Apprentissage automatique avancé

Représentations parcimonieuses et optimisation

Séance 1 — Objectifs, état de l'art, règles du jeu

INSA Rouen Normandie — Département ITI / Laboratoire LITIS

Année universitaire 2026–2027

Page du cours : moodle.insa-rouen.fr/course/view.php?id=848

Fil conducteur

1. **Où en est l'IA** en septembre 2026 : capacités, usages, ce qui a basculé en 2025–2026.
2. **Où va ce cours** : données tabulaires, parcimonie, et pourquoi cet objet résiste au « tout-LLM ».
3. **Comment on travaille ensemble** : politique d'usage de l'IA, évaluation, examen en deux temps.

1. Objet
données tabulaires,
séries temporelles

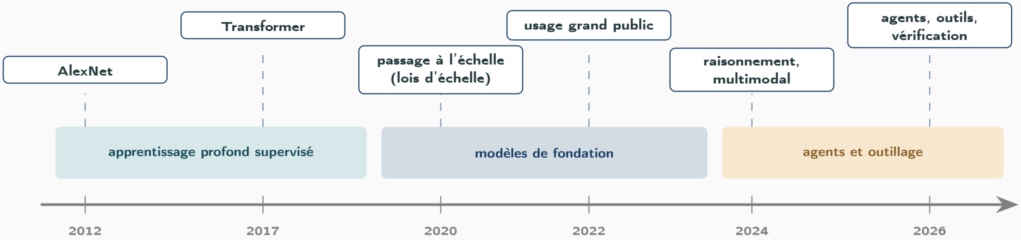
2. Méthode
parcimonie,
dictionnaires, sélection

3. Contexte
modèles de
fondation, agents

4. Posture
savoir vérifier, pas
seulement produire

Où en est l’IA en septembre 2026

Trois vagues en quatorze ans



Ce qui change — on ne conçoit plus une architecture par tâche ; on adapte un modèle générique.

Ce qui ne change pas — validation, fuite de données, dérive, coût : les erreurs de 2012 se commettent toujours.

Ce qui nous concerne — la troisième vague arrive *aussi* sur les données tabulaires et les séries temporelles.

Quatre chiffres pour cadrer le débat

$\pm$25 points d'Elo

C'est l'écart qui sépare, en mars 2026, six laboratoires (Anthropic, Google, OpenAI, xAI, Alibaba, DeepSeek) au sommet du classement Arena. La compétition s'est déplacée vers le **coût** et la **fiabilité**.

$\div$280 en 18 mois

Le coût d'inférence à qualité constante (niveau GPT-3.5) est passé d'environ 20 $ à 0,07 $ le million de tokens. Epoch AI mesure une division médiane par ∼50 par an, et par ∼200 depuis 2024.

53 % en 3 ans

Taux d'adoption de l'IA générative dans la population — plus rapide que le PC ou l'internet. Très corrélé au PIB par habitant, donc très inégal.

40 sur 100

Score moyen de l'indice de transparence des modèles de fondation, en baisse (58 auparavant) : code d'entraînement, taille des jeux de données et nombre de paramètres sont de moins en moins publiés. **Les modèles les plus capables sont les moins documentés.**

Source : Stanford HAI, *AI Index Report 2026* (13 avril 2026) — hai.stanford.edu/ai-index ; Epoch AI, *LLM Inference Price Trends*.

Où ces modèles sont réellement utilisés

Code
génération, revue, migration, tests, rétro-ingénierie, cybersécurité

Mathématiques
exploration, contre-exemples, formalisation en Lean

Langue
traduction, résumé, rédaction, accessibilité

Perception
vision, parole, visioconférence, sous-titrage temps réel

Science
lecture de corpus, conception d'expériences, analyse de données

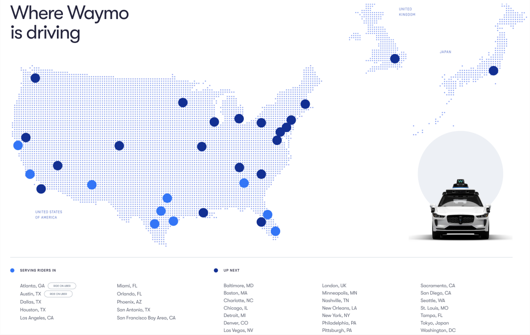
Décision
prévision, détection d'anomalies, maintenance, tarification

Notre créneau
les deux cases du bas reposent sur des données *structurées* : c'est là que le cours intervient

Remarque de méthode : les capacités sont *irrégulières* (« jagged intelligence »). Un modèle qui résout un problème de niveau olympiade peut échouer sur une jointure de tables mal spécifiée. Ne généralisez jamais une réussite spectaculaire en compétence stable.

Source : Stanford HAI, *AI Index Report 2026*.

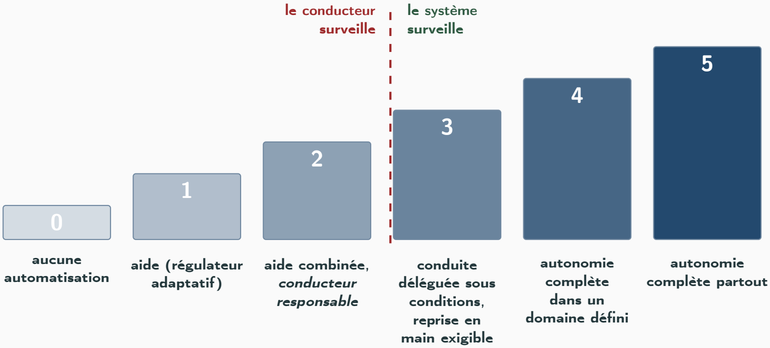
Cas d’école : le véhicule autonome



- Flotte de l’ordre de **4 000 véhicules** Aout 2026, en cours de renouvellement vers la 6ᵉ génération.
- Cap des **500 000 trajets payants par semaine** franchi début 2026 ; une dizaine de villes desservies, une vingtaine de marchés visés, première implantation hors des États-Unis à Londres et Tokyo.
- Les incidents résiduels sont des **cas limites** : chaussée inondée, feux éteints, bus scolaire à l’arrêt, panne électrique généralisée.

