**Contexte :**

Nous réalisons de façon continuelle et itérative des POCs autour des problématiques de la donnée et une des problématiques actuelles concerne le “**Product Matching**” (détection de produits à travers différentes **descriptions littérales / attributs** et à travers leurs **images** collectés sur plusieurs sites internet distincts) .

Le projet s’inscrit dans le cadre d’un programme de développement à long terme avec un objectif d’amélioration continu et d’enrichissement de nouvelles sources et fonctionnalités.

**Contenu du projet :**

- Réaliser une analyse sémantique (lexicale / syntaxique ) du corpus technique caractéristique des produits de nos clients, faire du Feature engineering grâce aux différentes techniques et **approches d’apprentissage Statistique / Machine Learning supervisées, semi ou non supervisées (NLP, NLTK, CRF, TF-IDF, Word Embedding)** afin de construire des modèles prédictifs : Classifications / Regroupement / Clustering.

- Améliorer la performance des modèles prédictifs grâce à la prise en compte des images des produits à travers des approches de **Deep Learning**

- Analyse / évaluer la similarité des produits par des techniques classiques ( Similarity & metrics learning).

- Prédiction de référence d’un produit suivant leurs descriptions / images.

- Réaliser une analyse sémantique afin de construire une base de connaissance lexicale (verbatims) des produits ainsi que leurs liens grâce aux techniques d’extraction d’information (Information retrieval) : NER, CRF etc…

*Les modèles pré-entrainés de reconnaissance d’images et/ou de Word embedding de Google, Microsoft Azure, AWS pourront être expérimentés et enrichi en première approche.*

**Outils / techniques** : Python / Pyspark, scikit-learn, Numpy, Pandas, Gensim, Machine learning, NER, NLTK, NLP, Word Embedding, TF-IDF, Tensorflow, Keras, Deep Learning