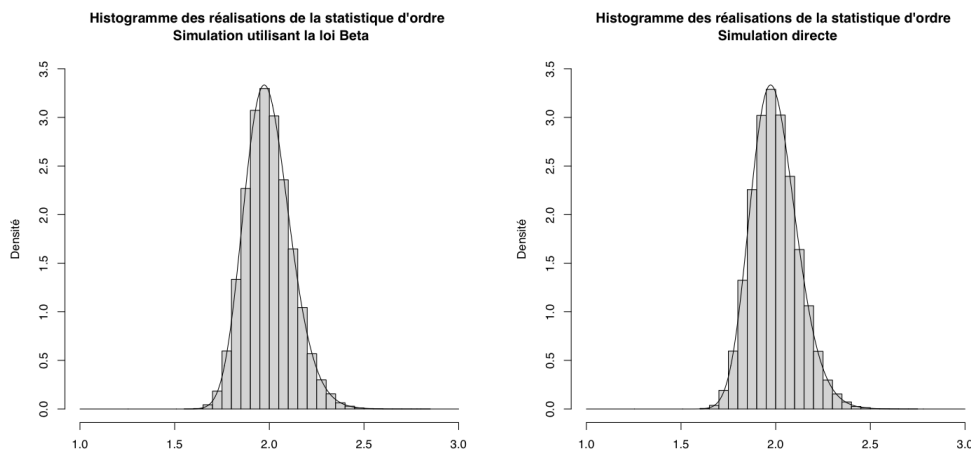


(RE)PENSER AUX MÉTHODES ANALYTIQUES

Mitchell Watnik (2011) Early Computational Statistics, Journal of Computational and Graphical Statistics, 20:4, 811-817, DOI: 10.1198/jcgs.2011.204b

- **Tukey (1962)** a noté que l'ordinateur avait parfois « stimulé le développement d'une méthode qui s'est ensuite révélée tout à fait applicable sans lui ». Il existe cependant d'autres situations « où l'ordinateur rend possible ce qui aurait été totalement irréalisable »
- Ex. en simulation : échantillonnage d'importance
- Ex. en validation croisée : des solutions exactes existent parfois
- Ex. en statistique d'ordre : simulation « directe »

Celisse, Alain; Mary-Huard, Tristan. Exact Cross-Validation for k NN : application to passive and active learning in classification. Journal de la société française de statistique, Tome 152 (2011) no. 3, pp. 83-97.



L. Gardes & S. Girard

Introduction à la statistique des valeurs extrêmes

FIG. 3.1 – Histogrammes des $N = 10^5$ réalisations de la k -ième statistique d'ordre de $n = 200$ variables aléatoires de loi de Pareto. La courbe en trait continu représente la densité théorique de la Proposition 3.2. A gauche, simulation utilisant le Corollaire 3.1 (temps d'exécution 0,2 seconde). A droite, simulation « directe » (temps d'exécution 2,6 secondes).

QUAND LA FRUGALITÉ AIDE À L'APPRENTISSAGE

Une préoccupation essentielle du Machine Learning est d'**éviter le surajustement**, avec pour conséquence la frugalité des hypothèses : plus l'hypothèse est frugale, moins elle risque de surajuster les données.

Evchenko, Mikhail & Vanschoren, Joaquin & Hoos, Holger & Schoenauer, Marc & Sebag, Michèle. (2021). Frugal Machine Learning. 10.48550/arXiv.2111.03731.

RÔLE DES LIMITES THÉORIQUES SUR LE CALCUL

Ce qu'on a appris par l'algorithme (connu)

La vérité (inconnue)

Besoin de garanties

- Algorithmes avec des garanties **démonstrables**
- Avec un coût de calcul **réduit** (minimal si possible)
- Applicables en pratique sur des **architectures existantes/classiques**

Limites théoriques en apprentissage

- Minimax lower bounds (example of multivariate regression):
- Given n examples in the dataset.
- f^* is s - times differentiable, $X = \mathbb{R}^d$ and $L(y, y') = (y - y')^2$

Theorem: For any algorithm there exists a learning problem for which

$$\mathbb{E}L(\hat{f}(x), y) - \mathbb{E}L(f^*(x), y) \geq cn^{-\frac{2s}{2s+d}}$$

A Distribution-Free Theory of Nonparametric Regression, Györfi et al. 2001

Peut-on construire un algorithme qui atteint ce type de borne optimale (en stockage, en calculs) ?