Leçon transférable : le passage à l’échelle ne bute plus sur le cas moyen mais sur la *queue de distribution*. C’est un problème statistique, pas un problème de puissance de calcul.

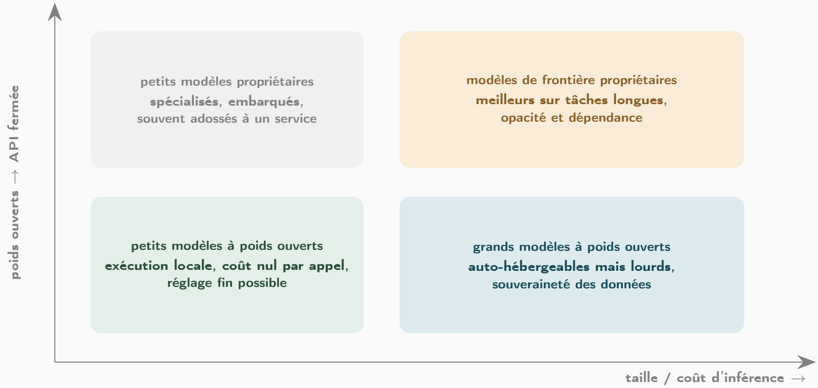
Les six niveaux d'automatisation de la conduite (SAE J3016)



La rupture n'est pas entre 4 et 5, elle est **entre 2 et 3** : c'est là que la responsabilité de la surveillance bascule de l'humain vers la machine. La même question se posera pour vos modèles : *qui vérifie, et avec quelles garanties ?*

Source : SAE International, norme J3016 (taxonomie des niveaux de conduite automatisée).

Lire le paysage des modèles : deux axes suffisent



Les noms de modèles changent tous les trimestres ; les **axes**, eux, sont stables. Choisir, c'est arbitrer entre performance, coût, confidentialité, reproductibilité et dépendance à un fournisseur. Pour un état des lieux daté, consultez un classement public plutôt qu'une diapositive : elle sera fausse dans trois mois.

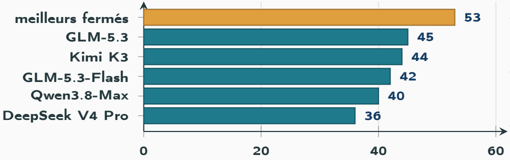
Sources : Classements publics : LMArena, Artificial Analysis, Epoch AI ; Hugging Face pour les poids ouverts. Stanford HAI, *AI Index 2026*, pour les tendances agrégées.

Poids ouverts : l'état de l'art est chinois, et l'écart se referme

Modèle	Éditeur	Paramètres (actifs)	Licence des poids	Poids publiés
GLM-5.3	Z.ai (ex-Zhipu)	753 Md	propriétaire	mi-août 2026
Kimi K3	Moonshot AI	2 800 Md (104 Md)	propriétaire*	27 juillet 2026
Qwen3.8-Max	Alibaba	2 400 Md (95 Md)	propriétaire†	12–13 août 2026
DeepSeek V4 Pro	DeepSeek	1 600 Md (49 Md)	**MIT**	13 août 2026

Fenêtre de contexte de 1 M de tokens pour les quatre. * « Kimi K3 License ». † variante 27B sous Apache 2.0.

Face aux meilleurs modèles fermés (Claude Fable 5.1, GPT-6 Astra : 53), l'écart n'est plus que de neuf points.



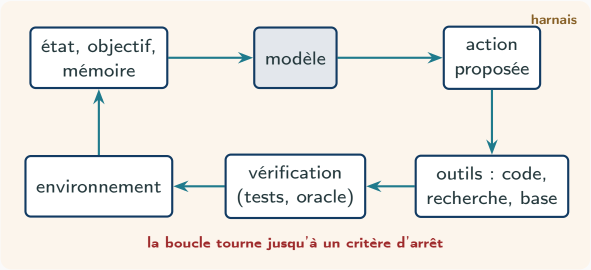
Intelligence Index, Artificial Analysis, sept. 2026.

- **Kimi K3** — le plus grand modèle ouvert publié, et le mieux noté des ouverts sur GPQA (93,5 %).
- **Qwen3.8-Max-0902** (2 sept.) — tête de *Code Arena WebDev*, devant Claude Opus 5 Max.
- **GLM-5.3** — long horizon (Terminal-Bench, SWE-bench) ; **DeepSeek V4 Pro**, seul en MIT, recule sur l'indice global.

Deux réserves. **Ce classement change d'une semaine à l'autre. Et « open source » recouvre des réalités distinctes : seuls DeepSeek V4 (MIT) et Qwen3.8-27B (Apache 2.0) relèvent de licences reconnues par l'OSI.**

Sources : Artificial Analysis, *Intelligence Index* (relevé de septembre 2026) ; cartes de modèles des éditeurs ; Code Arena *WebDev*. À réactualiser chaque année.

Un agent, c'est un modèle *plus* un harnais



Les progrès récents viennent autant du **harnais** que du modèle : boucle d'essai-erreur, accès aux outils, mémoire, et surtout **vérification automatique** (le code compile, le test passe, le score s'améliore).

> **Conséquence directe pour vous** : un agent est performant là où le succès est *mesurable automatiquement*. En science des données, cette mesure, c'est vous qui la définissez — et une métrique mal posée produit un agent efficacement faux.

Source : J. S. Chan *et al.*, *MLE-bench*, arXiv :2410.07095 (2024) — protocole et résultats détaillés plus loin.

La vague atteint nos données : modèles de fondation tabulaires

Prior synthétique
millions de jeux de données *engendrés* : modèles causaux, arbres, bruit, manquants

→

Pré-entraînement
un transformeur apprend à *inférer* plutôt qu'à ajuster (attention lignes + colonnes)

→

Inférence en contexte
on donne (X_{app}, y_{app}) et X_{test} en entrée : *une seule passe*

→

Prédiction
pas de descente de gradient, pas de réglage d'hyperparamètres

l'apprentissage a été déplacé du *temps de la tâche* vers le *temps du pré-entraînement*

Tabulaire

TabPFN v2 (*Nature*, 2025), TabICL, TabPFN-2.5, Mitra, LimiX. Domaine de validité annoncé : quelques milliers à 10^5 lignes, jusqu'à 10^3 variables. Évaluation : TabArena (51 jeux de données curés).

