

# Challenges Science de Données

## L2 IASO

Encadrants : Tony Bonnaire ([tony.bonnaire@ens.fr](mailto:tony.bonnaire@ens.fr))

Conçu avec : Nicolas Schreuder, Kimia Nadjahi, Alexandre Allauzen

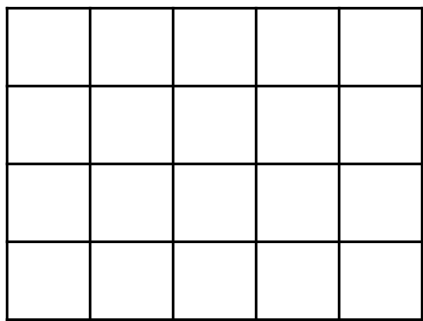
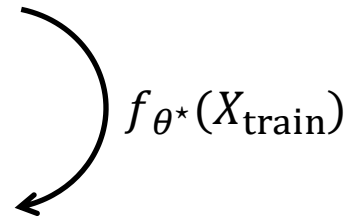
Paris Santé Campus, 2-6 juin 2025



Répondre à une problématique industrielle ou scientifique à **partir de données** en battant l'**algorithme de benchmark** selon une **métrique** choisie par l'organisateur  $\mathcal{L}$

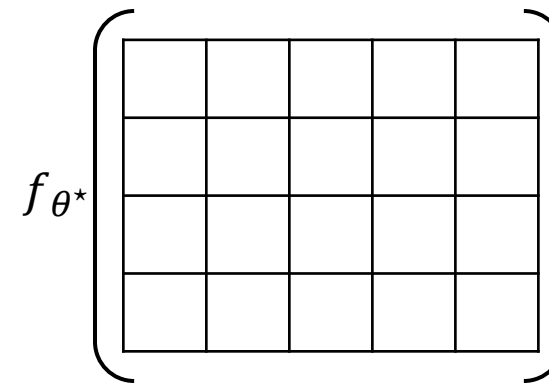
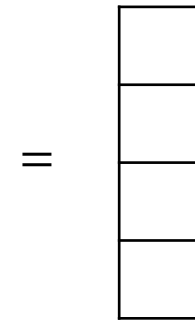
1

**Ensemble d'entraînement** pour créer et évaluer des modèles  $f_{\theta}$

 $X_{\text{train}}$  $y_{\text{train}}$ 

2

**Ensemble de test** (labels inconnus) pour comparer les participants

 $X_{\text{test}}$  $=$  $\hat{y}_{\text{test}}$ 

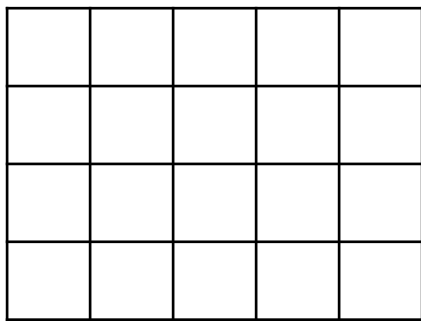
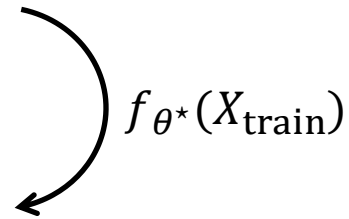
Calcul  $\mathcal{L}(\hat{y}_{\text{test}}, y_{\text{test}})$



Répondre à une problématique industrielle ou scientifique à **partir de données** en battant l'**algorithme de benchmark** selon une **métrique** choisie par l'organisateur  $\mathcal{L}$

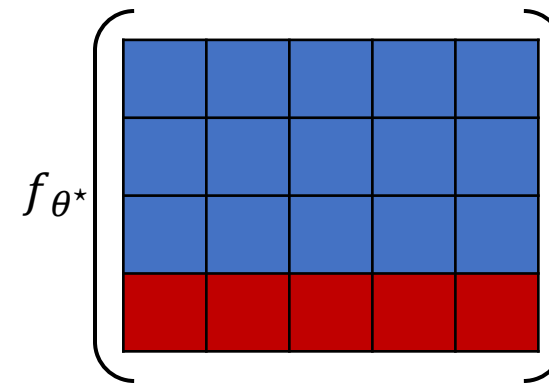
1

**Ensemble d'entraînement** pour créer et évaluer des modèles  $f_\theta$

 $X_{\text{train}}$  $y_{\text{train}}$  $f_{\theta^*}(X_{\text{train}})$ 

2

**Ensemble de test** (labels inconnus) pour comparer les participants

 $X_{\text{test}}$ 

=

 $\hat{y}_{\text{test}}$ 

Calcul  $\mathcal{L}(\hat{y}_{\text{test}}, y_{\text{test}})$

Séparation de l'ensemble de test en un **ensemble public** et un **ensemble privé** (sur lequel vous aurez les résultats une fois par jour) pour **éviter le sur-apprentissage**



## Prédiction du prix de l'immobilier

<https://challengedata.ens.fr/challenges/68>



Prédire le prix d'un bien immobilier à partir de ses caractéristiques visibles (surface, position, exposition, nombre de chambres, de salles de bain, etc.)



