

## Sujet de stage de fin d'étude

### Conception et développement de métriques de fidélité pour les réseaux de neurones interprétables

**Mots-clés :** *Machine learning*, *Deep learning*, Interprétabilité intrinsèque, *Concept learning*, Méta-variables (*meta-features*), Métriques d'évaluation, Biais statistique.

#### L'équipe que vous allez rejoindre

Le stage aura lieu au sein du laboratoire **IBISC** (Informatique, Bioinformatique, Systèmes Complexes) de l'Université Paris-Saclay (Univ. Évry), au sein de l'équipe **AROBAS**. Vous intégrerez un noyau de chercheurs spécialisés en apprentissage statistique et *deep learning*, développant des travaux théoriques et méthodologiques reconnus au plan international.

**Encadrement :** Pr. Blaise Hanczar

#### Contexte général

L'interprétabilité des modèles d'apprentissage profond est devenue un axe de recherche central pour fiabiliser la décision algorithmique. Au-delà des méthodes d'explication *post-hoc*, l'interprétabilité dite *intrinsèque* vise à concevoir des architectures dont certaines couches latentes représentent explicitement des entités intelligibles par construction.

Une approche récente repose sur la notion de **méta-variables** (*meta-features*), où chaque neurone latent est contraint pour représenter le comportement d'un sous-ensemble spécifique de variables d'entrée. Pour garantir la fiabilité de ces représentations, il est crucial d'évaluer fidèlement la correspondance entre le neurone et son groupe de variables associé, sans que des corrélations parasites ne viennent fausser la représentation.

Les métriques traditionnelles du domaine (issues du *concept learning*) montrent d'importantes limites dans ce cadre. Pour y répondre, de premières métriques dédiées, fondées notamment sur des simulations de perturbations ciblées et sur la préservation de structures géométriques, ont été introduites au laboratoire.

## Objectifs du stage

Le stage s'inscrit au cœur des fondements méthodologiques du *machine learning*. L'enjeu principal sera de consolider le cadre d'évaluation des représentations latentes interprétables en levant les verrous actuels des métriques existantes :

1. **Analyse critique et correction des biais** : Identifier et quantifier les biais inhérents aux métriques actuelles (sensibilité au chevauchement entre sous-ensembles de variables, effets d'échelle et dépendances résiduelles). Proposer des formulations corrigées et robustes.
2. **Conception et extension de nouvelles métriques** : Élargir le panel d'outils statistiques pour mesurer plus fidèlement le confinement de l'information et le démêlement des variables au sein de l'espace latent.
3. **Validation expérimentale** : Comparer rigoureusement les métriques proposées face aux approches de l'état de l'art sur des jeux de données synthétiques modulables (contrôle précis de l'enchevêtrement et des dimensions), complétés par une validation finale sur des données réelles de grande dimension (transcriptomique unicellulaire).

## Profil recherché

- Étudiant en Master 2 Recherche ou dernière année d'école d'ingénieur en informatique, machine learning ou mathématiques appliquées.
- Solides bases théoriques en apprentissage automatique et compétences pratiques en Python / PyTorch.
- Goût prononcé pour la recherche scientifique.

## Modalités de candidature

- **Type de contrat** : Stage de 5 à 6 mois (gratification légale).
- **Lieu** : Laboratoire IBISC, Université Paris-Saclay, Évry-Courcouronnes.
- **Contact** : Envoyer CV, lettre de motivation et relevés de notes récents à [blaise.hanczar@univ-evry.fr](mailto:blaise.hanczar@univ-evry.fr).