Séries temporelles

Chronos / Chronos-2, TimesFM, Moirai, TiRex, Toto, Lag-Llama, Sundial. Prévision *zéro-coup*, sans réentraînement par série. Évaluation : GIFT-Eval (97 tâches, 55 jeux de données).

Attention : ces classements sont animés par des équipes qui publient aussi les modèles évalués, et les jeux de données de pré-entraînement recouvrent parfois ceux d'évaluation. **Savoir auditer un protocole d'évaluation fait partie des compétences visées par ce cours.**

Sources : Hollmann *et al.*, *Nature* 637 :319–326 (2025) — Erickson *et al.*, *TabArena*, arXiv :2506.16791 — Aksu *et al.*, *GIFT-Eval*, NeurIPS-W 2024 — Ansari *et al.*, *Chronos*, TMLR 2024.

Où faire tourner tout cela : local, sur site, ou dans le nuage

Local (poste, GPU unique)

+ confidentialité totale, coût marginal nul, reproductible hors ligne

— modèles plus petits, débit limité, maintenance à votre charge

Sur site / mésocentre

+ grands modèles à poids ouverts, données qui ne sortent pas, mutualisation

— file d'attente, compétences d'exploitation, investissement

Nuage (API)

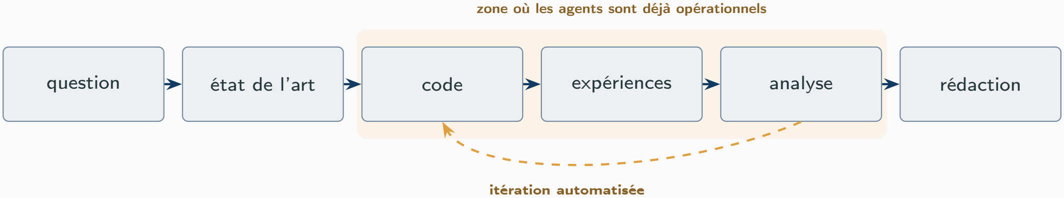
+ accès immédiat aux modèles de frontière, aucune infrastructure

— données exposées, coût à l'usage, dépendance et versions mouvantes

Règle de conduite pour les TP : aucune donnée à caractère personnel ni donnée d'entreprise sous accord de confidentialité ne sort vers un service tiers. Les jeux de données du cours sont publics : vous pouvez utiliser des services en ligne, mais vous devez pouvoir **rejouer votre chaîne de traitement hors ligne**.

Source : Recommandations RGPD / EU AI Act pour les traitements de données structurées ; charte numérique de l'établissement.

Automatiser le développement... et la recherche elle-même



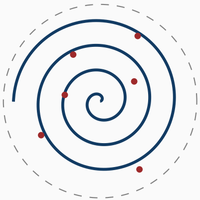
Ce qui est automatisable

Ce dont le succès se *mesure* : un test unitaire qui passe, une métrique qui monte, une preuve que le noyau Lean accepte.

Ce qui ne l'est pas

Poser la bonne question, choisir la métrique, décider si le protocole a du sens, assumer la conclusion. C'est votre valeur ajoutée — et c'est ce que l'examen évaluera.

Exemple 1 — le Vesuvius Challenge : un concours qui fait avancer la science



coupe d'un rouleau carbonisé :
l'encre (•) est presque invisible aux rayons X

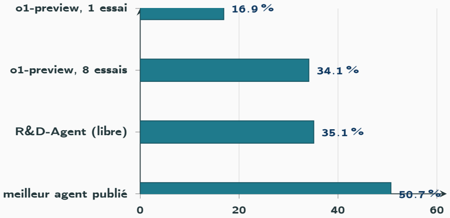
scrollprize.org

Lire les rouleaux d'Herculanum carbonisés en 79 apr. J.-C. **sans les ouvrir**. Chaîne complète : tomographie X → segmentation 3D de la surface de papyrus → « déroulement » virtuel → détection d'encre par apprentissage → lecture par des papyrologues.

- **2023** : premiers passages de grec lus dans un rouleau scellé.
- **2026** : PHerc. 1667 devient le **premier rouleau lu intégralement**, de bout en bout, annoncé le 25 juin 2026.
- Il reste **1 M$** de Grand Prix 2027 et 500 k$ de prix « premières lettres », ouverts jusqu'au 25 juin 2027.

Le verrou est un problème de généralisation : les modèles de détection d'encre entraînés sur les fragments lisibles ne transfèrent pas aux autres rouleaux. C'est un problème d'*adaptation de domaine*, pas de puissance de calcul.

Exemple 2 — les compétitions comme protocole d'évaluation



% de compétitions avec médaille (MLE-bench, 75 épreuves Kaggle)

Un agent enfermé 24 h sur une machine, sans accès au web, doit produire un fichier de prédictions. Le score est comparé au classement humain réel de la compétition.

- Les agents **obtiennent des médailles**, de plus en plus souvent.
- Ils restent **très inégaux** : le passage de 1 à 8 essais double le score, signe d'une forte variance.
- Des bancs d'essai spécifiquement **tabulaires** existent désormais (TML-bench, 2026).

Ce qu'une médaille ne dit pas : si la validation était saine, si le résultat est reproductible, si le modèle tiendra en production. C'est précisément ce qu'on vous demandera de juger.

