

---

# TP Clustering

## 5e ISS

---

MJ. HUGUET  
`homepages.laas.fr/huguet`

### Objectifs

Le but de ces TP est mettre en oeuvre et de comparer différents algorithmes de clustering tout d'abord à partir de quelques méthodes fournies par `scikit-learn` puis en utilisant une méthode externe :

- $k$ -Means
- clustering hiérarchique (agglomératif)
- DBSCAN
- HDBSCAN

### Encadrants

Marie-José Huguet, Laure Feuillet, Hieu Trong Tran

## 1 Environnement de travail

Durant ces TP, nous allons utiliser le langage Python et la distribution Anaconda (installée à l'INSA). Cette distribution fournit la librairie Python `scikit-learn` contenant un certain nombre de méthodes d'analyse de données et d'apprentissage.

Documentation : <https://scikit-learn.org/stable/>

## 2 Jeux de données

Nous allons utiliser des jeux de données disponibles et visualisables sur le site <https://github.com/deric/clustering-benchmark>. Seuls les jeux de données "artificiels" sont considérés dans ces TP. Ces jeux de données sont principalement en 2 dimensions pour des raisons pédagogiques. Vous trouverez une archive des jeux de données artificiels sur la page moodle de cet enseignement.

Note : ne prenez pas en compte les jeux de données en gris.

La lecture de ces jeux de données (format `arff`), s'appuie sur le package python `arff : from scipy.io import arff`. Le jeu de données est ensuite mémorisé sous la forme d'un tableau à  $d$  dimensions ( $d = 2$  pour la plupart des jeux de données proposés) puis affiché sous la forme d'un nuage de points.

Note : dans ces jeux de données, pour chaque exemple, la dernière colonne fournit un numéro de cluster (sans précision sur la méthode utilisée pour l'obtenir). En pratique, **vous ne devez pas utiliser cette colonne** car on suppose que les clusters ne sont pas connus.

### A faire

- Récupérez le code fourni pour lire un jeu de données `arff` et l'afficher. Profitez en pour lire et afficher un jeu de données en dimension 3, des jeux de données avec des nombres variables de points, et pour tester des paramètres de la visualisation.
  - Objectif : Découvrir quelques outils de visualisation (consultez la librairie `matplotlib`)
- Appliquez des pré-traitements pour standardiser les données et comparez les visualisations des données initiales et des données standardisées. Choisissez pour cela des jeux de données avec des caractéristiques différentes.
  - Objectif : Apprendre des méthodes de pré-traitement. (consultez la documentation de `scikit-learn` sur le pre-processing)
- Identifiez les méthodes de clustering disponibles sur `scikit-learn` et leur caractéristiques générales. Consultez la comparaison des méthodes de clustering de `scikit-learn`.
  - Objectif : Avoir un premier aperçu (par l'exemple) des avantages et limites des méthodes disponibles.

### 3 Clustering $k$ -Means

Documentation : <https://scikit-learn.org/stable/modules/clustering.html#k-means>.

#### A faire

- Récupérez le code fourni qui applique la méthode  $k$ -means pour un nombre fixé de clusters sur le jeu de données `xclara.arff`.
- Testez des paramètres de la méthode.
- Évaluez le résultat avec d'autres métriques de `scikit-learn`.

#### 3.1 Evaluation de la méthode

On va chercher à déterminer "automatiquement" le bon nombre de clusters en appliquant la méthode  $k$ -means pour plusieurs valeurs de  $k$  et en évaluant les résultats obtenus.

#### A faire

- Codez des procédures d'évaluation de la méthode  $k$ -means (méthode du "coude" basée sur les valeurs d'inertie, valeurs du coefficient de silhouette et valeurs des autres métriques disponibles).
- Présentez des graphiques pour les différentes métriques mesurées.
- Sélectionnez 2 ou 3 jeux de données pour lesquels il vous semble que la méthode  $k$ -means devrait bien fonctionner et 2 ou 3 exemples pour lesquels il vous semble que cette méthode ne fonctionnera pas correctement.

- Inspiration possible sur le web ... (attention à la qualité du code ...)
- N'oubliez pas de relever les temps de calcul.
- Arrivez-vous à déterminer correctement le bon nombre de clusters à l'aide des différentes métriques utilisées ?
- Avez-vous identifié les avantages et limites de la méthode  $k$ -means ?