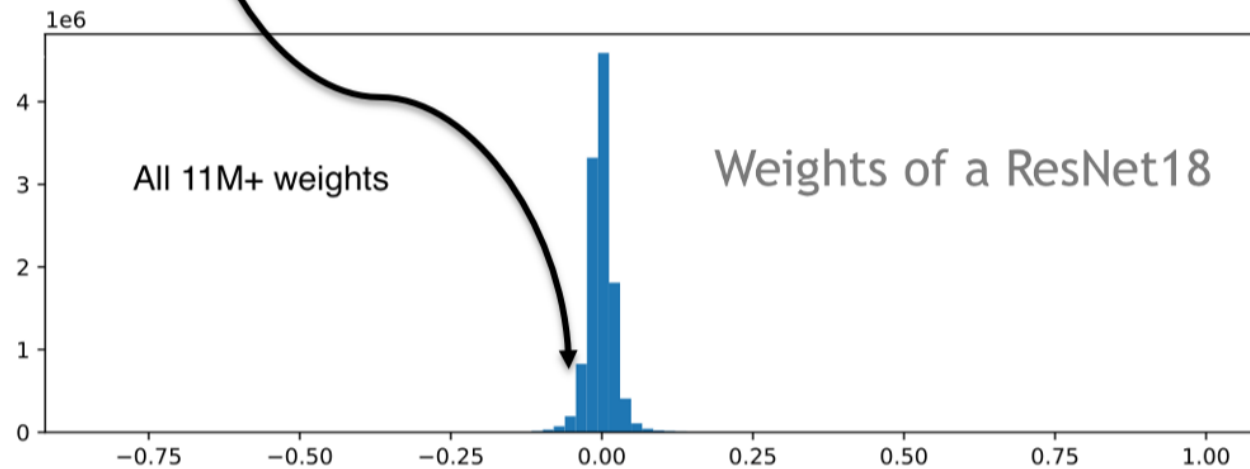
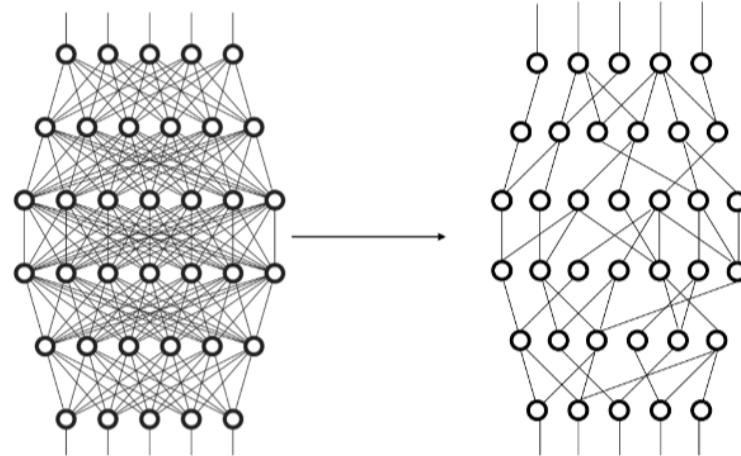
Les défis d'une IA frugale

24.11.2023, par [Martin Koppe](#)

“Il ne faut pas se leurrer, l'IA est considérée comme un vecteur de croissance économique par beaucoup de secteurs d'activité.”

Divers systèmes, aux frontières des neurosciences, des mathématiques et de la physique fondamentale ouvrent des perspectives intéressantes. Le cerveau humain nous montre ainsi que les possibilités de progrès sont énormes car il parvient à accomplir toutes ses tâches avec seulement une dizaine de watts, soit moins que l'énergie nécessaire à une lampe de chevet.

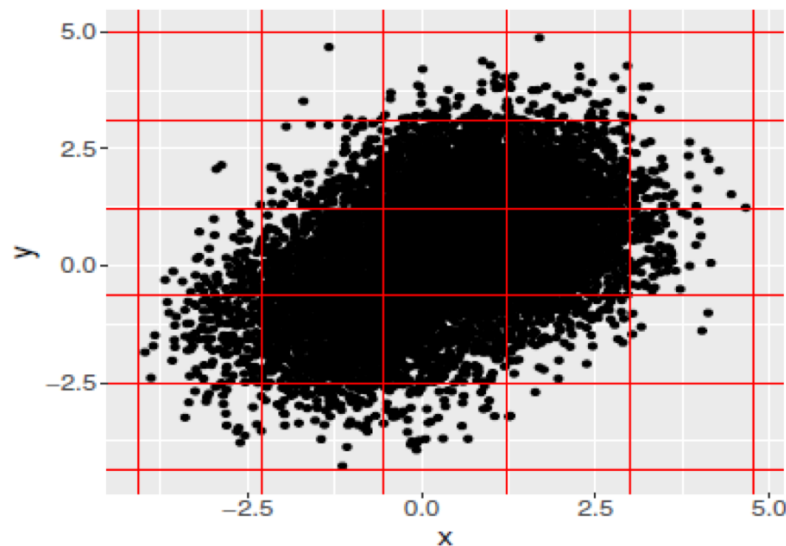
Beaucoup de (très)
petits coefficients...