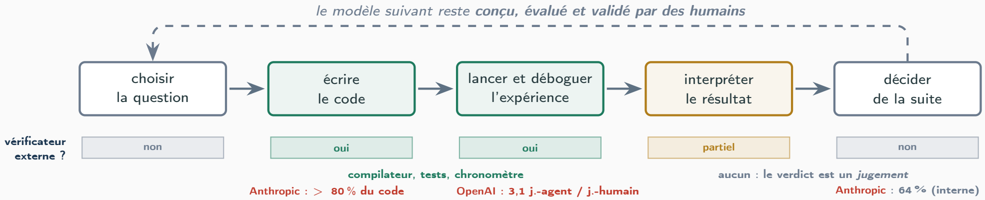
Exemple 3 — les mathématiques, ou l'art de vérifier

Depuis 2025, des problèmes ouverts sont résolus *avec un modèle dans la boucle*. Le site communautaire **VibeMathed** (en ligne depuis le 20 juillet 2026) recense ces résultats avec, pour chacun, une source vérifiable, un statut, un **niveau de vérification** et le rôle exact joué par le modèle (découverte, co-développement, assistance).

La communauté mathématique a réglé en dix-huit mois une question que la science des données n'a pas encore tranchée : **quel niveau de preuve exige-t-on d'un résultat produit avec une IA ?** Inspirez-vous-en pour vos rapports.

Sources : vibemathed.com (méthodologie et jeu de données CC BY 4.0) — T. Tao, *Crowdsourcing a list of general resources on AI and mathematics*, 10 septembre 2026.

Exemple 4 — Recursive self improvement



OpenAI, septembre 2026 — jalon déclaré atteint : un « *intern* de recherche automatisé », soit des tâches bien définies sous direction humaine. Suivant : « chercheur automatisé », mars 2028.

> **Les réserves sont écrites par les laboratoires eux-mêmes.** OpenAI : plus de la moitié des tâches de 4 à 8 h *réussies* ont demandé une intervention. Anthropic : le problème avait été *posé par des humains.*

Anthropic, avril 2026 — sur un problème *ouvert* d'alignement, des agents récupèrent 97 % de l'écart de performance, contre 23 % pour des chercheurs humains en une semaine.

> Même loi que la carte précédente : la boucle se ferme là où un vérificateur externe tranche. Chiffres auto-rapportés, sans jeu étalon indépendant : la partie 6 vaut aussi pour qui écrit la mesure.

Sources : OpenAI, *Research acceleration : the view inside OpenAI*, sept. 2026 — Anthropic, *When AI builds itself*, juin 2026.

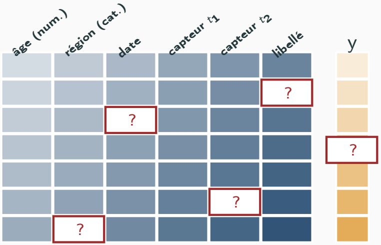
Objectifs du cours

Traiter des « donnée tabulaire »

Un tableau : n individus et p variables.

Un tableau (infromatique) = une matrice (mathématiques) = physique

Variables explicatives (observées : $X \in \mathbb{R}^{n \times p}$)



$X \in \mathbb{R}^{n \times p}$: les colonnes sont **hétérogènes** et sont permutables sans perte de sens, comme les lignes — contrairement aux pixels d'une image.

Ce qui pose problème

- p grand devant n
- données manquantes
- numérique / catégoriel / temporel mêlés
- OneHot encodeur = codage disjonctif complet
- y ou pas y ou peu. . .

Où on les trouve

Traces et journaux, réseaux sociaux, astronomie, génomique, procédés industriels, santé, finance. . .

Les données *tabulaire* sont structurées

n'est pas une image : les colonnes sont **hétérogènes**, il n'y a **ni invariance par translation, ni voisinage naturel**, les valeurs manquantes sont porteuses de sens et l'échantillon est souvent petit devant le nombre de variables.

C'est exactement ce qui fait échouer les recettes importées de la vision ou du texte — et ce qui rend la **parcimonie** utile.

Structuré

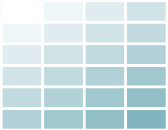
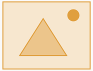
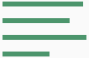


tableau $n \times p$
séries temporelles

Non structuré



image



texte

signal, audio

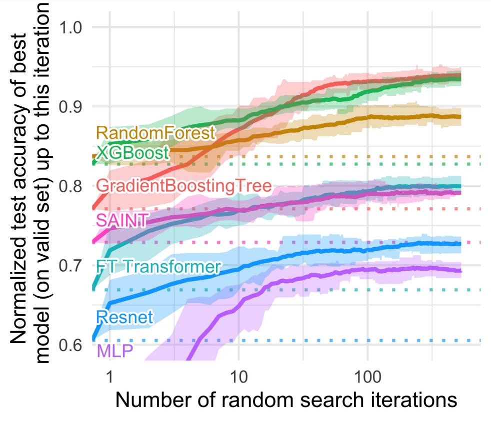
Source : Mark Ryan and Massaron Luca Machine Learning for Tabular Data (XGBoost, Deep Learning, and AI). Simon and Schuster, 2025.

Le tabulaire a longtemps résisté au deep learning

2022 — sur données tabulaires typiques, les arbres boostés (XGBoost, CatBoost, LightGBM) battent les réseaux profonds
→ Grinsztajn *et al.*, Neurips 2022

2025–2026 — bascule : les *modèles de fondation tabulaires* pré-entraînés sur des données *synthétiques* (TabPFN v2, publié dans *Nature*) rattrapent puis dépassent les arbres sur les tailles petites et moyennes, **sans entraînement** sur la tâche cible.

Question : que reste-t-il à modéliser explicitement quand un modèle générique prédit « sans apprendre » ?



SAINT : a Transformer model with an embedding module and an inter-samples attention mechanism, Somepalli et al. [2021].

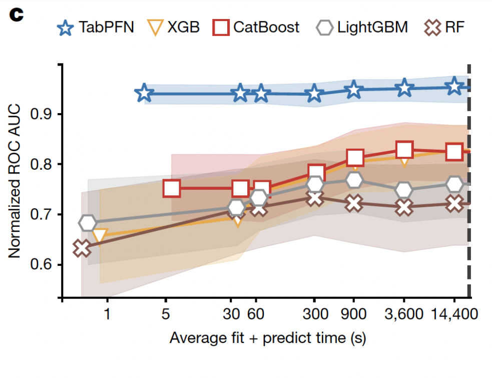
Sources : Grinsztajn *et al.*, Why do tree-based models still outperform deeplearning on typical tabular data ? NeurIPS 2022 — Hollmann *et al.*, *Accurate predictions on small data with a tabular foundation model*, *Nature* 637 :319–326, 2025

Le tabulaire a longtemps résisté au deep learning . . . jusqu'à 2025

2022 — sur données tabulaires typiques, les arbres boostés (XGBoost, CatBoost, LightGBM) battent les réseaux profonds
→ Grinsztajn *et al.*, Neurips 2022

2025 — bascule : les *modèles de fondation tabulaires* pré-entraînés sur des données *synthétiques* (TabPFN v2, publié dans *Nature*) rattrapent puis dépassent les arbres sur les tailles petites et moyennes, **sans entraînement** sur la tâche cible.