50,000 offres immobilières (40,000 pour l'entraînement et 10,000 pour le test)

**Données tabulaires** qui incluent 27 variables réparties dans trois principales catégories :

- Description du bien : taille, type, nombre de chambres, etc.,
- Localisation : latitude/longitude (bruitées), ville, code postale, performance énergétique, etc,

**Données d'images** (optionnel !) : entre 1 et 6 images pour chaque bien

**Métrique** Mean Absolute Percentage Error (MAPE)

$$\mathcal{L}(\mathbf{y}) = \frac{100}{n} \sum_{i=1}^n \left| \frac{y_i - \hat{y}_i}{y_i} \right|$$

**Benchmark** Modèle de régression XGBoost sur les caractéristiques tabulaires et sur un embedding simple des images ( $\mathcal{L}_{\text{test}} = 36.78$ )



## Prédiction de la prochaine place boursière d'une transaction

<https://challengedata.ens.fr/challenges/40>



**Classification (parmi six)** de la prochaine place boursière d'une transaction à partir de carnets d'ordre et d'historique d'échanges pour l'actif en question.



959,505 transactions avec pour chacune les données du **carnet d'ordre** et un **historique** pour l'actif d'intérêt.  
(Environ 400 actifs sur 250 jours)

**Carnet d'ordre** : Information sur les deux prix le plus haut des acheteurs et les deux prix les plus bas des vendeurs et des fonctions de ces prix moyennés (différences, etc.),

**Historique** : Description des 10 derniers échanges pour cet actif (prix, nombre, place boursière).

**Métrique** Précision de la classification mutli-classe

**Benchmark** Place du précédent échange dans l'historique ( $\mathcal{L}_{\text{test}} = 0.36$ )



## Relevé automatique de compteur d'eau

<https://challengedata.ens.fr/challenges/30>



**Classification** sur les entiers de l'indice d'un compteur d'eau à partir de photos clients (3 derniers chiffres)



L'ensemble d'entraînement dispose de  $n = 793$  images de tailles variables et de qualités hétérogènes de compteurs d'eau prises par des clients. Pour chaque image, on dispose de l'annotation humaine associée (l'indice du compteur d'eau).

**Problèmes :** Compteurs pas toujours verticaux, présence de terre, différents compteurs, etc.



**Métrique**

Coût binaire 0-1

$$\mathcal{L} = \frac{1}{n} \sum_{i=1}^n \delta(\hat{y}_i = y_i \bmod 1000)$$

**Benchmark**

CNN (<https://arxiv.org/abs/1902.09600>) ( $\mathcal{L}_{\text{test}} = 0.22$ )



Exemple d'une photo client d'un compteur d'eau. En noir les m<sup>3</sup>, en rouge les L. On cherche à prédire uniquement les m<sup>3</sup>.





Proposer et présenter vos solutions de ML à l'un des problèmes précédents



**Ne vous lancez pas dans un algorithme compliqué dès le départ** : analysez les données, construisez des modèles simples et comprenez pourquoi ils fonctionnent/ne fonctionnent pas, puis améliorez-les !

## ETAPES

1 Créer un compte sur le site <https://challengedata.ens.fr>

2 S'inscrire au cours « Data Challenge L2 IASO Dauphine – Juin 2025 » via ce lien



Pas de lien pour l'instant (site mort)

3 Constituer des **groupes de 3** pour travailler

4 Lire la page du challenge et télécharger les données

5 Chercher des solutions !

## A PROPOS DE L'EXAMEN

- Dates et heures du challenge : 2 au 6 juin, salle réservée de 9h30/17h (mardi 3 : salle rotonde 2<sup>e</sup> étage)
- **Date de l'examen : vendredi 6 juin de 9h30 à 14h30**
- **Présentation orale de 15 minutes de la/les solution(s) retenue(s) + 5 minutes de questions**

## QUELQUES CONSIGNES

- Battre le benchmark n'est pas l'objectif principal : il faut comprendre, analyser et justifier votre algorithme
- Vous pouvez bien sûr vous inspirer de méthodes trouvées sur internet, dans des articles, etc.
- Expliquez les résultats numériques : avis sur les sources d'erreurs, signes de sur-apprentissage (validation vs test public vs test privé), les avantages et inconvénients de votre approche
- Proposez des pistes et idées d'amélioration



- Sujet 1 : <https://web.archive.org/web/20250211161335/https://challengedata.ens.fr/challenges/68>
- Sujet 2 : <https://web.archive.org/web/20250117180702/https://challengedata.ens.fr/challenges/40>
- Sujet 3 : <https://web.archive.org/web/20250211161335/https://challengedata.ens.fr/challenges/30>