Quantmetry
Building AI with pionneers

# Stage : Data Scientist (H/F)
# Active learning

## Qui sommes-nous ?

Nous pourrions commencer en nous présentant comme l'un des leaders sur le marché international du conseil en data, avec un chiffre d'affaires défiant ceux des G.A.F.A. Mais non. Nous préférons mettre en avant notre cadre de travail, nos réussites et surtout nos consultants. Allez venez, c'est par là…

Quantmetry c'est une centaine de personnes qui travaille de concert pour accompagner nos clients dans leurs réflexions, P.O.C., industrialisation autour de la Data, l'Intelligence Artificielle, le Machine Learning et la Big Data. Nous sommes ce qu'on peut appeler communément un cabinet de conseil pure player en Data.
N'étant pas sectorisés, nous intervenons sur différents sujets (voici une petite liste non-exhaustive) :
- La maintenance prédictive,
- La segmentation clients, le scoring,
- La détection de fraude, de corruption, de blanchiment d'argent,
- Le développement des architectures de plateformes Data,
- L'analyse de textes et d'images dans le cadre de la lutte contre le cancer du sein.

Dans le cadre de notre incessante croissance, nous recherchons des personnes motivées pour nous accompagner et ainsi travailler sur des projets de recherche.

## Préalable

Quantmetry propose ci-dessous le volet R&D d'une offre de stage. Tout stagiaire entrant chez Quantmetry, outre le travail de R&D qui lui est proposé et pour lequel il sera encadré, aura aussi pour objectif de participer à certaines missions de conseil chez des clients variés, lui permettant d'aborder le monde du consulting.

Quantmetry
Building AI with pionneers

# Contexte

L'active learning est un domaine à part entière du machine learning. En effet, l'apprentissage supervisé nécessite un carburant : les labels. Par exemple, les transactions frauduleuses, les défauts sur les rails de train ou les spams doivent être labellisés (fraude/pas fraude, spam/pas spam) par un opérateur humain avant d'entraîner un modèle d'intelligence artificielle. La qualité des modèles de machine learning est grandement impactée par la quantité de données labellisées disponibles. Or, labelliser des données coûte cher pour plusieurs raisons :

- Il nécessite des opérateurs humains effectuant un travail long, répétitif, et souvent délocalisé dans des pays à faible coût de main d'œuvre.
- Pour certains cas d'usage, la labellisation nécessite une expertise rare (par exemple en santé) et peu de temps peut être alloué à cette tâche.

L'active learning est la science qui permet de minimiser le nombre de labels requis pour optimiser la performance d'un algorithme, dans une démarche à la fois éthique et pragmatique. C'est aussi un outil d'adaptation à la dérive des modèles, qui permet d'affiner la frontière de décision en fonction des évolutions du signal. Plusieurs outils sont disponibles en accès libre, en particulier :

- modAL, développé par l'académie hongroise des sciences
- libact, développé par l'université de Taiwan
- AliPy, développé par l'université de Nanjing (Chine)

La promesse de ces outils est de fournir des outils simples d'utilisation qui permettent d'implémenter des algorithmes d'active learning facilement.

Les objectifs de ce stage sont :

- Prendre en main les 3 outils et les appliquer à des jeux de données simples en accès libre
- Comparer leurs fonctionnalités, évaluer leur complétude vis-à-vis de l'état de l'art et lister leurs avantages/inconvénients
- Construire un guide méthodologique et un tutoriel d'active learning avec les outils jugés les plus pertinents
- Animer une formation en interne sur l'active learning