Question : que reste-t-il à modéliser explicitement quand un modèle générique prédit « sans apprendre » ?



The impact of hyperparameter tuning forthe considered methods. The x-axis shows the average time required to fit andpredict with the algorithm..

Sources : Grinsztajn *et al.*, Why do tree-based models still outperform deeplearning on typical tabular data ? NeurIPS 2022 — Hollmann *et al.*, *Accurate predictions on small data with a tabular foundation model*, *Nature* 637 :319–326, 2025

Données tabulaire et parcimonie : programmation par l'exemple

Version informatique de l'apprentissage statistique (statistical machine learning)

E — expérience
ce que l'on *montre*
$\rightarrow$ le jeu de données, son support, ses biais

T — tâche
ce que l'on *demande*
$\rightarrow$ entrée, sortie, moment de la prédiction

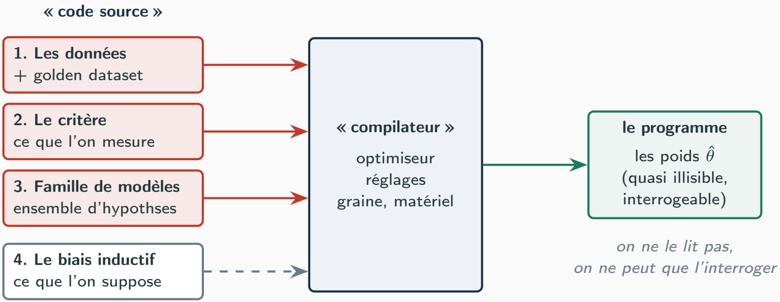
P — mesure
ce que l'on *mesure*
$\rightarrow$ la perte, la métrique, le *jeu étalon*

la mesure dicte la prochaine collecte

La spécification n'est pas « faire de la classification » : c'est le triplet complet, écrit et versionné.

Ce qui reste implicite est choisi *par défaut* — par l'outil, par l'échantillon disponible, ou par la métrique la plus facile à calculer.

Programmation par l'exemple : Software 2.0



Karpathy, *Software 2.0* (2017). Le programme, ce sont les poids ; le code source, c'est le jeu de données, la fonction objectif et la famille de modèles. Ce sont donc eux qu'il faut versionner, relire et tester.

Source : `https://karpathy.medium.com/software-2-0-a64152b37c35`

Données, critére, famille de modèles et biais

Données (X, y), Attention, Il y a un golden data set

Famille $\mathcal{F}$: Linéaire (simple) $f(x) = x^t\beta$, ou transformer...

Biais Parcimonie : le modèle doit être de petite taille ou autre

Critére Modèle $y = f(x) + \varepsilon$ Minimise l'erreur + biais

$$\min_{f \in \mathcal{F}} \|y - X\beta\|^2 \text{ avec } \|\beta\|_0 \leq s$$

Données tabulaire et parcimonie ?

Modèle statistique

$$y = f(x) + \varepsilon$$

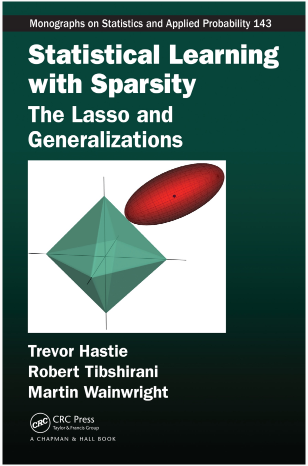
- du point de vue informatique : économie de moyens (efficacité en termes de place mémoire et vitesse d'exécution) $\rightarrow f$ de petite taille : principe d'Okam
- du point de vue du statisticien : améliorer les performances et les garanties de performance
  - quand $n > p$
    - moins de variables ($q < p$)
    - moins d'individus ($m < n$)
    - moins de paramètres pour le modèle $\rightarrow f$ petit
  - quand $p \geq n$: « peu de données » n'est pas un *bénéfice* de la parcimonie, c'est sa *condition d'applicabilité*
- du point de vue du client :
  - obtenir un modèle explicable
  - moins de mesures : quels sont les capteurs/mesures inutiles ?

Comment obtenir la parcimonie ?

Le meilleur modèle avec s paramètres = le meilleur modèle avec $\|\beta\|_0 \leq s$
Le problème naturel est combinatoire :

$$\min_{\beta} \ \|y - X\beta\|^2 \quad \text{s.c.} \quad \|\beta\|_0 \leq s$$

- ℓ_0 compte les coefficients non nuls : **NP-difficile**.
- ℓ_1 en est la **relaxation convexe** — la plus petite enveloppe convexe de ℓ_0 sur la boule unité. : Le lasso (Tibshirani, 1996)



Sources : Hastie, Tibshirani & Wainwright, *Statistical Learning with Sparsity*, CRC Press, 2015.