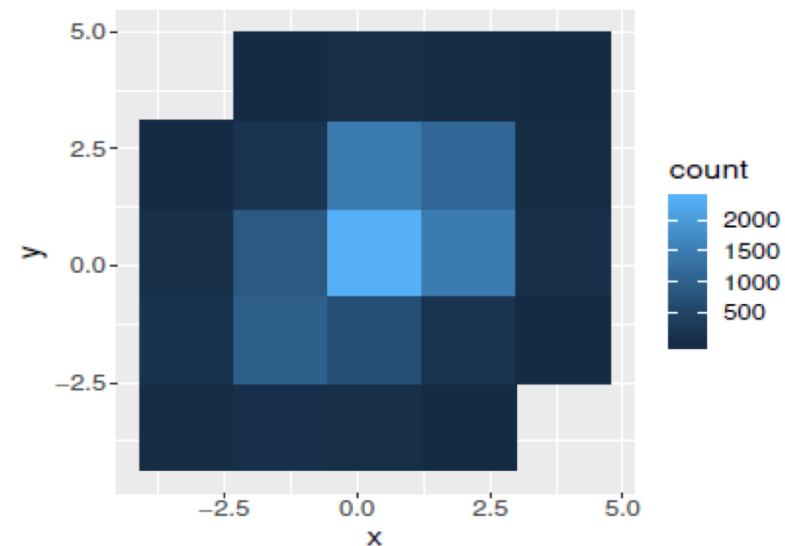
MÊME DANS
LES RNN ILY
A FRUGALITÉ
POTENTIELLE

EX. : COMPRESSION DE DONNÉES

- Conserve la convergence et l'identifiabilité
- Préserve les petites structures (au contraire de l'échantillonnage)