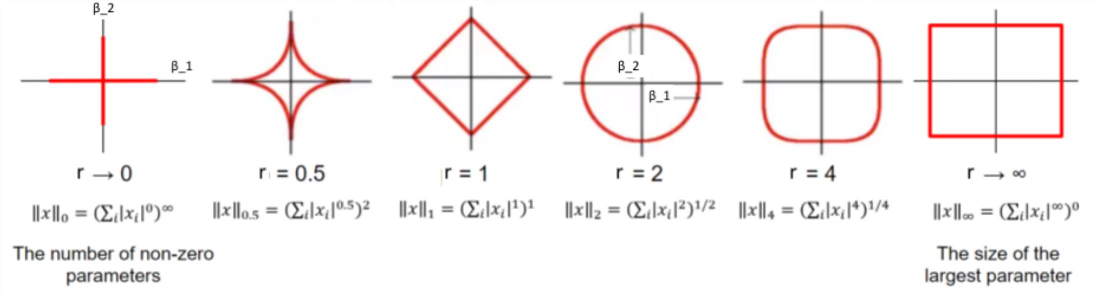
Boule unité

La norme :

$$\|\beta\|_r = (|\beta_1|^r + |\beta_2|^r)^{1/r}$$

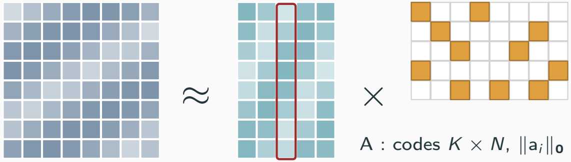
Le boulle unité est un ensemble

$$\mathcal{B}_r = \{\beta = (\beta_1, \beta_2) \mid \|\beta\|_r \leq 1\}$$



Sources : `https://theamyma101.substack.com/p/courage-to-learn-ml-demystifying`

Codage parcimonieux : X ≈ DA, avec A creuse



A : codes $K \times N$, $\|a_i\|_0 \ll K$

atome $\mathbf{d}_k$

X : $d \times N$ signaux   D : dictionnaire $d \times K$

Deux inconnues, deux régimes

D fixé (Fourier, ondelettes) ⇒ problème convexe en A.
D appris ⇒ problème biconvexe, alternance (MOD, K-SVD, descente en ligne).

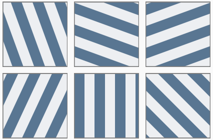
Ce qu'on gagne

débruitage, complétion de matrice, compression, séparation de sources, et surtout des *représentations interprétables*.
C'est le principe de certaines méthodes de fine tuning comme LoRA

Sources : B. Olshausen & D. Field, *Nature* 381 :607–609, 1996 — M. Aharon, M. Elad, A. Bruckstein, *K-SVD*, IEEE TSP, 2006 — J. Mairal *et al.*, *Online dictionary learning*, ICML 2009.

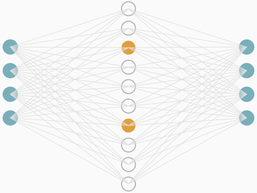
Le codage parcimonieux n'est pas un chapitre d'histoire

1996 — champs récepteurs V1



Bases apprises sur images naturelles : localisées, orientées, multi-échelles.

2023–2026 — autoencodeurs parcimonieux (SAE) dans les LLM



Même formulation : $\min \|x - \mathbf{D}a\|^2 + \lambda\|a\|_1$. Objectif : extraire des *directions monosémantiques* dans un réseau.

Trente ans plus tard, l'interprétabilité des grands modèles réutilise **exactement** l'apprentissage de dictionnaire parcimonieux. Les outils de ce cours ne sont pas périphériques au sujet du moment : ils en sont un des instruments.

Sources : Gao *et al.*, *Scaling and evaluating sparse autoencoders*, 2024.

Le codage parcimonieux n’est pas un chapitre d’histoire

WORKSHOP

Ended

Workshop on Sparsity in LLMs (SLLM): Deep Dive into Mixture of Experts, Quantization, Hardware, and Inference

Tianlong Chen · Utku Evci · Yani Ioannou · Berivan Isik · Shiwei Liu · Mohammed Adnan · Aleksandra I. Nowak · Ashwinee Panda

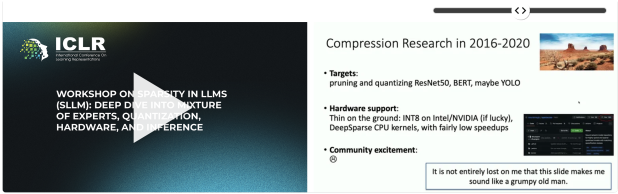
Project Page

Abstract

Large Language Models (LLMs) have emerged as transformative tools in both research and industry, excelling across a wide array of tasks. However, their growing computational demands especially during inference—raise significant concerns about accessibility, environmental sustainability, and deployment feasibility. At the same time, sparsity-based techniques are proving critical not just for improving efficiency but also for enhancing interpretability, modularity, and adaptability in AI systems. This workshop aims to bring together researchers and practitioners from academia and industry who are advancing the frontiers of sparsity in deep learning. Our scope spans several interrelated topics, including Mixture of Experts (MoEs), LLM inference and serving, network pruning, sparse training, distillation, activation sparsity, low-rank adapters, hardware innovations and quantization. A key objective is to foster connections and unlock synergies between traditionally independent yet highly related research areas, such as activation sparsity and sparse autoencoders (SAEs), or quantization and KV cache compression. Rather than focusing solely on efficiency, we aim to explore how sparsity can serve as a unifying framework across multiple dimensions of AI—driving advances in interpretability, generalization, and system design. By facilitating the fusion of ideas from different topics, the workshop will create new opportunities for innovation. We encourage participants to think beyond traditional constraints, exploring how different forms of sparsity can inform each other and yield new

Show more

Video



Sources : https://iclr.cc/virtual/2025/workshop/23996

Le prix de la parcimonie

Type de problème	Ce qu'il implique en pratique
Biais	Si l'hypothèse est fausse, c'est de la variance échangée contre du biais, *à perte*. Tester la densité du problème avant d'imposer la parcimonie.
Combinatoire	ℓ_0 est NP-difficile ; ℓ_1 n'est valide que sous conditions (RIP, irreprésentabilité) que les données réelles satisfont rarement.
Instabilité du support	Sous corrélation, la sélection n'est pas reproductible. Ne jamais interpréter le support comme causal.
Écart idéal / réalisable	Le matériel n'accélère que des motifs contraints. Le taux de parcimonie annoncé n'est pas le gain obtenu.
Évaluation trompeuse	*The Sparse Frontier*, des mécanismes d'attention parcimonieux. Une méthode validée sur des tâches à faible dispersion (trouver une date) paraît sans perte et s'effondre ailleurs : le choix des tâches décide du résultat.