(a) Raw data



(b) Binned data

EX.: TRANSFER LEARNING

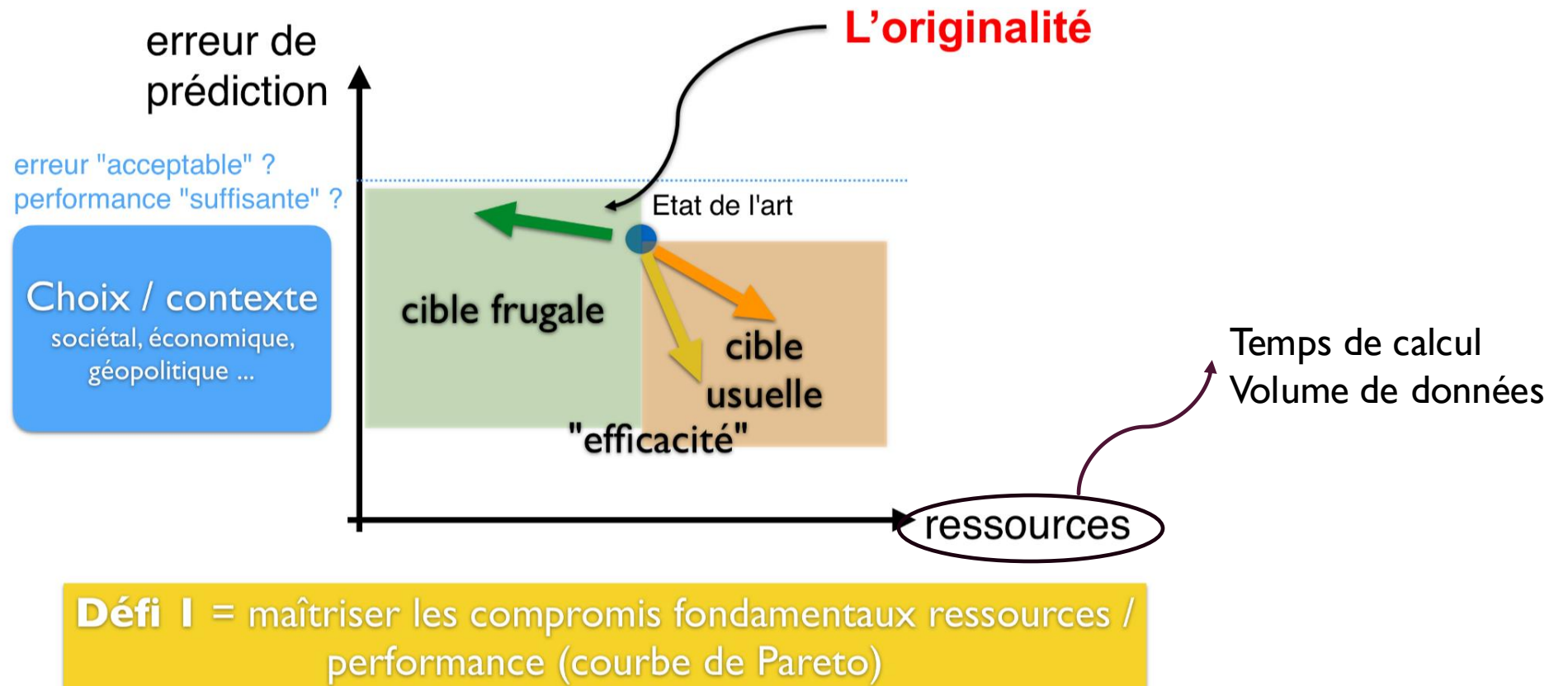
Quantmetry
Part of Capgemini Invent

Intelligence artificielle frugale

Le transfer learning, un moyen de rendre l'IA plus « verte »

L'une des solutions explorées est l'apprentissage par transfert (transfer learning) où il est possible d'exploiter les connaissances (caractéristiques, poids, etc.) des modèles précédemment formés pour former de nouveaux modèles, et même résoudre des problèmes tels que le manque de jeux de données. Par conséquent, avec l'apprentissage par transfert, au lieu de former un réseau de neurones à partir de zéro pour exécuter une certaine tâche, nous pouvons prendre un réseau déjà entraîné sur un domaine différent pour exécuter une tâche source différente et l'adapter au domaine dont nous avons besoin et à notre tâche cible.

EX. : COMPROMIS RESSOURCES / PRECISION





CONCLUSION



LES ACTEURS OFFICIELS S'EN MÊLENT



Ministère du Partenariat avec les territoires et de la Décentralisation
Ministère de la Transition écologique, de l'Énergie,
du Climat et de la Prévention des risques
Ministère du Logement et de la Rénovation urbaine

Publié le 28 juin 2024

Publication du référentiel général pour l'IA frugale : s'attaquer à l'impact environnemental de l'IA et défendre la diffusion de l'IA frugale

- **Données** : moins de données, de meilleure qualité (nettoyage), compressées
- **Modèles** : plus simples à apprendre, réutilisation
- **Réseaux** : limiter les échanges de données (« *edge computing* »)
- **Matériel** : réutiliser, plus faible précision

PETIT RÉSUMÉ DE SOLUTIONS

- Au coeur des enjeux de frugalité, la **disponibilité des données**
 - Utiliser les ressources publiques et *open source*
 - Augmenter son jeu de données grâce aux données synthétiques
- Comment **améliorer l'efficacité énergétique** des grands modèles d'IA sans réduire leur complexité ou leur performance ?
 - *Transfer learning* : recyclage des modèles et *knowledge distillation*
 - Optimiser le ré-entraînement des modèles
 - L'art de trouver les meilleurs hyperparamètres
- L'émergence des **modèles compacts**, alliés de la frugalité
 - Compression des réseaux de neurones
 - Un apprentissage avec peu de données : les approches *few-shot*, *one-shot* et *zero-shot learning*
- L'IA pour les **systèmes embarqués : une IA frugale par nature**
 - Comprendre le concept de l'**edge AI** et ses différentes composantes (**tiny-ML**)
 - Les nouvelles technologies **hardware** frugales compatibles avec l'*edge AI*

L'IA Frugale bouleverse les codes technologiques :
décryptage des solutions techniques innovantes depuis la
Silicon Valley

décembre 19, 2023



france
science

DES SOLUTIONS MAIS ATTENTION À L'EFFET REBOND...



Src: Idris, Rafael Medeiros – ORAP – 2021

L'effet rebond caractérise un effet pervers et paradoxal des progrès en matière d'efficacité énergétique.



Merci
et bon workshop !