Sources : Nawrot, Piotr, et al. "The sparse frontier : Sparse attention trade-offs in transformer llms." Findings of the Association for Computational Linguistics : ACL 2026. 2026.

Périmètre : ce que ce cours est, et ce qu'il n'est pas

Ce n'est pas ce cours

- × vision par ordinateur, traitement d'images
- × construire un LLM depuis zéro
- × apprentissage profond généraliste
- × ingénierie de prompts
- × panorama d'actualité de l'IA

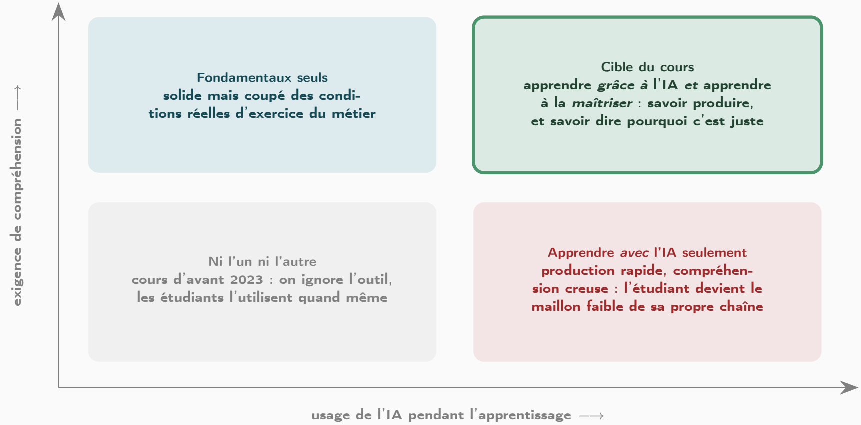
C'est ce cours

- ✓ parcimonie : ℓ_0, ℓ_1, seuillage, poursuite
- ✓ apprentissage de dictionnaire, factorisations
- ✓ grande dimension, sélection, généralisation
- ✓ données tabulaires et séries temporelles
- ✓ protocoles d'évaluation qui tiennent

La séance d'aujourd'hui parle abondamment de l'état de l'IA parce qu'il détermine *vos conditions de travail* et *le niveau d'exigence attendu de vous*. Le reste du semestre revient aux mathématiques et aux données.

Quelle politique vis-à-vis de l’IA dans ce cours

Deux objectifs distincts, souvent confondus



Maîtriser suppose de pouvoir **reproduire sans l'outil** ce qu'on lui a fait produire, ou à défaut de pouvoir **démontrer que le résultat est correct**. Tout le dispositif d'évaluation découle de cette phrase.

Quid de l'existant : trois régimes dans les établissements

Régime	**Défaut**	**Exemples documentés**
Interdiction sauf autorisation explicite de l'enseignant	usage nul	Cambridge : pas d'usage en évaluation sommative sans autorisation
Seuil d'usage explicite, par type de tâche	usage encadré	Stanford CS336 : aide autorisée sur des questions de bas niveau ou conceptuelles, résolution directe interdite
Délégation à l'enseignant, avec obligation de déclaration	dépend du cours	ETH Zurich, EPFL (*Lex* 1.3.3, art. 4)

Le point commun des trois régimes : **la charge de la transparence pèse sur l'étudiant**, pas sur un détecteur. La détection automatique de texte généré n'est pas fiable et n'est utilisée nulle part comme preuve.

CS336 ajoute une recommandation pratique que je reprends : **désactiver l'autocomplétion par IA** dans l'éditeur pendant les exercices d'implémentation. L'autocomplétion supprime précisément le moment où l'on apprend : celui où l'on est bloqué.

Sources : Stanford CS336, *Language Modeling from Scratch*, section *Policies* — ETH Zurich, *Generative AI in Teaching and Learning : Guidelines* — EPFL, guide de l'enseignement, *Lex* 1.3.3 art. 4 — University of Cambridge, orientations sur

L'examen : deux fois 1h30, deux compétences différentes

Partie A — 1h30, sans IA

Papier, sans machine. On y évalue ce qui doit être disponible dans votre tête : géométrie de ℓ_1, conditions d'optimalité, seuillage, complexité, lecture d'un protocole expérimental, calcul d'ordre de grandeur.

Partie B — 1h30, avec IA

Machine et modèle au choix. On y évalue ce que vous faites de l'outil : cadrer un problème mal posé, détecter une erreur produite par le modèle, justifier un choix, documenter votre démarche.

aucun appareil → appareils autorisés, trace conservée

Rendu de la partie B : **votre réponse *et* la trace de vos échanges avec le modèle, *et* un paragraphe de critique (« ce que le modèle a proposé, ce que j'ai gardé, ce que j'ai corrigé et pourquoi »). Une réponse juste sans critique argumentée ne vaut pas les points de la critique.**

L'intention n'est pas de vous piéger : les deux parties sont annoncées, les attendus sont explicites, et des annales de chaque type seront publiées sur Moodle.

Charte pratique du cours

Autorisé sans réserve

- expliquer un concept, reformuler un énoncé
- aide de syntaxe, messages d'erreur, documentation
- traduction, relecture de forme
- exploration bibliographique *à vérifier*

Autorisé, à déclarer

- génération de code intégré à un rendu
- génération de figures ou de tableaux
- rédaction de sections de rapport

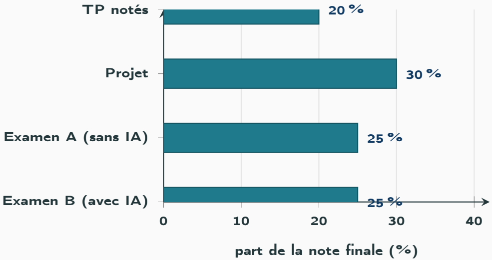
Déclaration : outil, version, à quoi il a servi, ce que vous avez vérifié. Trois lignes en annexe suffisent.

Interdit

- faire résoudre l'exercice d'implémentation par le modèle
- autocomplétion IA pendant les TP d'implémentation
- rendre un résultat que vous ne savez pas expliquer
- citer une référence que vous n'avez pas ouverte

Règle unique en cas de doute : si vous ne pouvez pas répondre au tableau à la question « pourquoi cette ligne ? », le rendu n'est pas le vôtre. Une référence inventée par un modèle et recopiée est traitée comme une faute de citation, indépendamment de l'IA.

Évaluation et compétences visées

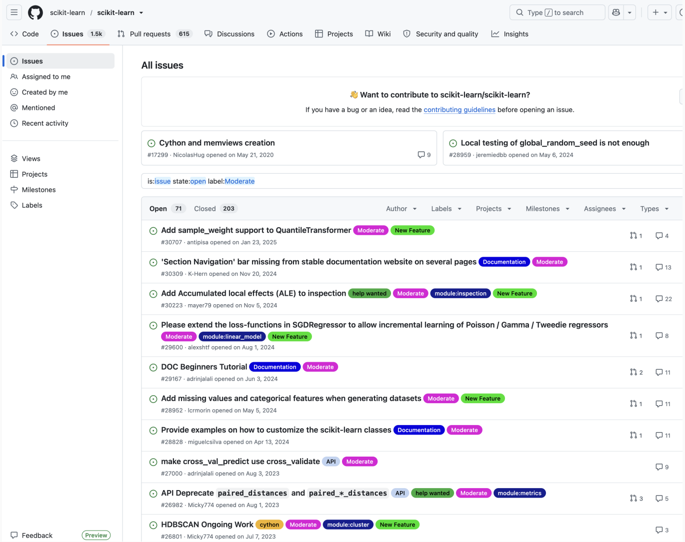


Barème indicatif

À la fin du semestre, vous devez savoir :

1. poser un problème de reconstruction ou de prédiction sous contrainte de parcimonie, et justifier le choix de la pénalité ;
2. implémenter et diagnostiquer un algorithme d'optimisation parcimonieuse (seuillage itératif, poursuite, alternance) ;
3. construire un protocole d'évaluation résistant aux fuites, à la dérive temporelle et au sur-ajustement de validation ;
4. situer une méthode récente (modèle de fondation tabulaire) par rapport à une ligne de base solide, chiffres à l'appui ;
5. **auditer un résultat produit par un tiers**, humain ou machine, et énoncer ce qui reste à vérifier.

Projet = Contribution à un logiciel open source



Organisation pratique

- **Supports, énoncés, jeux de données et bibliographie** : page Moodle du cours, mise à jour après chaque séance.
- **Environnement** : Python, `scikit-learn`, `numpy`/`scipy` ; versions figées et fournies. Tout doit pouvoir tourner sur votre machine, hors ligne.
- **Données** : publiques et redistribuables. Aucune donnée personnelle n'est manipulée dans le cours.
- **Prérequis** : algèbre linéaire, optimisation convexe de base, probabilités, Python scientifique.
- **Questions** : forum Moodle de préférence — une question posée profite à toute la promotion.

moodle.insa-rouen.fr/course/view.php?id=848

Sources citées — méthodes

Parcimonie et dictionnaires

- B. Olshausen, D. Field, *Emergence of simple-cell receptive field properties by learning a sparse code for natural images*, *Nature* 381 :607–609, 1996.
- R. Tibshirani, *Regression shrinkage and selection via the Lasso*, *JRSS-B* 58(1) :267–288, 1996.
- M. Aharon, M. Elad, A. Bruckstein, *K-SVD*, *IEEE TSP* 54(11), 2006.
- J. Mairal, F. Bach, J. Ponce, G. Sapiro, *Online dictionary learning for sparse coding*, ICML 2009.
- T. Bricken *et al.*, *Towards Monosemanticity*, Transformer Circuits, 2023 ; A. Templeton *et al.*, *Scaling Monosemanticity*, 2024 ; L. Gao *et al.*, *Scaling and evaluating sparse autoencoders*, 2024.

Données tabulaires et séries temporelles

- L. Grinsztajn, E. Oyallon, G. Varoquaux, *Why do tree-based models still outperform deep learning on typical tabular data ?*, NeurIPS 2022.
- N. Hollmann *et al.*, *Accurate predictions on small data with a tabular foundation model*, *Nature* 637(8045) :319–326, 2025.
- L. Grinsztajn *et al.*, *TabPFN-2.5*, arXiv :2511.08667, 2025.
- N. Erickson *et al.*, *TabArena*, arXiv :2506.16791, NeurIPS Datasets & Benchmarks 2025.
- A. F. Ansari *et al.*, *Chronos*, TMLR 2024 ; T. Aksu *et al.*, *GIFT-Eval*, NeurIPS-W 2024.
- J. S. Chan *et al.*, *MLE-bench*, arXiv :2410.07095, 2024 ; M. Pinchuk, *TML-bench*, arXiv :2603.05764, 2026.

Sources citées — état de l'IA, usages, politiques

Chiffres et tendances

- Stanford HAI, *Artificial Intelligence Index Report 2026*, 13 avril 2026 — hai.stanford.edu/ai-index.
- Epoch AI, *LLM Inference Price Trends* — epoch.ai.
- Classements publics : LMArena, Artificial Analysis, Hugging Face.
- SAE International, norme *J3016* — niveaux d'automatisation de la conduite. Chiffres Waymo : déclarations NHTSA et communications de l'entreprise, début 2026.

Exemples et politiques d'établissement

- Vesuvius Challenge — scrollprize.org (pages *Prizes*, *Master Plan* ; annonce du 25 juin 2026).
- VibeMathed — vibemathed.com (méthodologie ; jeu de données CC BY 4.0 ; en ligne depuis le 20 juillet 2026).
- Stanford CS336, *Language Modeling from Scratch*, section *Policies* — stanford-cs336.github.io.
- ETH Zurich, *Generative AI in Teaching and Learning : Guidelines* ; EPFL, guide de l'enseignement, *Lex* 1.3.3 art. 4 ; University of Cambridge, orientations sur l'IA générative.

Les chiffres datés (flottes, prix, classements) sont ceux disponibles en septembre 2026 et doivent être revérifiés avant chaque édition du cours. Les liens actifs sont regroupés sur la page Moodle du cours.